The social components of innovation: from data analysis to mathematical modelling

Gabriele Di Bona

PhD thesis

School of Mathematical Sciences
Queen Mary University of London
United Kingdom

October 2023

Abstract

Novelties are a key driver of societal progress, yet we lack a comprehensive understanding of the factors that generate them. Recent evidence suggests that innovation emerges from the balance between exploiting past discoveries and exploring new possibilities, the so-called “adjacent possible". This thesis aims at developing new analysis tools and models to study how people navigate the seemingly infinite space of possibilities.

Firstly, I extend the notion of the adjacent possible to account for novelties as combinations of existing elements. In particular, I model innovation as a random walk on an expanding complex network of content, in which novelties correspond not only to the first visit of nodes, but also of links. The model correctly reproduces how novelties emerge in empirical data, highlighting the importance of the exploration process in shaping the growth of the network.

Secondly, since people continuously interact and exchange information with each other, I investigate the role of social interactions in enhancing discoveries. I hence propose a model where multiple agents extend their adjacent possible through the links of a complex social network, exploiting in this way opportunities coming from their contacts. By adding a social dimension to the adjacent possible, I prove that the discovery potential of an individual is influenced by its position on the social network.

Finally, I combine the two concepts of the adjacent possible in the content and social dimension to develop a data-driven model of music exploration on online platforms. In such a model, multiple agents grow their individual space of possibilities by exploring a network of similarity between artists, while exploiting suggestions from their friends on the social network. The comparison with the empirical data indicates that the adjacent possible, in both the content and the social space, plays a crucial role in determining the individual propensity to innovate.

Acknowledgements

When I decided I wanted to pursue a PhD degree, I had no idea about what a PhD really was. Now that I have finished writing this PhD thesis, well, I still have no idea. Maybe the best way I can describe it is “an opportunity". Let me explain.

It all started when I met someone who is now much more than a supervisor, Prof. Vito Latora. When we first met, in a seminar at the Scuola Superiore di Catania (SSC), I instantly knew I could not miss the opportunity to talk to him, and I am so glad I did. Well, I did not do much during that first discussion; after all, I was just a mere student with a lot of enthusiasm and little knowledge of what it really means to do research. In return, he gave me the opportunity to do research, and, most importantly, be a researcher. I cannot be more proud and honored to have been one of his PhD students. He managed to channel all my confused enthusiasm into clear research ideas and methodical work. Just to give an example, it is all thanks to his patience and time spent side by side writing our research papers that (I hope) I have stopped writing in *Gabrielese*, as he used to say (probably these acknowledgements are a good example of what I mean). These few lines are definitely not enough to explain how grateful I am for having spent these last years with such a person, or as I say about him, so “duci" (the closest translation is “sweet", but it is not just that). I hope to inherit all his amazing qualities, being a volcano of ideas and research projects, without losing humility and sweetness, and always ready with a smile.

I have to equally thank Prof. Vittorio Loreto. Probably he is the one that really sparked in me the interest for complex systems, innovation and creativity. He gave me the opportunity to start to do research well before my PhD. He truly believed in me; his support and supervision have been fundamental in the last (I lost count of how many) years. He was the first one to give me the opportunity to move out from my beloved Sicily, but most importantly, he gave me the opportunity to see how much impact we can do on our society using some maths and physics, and how entertaining it can be.

I hope this thesis will be convincing enough that there is no innovation without social interactions. If I have learned anything from this PhD, it is that collaborations are the building blocks of how science progresses. And it is so much fun! During these years I have had the opportunity to engage and collaborate with a bunch of extraordi-

nary scientists scattered around the world: Prof. Andrea Giacobbe, Prof. Lucas Lacasa, Prof. Nicola Perra, Prof. Andrea Baronchelli, Prof. Francesca Tria, Prof. Caroline Di Bernardi Luft, Prof. Federico Battiston, Dr. Iacopo Iacopini, Dr. Enrico Ubaldi, Dr. Bernardo Monechi, Dr. Alberto Bracci, Dr. Andrea Civilini, Dr. Angelo Petralia, as well as all the other scientists that I have had the chance to meet and talk to, online or in presence, working or joking in front of a drink. I wish I could say something about each of you, but for the sake of length of these acknowledgements, I will just say that you have all been a role model for me, and each of you has taught me something I will always cherish. In addition, I would like to acknowledge the examiners, Prof. Renaud Lambiotte and Prof. Rainer Klages. It does not happen often to have someone to spend so much time reading and commenting on your own research work, so I am really glad that they have agreed to examine my thesis in the first place.

Without this PhD, I would not have had the opportunity to meet this incredible long list of friends who have made my staying in London a lot less gray and rainy (I did not expect it so long, I feel truly blessed!). It is really hard to sort this list, but in the end it does not really make sense to sort it. You have all been special to me, even those who have not actually been physically in London with me. If I have made it through the many hard times, it is only thanks to you. I wish to say something about each of you, but I will save it for the next time we will meet, I hope very soon. Anyway, this is a randomized (yes, I actually did it) version of the list (I know I have forgotten someone, please forgive me!): Cinzia, Muthu, Hanlin, Lisa, Sarah, Theo, Ana Karen, Mathieu, Luca, Tim, Carola, Addie, Lollo, Benny, Gallo, Martina, Leo, Halime, Sunny, Angi, Eshti, Adi, Elisa Babes, Grace, William, Zio Mollica, Jevon, Ester, Houda, Helen, Albi, Emma, Daniela, Silvia, Andy, Dani, Dom, Esma, Sofi, Marco, Diego, Evelyn, Pietro, Marcia, Timur, JP, Samuel, Bea, Enrico, Antoine, Chino, Francisco, Fra, Giordano, Ocean, Alessandro, Andrea, Fratello Lorenzo, Patrick, Jerome, Vicio, Giac, Fede, Julio, British Marco, Clementina, and Oliver. Friendships are hard to build and maintain, but you have truly helped me a lot. I hope I will never forget all the incredible moments we have spent (and will spend!) together, from the super sad to the super happy ones. I also hope I can still be there for all of you, wherever and whenever you will need me. I am sorry if sometimes I did not meet your expectations and could not be there for you, I hope you will understand.

Finally, I wish to thank my whole family, who has been my first and foremost supporter and to whom I wish to dedicate this thesis. Without their love, I could have never made it. I know I have been very far away, I know I did not call as much as you wanted to, I know you wish I could have stayed all the time at home with you, I know how much you miss me. And this is why I cannot be more grateful to you. But I hope you can see what an incredible opportunity you have given me. We have had our ups and downs, but I will never stop loving you, whatever the next steps will be. Everyone I have met knows how much I love my Sicily, my town. You probably do not know

this, but whenever people I meet ask me where I am from, I always say I am from Sicily (no offense for the rest of Italy, I also love you). If I am the person I am now, it is all thanks to you and your love. People also say that the Covid pandemic has been a disaster. And it is true, we have lost so much. At the same time, however, it has given me the opportunity to spend so many months at home, bringing us closer, probably the closest I have ever felt to you, after so long we have been apart. Thank you mom, dad, Sara, Piera, Giovanni, Daniele, Marialuisa and all the little ones, the most “duci" of all (sorry Vito).

And thanks to you who have come across this thesis, I hope you will enjoy it! I do not know where or when you are, but if you want or need, please, do not hesitate to get in touch with me!

Declaration and publications

I hereby declare that, except where specific reference is made to the work of others, the contents of this thesis are original. This dissertation is my own work or is the outcome of work done in collaboration with others. This work is copyright © 2023 Gabriele Di Bona, and is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported Licence. To view a copy of this licence, visit `http://creativecommons.org/licenses/by-sa/3.0/`. Ideas, results, and figures appearing in this dissertation are based on the publications listed below.

(I) I. Iacopini[1], **G. Di Bona**[1], E. Ubaldi, V. Loreto, and V. Latora. "Interacting discovery processes on complex networks". In: *Physical Review Letters* (2020), 125.24, p. 248301 [1].

(II) **G. Di Bona**, E. Ubaldi, I. Iacopini, B. Monechi, V. Latora, and V. Loreto. "Social interactions affect discovery processes". In: *arXiv preprint* 2202.05099 (2022) [2].

(III) **G. Di Bona**[1], A. Bracci[1], N. Perra, V. Latora, and A. Baronchelli. "The decentralized evolution of decentralization across fields: from Governance to Blockchain". In: *arXiv preprint* 2207.14260 (2022) [3].

(IV) **G. Di Bona**, L. Di Gaetano, V. Latora, and F. Coghi. "Maximal dispersion of adaptive random walks". In: *Physical Review Research* (2022), 4, L042051 [4].

(V) C. D. B. Luft, I. Zioga, E. Giannopoulos, **G. Di Bona**, N. Binetti, A. Civilini, V. Latora, I. Mareschal. "Social synchronization of brain activity increases during eye-contact". In: *Communications biology* (2022), 5.1, pp. 1–15 [5].

(VI) **G. Di Bona**, A. Bellina, G. De Marzo, A. Petralia, I. Iacopini, and V. Latora. The dynamics of higher-order novelties". In: *arXiv preprint* 2307.06147 (2023) [6].

Other publications not covered in the thesis are listed below.

(VII) **G. Di Bona** and A. Giacobbe. "A Simple Theoretical Model for Lags and Asymmetries of Surface Temperature". In: *Climate* (2021), 9.5, p. 78 [7].

[1] These authors contributed equally.

Contents

List of Figures

List of Tables

List of Abbreviations

ABM	Agent-based Model
AP	Adjacent Possible
APS	American Physical Society
BA	Barabási-Albert
ciPLV	corrected imaginary Phase-Locking Value
EEG	Electroencephalogram
ER	Erdős Rényi
ERRW	Edge-Reinforced Random Walk
ERRWT	Edge-Reinforced Random Walk with Triggering
MPN	Mobile Phone Network
RW	Random Walk
SD	Standard Deviation
SEM	Standard Error of the Mean
SW	Small World
TMN	Twitter Mention Network
UM	Urn Model
UMT	Urn Model with Triggering
UMST	Urn Model with Semantic Triggering
ZKC	Zachary Karate Club

Chapter 1

Introduction

Innovation is the engine of societal progress, pushing us towards uncharted territories of knowledge and boundless possibilities. The ability to come up with novel ideas and concepts, or make new technologies and other breakthroughs, holds the key to foster progress and change across diverse domains, from science and technology to arts and humanities, to cite a few. Therefore, this thesis aims to establish a comprehensive mathematical framework capable to address pivotal questions regarding innovation. Specifically, what are the underlying mechanisms that shape the appearance and diffusion of innovative ideas, products, technologies, or artistic creations? How do various factors impact the delicate balance between exploiting existing knowledge and resources versus exploring new possibilities and discoveries? Additionally, what role do networks and social interactions play in the dissemination and adoption of innovative concepts and products across different domains?

To answer these questions, we first need to acknowledge the inherent sequentiality in the process of innovation. Far from being just the consequence of random occurrences, this process can indeed be described as a structured sequence of events, characterized by distinct steps and stages that build upon one another. On the one hand, innovation involves the continuous exploitation and refinement of existing knowledge, reinforcing their presence and influence in various aspects of society. On the other hand, it encompasses the exploration and discovery of adjacent possibilities, pushing the boundaries of what is currently feasible. The combination of exploitation and exploration forms a progressive series of discoveries that seamlessly integrate with the established knowledge, fostering innovation.

To have a practical idea of how this process unfolds, let us take into consideration the development of a remarkable technology that has transformed our daily lives, the smartphone. Initially, the concept of a portable telephone emerged as a novel idea, driven by the need for enhanced communication capabilities. As it became more accessible to a broader number of people around the globe, engineers and inventors started to

explore various technologies and components that could be integrated into the evolving design. Through iterative steps combining existing knowledge with new discoveries, the realization of a fully functional smartphone was approached.

In this example, the concept of a smartphone appeared distant and far-fetched initially. However, through a gradual progression and the convergence of diverse technologies and attainable adjacent possibilities, it eventually became a reality. The concept of the *adjacent possible* and its exploration indeed offers a fascinating perspective on innovation [8]. Pioneered by scientists such as N. Packard [9], C. Langton [10, 11] and S. Kauffman [12, 13], it posits that, at any given moment, there exists a realm of potential discoveries that can emerge from the existing state of the system. In particular, the adjacent possible refers to the set of potential opportunities, ideas, and other possibilities that are only one step away and immediately accessible from the current state of knowledge, technology, and resources. It hence represents the range of potential developments that can emerge from the existing conditions and serve as foundation for further exploration and innovation, as it has been the case of the smartphone.

We have started to understand that representing a discovery process as a sequence of exploration of new elements and exploitation of already known ones is the key to answer the questions posed at the beginning of this chapter. The second step we need is finding measures and empirical laws to analyse these processes through their sequences. The recent availability of vast amounts of digital data has revolutionized our ability to track and analyze innovation processes. This wealth of data has opened up new possibilities for understanding and uncovering empirical laws that govern the dynamics of innovation, such as the *Heaps' law* [14, 15], which has been observed in a variety of contexts [8].

In general, given a growing collection of items, such as words in a text or concepts developed in a scientific field, the Heaps' law describes the relationship between the total size of the collection and the number of unique items in it. In particular, the growing collection follows the Heaps' law if the number of unique elements increases grows as a sublinear power-law function of the total size of the collection. This law has been successfully exploited in discovery processes to characterize the pace at which novelties emerge along the process. In such a process the growing collection of items can indeed be represented by the related temporally-ordered sequence containing the concepts, technologies, artworks, or any other items explored in the process under study. Then, counting the number of items N in the sequence, the Heaps' law states that the number of different elements $D(N)$ in the sequence grows as a sublinear power-law function of the size N of the sequence, i.e., $D(N) \sim N^{\beta}$, with $0 \leq \beta \leq 1$. For example, considering as a discovery process the listening activity of a person, the ordered sequence is made of all the songs listened by the person. Here, a discovery is made whenever a song is listened for the first time. The Heaps' law hence states

that the number of discoveries, i.e., the number of unique songs listened, grows as a power-law function of the total size of the sequence, i.e., the total number of listening records.

The Heaps' law provides valuable insights into the patterns and dynamics underlying the emergence of novelties. Notice in fact that, as a consequence of the Heaps' law, the average time between two consecutive novelties increases as the system evolves. Drawing upon the previous example of the smartphone development, during its early stages the collection of ideas and possibilities to apply experienced a rapid growth, expanding the adjacent possible at a fast pace and giving rise to a plethora of potential features, functionalities, and design variations. However, as smartphones evolved and matured, the pace at which entirely new and distinct ideas emerged gradually slowed down, leading to longer intervals between impactful discoveries. In other words, there has been a gradual change from a phase of rapid exploration to one of incremental refinement, or exploitation. Such change is controlled by the Heaps' exponent: the lower the exponent, the longer the intervals between two novelties. In particular, when the exponent is close to its maximum value of one, discoveries occur at a nearly linear pace, with a continuous stream of new findings. Conversely, an exponent closer to zero indicates that after an initial burst of discoveries, the number of new breakthroughs significantly diminishes, leaving little room for further exploration.

The exploration of the adjacent possible and the Heaps' law discussed so far recognize that breakthroughs and novel ideas do not materialize out of thin air, and highlight the sophisticated interplay between existing knowledge and adjacent possibilities. Various approaches have emerged from diverse disciplines to elucidate the underlying mechanisms of this process. For instance, the idea from evolutionary biology that novel species emerge through genetic mutations and subsequent selection processes [16] has then lead to a deeper understanding of the concept of the adjacent possible in more recent years [8, 13]. Similarly, cognitive psychology has studied the exploration-exploitation dilemma to understand how individuals balance between exploiting familiar options and exploring new alternatives [17, 18], advancing the theory of reinforcement learning [19]. Furthermore, in the context of innovation economics, the emergence of novelties is seen as the consequence of a dual process of creation and destruction [20], where new associations of existing factors may give rise to innovations (creation) and rule out of the market obsolete products and services (destruction), thereby increasing the probability of reaching further novelties [21, 22].

Various mathematicians and physicists, among other researchers, have also proposed generative models of innovation. Based on extractions from urns or on the movement of random walkers on complex networks, such models are able to capture the key features of innovation. In particular, *Pólya's urn model* translates the concept of reinforcement learning into the process of drawing colored balls from an urn and

adding more balls of the same extracted color into the urn [23, 24]. Kauffman's idea of the adjacent possible has also been integrated in the context of urns thanks to the *Urn Model with Triggering* [8], or UMT. This model extends Pólya's urn model by adding a triggering mechanism of new colors: whenever a color never observed before is extracted, new colors are added in the urn. In other words, the exploration of a novelty gives rise to the expansion of the colors present in the urn, opening up more possibilities of further discoveries. Through the reinforcement and triggering mechanisms, this model is able to reproduce the statistical patterns observed in real-world data of exploration and innovation processes, such as the characteristic emergence of novelties as captured by the Heaps' law [25–27]. Since its introduction, the UMT has been applied in a variety of contexts, for example, to study the rise and fall of popularity in technological and artistic productions [28], the cognitive growth of knowledge in scientific disciplines [29], the emergence and evolution of social networks [30], or the evolution of the cryptocurrency ecosystem [31].

In addition to urn models, complex networks and random walks have provided a powerful framework to understand how the adjacent possible unfolds, grows and is explored. By representing concepts, ideas, or technological advancements as nodes, and the connections between them as edges, a network offers a visual and mathematical representation of the adjacent possible. Thanks to the extensive research conducted on the properties of complex networks [32–36], such representation enables to investigate how the interplay of ideas and their combinations can lead to the emergence of novel and innovative elements in the system under study. For example, the preferential attachment mechanism proposed in the Barabási-Albert model [37] illustrates how the rich-get-richer phenomenon influences the growth of networks, where nodes with more connections have a higher probability of acquiring additional connections from new nodes in the network.

If complex networks encrypt the topological properties of the growing space of possibilities in their structure, random walks have proven to be a valuable and versatile tool for modelling and understanding how such space is explored and discoveries are made [29]. Random walks consist in exploring a network by randomly moving from one node to another based on certain rules, from the simplest unbiased random walk in which the next node is chosen uniformly at random among the neighbors, to the most sophisticated ones. Thanks to their versatility, random walks have recently received significant attention and have been extensively studied [38–40]. They have been used to build exploration models for social annotation [41], music album popularity [42], knowledge acquisition [43], human language complexity [44], and evolution in research interests [45]. Therefore, by capturing the movement and interaction of entities within such networks, random walks can offer valuable insights into the emergence and propagation of ideas, the acquisition of knowledge, and the evolution of complex systems.

Building upon the insights provided by the Heaps' law and the mathematical models described above, in this thesis we take a step further to understand how the space of possibilities grows in an innovation process. The **first major contribution of this thesis** is to associate innovation to the emergence and exploration of a more general definition of novelties [6]. In particular, we consider novelties as new combinations or associations of existing elements, which can impact on how the space of possibilities being explored grows. For instance, in the context of the smartphone, various components, features, or technologies, such as the camera, processor, display, battery, operating system, and so on, can be combined together to obtain more complex functionalities of the smartphone. Similarly in other contexts, the first association of already known items can generate novelties, as it is the case of words to form a poem, or of nucleic bases in a DNA sequence.

To quantify the pace at which novel associations of n elements appear in a sequence, we extend the Heaps' law to what we call the *n-th order Heaps' law*. Such law characterizes the time distribution of novelties of size n, i.e., the first time n elements are associated together, through a power law. We refer to the exponent of such power law as the *n^{th}-order Heaps' exponent*. The analysis of higher-order Heaps' laws in different contexts allows us to distinguish processes that generate the same amount of new items, but that exhibit different rates of new associations between them. In light of this, we envision innovation as the process of exploring a network, where each node represents a concept or item, and the edges symbolize the connections between them. This network represents the space of possibilities of the system, growing along with the process either by adding new nodes, making new connections, or reinforcing existing ones. By integrating the mechanisms of reinforcement learning and triggering of adjacent possible nodes and links, we hence model innovation as a generative process in which a weighted network grows in time while it is being explored. Building on the *Edge-Reinforced Random Walk* model [29] and the *Urn Model with Triggering* [8], we introduce a new model, which we call the *Edge-Reinforced Random Walk with Triggering*, that is able to capture not only the dynamic nature of innovation as a sequential exploration of old and new nodes or links, but also the expansion of the adjacent possible in conjunction with the exploration process. In fact, as the random walker explores the growing network, it not only exploits and reinforces existing connections and knowledge, but also actively triggers the addition of new elements and associations in the network, fostering the emergence of novelties at various orders.

Another pivotal aspect of innovation lies in the power of social interactions and human connections. Competition and cooperation are indeed two of the key drivers behind the evolutionary success of the human species [46–48]. Such interactions can be encoded in social networks [49], which can be studied to answer various research questions at different levels of analysis, from the individual to the collective ones [36,

50, 51]. In particular, social networks have been extensively used as a substrate on top of which dynamical processes take place [35, 52]. In the context of innovation, social interactions exert a profound influence on the individual's propensity to generate and diffuse novel ideas and content. As a matter of fact, even if someone does not invent a technology or writes a song, the first time the individual uses such technology or listens to this song can still be considered a novelty for the person. This novelty at the individual level can then open up more possibilities to discover something new, as it enlarges the personal knowledge space. Recent empirical evidence shows that peer influence does not necessarily need active and engaging interactions to make an impact. For example, it has been demonstrated that even simple eye contact enhances synchronization between the brains of two individuals [5]. This synchronization directly impacts our choices in simple tasks, and highlights the importance of social dynamics in more complex processes. Thus, recognizing the collective nature of innovation, it becomes essential to unveil the interplay between social interactions and the emergence of novelties.

Graph theory and network science provide valuable tools to study these intricate networks of social interactions, uncovering patterns and structures that facilitate discoveries. Network analysis indeed enables the identification of key players, influential nodes, and hidden pathways in the social network. For example, going back to the previous example of how smartphones developed, these analyses can shed light on how multidisciplinary teams have collaborated, leveraging their different expertise. Through the exchange of ideas and integration of different perspectives, these teams have expanded each other's adjacent possible, thus creating something that individually would have taken much longer to develop. Therefore, the **second major contribution of this thesis** is to theorize the existence of an *adjacent possible in the social space*, which enlarges the *adjacent possible in the content space* of multiple individuals through their social interactions [1]. We incorporate this new concept in the modelling scheme provided by the UMT, extending the model dynamics to account for multi-agent collaborative exploration. In the model we propose, called the *UrNet* model, multiple urns represent different explorers who are interconnected through a social network. This coupling allows such individuals to exploit opportunities that arise from their social contacts. Overall, the UrNet model contributes to a more comprehensive understanding of innovation as a social process, emphasizing the importance of collaboration to push the boundaries of what is possible. In particular, analyzing various network structures, from small synthetic graphs to large real-world ones, we demonstrate that an explorer's pace of discovery depends on their centrality within the social network, underscoring the crucial role that social connections play.

To further explore the influence of peers on the discovery dynamics, we employ a real-world data set obtained from *Last.fm*, that is an online platform and music stream-

ing service that focuses on music discovery, personalized recommendations, and social interactions between its users. Last.fm was originally launched in 2002 as a music-oriented social networking site, and has since evolved into a popular music-recommendation service. One of the key features of Last.fm is its scrobbling functionality. Scrobbling means that the service automatically keeps track of the songs a user listens to on various music streaming platforms or media players connected to the Last.fm account, creating a comprehensive history of listening records. This information is used to generate charts, statistics, and recommendations based on their music tastes. Last.fm also provides users with the ability to create and join groups, share music-related content, and connect with like-minded individuals. The platform fosters a sense of community by allowing users to interact with each other, discover new music based on their friends' recommendations, and participate in various discussions and events.

Overall, Last.fm offers a combination of music discovery and social networking, making it a popular choice for music enthusiasts looking to explore new artists, connect with others who share their musical interests, and expand their music library. Moreover, this platform provides open access to the sequences of songs listened by its users, as well as their social connections. Therefore, in the context of this thesis, Last.fm is the perfect test bed to analyze how social interactions affect the way the adjacent possible of each person grows. Using Last.fm APIs, we obtain a connected and representative sample of users along with their entire listening history and their social connections. For each user, we can hence measure their pace of discovery of new music, as well as the structure of their space of possibilities. Moreover, thanks to the information on their social connections, we can study how these measures relate to those of their friends. For example, we observe that individuals with a high discovery rate tend to have a high number of connections with others who also discover new music frequently. This finding indicates the presence of a positive social influence, where being exposed to peers with a strong inclination for exploration and discovery fosters an individual's own propensity to encounter new music.

The **third major contribution of this thesis** is hence to adapt and combine the mechanisms introduced above to replicate and explain the effects of social interactions on individual music discovery on Last.fm [2]. Specifically, we develop an agent-based model with two main dynamics. On the one hand, agents explore the music network and grow their individual space of possibilities as a weighted subnetwork; on the other hand, they interact with their friends over the social network, further expanding their adjacent-possible space. The first crucial element in our model is the representation of the musical space as a similarity network between artists, where links signify meaningful associations between them. To obtain this network, we use the same approach of the edge-reinforced random walk with triggering, that is, we imagine that the sequences of listening records are the result of a random walk over a network of artists. Secondly, we

build each agent's dynamics using the discussed mechanisms at the base of the balance between exploitation and reinforcement of existing knowledge versus exploration and discovery of adjacent possibilities. Finally, drawing inspiration from the UrNet model, we also include the expansion of the adjacent possible through the social space. Specifically, mimicking realistic interactions on Last.fm, agents randomly observe what their friends have recently explored in the network, allowing for the discovery of new artists through suggestions from their peers. This modelling approach enables us to simulate and study the intricate process of exploration and discovery, gaining valuable insights into the effect of social influence on the emergence of novelties. By capturing how social interactions affect the individual exploit–explore process and replicating the dynamics observed in the empirical data, our model, named the *ExploNet* model, paves the way for a deeper understanding of the underlying mechanisms in music exploration.

In summary, in this thesis, we present a multidimensional study of innovation and the emergence of novelties through social interactions. By integrating empirical data analysis, various modelling techniques, and an interdisciplinary perspective, we are able to unveil the hidden mechanisms driving the process of innovation. Ultimately, this work not only sheds light on the impact of social interactions on the generation and diffusion of novelties, but also contributes to the development of powerful tools and versatile mathematical models, based on different aspects of the concept of the adjacent possible, to investigate how novelties emerge and how the creative process unfolds. This general framework spans over multiple dimensions and mechanisms of innovation, and can be potentially adapted and applied to various specific cases, as exemplified by the application to the realm of music exploration.

1.1 Outline of the thesis

More in details, this thesis is organized in six Chapters as follows.

In **Chapter 2**, we conduct an extensive analysis of existing literature on innovation dynamics. Our objective here is to review existing models that simulate human exploration processes, with a specific focus on identifying the key drivers of innovation that have been proposed thus far. The first set of models examined refers to urn models (UMs), in which the dynamics is based on drawing colored balls from an urn [25]. By recording the sequence of colors drawn, these models allow to study the impact of various mechanisms on the innovation dynamics [8]. Among these, the Urn Model with Triggering (UMT) captures important aspects of real-world discovery processes thanks to its mechanisms of reinforcement and expansion of the adjacent possible [26, 27], as discussed in the introduction above. The second set of models reviewed in this chapter pertains to random walks (RWs), which simulate exploration processes on complex networks [40]. Contrarily to UMs where colors are mixed in the urn from which they

are randomly extracted, RWs can naturally incorporate semantic relationships between the elements of the network being explored. A particular set of RWs we consider here are those with reinforcement, which have been employed to simulate innovation processes, as demonstrated by the edge-reinforced random walk [29]. An equivalent random-walk version of the UMT is also discussed [8]. Furthermore, we review other models of innovation developed in other contexts. Specifically, we examine the Bak–Sneppen model [53], which focuses on the dynamics of evolving ecosystems, where each species is associated with a fitness value representing its competitive advantage. We further consider Thurner's model [20], where innovation is seen as the consequence of a process of creation and destruction, according to Schumpeter's theory in innovation economics [21, 22]. Finally, we present an interesting application of the UMT for the generation and growth of a social network. We also show how network theory can be used in the context of an eye-contact experiment, highlighting, from a neurological perspective, the importance and impact of social interactions in our every-day life.

In **Chapter 3**, we present a new and more general definition of novelty, representing the first major contribution of this thesis discussed in the introduction. Acknowledging that novelties can also stem from new combinations of existing elements, we define a novelty of order n as the first appearance of n consecutive elements in a sequence. To quantify the pace of discovery of novelties at different orders, we define the *n-th order Heaps' law*. Through extensive analysis of real-world sequences, we show that processes that exhibit similar Heaps' exponents at the first order can significantly diverge at higher orders. Since the models described in Chapter 2 do not account for the emergence of different paces of discovery at higher orders, we propose a novel modelling approach, called the *Edge-Reinforced Random Walk with Triggering*. In this model a random walker navigates a growing network of contents, reinforcing the edges traversed, and triggering the addition of new nodes and edges whenever novelties are experienced. This model captures the dynamic nature of innovation, highlighting the complexity of the expansion of the adjacent possible.

In **Chapter 4**, in which we present the second major contribution of this thesis, we shift our attention towards understanding the impact of social connections onto the discovery process. We specifically propose a model, the *UrNet* model, in which many urns, representing different explorers, are coupled through the links of a social network and exploit opportunities coming from their contacts. Individually, the dynamics of an explorer is governed by an UMT, which accounts for the adjacent possible in the content space. In the UrNet, instead we extend each individual urn to a socially-enriched one, adding other colors coming from the urns of their friends, thus expanding the adjacent possible through the social space. We numerically and analytically investigate different social network structures, revealing that the pace of discovery of an agent is influenced by its centrality in the social network. This highlights the dual nature of each individual's adjacent possible, encompassing both their personal exploration space and

the opportunities arising from their social connections.

In **Chapter 5**, where we provide the third major contribution of this thesis, we analyze a unique data set of music exploration to better understand the impact of social interactions on individual and collective exploration with a more data-driven approach. In particular, our data set contains the complete listening history of a large sample of users, as well as their social connections. Firstly, analyzing this data we uncover the heterogeneous nature of user exploration behaviors and the presence of homophily among explorers. We thus introduce an agent-based model, the *ExploNet* model, drawing upon the insights gained in the previous chapters regarding the expansion of the adjacent possible in both the content and social space. In this model, each agent, representing a user on Last.fm, explores the universal network of artists, growing its space of possibilities. At the same time, it receives recommendations from its friends, based on their recent listened songs. Starting from a uniform population of agents, the model reproduce an heterogeneous distribution of pace of discoveries. We find that this heterogeneity depends on stochastic opportunities coming from the adjacent possible in both the content and social space. In particular, the presence of a social dimension creates an assortative arrangement of the explorers at the local level and the emergence of communities with similar music tastes at the global level. These findings contribute to a deeper understanding of the role of social interactions in shaping exploration patterns in online music-streaming platforms. They also provide empirical evidence of the effectiveness and applicability of our modelling framework, highlighting its potential application to other systems.

Finally, in **Chapter 6**, we summarize the main results of the thesis. Further research ideas are also discussed, based on the findings and contributions of this thesis.

Chapter 2

Background

2.1 Introduction and outline

Dynamical processes that emulate the way humans can explore new elements have been the focus of many works in the last decades [25]. In this chapter, we review some of the existing models of innovation dynamics, and analyse their main features. In particular, we focus on understanding the key ingredients that are at the base of a discovery process. These are indeed the fundamental building blocks of a proper model for such processes.

The first set of models we consider is made by the so called urn models, where the general dynamics is based on extractions of colored balls from an urn. In these models, one can record the color of the extracted ball at each time step, creating an ordered sequence of colors. Representing the series of items consumed, technologies used or artworks explored, this sequence can be studied in terms of discovery rates and other innovation footprints. In this thesis, we refer to this sequence as the *sequences of events* $\mathcal{S}$ in the system under study.

The simplicity of these models is ideal to understand how different fundamental mechanisms affect the discovery dynamics simulated. In particular, we check the validity of some empirical laws found across different real-world systems, such as the *Heaps' law* [14, 15] and the *Zipf's law* [54–57]. The Heaps' law describes the way novelties appear in a sequence, stating that the number of different elements $D(N)$ grows as a power-law of the total number of items N in the sequence, i.e., $D(N) \sim N^{\beta}$, with $0 \leq \beta \leq 1$. The Heaps' law is directly related to the Zipf's law [58–60], which instead characterises the frequency distribution of the items in the sequence. Decreasingly ordering the unique elements of the sequence based on their frequency, the Zipf's law states that such frequency f is related to the rank R as a power-law, i.e., $f(R) \sim R^{-\alpha}$, with $\alpha \geq 1$.

Differently from previous urn models, the urn model with triggering (UMT) cap-

tures both the Heaps' and Zipf's laws [8]. This is due to the presence of two key mechanisms in the UMT, namely, the reinforcement of the drawn colors, and the expansion of the adjacent possible (AP) through the triggering of new elements in the urn when a novelty is found. Notice that the UMT focuses on the dynamics of a single entity, for example an individual or the society as a whole. It hence disregards the importance of social interactions between different individuals. We will fill this gap in Chapter 4 by coupling multiple UMTs over a social network.

The second set of models we review in this chapter concerns random walks (RWs) for their ability to simulate exploration processes on complex networks. As a matter of fact, even if semantic relationships between colors can be introduced in the UMT, as done in the urn model with semantic triggering (UMST), urn models fail to properly incorporate relationships between elements coming from empirical data. This problem is overcome by RWs on complex networks, which have been widely used to build different exploration models [40], for example for social annotation [41], music album popularity [42], knowledge acquisition [43], human language complexity [44], and evolution in research interests [45], among others. Interestingly, the key mechanisms of the UMT can also be introduced through RWs, as done for example in the edge-reinforced random walk (ERRW) where the reinforcement mechanism acts on the edges [29]. We will extend the ERRW to include also the expansion of the adjacent possible in Chapter 3, where we introduce the edge-reinforced random walk with triggering (ERRWT), thanks to a broader definition of novelty. Moreover, in Chapter 5, we will consider multiple interacting agents exploring a network of artists, taking inspiration from the key mechanisms of discovery processes analysed in this chapter and the advancements of Chapter 3 and Chapter 4.

Finally, we review other models coming from different fields, from biology and economics to neuroscience and computational social science, which can help us to understand other important factors that are not taken into consideration in the previously discussed models. On the one hand, the analysis of such models reveals the importance of the structure of the space of possibilities in the innovation process. In particular, we highlight how a novelty can be also a new combination of existing (or already explored) elements [20, 21]. This will be the starting point for a new measure of the pace of discovery and a more complex representation of the space of possibilities, discussed in Chapter 3, which relates to the appearance of novelties of higher order, i.e., combinations of more than one element. On the other hand, these works highlight the impact of social interactions in the innovation process. For example, in the context of evolutionary biology, a mutation of a species influences the fitness landscape of other closely connected species [53]. Moreover, thinking about human interactions, we show that even a simple eye contact between two individuals can have a strong impact on the brain activity and choices of the two, even for very simple tasks. In Chapter 4 and Chapter 5, we will hence study the effect of more complex social interactions in

the innovation process, incorporating collaboration and diffusion of discoveries in our modelling framework of innovation.

This chapter is structured as follows. Firstly, in Sec. 2.2 we analyse different urn models in literature. Because of their simplicity, the analysis of the pace of discovery in these models can be done analytically. The order of the models presented in this section is merely chronological. More in details, we introduce Polya's urn model in Sec. 2.2.1, Yule–Simon's model in Sec. 2.2.2, Hoppe's urn model in Sec. 2.2.3, and the urn model with triggering (UMT) in Sec. 2.2.4. Moreover, in Sec. 2.2.5 we go over an extension of the UMT, i.e., the urn model with semantic triggering (UMST), where the notions of semantics and correlated novelties are introduced in the context of urn models.

Secondly, in Sec. 2.3 we consider two families of RWs that include the core mechanisms introduced in the UMT in exploration processes of networks. In particular, in the RW version of the UMST, shown in Sec. 2.3.1, visited nodes are reinforced, while in the ERRW introduced in Sec. 2.3.2 the reinforcement acts on the traversed edges.

Then, in Sec. 2.4 we review other models of innovation coming from other fields of study. In particular, in Sec. 2.4.1 we take into consideration the Bak–Sneppen model, where innovation is seen from a biological point of view as a sequence of subsequent genetic mutations. In Sec. 2.4.2, instead, we analyze innovation from an economic perspective through Thurner's model. Here, innovation comes from the combination of goods creating new ones and making other ones obsolete.

Furthermore, in Sec. 2.5.1 we go over an application of the UMT to model the emergence and evolution of social networks. The importance of social interactions is also highlighted in Sec. 2.5.2, where we analyze the effect of eye contact on the choices and on the brain activity of the participants of a recent scientific experiment.

Finally, we summarize the main findings of this chapter in Sec. 2.6. Specifically, we discuss the impact of the various mechanisms on the innovation properties of discovery processes, highlighting their importance for the models proposed in the following chapters.

2.2 Modelling discovery processes with urn models

2.2.1 Polya's urn model

In 1930, George Polya proposed a model, that is now referred to as *Polya's urn* model (UM) [24]. In its classical version, N_0 balls of different colors are initially placed in an empty urn. Then, at each time step, a ball is drawn uniformly at random from the urn. It is hence put back in the urn along with other ρ balls of the same color. The adopted mechanism leads to a *rich-get-richer* type of dynamics. Indeed, the addition of ρ balls

of the extracted color increases the likelihood to draw this color again in the future. For this reason, ρ has also been referred to as the *reinforcement* parameter.

Notice that the *reinforcement* and the *rich-get-richer* mechanisms are present in many natural [59, 61–64] and human [65–71] processes, as well as in reinforcement learning algorithms [19]. Moreover, the *rich-get-richer* mechanism, reinvented as *preferential attachment* mechanism in the network science community, is a key feature of the growth of most real-world networks [36], as already pointed out in 1965 about scientific citation networks [72]. This has inspired various models of network growth, for example the Barabási-Albert (BA) model for network generation [32]. In this model, such mechanism is present in the form of a preferential attachment rule, where a new node connects to some existing nodes with probability proportional to their degree. This model leads to a family of possible networks which has many statistical properties in common with most real complex networks.

In order to analyse the appearance of novelties in Polya's urn, whenever we extract a ball from the urn we annotate its color on a sequence $\mathcal{S}$ of size N, which stores the whole history of the simulation. Therefore, each extraction increases of a unit the intrinsic time t of the system. We can thus suppose that the length of the sequence is equal to the number of time steps of the model, and use the variable t instead of N to measure various discovery properties of such sequence generated by the model.

Let us hence analyze what is the time evolution of the number of novelties in time in Polya's urn. In particular, we check the validity of the Heaps' law [15], which indicates an asymptotic power-law behaviour $D(t) \sim t^{\beta}$ for the number of different elements $D(t)$ in a sequence of length t, where $\beta \in [0, 1]$ is called the *Heaps' exponent*. For the simplicity of the model, we can study analytically the typical behaviour of $D(t)$. The following procedure, based on a master equation and its continuous approximation, is standard, and can be applied to other urn models. It relies on the estimation of the probability of drawing a new color, which is easily determined in this case. Recalling that the total number of colors inside the urn is fixed and equal to N_0, and that at each time step we add ρ balls in the urn as a reinforcement of the drawn color, the total number of balls in the urn after t extractions is $N_0 + \rho t$. Moreover, since there is only one ball for each color never extracted from the urn, the number of balls of never extracted colors is $N_0 - D(t)$. Therefore, at the next time step $(t + 1)$, the value of $D(t + 1)$ is

$$D(t+1) = \begin{cases} D(t) + 1 & \text{with probability } \mathbb{P}\left(\text{“Extract new color"}\right) = \frac{N_0 - D(t)}{N_0 + \rho t} \\ D(t) & \text{with probability } \mathbb{P}\left(\text{“Extract old color"}\right) = \frac{\rho t + D(t)}{N_0 + \rho t}. \end{cases} \tag{2.1}$$

With abuse of notation, denoting with $D(t)$ the expected value of the number of differ-

ent elements at time t, from Eq. (2.1) the expected value of $D(t+1) - D(t)$ is equal to the probability of extracting a new color, i.e., we obtain the following discrete master equation:

$$D(t+1) - D(t) = \mathbb{P}\left(\text{“Extract new color"}\right) = \frac{N_0 - D(t)}{N_0 + \rho t}. \tag{2.2}$$

For large times $t \gg 1$, we can approximate the finite difference $\frac{D(t+1)-D(t)}{1}$ with the derivative $\frac{dD}{dt}$. We thus obtain the continuous master equation:

$$\frac{dD}{dt} = \frac{N_0 - D(t)}{N_0 + \rho t}. \tag{2.3}$$

With the initial condition $D(0) = 0$ the previous equation has solution

$$D(t) = N_0 \left[1 - \left(1 + \frac{\rho t}{N_0}\right)^{-\frac{1}{\rho}}\right]. \tag{2.4}$$

We hence find out the number of different balls $D(t)$ does not grow as a power-law function, resulting in the absence of the Heaps' law, which is instead a key phenomenon of empirical processes [8]. This result comes with no surprise, since the Heaps' law foresees a sublinear power-law behaviour, which is infinite by definition, while here $D(t) \leq N_0 \ \forall t$. Continuing the comparison with the BA model, the problem is that there is no growth of the total number of colors in the urn, which is instead present in the BA model, where new nodes are added as the network grows.

2.2.2 Yule–Simon model

As we have seen in the last section, a classical Polya's urn has a fixed amount of different colors to choose from, which results in a pace of discovery too slow. Therefore, we need to find other mechanisms to add new colors in the urn. The Yule process [73] is one of the first mechanisms introduced that solve this issue, generating the empirically observed power-laws. Even if it is not exactly an urn model, we include it here because many processes with reinforcements have been extending the Yule model [39]. For example, as we will see in Sec. 2.2.4, this model can be reformulated as a particular case of the urn model with triggering.

In 1925, George Udny Yule [73] designed a probabilistic model generating mutations to explain the evolution and diversity of species over time [56, 74, 75]. The resulting distribution of species by genera is indeed a power law distribution, as it has been found for many datasets [57, 61]. Following the argument of the Nobel laureate Herbert Alexander Simon in 1955 [76], Yule's assumptions are too restrictive and fit only biological contexts. For this reason, we present the model generalised by Simon,

which relies on weaker assumptions.

Simon considered in fact a stochastic process of text generation [76]. Starting with a set of N_0 words, at each time step a random word is selected and written in a sequence $\mathcal{S}$. With probability p the chosen word is a new one, i.e., different from all other already-existing words. Notice that the existing words include the N_0 initial words and all the words in the sequence $\mathcal{S}$. Finally, with probability $1-p$ the word is chosen randomly from the set of existing words. Therefore, the probability to choose a word follows a *"rich-get-richer"* mechanism. In contrast to Polya's urn model, the total number of different words *grows* along with the process. Nevertheless, this model has some limitations in terms of Heaps' law, as we are about to show.

Similarly to what we have done for the Polya's urn in Sec. 2.2.1, we can easily derive the continuous form of the master equation and its solution for large times:

$$\frac{dD}{dt} = p \implies D(t) = D_0 + pt. \tag{2.5}$$

Therefore, the Heaps' law is present in its linear form, in contrast with what has been found in most real cases, where a sublinear scaling is present [8].

Moreover, in order to find the frequency-rank distribution at the base of the Zipf's law, let us call $n_i(t)$ the number of copies of the i-th word at time t, which can be obtained as a solution of the following differential equation:

$$\frac{dn_i}{dt} = (1-p)\frac{n_i}{t+N_0}, \quad n_i(t_i) = 1 \implies n_i(t) = \left(\frac{t+N_0}{t_i}\right)^{1-p}, \tag{2.6}$$

where t_i is the time step in which the word i has been first written in the sequence. Notice the power law functional form of $n_i(t)$, which confirms analytically the "rich-get-richer" mechanism: the sooner a word gets discovered, the more likely it is picked in the future. From Eq. (2.6) we can obtain the frequency-rank distribution:

$$P(n_i \leq n) = P\left(t_i \geq t\, n^{-\frac{1}{1-p}}\right) = 1 - P\left(t_i < t\, n^{-\frac{1}{1-p}}\right) \tag{2.7}$$

$$P\left(t_i < t\, n^{-\frac{1}{1-p}}\right) \simeq \frac{D(t\, n^{-\frac{1}{1-p}})}{D(t)} = \frac{D_0 + p\, t\, n^{-\frac{1}{1-p}}}{D_0 + pt} \sim n^{-\frac{1}{1-p}} \tag{2.8}$$

$$\implies p(n) = \frac{\partial P(n_i < n)}{\partial n} \sim n^{-\gamma} \implies f(R) \sim R^{-\alpha}, \tag{2.9}$$

with $\gamma = 1 - 1/(1-p) = (2-p)/(1-p)$, and $\alpha = 1/(\gamma - 1) = 1 - p$.

To sum up, we have found a linear Heaps' law, because of the constant rate of addition of new words, in contrast with what is displayed in many datasets where there is usually a sublinear growth rate [8]. Moreover, this model does not account for higher values of the Zipf's exponent α in the Zipf's law, since it is upper bounded to 1. In-

stead, a heavy tail with exponent higher than 1 is often found in the frequency-rank distribution. Nevertheless, this is a first remarkable result, considered its simplicity and generality for the time it was first invented.

As a matter of fact, the model of text generation described in this section can be extended to other contexts, substituting words with any other kinds of data, colors, songs, etc. For example, the Barabási-Albert model (BA) of network generation [25, 32] is based on the Yule–Simon model. As we have said in Sec. 2.2.1, BA is characterised by a growth of the network using a preferential attachment rule, which are the same ingredients of the Yule–Simon model. Moreover, as shown in [25], we can map the two models, so that they can be considered equivalent, even though with different interpretations. In the BA model a new node of the graph is introduced at time t by connecting it to m existing nodes chosen with probability proportional to the number of their first neighbours. This effectively corresponds to a deterministic Simon's process with the probability of extracting an old token set to $(1-p) = \frac{1}{2}$. It is deterministic because for a correct mapping at every even time step there is always a rich-gets-richer in action, while in the classical Simon's model the entrance of an old token in the stream at any time is conditioned to the extraction of a random number. The analogy between the two models is complete after identifying the number of occurrences of the tokens in Simon's stream with the connectivity of the nodes in the graph generated with the BA model.

2.2.3 Hoppe's urn model

Fred M. Hoppe introduced for the first time in 1984 [77] a mechanism to extend the number of possible novelties in the framework of urn models. He took inspiration from Ewens' sampling formula [78], which described the allelic partition of a random sample of n genes from an infinite population at equilibrium. Such sampling evolves according to a discrete time neutral Wright-Fisher (WF) process [79, 80] with constant mutation rate μ per gene. The WF model describes a population with discrete, non-overlapping generations. In each generation the entire population is replaced by the offspring from the previous generation. Parents are chosen via random sampling with replacement. The allelic distribution in the offspring is obtained randomly from the parents, with a fixed mutation rate μ for each gene. Hoppe applied the idea behind such stochastic mutation mechanism to Polya's urn model.

In Hoppe's urn there are two different types of balls: normal colored balls of weight 1, as in Polya's urn, and one special black ball of weight θ. The dynamical process taking place in this model can be considered an upgrade of that in Polya's urn model. The urn is initialized with a black ball, and at each time step a ball is extracted at random, with probabilities proportional to their weight. When a black ball is drawn, a ball of a brand new color is added to the urn; when a colored ball is drawn, another ball

of the same color is placed back in the urn as a reinforcement. In such dynamics, the extraction of the black ball represents the event of a mutation.

As we did for the previous models, we can derive analytically the number of different colors $D(t)$ after t extractions. The probability to draw a black ball at time t, and hence to add a new color in the urn, is $\frac{\theta}{\theta+t}$. The master equation is then

$$D(t+1) = D(t) + \frac{\theta}{\theta + t}. \tag{2.10}$$

This recurrence equation is continuously approximated by

$$\frac{dD}{dt} = \frac{\theta}{\theta + t}, \quad D(0) = 0, \tag{2.11}$$

with solution

$$D(t) = \theta \ln\left(1 + \frac{t}{\theta}\right). \tag{2.12}$$

Analogously, we can derive an analytic approximation for the Zipf's law, which is

$$f(R) = \frac{t}{\theta} \exp\left(-\frac{R-1}{\theta}\right). \tag{2.13}$$

As we can see, the model predicts a novelty rate of appearance too much slower than what is found in many real systems. Heuristically, this is because the probability of extracting the black ball decays too fast, since its weight is fixed whilst the urn grows linearly. As a consequence, the number $D(t)$ of different colors in the urn grows only logarithmically with the number of extractions t. This is due to the lack of a consistent growth process of the balls responsible for the innovation of the system.

2.2.4 Urn model with triggering

Model definition

In 2014 Francesca Tria, Vito D. P. Servedio, Steven Strogatz and Vittorio Loreto [8] generalised the Polya's urn model so that one novelty can trigger further ones, from which the name *"Urn model with triggering"* (UMT). In fact, the triggering of novel colors in this model is essential to overcome the growth limitations seen in the classical Polya's urn.

The idea behind the UMT is to maintain the valuable "rich-get-richer" mechanism of the precedent models, while consistently introducing new colors in the urn that are related to each other. To this end, they generalised the mathematical framework of Polya's urn model, adding what is called the *space of possibilities*. To explain this new concept, let us imagine that the colors inside an urn are nodes of a network, that is the urn is embedded in an ideal space with semantic relations between the objects. These

can be for example songs, which can be linked if they share the same artist, or any other real network we can think of. Notice that in Polya's urn model the number of colors is fixed, and therefore with this embedding we obtain a space where the number of elements is fixed. In Polya's urn model, exploring this space means jumping from an element to another, each time reinforcing the weight of the next element, making it easier to move to that element again in the future.

In the UMT, we go beyond the exploration of a limited space of possibilities, with a triggering mechanism responsible for adding new adjacent possible elements into the space. We define the *adjacent possible* as the set of elements in the space of possibilities that are only one step away from what has been explored so far [13], and so are considered within reach from the current state of knowledge. Then, each time an element is explored for the first time, such element is discovered and reveals all its neighbouring elements, or its adjacent possible, expanding the space of possibilities. Let us suppose that we have explored some part of the space of possibilities we have introduced before and triggered other adjacent possibilities yet to be explored. Therefore, in the next step we can either move to an element that has already explored before, or to an adjacent possible element of some previously explored one. If we decide to move to a new element, we continue triggering new elements and the space of possibilities grows more and more as we explore it.

In details, the UMT can be formulated as follows [8]:

(i) an urn is initialized with N_0 balls of different colors and same weight;

(ii) at each time step a ball is randomly extracted from the urn and its color analysed;

(iii) the drawn ball is put back in the urn with other ρ new copies of the same color (reinforcement mechanism);

(iv) moreover, if the drawn color has never been drawn in the previous extractions, $\nu + 1$ balls of new distinct colors are introduced inside the urn (triggering mechanism).

As we have done for the previous models, one can extract the sequence of extractions and make the usual analysis for the Heaps' and Zipf's laws [8]. The numerical simulations of the model show a power law distribution for both the Heaps' and Zipf's laws. In particular, the Heaps' exponent is found to be either linear or sublinear depending on the set of parameters. Moreover, the Zipf's exponent seems to be the anti-reciprocal of the Heaps' exponent, which is often seen in the data sets and theoretically reasonable under particular hypotheses [58, 81]. All this is shown in Figure 2.1 where the Heaps' and Zipf's laws for some realisations of the model are compared to those for some datasets studied in Ref. [8]. Further notice that the UMT is capable to reproduce another empirical finding, which is the Taylor's law [26, 27, 82, 83]. This

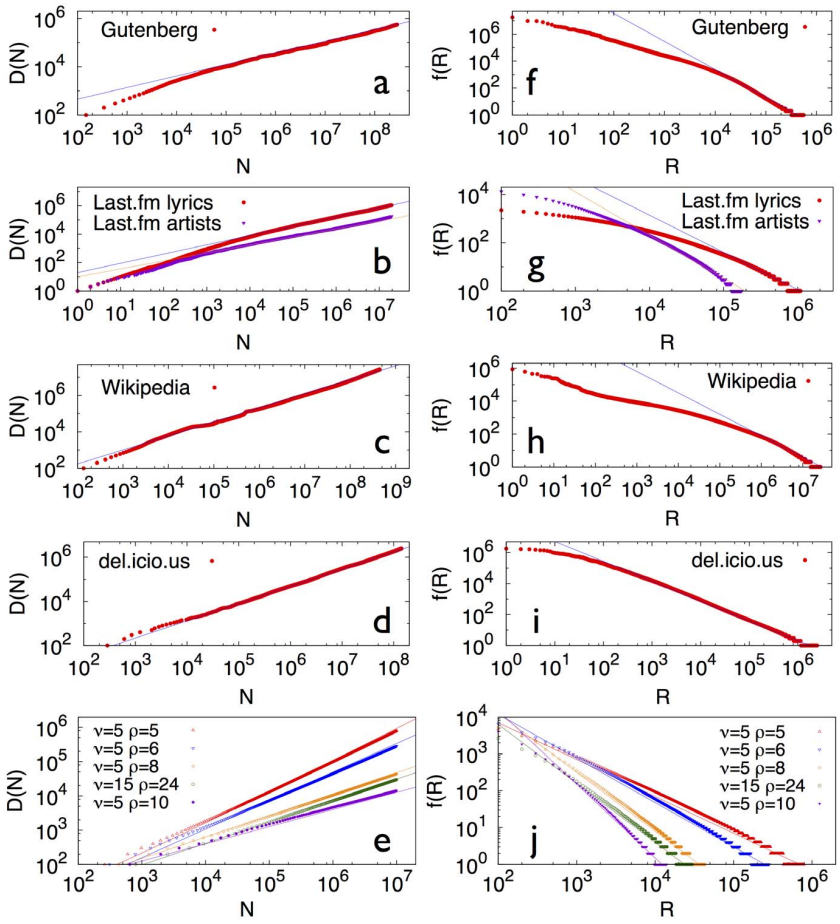


Figure 2.1: **Heaps' law and Zipf's law in real datasets and in the urn model with triggering.** Image courtesy of Ref. [8]. Heaps' law (a–e) and Zipf's law (f–l) in real datasets (a–d) and (f–i) and in the urn model with triggering (e,j). Gutenberg (a), (f), Last.fm (b), (g), Wikipedia (c), (h), del.icio.us (d), (i) datasets, and the urn model with triggering (e), (j). Straight lines in the Heaps' law plots show functions of the form $f(x) = ax^{\beta}$, with the exponent β equal respectively to $\beta = 0.45$ (Gutenberg), $\beta = 0.68$ (Last.fm lyrics), $\beta = 0.56$ (Last.fm artist), $\beta = 0.77$ (Wikipedia) and $\beta = 0.78$ (del.icio.us), and to the ratio ν/ρ in the urn model with triggering, showing that the exponents for the Heaps' law of the model predicted by the analytic results are confirmed in the simulations. Straight lines in the Zipf's law plots show functions of the form $f(x) = ax^{-\alpha}$, where the exponent α is equal to β^{-1} for the different β's considered above. Note that the frequency-rank plots in real data deviate from a pure power-law behavior and the correspondence between the β and α exponents is valid only asymptotically (for more details see [8]).

law quantifies the scaling properties of the fluctuations of the number of novelties in a discovery process, stating that the standard deviation of $D(t)$ grows as a power-law function of its mean. We refer the interested reader to Refs. [26, 27] for the analysis of Taylor's law in the UMT.

Thanks to these properties, the UMT has been used to study in various contexts, for example, to explain the rise and fall of popularity in technological and artistic productions [28], the emergence and evolution of social networks [30], and the evolution of the cryptocurrency ecosystem [31].

Analytical solution

Let us define $U(t)$ the total number of balls in the urn up to time t, and $U'(t)$ the number of unique elements in the urn at time t. Being $D(t)$ the number of different element extracted from the urn up to time t, we can write the following equation for the Heaps' law:

$$\frac{dD(t)}{dt} = \frac{U'(t) - D(t)}{U(t)}. \tag{2.14}$$

Eq. (2.14) can be rewritten as a function of the parameters of the model. In particular, we can write the total number of balls $U(t)$ as the initial number of balls N_0, plus the ρt balls added as reinforcement at each time step, plus the $(\nu+1)D(t)$ balls added due to the triggering mechanism:

$$U(t) = N_0 + \rho t + (\nu + 1)D(t). \tag{2.15}$$

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:

$$U'(t) - D(t) = [U(t) - \rho t] - D(t) = N_0 + \nu D(t). \tag{2.16}$$

Thus, using Eq. (2.15) and Eq. (2.16) in Eq. (2.14) we obtain:

$$\frac{dD(t)}{dt} = \frac{N_0 + \nu D(t)}{N_0 + \rho t + (\nu + 1)D(t)}. \tag{2.17}$$

From now onwards we suppose that $t \gg N_0$, so that we can disregard N_0 in Eq. (2.17) and in the similar equations in the following sections. Therefore, after the introduction of the auxiliary variable $z(t) = \frac{D(t)}{t}$, Eq. (2.17) can be rewritten as:

$$\frac{dz(t)}{dt}t + z(t) = \frac{\nu z(t)t}{\rho t + (\nu + 1)z(t)t}, \tag{2.18}$$

which can be integrated as:

$$\int_{z(t_0)}^{z(t)} \frac{\rho + (\nu+1)z(t)}{z(t)[\nu - (\nu+1)z(t) - \rho]} dz(t) = \int_{t_0}^{t} \frac{1}{t} dt. \tag{2.19}$$

The asymptotic solution ($t \to \infty$) depends on the parameters ρ and ν. Starting from Eq. (2.19), it can be shown, as in the Supplemental Material of Ref. [8, 26], that the asymptotic solution for $D(t)$ is

$$\begin{cases} \rho > \nu & D(t) \underset{t\to\infty}{\approx} (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}, \\ \rho = \nu & D(t) \underset{t\to\infty}{\approx} \frac{\nu}{\nu+1} \frac{t}{\ln t}, \\ \rho < \nu & D(t) \underset{t\to\infty}{\approx} \frac{\nu-\rho}{\nu+1} t. \end{cases} \tag{2.20}$$

Notice how we obtain precisely the Heaps' law with sublinear growth for $\rho > \nu$, as seen in empirical processes [8, 15].

It can also be shown analytically that the asymptotic power-law behavior of the frequency-rank distribution typical Zipf's law is valid for the UMT. In fact, we can set up the differential equation for the number of elements n_i of a specific color i, that reads:

$$\frac{dn_i}{dt} = \frac{n_i \rho + 1}{N_0 + (\nu+1)D + \rho t}. \tag{2.21}$$

The solution of this equation is given by

$$\begin{cases} \rho > \nu & n_i \underset{t\to\infty}{\approx} \dfrac{t}{t_i} \\ \rho \leq \nu & n_i \underset{t\to\infty}{\approx} \left(\dfrac{t}{t_i}\right)^{\frac{\rho}{\nu}} \end{cases} \tag{2.22}$$

where t_i is the time at which the element i occurred for the first time in the sequence. From this, one can obtain the Zipf's law [8], which in this case is

$$f(R) \underset{t\to\infty}{\approx} R^{-\frac{\rho}{\nu}}, \tag{2.23}$$

Comparison with Yule–Simon model

Finally, let us notice that the Yule–Simon model can be obtained from the UMT in the case $\rho < \nu$, where in fact for the UMT we have a linear growth for the Heaps' law, like in the Yule–Simon model. In particular, we have $p = \frac{\nu-\rho}{\nu+1}$. However, the dynamics taking place in the two models are very different in nature. In the Yule–Simon model, in fact, the linear growth comes from a constant rate of innovation p, while in the UMT, for $\rho < \nu$ the linear growth is due to the presence in the urn of too many colors to be discovered, which keeps triggering new ones. Finally, in both cases we have a similar

Zipf's law. In fact, for the UMT the Zipf's exponent is ρ/ν, while for the Yule–Simon model it is $1 - p = 1 - \frac{\nu-\rho}{\nu+1} = \frac{\rho+1}{\nu+1}$, having used the correspondence between the two models given by the Heaps' laws.

2.2.5 Urn model with semantic triggering

In the datasets studied in Ref. [8] and first analyzed in Fig. 2.1 we can also find significant semantic correlations between the appearance of elements and novelties in the sequence. Semantic correlations in a sequence of elements explored refer to the relationships or associations between such elements, based on their meaning or semantics. For example, it has been seen how a novelty may trigger another one, thus correlating an earlier novelty with a later one. For example, discovering an interesting song may conduct to search for other music by the same artist. In the context of language or text, semantic correlations capture the semantic connections between words, phrases, or sentences that share related or similar meanings. When analyzing a sequence of words or text, identifying semantic correlations can provide insights into the underlying structure, context, or thematic content of the sequence. These correlations can then be used to extract meaningful patterns, infer relationships between elements, or enable various natural language processing tasks.

For a generic sequence of events made by words read, technologies used, songs listened, etc., one can observe semantic correlations by studying the semantic distribution of the elements inside the sequence, for example measuring the Shannon entropy of the events associated to the same semantic group, or label. Roughly speaking, this entropy measures the extent of clustering among the events associated to the given semantic group, with a larger clustering denoting stronger correlations among their occurrences, and thus a stronger triggering effect of the adjacent possible. In order to compute the Shannon entropy of a semantic group, we first need to introduce the notion of semantics, identifying a semantic label to each element of the system under study. For example, in a dataset of music listening records, where the sequence is made of songs, the assigned label can be the corresponding artist. In the UMT, we can also identify the labels using the triggering events. We divide the initial N_0 balls of the urn into $N_0/(\nu + 1)$ groups, each ball in the same group sharing the same label. Then, whenever a ball is drawn, the balls of the same color added to the urn get the same label of the drawn ball. If the ball is new, a triggering event takes place, and all the new $\nu + 1$ balls get the same new label, since they all share the same "mother" color. Such semantics will be used to create a more refined version of the model, called the *Urn Model with Semantic Triggering* (UMST), as defined below.

Let us first show how we calculate the Shannon entropy of a the distribution of the elements of a certain label for a given sequence with semantics. Let us consider a label A appeared in the considered sequence and let k be the number of occurrences

of label A in the sequence. Let us consider the subsequence beginning from the first appearance of A to the end of the sequence and let us divide it in k equal intervals and let f_i be the number of occurrences of label A in the i-th interval. Then, the Shannon entropy of label A is given by

$$S_A(k) = -\sum_{i=1}^{k} \frac{f_i}{k} \log \frac{f_i}{k}. \tag{2.24}$$

Let us remark that even if there is a minus sign, this number is always non-negative. Let us show two clarifying examples. If $f_i = 1$ for all $i = 1, \ldots, k$, that is the occurrences of A are equally distributed among the intervals, then $S_A(k)$ would get its maximum value $\log k$. On the contrary, if all the occurrences are in the same interval, for example the first one, we have $f_1 = k$, $f_i = 0$ for all $i = 2, \ldots, k$, and $S_A(k) = 0$. This means that low values of this entropy are related to a higher extent of clustering of the elements sharing the label A in the sequence. However, for a more significant evaluation of this measure, we need to compare the obtained entropy value with a random case, where we disrupt the semantic correlations of elements in the sequence. To do so, we globally reshuffle the sequence of labels and recalculate the entropy. Another null case can be obtained with a local reshuffling, as explained in [8]. In this case, for each label A, we consider the subsequence of the original sequence starting from the first appearance of A. Then we reshuffle the subsequence obtained, keeping fixed just the first element A. Then we calculate the entropy of A with Eq. 2.24.

From the simulations of the UMT, we observe no difference between the entropy distribution of the simulated sequence and the distribution of the reshuffled one. This means that in the UMT the extracted labels are almost random and show no correlations among themselves and no clustering, which is instead observed in the datasets. To see significant changes, the authors have introduced a new part to the model, giving rise to the urn model with semantic triggering (UMST). Let us hence introduce a new parameter $\eta \in]0,\, 1]$ responsible for the rise of semantic correlations. Let us save the color of the last drawn ball and let us suppose it has color x and label Y. Let us suppose that the first ball of color x has been introduced in the urn when extracting the color a (i.e., a is "mother" of x), and that the new colors introduced when picking x for the first time (the "sons" of x) have label Z. Then, at the next time step, we make an extraction from the urn after having changed the weights of the ball according to the following scheme:

- the balls of color a, the balls of label Y, and those of label Z get weight 1. Let us note that these balls are all semantically correlated with the ball extracted in the previous time step. In fact, they are in order the "mother" (the color generating x), the "brothers" (the colors generated with x), and the "sons" (the colors generated by x).

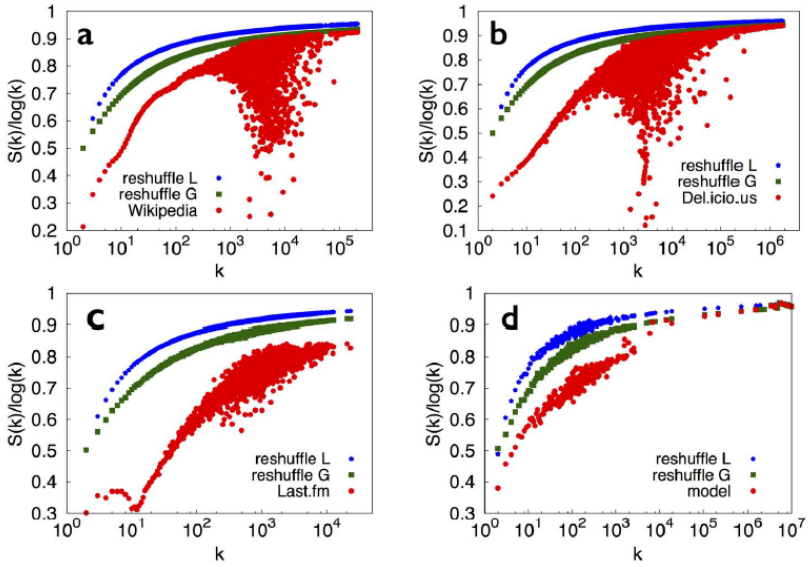


Figure 2.2: **Entropy distribution comparison between datasets and the urn model with semantic triggering.** Image courtesy of Ref. [8]. Normalised entropy of a sequence associated to a specific label A vs. the number of events, k, with that label. The entropy is averaged for each k over the labels with the same number of occurrences. The datasets used are (a) Wikipedia, (b) Del.icio.us, (c) Last.fm. For more insight on those datasets see [8]. In (d) the plot for the model is an average over 10 realizations of the process, with parameters $\rho = 8$, $\nu = 10$, $\eta = 0.3$, and $N_0 = \nu + 1$.

- All the other balls in the urn get weight $\eta \leq 1$.

With low values of η one can obtain an entropy distribution (as measured in Equation 2.24) significantly lower than that of the random sequence, and therefore more similar to the real ones, as shown in Figure 2.2, where the entropy distribution of a realisation of the UMST is shown in with comparison with other datasets.

Finally, let us note that for the UMST the Heaps' and Zipf's exponents are different from the standard case with $\eta = 1$ of the previous section. For example, the authors have found the following bounds for the Heaps' exponent β:

$$\min\left(\frac{\nu\eta}{\rho},\, 1\right) \leq \beta \leq \min\left(\frac{\nu}{\rho},\, 1\right). \tag{2.25}$$

Even though the Heaps' law is valid, the UMST is not able to properly model the semantic associations between the elements explored, as we are going to show in Chapter 3.

2.3 Modelling discovery processes as random walks on networks

2.3.1 Random walk model with semantic triggering

In the previous Sec. 2.2 we have reviewed and analyzed the properties of some discovery models based on extractions from an urn. In particular, we have seen how the urn model with triggering (UMT) can reproduce the empirical footprints of innovation and exploration processes, such the Heaps' law, the Zipf's law, and the semantic correlations [8]. Authors of Ref. [8] further introduce an equivalent version of UMT and UMST in the form of a random walker (RW). This version builds upon the idea of the space of possibilities as a network to be explored, where the nodes in the network represent the colors in the urn and the weight of a node indicates number of balls of the corresponding color. Similarly to the UMST, links between the nodes are made depending on their semantic information (their label), based on the concept of the adjacent possible.

Let us start with a graph $\mathcal{G}^0$ made of N_0 nodes of same weight $w_i = 1$, divided into $N_0/(\nu + 1)$ cliques, each node in the same clique sharing a common label and connected to all other nodes in the clique. We then draw a link between each pair of nodes belonging to different cliques with probability $\eta \leq 1$. Let us also define the graph $\mathcal{G}^t$ as the instance of the graph after t time steps. Notice that we can encode the topological structure of the graph into its adjacency matrix $\boldsymbol{A}^t = \{a_{ij}^t\}$. If two nodes i and j are connected, then the corresponding value in the adjacency matrix is $a_{ij}^t = 1$, otherwise it is $a_{ij}^t = 0$. Moreover, to replicate the dynamics of the UMST, we assume that the RW can remain in the same node, i.e., each node is connected to itself, or, mathematically, $a_{ii}^t = 1$ for all nodes i in $\mathcal{G}^t$.

After initializing the network, the exploration process starts from a random node of the network according to the following steps.

1. At each (discrete) time step t, the RW moves to a neighbour node or stays in the previous node with a weight-dependent probability. In other words, assuming that the RW is in node i in the previous time step, the probability that the RW moves from node i to node j in the network is

$$\pi_{ij}^t = \frac{a_{ij}^t w_j^t}{\sum_{l \in \mathcal{G}^t} a_{il}^t w_l^t}. \tag{2.26}$$

   Notice that here j can be the same as i.

2. Assuming that the RW moves to node j, the selected node weight is reinforced by ρ, i.e., $w_j^{t+1} = w_j^t + \rho$.

3. If the visited node j is new, i.e., it is visited for the first time by the RW, then the adjacent possible expands, with more nodes entering in the network. In particular, a fully-connected clique of $\nu+1$ new nodes with unitary weight and common new label are added in $\mathcal{G}_{t+1}$. These new nodes are connected to the visited node j, the *mother* node, who triggered their appearance in the network. Moreover, for each node in the clique and each other node in the graph, we draw a link between them with probability $\eta \leq 1$.

If $\eta = 1$, this model perfectly maps the UMT. In fact, for each new node, all possible links are drawn, making the graph $\mathcal{G}^t$ always fully connected. Therefore, at each time step the RW can move to any node in the graph, reinforcing its weight and adding $\nu+1$ new nodes if the visited node is new, like in the corresponding urn model.

On the contrary, when $\eta < 1$ the correspondence with the UMST is not one-to-one. In fact, in the case of the graph the connections between two nodes are fixed (or *quenched*) and the graph is not fully connected. This means that in the RW version of the UMST there are some nodes that are not reachable with one step only. Instead, in the urn version, one can go from one color to any other one, with the parameter η, responsible for the semantic triggering, affecting the weight of the other colors dynamically at each step. Despite this difference, the statistical properties of the two models turn out to be equivalent from a qualitative point of view, also in the case $\eta < 1$ [8]. In Fig. 2.3, we report some examples of the Heaps' and Zipf's laws obtained simulating this RW model with triggering. Here you can see how the change of the parameter η influences the appearance of novelties. A higher value of η indicates a higher number of nodes accessible from any node, which translates into a higher Heaps' exponent, i.e., a higher rate of discovery.

Finally, notice that the initial topology of the network can be naturally extended to deal with a more realistic network. The semantic relations are in fact encoded in the growing graph topology, and one can imagine different ways of linking the new nodes, corresponding to more complex and realistic semantic structures, as we will do in Chapter 3 and Chapter 5.

2.3.2 Edge-reinforced random walk model

In Sec. 2.3.1 we have seen a first example of random walk (RW) with reinforcement to model the exploration of a space of possibilities typical of innovation processes. A lot of attention has indeed been given to the class of RWs with reinforcement [39, 84, 85], which have been successfully applied to biology [86] and mobility [87, 88], to name a couple of examples. In the adaptation of the UMST to a RW, the reinforcement is introduced at the level of the node. In other words, the probability to move to a neighboring node depends on the weight of the adjacent nodes, which is reinforced whenever explored. However, in literature there are other examples of RWs, where the

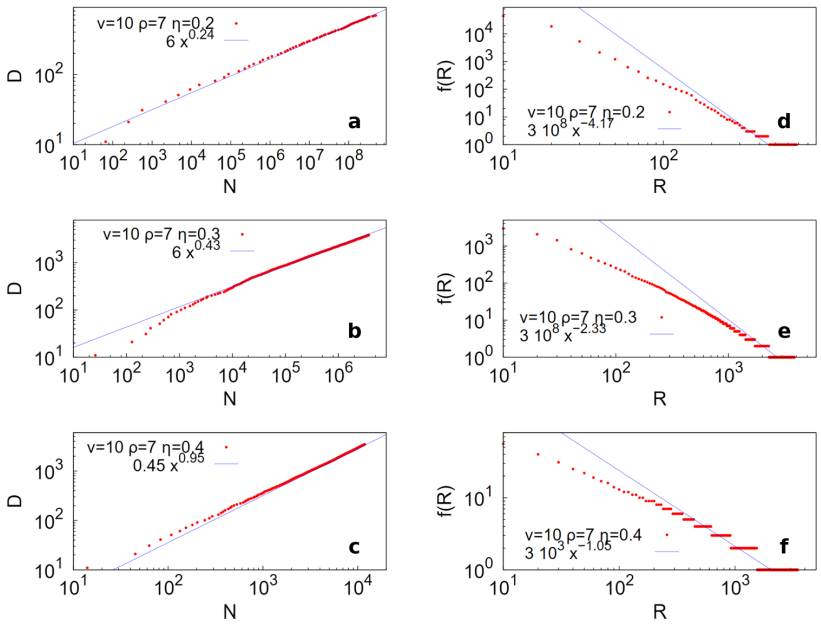


Figure 2.3: **Heaps' law and Zipf's law for the RW model with semantic triggering.** Image courtesy of Ref. [8]. Heaps' law (a–c) and Zipf's law (d-f) in simulations of the Random walk model with semantic triggering with parameters $\rho = 7$, $\nu = 10$, and $\eta = 0.2$ (a,d), 0.3 (b,e), and 0.4 (c,f). Straight lines in the Heaps' law plots (a-c) show functions of the form $f(x) = ax^{\beta}$, with the exponent β representing the Heaps' exponent. Straight lines in the Zipf's law plots (d-f) show functions of the form $f(x) = ax^{-\alpha}$, where the exponent α is approximately equal to β^{-1} for the different β's considered above. Notice how the Heaps' exponents β increase with increasing values of η, while the value of α decrease. Further note that the frequency-rank plots in real data deviate from a pure power-law behavior and the correspondence between the β and α exponents is valid only asymptotically (for more details see [8]).

reinforcements acts at the level of edges instead of nodes. Specifically, the concept of edge reinforcement [89, 90] was introduced in the mathematical literature by Coppersmith and Diaconis in 1987 [91]. Interestingly for the scope of this thesis, in 2018, Iacopo Iacopini, Staša Milojević and Vito Latora have developed an edge-reinforced random walk (ERRW) model that has all the ingredients of an innovation model [29]. For instance, it can reproduce the emergence of novelties and the pace of discovery as captured by the Heaps' law [15].

Similarly to the RW version of the UMST, the ERRW moves from a node to a neighboring one on an underlying network of relations among concepts, technological advancements, or other items explored in the discovery process. However, differently from the UMST networked model, whenever the ERRW traverses a link, the link itself is reinforced instead of the chosen following node. More in details, the ERRW model is initialized over a weighted connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, with a fixed number of nodes

$N = |\mathcal{V}|$ and links $M = |\mathcal{E}|$. In the context of the exploration of concepts, each node of the graph represents a concept, and the presence of a link (i, j) denotes the existence of a direct relation between the two concepts i and j. In this model, the structure of the graph is assumed to be fixed, while the weights of the edges can change in time according to the dynamics of the RW. The evolution of the weights of the graph is fully described by the non-negative time-dependent adjacency matrix $\boldsymbol{W}^t = \left\{w_{ij}^t\right\}$. At the beginning of the simulation, all existing links in $\mathcal{E}$ are initialized with unitary weight, i.e., $w_{ij}^0 = 1\ \forall (i, j) \in \mathcal{E}$. Then, the dynamics of the walker is defined as follows.

1. At each time step t, assuming that the walker is positioned on node i, it moves from node i to a neighboring node j with probability proportional to the weight of the connecting edge. Formally, such probability is

$$\pi_{ij}^t = \frac{w_{ij}^t}{\sum_{l \in \mathcal{V}} w_{il}^t}. \tag{2.27}$$

   Notice that here we assume that $\mathcal{G}$ has no self-loops, so that the walker changes position at each time step.

2. Whenever the walker moves from node i to node j, the corresponding edge (i, j) is reinforced by a quantity ρ, i.e., $w_{ij}^{t+1} = w_{ij}^t + \rho$.

Similarly to the reinforcement mechanism of Polya's urns, the edge-reinforcement mechanism mimics the fact that the relation between two concepts is reinforced every time the two concepts are associated in the cognitive process. 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. Continuing the comparison between the ERRW model and the UMST networked model, notice how in the ERRW model there is no triggering of new nodes or links entering the network. Instead, the concept of the adjacent possible is naturally included through the underlying network, without the need of a triggering mechanism and further parameters. A graphical representation of the ERRW model is shown in Fig. 2.4.

In Ref. [29], large small-world (SW) synthetic networks and extensive real-world networks have been used. In particular, the SW networks have been obtained using the Watts-Strogatz model [92, 93], while the empirical network representing the relations among the keywords of peer-reviewed articles. In particular, the scientific articles were taken from core journals of four fields between 1991 and 2010, using the Web of Science database [94]. For each field, a real temporal sequence $\mathcal{S}$ has been created using the relevant concepts extracted from a text analysis of each abstract. Analyzing such sequences, the Heaps' law has been empirically observed. For example, the number of different concepts in astronomy grows as a power-law function of the size of the sequence, with an Heaps' exponent $\beta = 0.82$. Moreover, these concepts have then

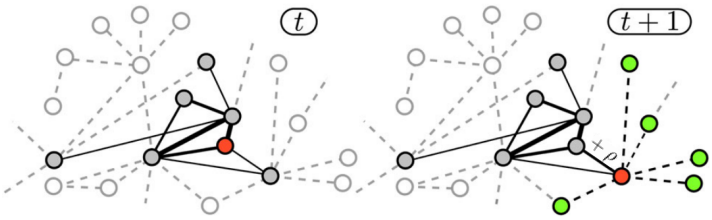


Figure 2.4: **The Edge-Reinforced Random Walk (ERRW)** Table courtesy of Ref. [29]. The ERRW produces a coevolution of the network with the dynamics of the walker. At time t the walker is on the red node and has already visited the gray nodes, while the shaded nodes are still unexplored. The widths of edges are proportional to their weights. At time $t + 1$. the walker has moved to a neighbor (red) with probability as in Eq. (2.27), and the weight of the used edge has been reinforced by ρ. At this point, the walker will preferentially go back, although it can also access the set of "adjacent possible" (green).

Research field	Papers	N	$\langle k \rangle$	C	L	β	ρ
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
Economy	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 2.1: Table courtesy of Ref. [29] Statistics of the network of concepts in four research fields from Web of Science, together with the empirical Heaps exponent β and the value of ρ that reproduces it. N is the number of concepts, $\langle k \rangle$ is the average degree, C is the clustering coefficient, and L is the characteristic path length.

been used to create the underlying networks of relations among concepts from their co-occurrence in the abstracts.

The ERRW model manages to reproduce the Heaps' law, as well as the semantic correlations among novelties, either exploring the synthetic networks or the network of concepts from Web of Science. Interestingly, since the adjacent possible is naturally accounted for inside the underlying structure, we can fit the value of the reinforcement parameter ρ to obtain information on how strong the reinforcement mechanism has to be in the discovery dynamics to reproduce the same pace of discovery in each field. The statistics of the empirical networks, together with the related Heaps' exponents and the fitted reinforcement parameters, are shown in Table 2.1.

Finally, let us point out that there is only one tunable parameter in the ERRW, the edge-reinforcement parameter. All the other parameters and mechanisms in the UMST are instead accounted for in the underlying network. For example, such network encodes the space of possibilities and takes the role of the semantic triggering mechanisms controlled by the parameters ν and η in the UMST. Furthermore, the pres-

ence of semantic correlations of the elements explored in the ERRW is ensured by the chosen dynamics, since the RW moves from a node to an adjacent one through a link the network, where the link itself represents a semantic relation. Nevertheless, we note that this comes at a cost: the network needs to be sufficiently large compared to the length of the random walk. In fact, if the walker moves indefinitely, eventually it might explore the whole network, since there is no growth mechanism in place.

2.4 Other models

The emergence of novelties has been empirically observed in a wide range of fields, such as science [95], gastronomy [96], goods and product [69], network science [97], language [25, 65], information [98] and cinema [99]. For such reason, in literature we can find other models of innovation dynamics inspired by these fields. In this section we review two of these, with the scope of understanding how this dynamics can be modelled in other frameworks. The first one in Sec. 2.4.1 is the Bak and Snippen model, an evolutionary model with cascading mutations, while the second is the Thurner model, introduced in Sec. 2.4.2, where innovation is seen as a process of creation and destruction.

2.4.1 Bak–Sneppen model

The Bak–Sneppen model, proposed by Per Bak and Kim Sneppen in 1993 [53], offers an interesting perspective on how the innovation process unfolds in a biological context. This model provides a framework that explores the connection between evolutionary dynamics and the emergence of innovation within biological systems. By simulating co-evolutionary processes of multiple species, the Bak–Sneppen model sheds light on the role of disruptive events and the subsequent adaptive responses in driving species evolution and innovation through the adjacent possible.

At the core of the Bak–Sneppen model is the concept of *punctuated equilibrium*, which suggests that evolutionary changes occur in bursts rather than through gradual, continuous processes. In the model, N species are represented as agents, arranged in a one-dimensional line. Each species is associated to a fitness parameter, called “barrier" and denoted with B_i, randomly chosen from an uniform distribution between 0 and 1. Such fitness represents their ability to survive and reproduce. In a biological context, the barrier height is a measure of the amount of genetic code that has to be changed to have significant mutations in the species, such as developing wings to allow a creature to fly. The model introduces occasional disruptive events known as “extinctions", which can start drastic cascading changes in the environment. In particular, at each time-step, the species with the lowest barrier goes extinct, and is replaced by a new species with a randomly assigned barrier value. This change also affects the barriers of

their right and left neighbors: a new barrier is hence randomly selected from the same uniform distribution, and assigned to them. Such cascading mechanism is a fundamental one in biology, representing in a simplistic way how species interact with each other. For instance, the interaction could represent the fact the two species are consecutive links of a food chain [53]. Therefore, a change in a species might eventually affect the probability that a related species mutates in the future as well.

In the first steps of the evolutionary process simulated with the Bak–Sneppen model, mutations are uncorrelated in the space (the one dimensional line). However, when the general barrier level increases, then the exploration of the *adjacent possible* is triggered. In fact, neighbors of mutating species are very likely to mutate, in a cascading effect similar to the triggering mechanism in the UMT. It has been shown that the frequency distribution $C(x)$ of subsequent mutations having distance x follows a power-law, i.e., $x^{-\alpha}$, where $\alpha \simeq 3.15$ [53]. Interestingly, this result is achieved whatever initial conditions are set. In other words, the system is self-organized. Moreover, the distribution of barrier values B at the critical state, when $C(x)$ becomes stationary, unveils how mutations happen at a much faster pace for species with barrier below a critical value $B^* = 0.67$. Furthermore, the unitary fitness state in which each species has fitness 1 and no more mutations take place is never achieved. Instead, if we remove the interaction between species—and thus the cascading effect—, the unitary state is reached very slowly, due to the need for coordinated mutations.

The Bak–Sneppen model demonstrates that innovation, represented by the emergence of new species with different fitness values, is crucial in driving species evolution. The new species might indeed bring forth new traits and characteristics that could have a higher fitness value. These disruptive events trigger environmental changes that challenge the status quo in neighboring species, leading to a cascade of adaptive responses, creating opportunities for new variations to arise. Such mechanism opens avenues for exploring similar dynamics in other contexts, such as social or technological innovation. In fact, even though the Bak–Sneppen model is a simplification of the complex biological evolutionary process, it provides a valuable starting point to investigate the relationship between innovation and evolutionary dynamics. In particular, drawing upon the parallelism between biological evolution and the dynamics of adaptive systems, the model showcases how adaptive changes triggered by innovative events play a pivotal role in driving evolutionary advancements.

2.4.2 Thurner's model

Another interesting perspective on innovation comes from economy, where innovation has been seen as the consequence of a process of creation and destruction. This is typical of the Schumpeterian approach [21, 22], whose theory is still today considered one of the most relevant contributions in innovation economics. According to Joseph

Schumpeter, the appearance of new products and services is the natural consequence of market behaviors, in which each participant maximizes its utility function, allowing technology progress and the rise of innovations. In particular he proposed the concept of *creative destruction*, i.e., the creation of new products and services, born from the combination of already existing goods, which drives obsolete goods out of the market.

This idea is fully represented in the model by Thurner et al. [20], characterized by a generative process of novelties based on deterministic coupling of elements. In Thurner's model, creation and destruction are considered as association of existing elements, respectively determined by production and destruction tables. In their work, they have studied various time series related to innovation economics, for example the frequency distributions of percentage increase of GDP (gross domestic product) of several nations, or of business failures, or of patents issued. They have found that such time series assume the form of power-laws, and have hence tried to provide plausible mechanisms responsible for this peculiar finding, using a Schumpeterian approach. Their first assumption was that discoveries are obtained through combinations of existing elements. Once an innovation arises, it can influence the market in three ways: (i) it can become a component of a new innovation, (ii) it can spread its destructive effect by leading out of the market other existing products, or, (iii) it does not produce any effect on the market at all.

In the model all possibly existing goods at time t are listed in a $N-$dimensional vector $\sigma(t) = (\sigma_1(t), .., \sigma_N(t))$ where N could also be infinite. With a comparison with the urn models of Sec. 2.2, such goods can also be imagined as "balls" of an imaginary urn. Each component $\sigma_i(t)$ of the vector $\sigma(t)$ is 0 if the element at time t does not exist, which can happen in two cases: either it has never appeared before, or it has been already removed from the market. Contrarily, if the element i exists at time t, then $\sigma_i(t) = 1$. Novelties in the model arise from a combination that is already encoded in a *production table*, a tensor $A^+ = \{A_1^+, .., A_N^+\}$, such that for each good k, the matrix A_k^+ has entries $a_{ijk}^+ = 1$ if the combination of goods i and j produce k; if this happens, then $\{i, j\}$ is a *productive set* for k. Instead, whenever the combination of goods i and j does not allow the production of good k, we have $a_{ijk}^+ = 0$. We can hence estimate the number of possible ways to produce a given good k as

$$N_k^{prod}(t) = \sum_{ij} a_{ijk}^+ \sigma_i(t)\sigma_j(t). \tag{2.28}$$

At the same time, a combination of elements can also produce a *destructive* innovation, i.e., it can lead some products out of the market due to a technological obsolescence. Also the destructive power of innovation is encoded in a so-called *destruction table* $A^- = \{A_1^-, .., A_N^-\}$. For each A_k^-, the entry a_{ijk}^- is 1 if the combination of i and j lead k out of the market, whereas $a_{ijk}^- = 0$ if the combination i and j does

not influence the behavior of element k in the market. In the first case, we call $\{i, j\}$ a *destructive set* for k. Similarly to production processes, there could exist several destructive processes for the same element. We define the number of destructive processes for a given element k as

$$N_k^{destr}(t) = \sum_{ij} a^-_{ijk} \sigma_i(t) \sigma_j(t). \tag{2.29}$$

At each time t, the state of the good k will depend on the number of productive and destructive processes on it, with three different outcomes. If the number of productive sets on k is greater than the number of destructive sets on the same k, then k will be produced. If the family of destructive sets is strictly larger than those of productive sets, the good k will not be produced, or it will be expelled from the market if it previously existed. Finally, if the number of productive sets equals the one of destructive sets, then the status of k at time t will be equal to the status of the good at time $t-1$. More precisely we have:

$$\begin{cases} N_k^{prod}(t) > N_k^{destr}(t) & \implies \quad \sigma_k(t) = 1 \\ N_k^{prod}(t) < N_k^{destr}(t) & \implies \quad \sigma_k(t) = 0 \\ N_k^{prod}(t) = N_k^{destr}(t) & \implies \quad \sigma_k(t) = \sigma_k(t-1). \end{cases} \tag{2.30}$$

In summary, a production/destruction process will be activated only if each good of the corresponding productive/destructive set is available at time t. In this way, changes of state of a given good can have a relevant impact on productive/destructive sets of other connected goods, in a cascading effect. In the original work by Stefan Thurner et al. [20], scale-free versions of production/destruction topologies have been used for the related tables. In fact, even though it is possible to empirically assess production or destruction networks in the real economy, in practice, this is unrealistic and would involve tremendous effort. Moreover, for a systemic understanding of Schumpeterian dynamics, a detailed knowledge of these networks may not be necessary, as argued in Ref. [20].

The results of Thurner's model simulations are characterized by phases of boosts in economic development (measured in the diversity of available goods), phases of crashes and phases of relative stability followed by turbulent restructuring of the entire economic market. It is possible to interpret the successions of characteristic phases of construction, destruction and relative stability as Schumpeterian business "cycles', leading to innovation. Moreover, we observe clustered volatility and power laws distributions, both in the number of goods that are produced or destroyed in a single event, and in the lengths of the cycles. Overall, these results are aligned with those of the urn model with triggering discussed in Sec. 2.2.4, where novelties are also correlated and

emerge according to power-laws, but provide another perspective on the way discoveries can be made, i.e., through combinations of existing elements.

2.5 The influence of social interactions

2.5.1 Application of UMT to the emergence and evolution of social networks

An interesting application of the UMT framework is given in Ref. [30], where the model is used to generate a social network with similar properties to the way empirical social networks grow. Here, a multi-agent version of the UMT is considered, where agents explore a particular version of the adjacent possible space, formed by the same agents that explore the space and can be encountered during the exploration. In other words, each urn refers to a different agent, and the colors of the balls inside the urn correspond to other agents. With this framework, a social network emerges and evolves with the dynamics. In particular, the connections of an individual i are represented by the "discovered" agents during the process, i.e., the set of agents randomly drawn from i's urn.

We report here the model scheme as defined in Ref [30]. The model is initialized with two urns, $\mathcal{U}_a$ and $\mathcal{U}_b$ related to the agents a and b, having a copy of each other's ID inside of them. The urns also contain a set of $\nu + 1$ distinct identities (IDs) of other urns that did not participate yet in any interaction. This set corresponds to their *memory buffer* at the initial stage. As in the urn models, this model can also be characterized by a sequence of events $\mathcal{S}$, the draws from the urns, which is initially empty. After the initialization stage, each evolutionary step of the model is defined as a repetition of the following steps.

1. A "calling" urn i is extracted with probability proportional to the size U_i of the urn $\mathcal{U}_i$ (the number of balls within the urn $\mathcal{U}_i$). We then draw a ball from the calling urn $\mathcal{U}_i$, say the ID j. This double extraction corresponds to a single event $(i,\, j)$ that we append to the main sequence $\mathcal{S}$.

2. Reinforcement: following the event $(i,\, j)$, we add ρ copies of i in j's urn and ρ copies of j in i's urn.

3. Novelty: if it is the first time that i and j interact, i and j exchange their memory buffer. With this mechanism, we add j's memory buffer into $\mathcal{U}_i$ and, vice-versa, i's memory buffer into $\mathcal{U}_j$.

4. If a node j is called for the first time by another node (i.e., $\mathcal{U}_j$ is an empty urn so that $U_j = 0$), $\nu + 1$ new agents (empty urns) are created and a ball for each of them is added into $\mathcal{U}_j$. These $\nu + 1$ IDs represent the initial memory buffer of

j. Notice here that the newly created agents are initially empty urns so that they can participate in the dynamics (they can be included in the social network) only if another urn (agent) calls them. Only after this first call, they become active. In this scheme, an agent cannot join the network “from outside", i.e., it needs to be engaged by another agent already belonging to the network.

Leveraging on these simple microscopic rules defining how people get in touch and interact, the model not only represents a network growth model, but also reproduces a sequence of interaction events that mimics the empirical network dynamics. In particular, the creation of new edges and the reinforcements of their weights in the model mimics the way each individual explores the space of possible connections, strengthening the relation with some friends more than other distant acquaintances.

In this work, the proposed model's parameters are fitted to three different social networks, minimizing a cost function including various observables, both local and global, topological and dynamical. These empirical networks are the American Physical Society (APS) co-authorship network, generated by all the papers published in all the APS journals from January 1970 to December 2006, the Twitter Mention Network (TMN), logging all the mentions between users recorded between January and September 2008, and the Mobile Phone Network (MPN), recording the calls between users of a national provider in an undisclosed European country between January and July 2008. Based on the exploit–explore mechanism typical of the UMT, the proposed model manages to predict both microscopic and macroscopic features of these social networks. For example, it captures the probability for an individual with already k connections to acquire new acquaintance, and reproduces the main topological and dynamical features of social networks: the broad distribution of degree and activity, the average clustering coefficient, and the discovery rate of new friends at the global and local levels. Moreover, the model offers a deeper understanding of the propensity of people to reinforce old contacts or establish new ones in various social systems. For instance, in APS network new connections massively expand the adjacent possible space of an individual, in the TMN the exploration and reinforcement processes are of equal importance, while in the MPN people reinforce their existing bonds more than they explore new ones.

2.5.2 Social synchronisation of brain activity by eye contact

Innovation is only one of the many dynamical processes in which human interactions play a pivotal role. When we engage with one another, the exchange of information and knowledge takes on diverse forms, giving rise to potentially countless novelties on an individual level. These novelties may manifest as learning something new, forging connections with new acquaintances, discovering captivating music, and so much more. Moreover, the collective knowledge emerging from these interactions has the potential

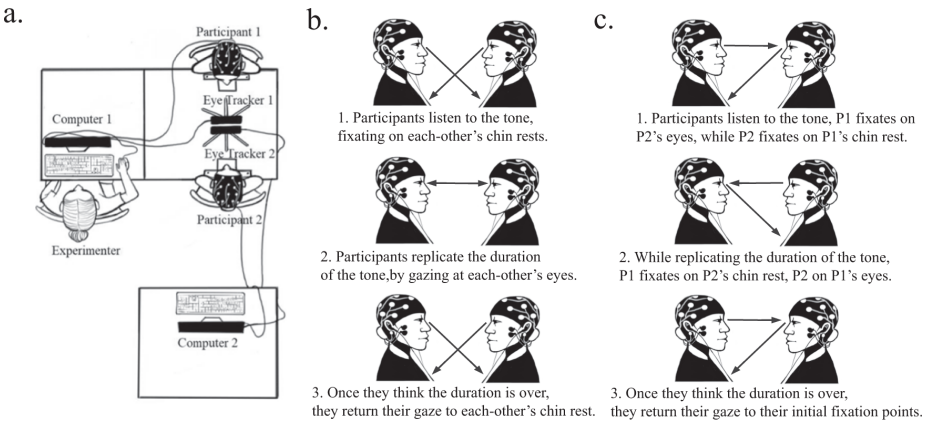


Figure 2.5: **Experimental setup and task. a.** Two computers were synchronized during the experiment and each computer collected EEG and eye-tracking data from one participant but received EEG and eye-tracking event markers with no delay (all hardware was centrally controlled through Matlab). **b. Eye contact time estimation task**: participants faced each other and were required to reproduce the duration of a tone delivered through headphones without speaking to their dyadic partner. **Eye-contact condition**: during the delivery of the tone, participants were instructed to fixate on a sticker on the other participant's chinrest. After the tone ended, the participants were instructed to look at other's eyes for the duration of the tone, looking back down to indicate that they finished reproducing the tone duration. The tone the participant heard was either long (2.5s) or short (1.5s) and in some trials, the participants heard different tone durations. **c. Non-eye-contact control condition**: participants replicated the duration of the tone but never made eye contact. Participant A listened to the tone while fixating on their partner's eyes while participant B listened to the tone while fixating on their partner's chinrest. After the tone, both participants replicated its duration, participant A by fixating their partner's chinrest and participant B by fixating their partner's eyes. To indicate the end of the tone interval, each participant reverted their gaze back to the starting position. In one block, participant A replicated the duration by looking in their partners' chinrest and the other by looking at the participant B's eyes (while participant B looked down to the chinrest), and in the other block, the roles reversed. The analysis of connectivity was restricted to the data from the period where both participants were performing the time reproduction task (step 2 of Fig.1B and C). Drawing credits to Tatiana Adamczewska.

to catalyze transformative innovations with far-reaching impacts on society. Direct eye contact, being an essential component of social interactions, plays a significant role in establishing interpersonal connections and conveying emotions and intentions [100, 101]. As such, we have contributed to the analysis of the data of a recent experiment by Caroline Di Bernardi Luft et al. [5], investigating the effects of eye contact on brain activity synchronization between pairs of individuals, offer valuable insights into the neural substrates of human social cognition.

In this experiment, whose setup and task is shown in Fig. 2.5, pairs of participants, either friends or strangers, sat facing each other in a quiet lab room. Each partici-

pant was connected to an electroencephalogram (EEG) machine and an eye tracker was positioned eye-level in front of each participant to capture their respective eye movements. The participants completed an eye contact time estimation task, where they were required to reproduce the duration of a tone delivered through headphones without speaking to their dyadic partner. During the eye contact condition, participants were instructed to fixate on a sticker on the other participant's chinrest while listening to the tone, and then look at the other's eyes for the duration of the tone. During the non-eye-contact control condition, participants replicated the duration of the tone but never made eye contact. The study investigated how eye contact affected the synchronization of brain activity between the participants.

In this study, we have investigated the synchronization of the various brain regions of the pair of participants synchronize their activity through eye contact. In particular, we have mapped such regions and their synchronization levels thanks to what is called the *hyperbrain*, which refers to the network of connections between the brains of two individuals engaging in a social interaction. We have measured the synchronization between two nodes of the hyperbrain using undirected ciPLV (corrected imaginary Phase-Locking Value) derived from the EEG data in the gamma frequency bands. The ciPLV provides a robust measure of undirected phase synchronization which is insensitive to volume conduction, with higher values indicating stronger connectivity. The ciPLV were calculated for each pair of EEG electrodes for each participant, and then averaged across participants to obtain a single ciPLV value for each pair of electrodes. The ciPLV were then used to construct the connectivity matrices that were used to generate the hyperbrain network. The connectivity matrices were further processed to generate both an unweighted and a weighted graph representation of the hyperbrain network.

Our analysis in Fig. 2.6 showed that making eye contact affects the general strength of the interbrain (between the two participants) connections of the hyperbrain more than the intrabrain (within the same participant) ones. Interestingly, the strength of brain activity synchronization was found to be significantly higher among individuals who shared preexisting bonds of friendship compared to pairs of strangers. This finding suggests that the familiarity and trust established in friendships might facilitate more pronounced neural resonance during eye contact. Furthermore, specific brain regions, such as the superior temporal sulcus and the inferior frontal gyrus, exhibited increased connectivity and acted as central hubs in the hyperbrain network during eye contact interactions.

Overall, these findings suggest that the strength and pattern of interbrain connectivity during eye contact is modulated by the social relationship between the individuals involved. The heightened synchronization observed during direct eye contact signifies a potential mechanism through which humans establish interpersonal connections and perceive shared intentions, emotions, and information. Moreover, the stronger brain

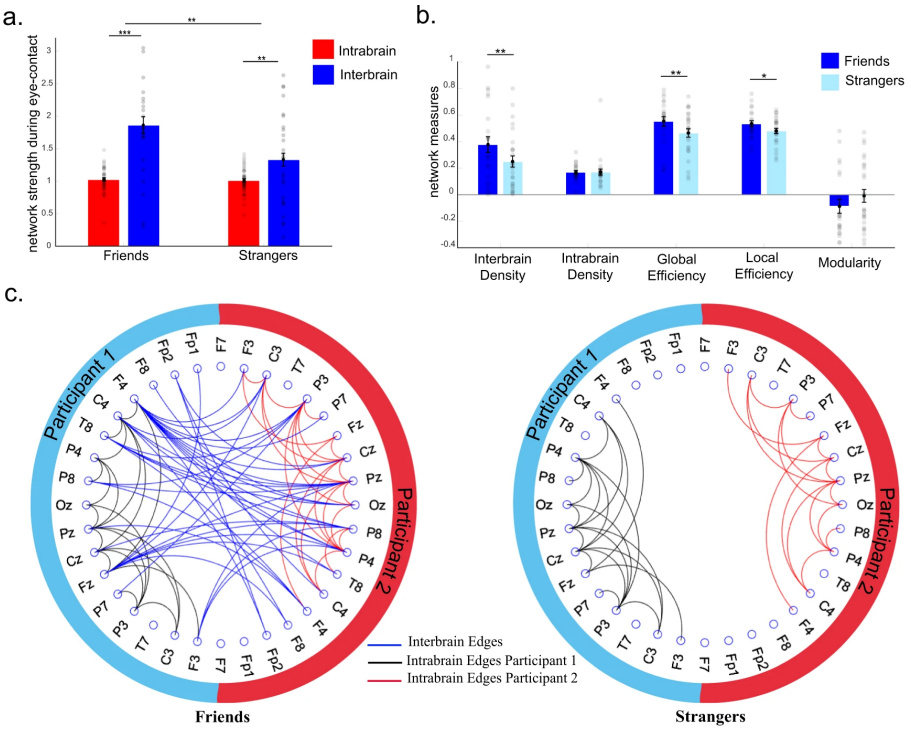


Figure 2.6: **Hyperbrain network during eye contact. a.** Intra- (red) and inter-brain (blue) network strength of friends and strangers measured as average of z-scores during eye-contact against control. The shaded dots represent each participant data ($n = 99$, the data of one participant was not plotted because its value was 3.90, which would increase the y-axis range making the error bars illegible). **b.** Network measures of friends (dark blue) and strangers (light blue) during eye contact. The measures include inter- and intra-brain density ($n = 100$), global and local efficiency and modularity ($n = 50$ dyads). These measures are based on the binary ciPLV networks. **c.** ciPLV networks during eye contact, the edges represent the phase synchronization values which increased on average by more than 1 standard deviation (SD) against the control task for friends (left) and strangers (right). Inter-brain edges are represented in blue, intra-brain edges of participants 1 are in black and 2 are in red. Error bars represent ± 1 standard error of the mean (SEM). '***' refers to $p < 0.001$, '**' to $p < 0.01$, '*' to $p < 0.05$.

activity synchronization among friends underscores the significance of social bonds in modulating neural responses during social interactions. Such findings suggest that the neural processes underlying innovation might be further enhanced within cohesive social networks. They emphasize that social interactions have a profound impact on the very dynamics of brain activity, underscoring the significance of these interactions to better grasp the intricacies of human behavior and social cognition. By unraveling the mechanisms through which social interactions foster synchronization and innovation, we hence wish to open doors to a deeper understanding of what drives creativity and progress in our society.

2.6 Summary and conclusions

In this chapter we have reviewed various models of innovation present in the literature. For each of these models, we have particularly focused on understanding how their key ingredients impact on the emergence of novelties in the innovation process. In this final section, we provide a brief overview of such models and their ingredients.

In Polya's urn model, analyzed in Sec. 2.2.1, we have observed the first important ingredient present in all innovation processes: the *reinforcement* of the elements explored, such that they are more likely to be exploited again in the future. Such mechanism is closely related to a *"rich-get-richer"* phenomenon, and often observed in many complex systems [32]. However, this mechanism alone is not enough, alone, to generate exploration sequences of events that replicate the empirical pace of discovery, as quantified by the Heaps' law. In fact, in this chapter we have seen how innovation can be seen as the exploration of a growing space of possibilities, where new elements are added over time.

Inspired by experimental data, different models have been proposed to add a *growth* mechanism of the space of possibilities. For example, Hoppe's attempt to add such a mechanism was moved by advancements in biology and genetics [77], but, as we have seen in Sec. 2.2.3, this urn model cannot reproduce the Heaps' law. The first model that has managed to reproduce the Heaps' and Zipf's laws, at least in some cases, is the Yule–Simon model, discussed in Sec. 2.2.2. Inspired by linguistic studies, this model combined both the reinforcement mechanism with a constant innovation rate in time, such that with probability p brand new elements are added at every time step. This way, however, the number of novelties grows linearly in time, and the model cannot catch sublinear Heaps' laws, which constitutes the majority of empirical cases.

The already cited models make it clear that innovation processes are to be considered as *complex systems*. We have indeed seen how the emergence of novelties creates further possibilities to obtain other novelties, i.e., what we have called the expansion of the *adjacent possible*, which refers to the set of all elements that are one step away from what already explored. In this direction, other recent models have had higher success, combining the reinforcement and growth mechanisms with the concept of the space of possibilities and its adjacent possible. Authors of Ref. [8] argue, indeed, that each person grows his space of possibilities by exploring the edge of it, with one novelty leading to another. This reasoning has lead to the formulation of the urn model with triggering (UMT), described in Sec. 2.2.4, which can reproduce the empirical Heaps' laws. This model includes Polya's reinforcement mechanism, and adds a *triggering* one, which expands the adjacent possible by triggering new colors into the urn whenever a color is drawn for the first time. Moreover, semantic correlations in the sequence of colors drawn are also reproduced in the more general urn model with semantic triggering (UMST), reviewed in Sec. 2.2.5.

Another set of models we have analysed makes use of networks and random walks. Indeed, a natural way to characterize the space of possibilities and its adjacent possible is to represent such space as a network. Random walks have hence been used to simulate various exploration and innovation processes, as shown for example by an equivalent random-walk version of the UMST introduced in Sec. 2.3.1. The edge-reinforced random walk (ERRW) model presented in Sec. 2.3.2 moves in a similar direction. The main difference between UMST and ERRW is that the reinforcement is not applied on the node but on the edge, and that the set of nodes and edges in the network is fixed along the process. In other words, there is no growth of the network in terms of size, but only in terms of edge weights. Notice that these RW models can also reproduce the Heaps' and Zipf's laws, as well as the presence of strong semantic correlations in the sequences of exploration.

In this chapter, we have also taken into consideration other models of innovation from different research areas. In particular, in Sec. 2.4.1 we have reviewed the Bak–Sneppen model for the evolution of species. Here, each species has a randomly assigned barrier which determines whether or not the species mutate. Novelties are seen as mutations of a given species, which affect the barriers of neighbours, thus often coming all together following a *cascade effect*. This model hence focuses on the effect of novelties on adjacent species in an evolutionary dynamics of species mutations, using a computational approach. Then, in Sec. 2.4.2 we have seen innovation from an economic perspective. With a Schumpeterian approach, in Thurner's model innovation is seen as the consequence of *creative destruction*, namely, the creation of new products and services, born from the combination of already existing goods, can drive obsolete goods out of the market. We will further investigate the role of combinations in Chapter 3, where we will define *higher-order Heaps' laws* to measure the pace of discovery of new combinations. We will also propose to model innovation as an *edge-reinforced random walk with triggering* (ERRWT) on a co-evolving network, rooted on the reinforcement and triggering mechanisms described in this chapter, but adapted to a more general definition of novelty, revealing the complexity of the growth of the underlying network of possibilities.

Finally, in this chapter we have highlighted how important *social interactions* are in dynamical processes. For example, in Sec. 2.5.1 we have investigated an application of the UMT to mimic how social networks emerge and evolve. In particular, we have considered many UMTs to represent different agents, while the colors of the balls in the urn identify the possible agents that can become friends. This model gives rise to a social network, which captures most of the microscopic and macroscopic features found in real-world ones. Notice, however, how this model is not properly a model of innovation, since there is no space of content being explored. We have further unveiled the importance of social interactions in Sec. 2.5.2, where we have analyzed the effect of eye contact in simple tasks. In fact, such interaction significantly influences the

brain activity and choices of the participants. We will hence combine the exploration of a space of content with social interactions in Chapter 4 and Chapter 5. In particular, in Chapter 4 we will couple many UMTs through the links of a social network, introducing the concept of *social expansion of the adjacent possible*. In that model, the urns will be enriched with the adjacent possible of their contacts in a cooperative way, enhancing their pace of discovery. The choice of the UMT will come natural from the capacity of the UMT to encode the fundamental ingredients of a discovery process into a simple and analytical tractable model, as we have seen in this chapter. We will further analyze the influence of social interactions on collective discovery processes in Chapter 5, where we will analyze extensive music exploration data. There, we will also model such process as a multi-agent exploration of an underlying network of music, where multiple agents share recommendations to each other, and thus influence each other's musical taste and pace of discovery.

Chapter 3

The adjacent possible in the content space

3.1 Introduction and outline

As humans, we experience novelties as part of our daily life. By the term *novelty* we generally indicate two apparently different things [8]. On the one hand, we can think of a novelty as the first time we visit a neighborhood, enter a newly launched pub, or listen to a song from an artist we previously did not know. In this case, the novelty represents a discovery for a single individual of a place, an artist or, more in general, an item. On the other hand, there are discoveries that are new to the entire population, as could be a technological advancement or the development of a new drug. However, these two cases are not entirely distinct, as the second set of novelties, those new to everyone, represent just as a subset of the first one. Analysing how novelties emerge both at the individual level, and at the level of the entire population, is key to understand human creativity and the neural and social mechanisms that can lead to new discoveries and innovation.

The increasing availability of data on human behavior and consumption habits has allowed to study how humans explore the world, how novelties emerge in different contexts, and how they are distributed in time [8, 28, 29]. Empirical investigations cover a broad range of different areas [102]ranging from science [95] and language [25, 65], to gastronomy [96], goods and products [69], network science [97], information [98], and cinema [99]. No matter the topic, one can always represent data coming from real-world exploration processes as sequences of items that are sequentially adopted or consumed [103]. In this way, the activity of a user of, for example, an online digital music platform is turned into a sequence of listened songs, and a novelty is defined as the first time a song, or an artist, appears in the sequence. Analogously, articles

published in a scientific journal can be turned into a time-ordered sequence of concepts or keywords discovered by the community, and a novelty can be defined, again, as the first-time appearance of a keyword [29]. Under this framework, evidence shows that novelties seem to obey the same statistical patterns on the way they are distributed and correlated in time, independently of the system they belong to [8]. In particular, most empirical sequences follow Heaps' [14, 15, 58], Zipf's [54–57, 104], and Taylor's laws [82].

Along with data-driven investigations, a relevant scientific problem is that of finding plausible mechanisms to reproduce and explain the empirical observations. What are the rules controlling the appearance of new items in a sequence? How do humans explore the seemingly infinite space of possibilities in search of novelties? As we have discussed in Chapter 2, an insightful answer comes from biology, when, in 1996, Stuart Kauffman introduced the concept of the *adjacent possible* [13] (AP), referring to "*all those molecular species that are not members of the actual, but are one reaction step away from the actual*". Inspired by previous works by Packard and Langton [9–11], the AP provides a fresh view on the problem, for which discoveries (the possible) can only be found among those items which are close (the adjacent) to what is already known (the actual). New discoveries would then generate an expanding space of opportunities that are only available to us in the moment we "unlock" what is adjacent to them.

Kauffman's AP has seen many interesting applications ranging from biology [13, 53] and economics [69, 105] to models of discovery and innovation processes. Among these, of particular interest is the recently proposed Urn Model with Triggering [8, 25, 106] (UMT). Building upon the work of Pólya [23, 77], the UMT adds to the traditional *reinforcement* mechanism of the Pólya urn's scheme a *triggering* mechanism that expands the space of possible discoveries upon the extraction of each novelty. As we have seen in Sec. 2.2.4, the UMT is able to reproduce the empirical laws by properly balancing these two mechanisms. The AP accounts for the emergence of the new starting from the "edge of what is known". In this view, one could also picture ideas, concepts, or items as the linked elements of an abstract network, as described in Sec. 2.3. Within this framework, the way we explore the world based on the association of different concepts can be naturally modelled as a random walk over this network. For example, this approach has been used to investigate the cognitive growth of knowledge in scientific disciplines [29].

There is, however, another important mechanism of creation of the new, which is neglected by the frameworks discussed above: novelties can arise from the combination of already-known elements. In fact, as we have seen in Sec. 2.4.2 in the context of innovation economics with Thurner's model [20], innovation can be seen as a process of "creative destruction", where the combination of goods may activate the production or destruction of other products. As originally discussed by Schumpeter [21, 22] and later confirmed by recent works on the generation of technologies [107–109], new

associations of existing factors may give rise to new discoveries, which rule out of the market obsolete products and services [110, 111], thus increasing the probability of reaching further novelties and innovation.

Such mechanism of combining "pre-existing" items to create something new also applies in various other contexts. For instance, a meaningless sequence of words, if ordered in a different way, may generate elegant poetry. Novel combinations of existing hashtags may lead to new social-media trends. Different orderings of the same musical notes may in principle generate an endless number of songs. Moreover, the mechanics of combination and association of existing elements has been studied in other fields too, e.g., in biology, where combinations are the keys to produce new entities and organisms. For instance, it has been shown that the immune system recombines existing segments of genes to produce new receptors [112, 113]. Furthermore, we can consider publications and collaborations in science [114] as combinations of multiple research ideas [115–117] and expertises [118–120].

In this chapter we hence explore a new and more general notion of novelty, defined as anovel combination of existing elements. We thus investigate the dynamics of "higher-order" novelties, i.e., novel combinations of pairs, triplets, etc., of consecutive items in a sequence [6]. In particular, we focus on the Heaps' law, which describes the growth in the number of novelties as a power-law, whose exponent is a proxy for the pace of discovery [15] in a system. Namely, in Sec. 3.2 we introduce higher-order Heaps' laws to characterize the pace at which novel combinations of two and more elements appear in a sequence. In Sec. 3.3 we then analyse various types of empirical sequences ranging from music listening records, to words in texts, and concepts in scientific articles, finding that Heaps' laws also holds at higher orders. We discover that individual processes with the same pace of discovery of single items, can instead display different rates of discovery at higher orders, and can hence be differentiated in this way. We also simulate and analyze some existing models of innovation in Sec. 3.4, finding that they can only reproduce higher-order Heaps' exponents equal to the 1^{st}-order ones.

Therefore, in Sec. 3.5, we propose a new model which is capable of reproducing all these empirically observed features of higher-order Heaps' laws. In our model the process of exploration is described as an edge-reinforced random walks with triggering (ERRWT) over a growing network. In our framework, the novelties at different orders (nodes and links visited for the first time by the walker) shape the explored network by reinforcing traversed links while, at the same time, triggering the expansion of the adjacent possible. This expansion can happen whenever a node is visited for the first time, making other nodes accessible to the explorer, but also whenever a link is firstly used. In this case, the newly established connection will trigger novel combinations between previously explored nodes. In other words, we propose a new mathematical framework which describes the process of innovation as the exploration of a space of

possibilities, which is seen as a growing network representing the relations between elements of various types of content. With this new formulation, novelties of various orders are the result of the exploration of a more definition of adjacent possible. At the first and second order, the adjacent possible is represented by the set of neighboring nodes and links, respectively, which have not been explored yet by the random walker. In our model, we also add a mechanism to trigger new parts of the adjacent possible whenever such adjacent possible is explored. By fitting the contributions of the two mechanisms of reinforcement and triggering, the ERRWT model is able to reproduces well the variety of scaling exponents found in real systems for the Heaps' laws at different orders. Further notice that the ERRWT model significantly differs from Thurner's model. The latter model, indeed, requires the presence of production and destruction networks, which are pre-determined. Moreover, the appearance of a new product is deterministic, and only depends on the presence (absence) of all the necessary productive (destructive) items related to the product. In our model, instead, new nodes and links appear in the network as the walker explores such network, which can grow indefinitely by expanding the adjacent possible. Additionally, the more a connection is used, the more it is reinforced and likely to be used again in the future.

Finally, we conclude this chapter with some analytical results on the urn model with triggering and on the ERRWT in Sec. 3.6, and summarize the main findings in Sec. 3.7.

3.2 Higher-order Heaps' laws

An exploration process can be represented as an ordered set of T symbols $\mathcal{S} = \{a_1, a_2, \ldots, a_T\}$. Such a set describes the sequence of "events" or "items" produced along the journey, e.g., the songs listened by a given individual over time, the list of hashtags posted on an online social network, the list of words in a text, or any other ordered list of items or ideas generated by single individuals or social groups [8, 26, 103]. Similarly, in the context of some recent modelling schemes of discovery (see Sec. 2.2 and Sec. 2.3), $\mathcal{S}$ can represent the balls extracted from an urn [8, 26], or the nodes visited over time by a random walker moving over a network [29]. Although real-world events have an associated time, here, for simplicity, we focus only on their sequence, i.e., the relative temporal order of the events, neglecting the precise time at which they happen. For instance, if a person listens to song a_1 at time t_1, song a_2 at time t_2, song a_i at time t_i, and so on, with $t_1 < t_2 < \cdots < t_i < \ldots$, we neglect these times and only retain the order of the songs in the sequence $\{a_1, a_2, \ldots, a_T\}$. In other words, we assume that a_1 is associated to the discrete time $t = 1$, a_2 is associated to time $t = 2$, and so forth.

Among the different ways to characterize the discovery rate of a given process, the Heaps' law, $D(t) \sim t^{\beta}$, describes the power-law growth of the number of novelties

as a function of time, i.e., how the number $D(t)$ of novel elements in the sequence $\mathcal{S}$ scale with the sequence length t [15]. The so-called (standard) Heaps' exponent β, that from now on we indicate as 1^{st}*-order Heaps' exponent* β_1, is thus a measure of the pace of discovery of the process that generated the considered sequence. Given that the number of different elements $D(t)$ is smaller (or equal) than the total length t of the sequence, the value of β_1 is always bounded in the interval $[0, 1]$, with the extreme case $\beta_1 = 1$ reached by a process that generates new elements at a linear rate.

Here, we propose to go one step beyond and look at novelties as novel pairs, triplets, and higher-order combinations of consecutive symbols in a sequence [121]. For instance, when exploring a network, a novel pair is represented by the first visit of a link. In order to measure the pace of discovery of these higher-order compounds starting from a sequence of events $\mathcal{S}$, we first create the surrogate sequence of overlapping pairs $\mathcal{S}_2 = \{(a_1, a_2), (a_2, a_3), \ldots, (a_{T-1}, a_T)\}$. Considering for example the sentence *"One ring to rule them all"*, from the sequence of events $\mathcal{S} = \{$*one*, *ring*, *to*, *rule*, *them*, *all* $\}$ we obtain the sequence of overlapping pairs $\mathcal{S}_2 = \{$(*one*, *ring*), (*ring*, *to*), (*to*, *rule*), (*rule*, *them*), (*them*, *all*)$\}$. From $\mathcal{S}_2$ we can then compute the number $D_2(t)$ of different pairs among the first t ones, with $t \leq T - 1$. Notice that, here, we consider the pairs $(one, ring)$ and $(ring, one)$ as two different pairs, i.e., order matters. By construction, we always have $D_1(t) \leq D_2(t) \leq t$, since, on the one hand, for each new element added to $\mathcal{S}$ there is a new pair in $\mathcal{S}_2$, and, on the other hand, there cannot be more than t different pairs among t items. From the power-law scaling $D_2(t) \sim t^{\beta_2}$, we can then extract the value of β_2, which we refer to as the 2^{nd}*-order Heaps' exponent*. This definition can be naturally extended to any order n, considering the sequence $\mathcal{S}_n$ of consecutive overlapping n-tuples present in $\mathcal{S}$. Notice that, if $|\mathcal{S}| = T$, then $|\mathcal{S}_n| = T - n + 1$. We can hence compute the number $D_n(t)$ of different tuples among the first t tuples in $\mathcal{S}_n$, and extract the n^{th}*-order Heaps' exponent* $\beta_n \in [0, 1]$ from $D_n(t) \sim t^{\beta_n}$. Notice also that the n^{th}-order Heaps' exponent can be also interpreted as the first order Heaps' exponent of a sequence whose events are the overlapping n-tuples of the original sequence. Finally, it is worth remarking that such an approach is close to the analysis of Zipf's law in linguistic data for n-grams or sentences [122, 123]. In this context, studies showed that as one moves from graphemes—representing individual phonemes (speech sounds)—, to words, sentences, and n-grams, the Zipf's exponent (reciprocal of the Zipf's for infinitely long sequences [58]) gradually diminishes. This implies that n-grams or sentences are characterized by a larger novelty rate than words, a behavior analogous to what we have discussed above.

3.2.1 Power-law fit

Fundamental for the estimation of the higher-order Heaps' exponent of a sequence is the power-law fitting procedure for the number of novel n-tuples $D_n(t)$ as a function

of the sequence length t, with $n \geq 1$. The sequences analyzed in this chapter come from very different contexts, from empirical data sets to model simulations. We thus need to take into consideration all those cases that show a transient regime—whose length might also depend on the system structure—in which the pace of discovery can fluctuate before reaching its stationary value. We fit each sequence according to the following procedure. To reduce computational times, we first logarithmically sample 1000 points from each sequence in the range $[1, T]$, where T is the length of the considered sequence. Considering their integer part and discarding all duplicates, we obtain a set of k integer times $\{t_i\}_{i=1,\dots,k}$ between 1 and T. If $T \geq 1000$, that is the case of all sequences used in this work, then this process results in $k \geq 424$ points. Taking into account that the associated sequence of n-tuples has length $T - n + 1$, we thus consider the points $\{(t_i - n + 1,\ D_n(t))\}_{i=1,\dots,k}$ in logarithmic scale, i.e.,

$$(x_i, y_i) = (\log_{10}(t_i - n + 1),\ \log_{10}(D_n(t)))\,, \tag{3.1}$$

with $i = 1, \dots, k$. In order to neglect the initial transient regime, but still have enough points for a sufficiently significant fit, we select only the last 100 of such points. We hence look for the best fit of $\{(x_i, y_i)\}_{i=k-100+1,\dots,k}$ by optimizing the linear function $y = a + b\,x$, with $a > 0$, using the tool `curve_fit` of the Python package `Scipy` [124]. If $\overline{a}$ and $\overline{b}$ are the best parameters, then the power-law fit of the Heaps' law is $D_n(t) \approx 10^{\overline{a}}\, t^{\overline{b}}$, that is, the n^{th}-order Heaps' exponent is approximated by the slope $\overline{b}$ of the fit.

3.3 Empirical analysis

3.3.1 Data

Let us consider three different data sets on music listening records (*Last.fm*), books (*Project Gutenberg*), and scientific articles (*Semantic Scholar*).

Last.fm is a digital platform for music born in 2002, famous for logging all listening activities of its users, providing both personal recommendations and a space to interact with other users interested in music [125]. Here, we use a data set presented in Ref. [126] and available at Ref. [127], containing all listening records of about 1000 users. In order to have sequences long enough for statistically relevant fits, only users with more than 1000 logs have been retained. The final data set contains 890 users having a median number of listened records of 13 985. Each record contains the timestamp at which a user listened to a given song. In the database, each song is associated to a title, the artist's name and a unique MusicBrainz Identifier (MBID), which can be used to obtain additional metadata [128]. Using this information, we are able to create, for each user, a temporally ordered sequence of songs together with the associated

sequence of artists.

Project Gutenberg is an open access text corpus containing more than 50 000 books of different nature. Here, we make use of the Standardized Project Gutenberg Corpus [129], which allows to download and process an updated version of the corpus. Using Google's Compact Language Detector 3 (`cld3` package in Python), we filter out all non-English texts. We then discard all texts with less than 1000 words, retaining a total of 19 637 books with a median number of 50 726 words. A sequence of events for each book is hence created with the lemmatized words, disregarding punctuation and putting all characters in lower case. We also extract stems from each word using the English Snowball stemmer [130]—a more accurate extension of the Porter stemmer [131]—, which is not as aggressive as the Lancaster stemmer [132].

Semantic Scholar is a recent project with the scope of facilitating scientific analysis of academic publications. It provides monthly snapshots of research papers published in all fields, publicly accessible through the *Semantic Scholar Academic Graph* (*S2AG*, pronounced "stag") [133]. This database (1$^{\text{st}}$ Jan. 2022 snapshot) contains about 203.6M papers, 76.4M authors, and 2B citations. It also classifies each paper into one or more fields of study [134], for a total of 19 different fields. For simplicity, we associate each paper to its first (and most relevant) field of study. To create the sequences to analyze, for each field we consider the first 1000 journals in terms of number of English papers. Then, for each journal, we order the published papers based on the respective year of publication, volume, issue, and first page. When some of this information is not available, the Semantic Scholar unique ID of the paper is also used in the ordering process. Thus, for each paper, we extract and lemmatize their title, similarly to what done for the Project Gutenberg. Finally, a sequence of events is created for each selected journal, concatenating the lemmatized words in the titles of each paper in their temporal order, for a total of 19 000 sequences with median length of 9 114.5. Associated to this sequence, we also consider the sequence of stemmed words for further analysis.

All the code used to download, process and analyse the data can be found at Ref. [135]. Finally, the distribution of the length T of each sequence in the three data sets is shown in Fig. 3.1.

3.3.2 Analysis of empirical sequences

We start investigating the emergence of novelties of different orders in empirical exploration processes associated to three different data sets described in the previous section. Notice that these data sets are substantially different in nature, since they refer, respectively, to songs listened by users of *Last.fm*, words in books collected in the *Project Gutenberg*, and words of titles of scientific journals from *Semantic Scholar*. In Fig. 3.2(**a-c**) we plot the average temporal evolution of the number $D_n(t)$ of novel-

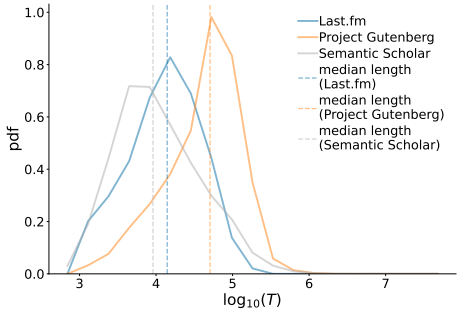


Figure 3.1: **Length T distribution of the sequences in the data sets.** Probability density function of the length T of all sequences in the three data sets (Last.fm in blue, Project Gutenberg in orange, Semantic Scholar in green). Moreover, the median lengths, respectively equal to 13 985, 50 726, and 9 114.5, are shown in the plot as vertical dashed lines, with the color corresponding to each data set.

ties of order n, with $n = 1, 2, 3$, in the three datasets (from left to right, respectively, Last.fm, Project Gutenberg, Semantic Scholar). In order to avoid spurious effects due to different lengths of the sequences, we restrict the averages to the sequences of length T greater than the median length $\tilde{T}$ in the corresponding data set (see Fig. 3.1 for their distribution). Each continuous curve, plotted up to time $\tilde{T}$, is obtained by averaging $D_n(t)$ over all such sequences, while the shaded area represents one standard deviation above and below the mean. We also perform power-law fits, as described in Sec. 3.2.1, and plot the resulting curves as dashed lines. Focusing first on the broadly-studied (1^{st}-order) Heaps' law, notice how the power-law fit is only accurate in the last part of the sequence. This highlights that the Heaps' law starts after a transient phase, where most of the events are new for the individual, as also reported in Ref. [8] and similarly reported in other contexts [136–140]. Secondly, notice how the n^{th}-order Heaps' law, with $n = 2, 3$, is valid across the data sets, but with different values of the fitted exponents, especially for $n = 2$. Finally, as expected from their definition, the fitted Heaps' exponents of order $n + 1$, i.e., β_{n+1}, are higher than the lower-order ones, that is, $\beta_{n+1} \geq \beta_n$.

To explore the gain in information brought by the higher-order Heaps' exponents with respect to the 1^{st}-order Heaps', we now look directly at individual sequences. Figure 3.2(**d**-**i**) shows the scatter plots of β_2 (**d**-**f**) and β_3 (**g**-**i**) against β_1, where each point refers to a single sequence from Last.fm (**d**,**g**), Project Gutenberg (**e**,**h**) or Semantic Scholar (**f**,**i**), with colors representing the density of points (see color bar at the bottom of the figure). Here, we have only considered sequences whose fitted exponent has a standard error below the 0.05 threshold (see Table 3.1 for more details). This filtering removes 30 (3.37%), 8 (0.04%), and 5 (0.03%) sequences in the three datasets, respectively. This shows that, in almost all cases, we can consider the Heaps'

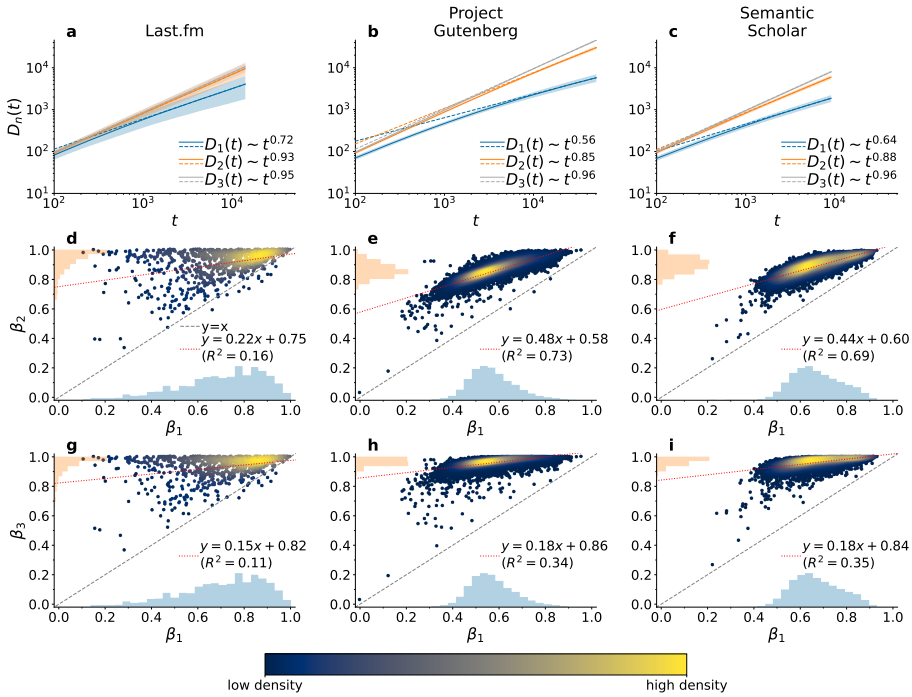


Figure 3.2: **Higher-order Heaps' exponents and their correlations in real-world data sets.** (**a-c**) Average number $D_n(t)$ of novelties of order n, with $n = 1, 2, 3$, as a function of the sequence length t, and fit of the associated Heaps' laws (dashed lines), with estimated exponents shown in the legend. Shaded area represents one standard deviation above and below the average. (**d-i**) Scatter plots between the (1^{st}-order) Heaps' exponents β_1 and the n^{th}-order exponents β_n, with $n = 2$ (**d-f**) and 3 (**g-i**). Each point refers to a different sequence, with colors representing the density of points (see color bar). Each panel also reports histograms of exponents distributions, the bisector $y = x$ (dashed gray line), as well as the fitted linear model (dotted red line) with the value of its coefficient of determination R^2. Each column refers to a different data set: (**a**,**d**,**g**) Last.fm, (**b**,**e**,**h**) Project Gutenberg and (**c**,**f**,**i**) Semantic Scholar, respectively.

law assumption to be valid. Looking at the plots, we notice that some cases have a higher density of points compared to others. For example, in (**d**), we see how users of Last.fm sharing the same value of β_1 can have very different values of β_2. Conversely, the other two data sets present stronger correlation between β_2 and β_1. To quantitatively characterize this, we fit a linear model with an ordinary least squares method, displayed in each plot as a red dotted line. In the legend we also report the value of the related coefficient of determination R^2, which represents the percentage of variance of the dependent variable explained by the linear fit with the independent variable.

For users of Last.fm, at both orders $n = 2$ and 3, we quantitatively confirm that points are much more spread around the linear fit, since the values of R^2 are very low, between 0.11 and 0.16. In the other two data sets there is instead a higher correlation

data set (β_n)	min	1^{st}%ile	25^{th}%ile	median	75^{th}%ile	99^{th}%ile	max
Last.fm (β_1)	0.0007	0.0016	0.0052	0.0079	0.0129	0.0693	0.1988
Last.fm (β_2)	0.0000	0.0001	0.0026	0.0047	0.0091	0.0510	0.1497
Last.fm (β_3)	0.0000	0.0000	0.0019	0.0038	0.0073	0.0388	0.1366
Gutenberg (β_1)	0.0000	0.0010	0.0021	0.0029	0.0043	0.0169	0.0727
Gutenberg (β_2)	0.0003	0.0005	0.0010	0.0014	0.0020	0.0087	0.0522
Gutenberg (β_3)	0.0001	0.0002	0.0004	0.0006	0.0009	0.0064	0.0484
S2AG (β_1)	0.0003	0.0008	0.0018	0.0025	0.0035	0.0115	0.1279
S2AG (β_2)	0.0002	0.0004	0.0010	0.0014	0.0021	0.0093	0.1677
S2AG (β_3)	0.0000	0.0002	0.0005	0.0008	0.0013	0.0078	0.1698

Table 3.1: **Statistics on the standard error of the fitted higher-order Heaps' exponents in the empirical data.** Various statistics on the standard error, or standard deviation of the estimator, of the fitted n^{th}-order Heaps' exponents β_n of the sequences in the three data sets, with $n = 1$, 2, and 3. Notice how the standard deviation of the distribution of the values of the exponents in the data sets (see Table. 3.2 for reference) is about two orders of magnitude higher than the median standard error and one order higher than its 99^{th} percentile. Moreover, the p-values of the fits are all zero (not shown in the table).

between β_1 and both β_2 (R^2 around 0.70) and β_3 (R^2 around 0.35). Moreover, the values of the parameters of the linear fit greatly change across datasets and orders. In particular, in (**d**) there is a much lower slope and higher intercept compared to the other data sets for the same order in (**e-f**). Furthermore, we notice how, for each data set, the higher the order, the lower the fitted slope —and the higher the intercept of the linear model. Finally, on an aggregate level, we observe that at all orders the distribution of the Heaps' exponents are very different across data sets (see Fig. 3.3 for a comparative figure, while further statistical information on the Heaps' exponents distribution can be found in Table 3.2). The exponents are more spread in Last.fm, which also shows a higher average of β_1 and β_2, but a lower one for β_3 compared to the other data sets.

Moreover, notice how the distributions for Project Gutenberg and Semantic Scholar, which are both related to linguistic data, are more peaked—at higher values for the latter dataset. This could be the result of how titles of scientific papers are written with respect to books or poems, that is, concentrating the whole message of a scientific work in a few significant words, avoiding stop-words and repetition. In addition, scientific advancements tend to favor the combinations of previously existing scientific concepts to form new ones, while the same does not apply to non-scientific literature in general, where instead similar constructions tend to be repeated across the piece. Finally, similar results are obtained also for more coarse-grained sequences generated using artists and stemmed words instead of songs and words, as shown in Fig 3.4.

data set (β_n)	mean	std
Last.fm (β_1)	0.7029	0.1797
Last.fm (β_2)	0.9048	0.1014
Last.fm (β_3)	0.9286	0.0862
Gutenberg (β_1)	0.5699	0.0973
Gutenberg (β_2)	0.8527	0.0547
Gutenberg (β_3)	0.9589	0.0300
S2AG (β_1)	0.6695	0.1019
S2AG (β_2)	0.8895	0.0536
S2AG (β_3)	0.9612	0.0305

data set (β_n)	min	1^{st}%ile	25^{th}%ile	median	75^{th}%ile	99^{th}%ile	max
Last.fm (β_1)	0.1063	0.2010	0.5965	0.7395	0.8436	0.9761	0.9959
Last.fm (β_2)	0.3342	0.5725	0.8699	0.9388	0.9754	0.9999	0.9999
Last.fm (β_3)	0.3664	0.6123	0.9041	0.9583	0.9861	0.9999	1.0000
Gutenberg (β_1)	0.0000	0.3678	0.5026	0.5594	0.6285	0.8302	0.9527
Gutenberg (β_2)	0.0304	0.7118	0.8191	0.8509	0.8883	0.9706	0.9919
Gutenberg (β_3)	0.0307	0.8648	0.9480	0.9627	0.9765	0.9968	0.9998
S2AG (β_1)	0.2225	0.4673	0.5923	0.6590	0.7436	0.8889	0.9293
S2AG (β_2)	0.2587	0.7509	0.8550	0.8936	0.9303	0.9803	0.9942
S2AG (β_3)	0.2665	0.8686	0.9478	0.9680	0.9825	0.9972	0.9999

Table 3.2: **Statistics of the fitted higher-order Heaps' exponents in the data.** Various statistics of the fitted n^{th}-order Heaps' exponents β_n of the sequences in the three data sets, with $n = 1$, 2, and 3.

3.4 Analysis of existing models

After studying higher-order Heaps' laws in real data, we check whether the observed patterns can be also reproduced by the available models for discovery processes. We start from the Urn Model with Triggering (UMT), where a sequence of events is generated by draws of coloured balls from an urn [8], different colours corresponding to different events/items being discovered/adopted and so on. In the UMT, for each extracted ball, the corresponding color is reinforced by adding ρ additional balls, of the same color, to the urn. At the same time, whenever a novel color is drawn, the discovery triggers the addition of $\nu + 1$ balls of new different colors to the urn (see detailed model definition in Sec. 2.2.4). Previous studies have shown that the 1^{st}-order Heaps' law is verified in sequences obtained with the UMT [8, 25]. In particular, the number of novelties in the model grows asymptotically as $D_1(t) \sim t^{\frac{\nu}{\rho}}$ when $\nu < \rho$, while a linear behaviour is found in the other cases.

We hence focus on the most interesting case, that is for $\nu \leq \rho$, studying how variations of the two parameters ρ and ν, respectively representing the reinforcement and the increase in size of the adjacent possible, affect the Heaps' law at various orders.

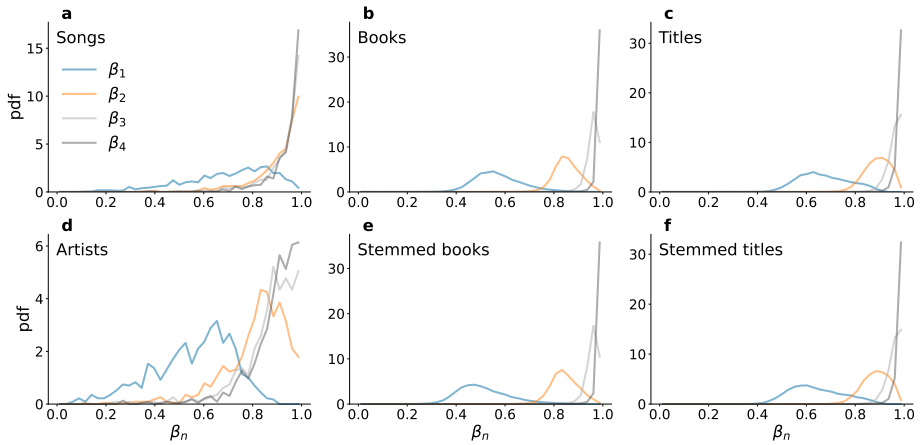


Figure 3.3: **Heaps' exponent distribution of the sequences in the data sets.** Probability density functions of the n^{th}-order Heaps' exponents β_n, with $n = 1, 2, 3, 4$, calculated from the empirical sequences (a-c) and respective sequences of labels (d-f). In particular, sequences contain songs (**a**) and artists (**d**) in Last.fm, words (**b**) and stemmed words (**e**) in Project Gutenberg books, words (**c**) and stemmed words (**f**) in Semantic Scholar journal titles.

Since the pace of discovery, at the first order, effectively depends only on the fraction ν/ρ, we fix $\rho = 20$ and numerically simulate the UMT with $\nu = 1, 2, 3, \ldots, 20$ for $T = 10^5$ time-steps. For each set of parameters we run 100 simulations, generating a total of 2×10^3 synthetic sequences. Then, for each generated sequence, we compute the temporal evolution of the number of novelties $D_n(t)$, and estimate a power-law fit, extracting the related n^{th}-order Heaps' exponent β_n. In Fig. 3.5(**a**), we show how the extracted values of β_2 change with respect to β_1 across simulations. The color represents the value of the parameter ν, as shown in the color bar. We observe that, although the exponents are distributed all across the interval $(0, 1)$, the points (β_1, β_2) are just above the bisector (gray dashed line). Moreover, for a certain value of β_1, the model produces very similar values of β_2 that do not vary much.

We can derive an analytical approximation of the higher-order Heaps' exponents for this model. As we show in Sec. 3.6.1, for the UMT the number of unique pairs grows as

$$D_2(t) \approx a\, t^{\beta_2}, \quad \text{with} \quad \beta_2 = \beta_1 + \frac{c}{d + \log(t)}, \tag{3.2}$$

where a, c, $d > 0$ depend on the parameters ρ and ν, and $\beta_1 = \nu/\rho$. Although the predicted 2^{nd}-order exponent is slightly higher than the 1^{st}-order one, their difference just depends on the sequence length, and vanishes at larger times. In other words, the increased value of the higher-order Heaps' exponent is only due to a finite time effect, and the UMT struggles in reproducing the empirical patterns discussed in Fig. 3.2.

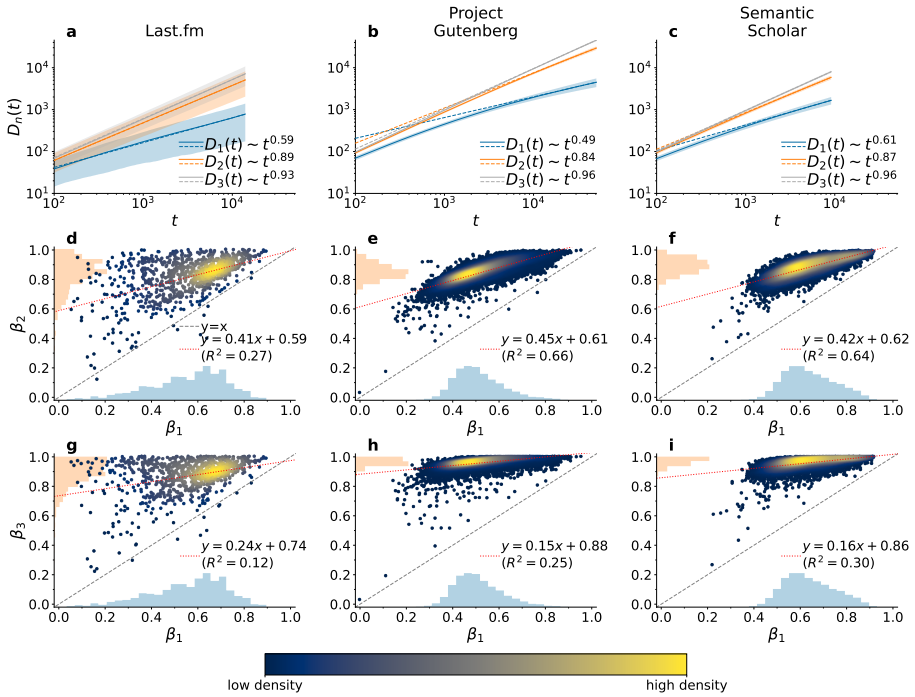


Figure 3.4: **Higher-order Heaps' exponents in more coarse-grained empirical sequences.** (**a-c**) Average number $D_n(t)$ of novelties of order n, with $n = 1, 2, 3$, as a function of the sequence length t, and fit of the associated Heaps' laws (dashed lines), with estimated exponents shown in the legend. Shaded area represents one standard deviation above and below the average. (**d-i**) Scatter plots between the (1^{st}-order) Heaps' exponents β_1 and the n^{th}-order exponents β_n, with $n = 2$ (**d-f**) and 3 (**g-i**). Each point refers to a different sequence, with colors representing the density of points (see color bar). Each panel also reports histograms of exponents distributions, the bisector $y = x$ (dashed gray line), as well as the fitted linear model (dotted red line) with the value of its coefficient of determination R^2. Each column refers to a different data set: (**a**,**d**,**g**) sequences of artists listened on Last.fm, (**b**,**e**,**h**) sequences of stemmed words of books from Project Gutenberg and (**c**,**f**,**i**) sequences of stemmed words of titles in journals of Semantic Scholar, respectively.

We repeat this analysis for the Urn Model with Semantic Triggering (UMST) [8] and the Edge-Reinforced Random Walk (ERRW) [29], which have also been proved to generate discovery sequences obeying to the Heaps' law. These models share the same foundations of the UMT, but with some crucial differences. The UMST builds on top of the UMT introducing also semantic groups for colors (topic common to different items). This addition effectively diminishes the probability to draw colors outside of the semantic group of the last extracted color by a factor η. The ERRW is formulated as a network exploration rather than a process of extractions from an urn. Instead of a sequence of extracted balls, the ERRW features a set of nodes sequentially visited

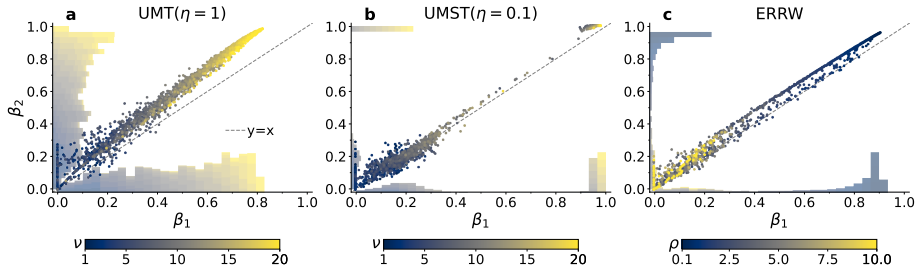


Figure 3.5: **Higher-order Heaps' exponents in existing models.** Scatter plots between the (1st-order) Heaps' exponent β_1 and the 2nd-order exponent β_2 in: (**a**) the urn model with triggering (UMT), no semantic correlations ($\eta = 1$), and $\rho = 20$, $\nu = 1, 2, \ldots, 20$; (**b**) the urn model with semantic triggering (UMST) with $\eta = 0.1$ and $\rho = 4$, $\nu = 1, 2, \ldots, 20$; (**c**) the edge-reinforced random walk (ERRW) on a small-world network (average degree $\langle k \rangle = 4$ and rewiring probability $p = 0.1$) with edge reinforcement ρ ranging geometrically from 0.1 to 10. Each point refers to a different simulation of the related model, with colors representing the value of the free parameter (see color bar). Each panel also reports histograms of exponent distributions on the respective axes, and the bisector $y = x$ (dashed gray line). All simulations have run for 10^5 time steps.

by a random walker over a weighted networks, where the weight of visited edges are reinforced at each time by ρ. A full description of the models can be found in Sec. 2.2.5 (UMST) and Sec. 2.3.2 (ERRW).

We simulate the UMST with parameters $\eta = 0.1$, $\rho = 4$, $\nu = 1, 2, \ldots, 20$, while the ERRW runs over a small-world network (with average degree $\langle k \rangle = 4$ and rewiring probability $p = 0.1$), with edge-reinforcement ρ ranging from 0.1 to 10. Similarly to the exploration of the UMT, we perform 100 simulations for each set of parameters and report the results in Fig. 3.5(**b-c**). For both UMST and ERRW, we find that the values of β_2 do not differ much from their corresponding value of β_1—as shown by the great proximity of the points (β_1, β_2) to the bisector. This means that also these models fail to reproduce the empirical variability of higher-order Heaps' exponents with respect to the 1st-order one. Moreover, we notice in (**b**) that for the UMST we only obtain exponents with either very low (up to 0.4) or very high (close to 1) values. It seems thus that there is an abrupt transition between the two cases, with the model not able to cover the values in-between. This is instead a crucial point when we are confronted with the empirical values reported in Fig. 3.2. A more detailed comparison with the analytical results of the UMT and UMST is discussed in Sec. 3.6.2.

Overall, with the analyses above we have just shown that while the existing models for discovery and innovation dynamics are able to reproduce the empirically observed pace of discovery of new items—as singletons—, they systematically fail when it comes to capturing the distributions of the Heaps' exponents of higher order and their

correlations.

3.5 ERRWT: a model for higher-order Heaps' laws

We now introduce a model that is able to generate synthetic sequences displaying different Heaps' exponents at various orders. As for the previously discussed ERRW, our novel model is formulated using a network framework in which: (*i*) the items to be explored correspond to the nodes of the network; (*ii*) links between nodes represent semantic associations between items that one can use to move from one to another; (*iii*) the exploration process is modelled as a random walk over the network, and the exploration sequence is given by the list of visited nodes.

Under these assumptions, the first visit of a node corresponds to a 1^{st}-order novelty, while a 2^{nd}-order novelty refers to the first exploration of a link. This definition can be trivially extended to higher orders, but here, for simplicity, we limit our attention to the first two orders. The ERRW proposed in Ref. [29] consists of a walker exploring a static network with a fixed topology, whose movements modify only the weights of the links. By contrast, in our model the network structure (not just the weights) co-evolves over time together with the exploration process such that new links can be triggered. Thus, blending together the ERRW and the UMT [8], we call the model *Edge-Reinforced Random Walk with Triggering* (ERRWT). More specifically, the model is based on two different triggering mechanisms that add new edges and new nodes every time a novelty appears. As per the UMT and the ERRW, exploring a node for the first time triggers the expansion of the adjacent possible, as new nodes become now accessible. For example, the invention of the transistor made it possible to create mobile phones, among other things. Concerning the triggering of new edges, the idea is that whenever two elements are associated for the first time, new possible combinations involving one of these elements are then triggered. For instance, once a camera and a mobile phone were firstly combined, this made clear that many more functions could be added to the latter, e.g., a music player, a game console, a GPS, etc.

The basic mechanisms of the ERRWT model are illustrated in Fig. 3.6. Suppose that, at a given time t, the walker is at node i of a network composed of some already visited nodes and links (filled nodes and continuous lines), and some others that belong to the adjacent possible (unfilled nodes and dashed lines). This is the starting point of Fig. 3.6(**a**). In Fig. 3.6(**b**), the walker crosses an already explored link, and its weight gets reinforced by a term ρ, meaning that the association of the two nodes becomes more likely. This is the same reinforcement process of the ERRW in Ref. [29]. If addition, if the edge is instead traversed for the first time, along with the edge-reinforcement the process triggers also the creation of new edges. In particular, as displayed in Fig. 3.6(**c**), $\nu_2 + 1$ new edges connecting the second node of the traversed link to other already-visited nodes are created. Finally, analogously to the triggering

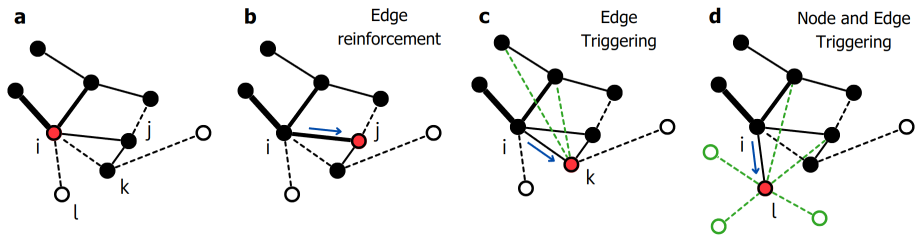


Figure 3.6: **The Edge-Reinforced Random Walk with Triggering (ERRWT) model.** An exploration process is modelled as a random walk on a growing weighted network. (**a**) At time t, the walker is at the red node i. Nodes that have been already visited by the walker are colored in black, in white those left to be visited. Similarly, traversed (old) and not-traversed (new) links are respectively depicted with continuous and dashed lines, whose widths represent their weights. At time $t+1$, the walker can move to each of the neighbours of i, e.g. j, k, or l, with a probability proportional to the weight of the respective link. (**b**) If the walker moves to j, since the link $(i,\, j)$ is old, its weight is reinforced by ρ (edge reinforcement); (**c**) if it moves to k, since link $(i,\, k)$ is new, but node k is old, in addition to the edge reinforcement, $\nu_2 + 1 = 2$ new edges (in green) between k and old nodes are added to the network (edge triggering); (**d**) if it moves to l, since both the link and the node are new, in addition to the edge reinforcement and the edge triggering, $\nu_1 + 1 = 3$ new nodes (in green) are added to the network and connected to l (node and edge triggering).

mechanism of the UMT, whenever a node is visited for the first time, it triggers the expansion of the node's adjacent possible with $\nu_1 + 1$ new nodes added to the network and connected to the node itself (Fig. 3.6(**d**)). Note that this also triggers the creation of other $\nu_2 + 1$ new links to already known elements, since whenever a node is explored for the first time, also the link leading to it is explored for the first time. A more mathematical description of the ERRWT can be found in the next section.

3.5.1 Model definition

Let us consider an initial connected network $G^0 = (\mathcal{V}^0,\, \mathcal{E}^0)$ with $N^0 = |\mathcal{V}^0| \geq 1$ nodes and $M^0 = |\mathcal{E}^0|$ links. Let us suppose that the nodes of the graph are indexed, that is, $\mathcal{V}^0 = \{1,\, 2,\, \ldots,\, N_0\}$. Similarly to the ERRW, in the ERRWT we assume that all initial links $(i, j) \in \mathcal{E}^0$ have weight $w_{ij}^0 = 1$. The initial node to start the exploration process is randomly selected from $\mathcal{V}^0$. We let the graph G^t evolve during the process, adding new nodes and links. The structure of the growing network is encrypted in the time-varying weight matrix $W^t = (w_{ij}^t)$. Then, supposing to be on node i of the graph G^t at time t, the model obeys to the following rules.

- *Choice of next node.* The ERRWT randomly moves to a neighbouring node j of the current node i. The probability to move to node j depends on the weight of

the outgoing links of i, i.e.,

$$\mathbb{P}(i \to j) = \frac{w_{ij}^t}{\sum_l w_{il}^t}. \tag{3.3}$$

- *Edge reinforcement.* The weight of the chosen edge (i, j) is reinforced by ρ, that is,

$$w_{ij}^{t+1} = w_{ij}^t + \rho. \tag{3.4}$$

- *Edge triggering.* If the walker never traversed the chosen edge (i, j) before this time, i.e., it is a new link, then ν_2+1 new possible links are added to the network. These links are connections of unitary weight between j and previously visited nodes $l = l_1, \ldots, l_{\nu_2}$ in $\mathcal{V}^t$, for which the link (j, l) has never been traversed by the walker. If one of these edges already exists in the space of possibilities, its weight is reinforced by one more unit, otherwise, it is added to $\mathcal{E}^{t+1}$. In other words, we have

$$w_{jl}^{t+1} = w_{jl}^t + 1, \quad l = l_1, \ldots, l_{\nu_2} \mid l \text{ old}, (j, l) \text{ new}. \tag{3.5}$$

- *Node triggering.* If the walker never visited the chosen node j before this time, i.e., it is a new node, then $\nu_1 + 1$ new nodes are added to the network; these are connected to node j with unitary weights. Mathematically, we have

$$\begin{aligned} \mathcal{V}^{t+1} &= \mathcal{V}^t + \left\{|\mathcal{V}^t| + 1, \ldots, |\mathcal{V}^t| + \nu_1 + 2\right\} \\ w_{jl}^{t+1} &= 1, \quad l = |\mathcal{V}^t| + 1, \ldots, |\mathcal{V}^t| + \nu_1 + 2. \end{aligned} \tag{3.6}$$

  Notice that if the chosen node j is new, then also the traversed edge (i, j) is necessarily new as well. Therefore, in this case there is also a triggering of $\nu_2 + 1$ edges from j to other previously visited nodes, as described before.

In the following sections, we let G_0 be a small graph that emulates the triggering mechanism introduced, shown in Fig. 3.7. This is a regular tree with branching parameter ν_1+1 and 2 levels, where the leaves are considered new, while all other nodes have already triggered. In other words, a root node has triggered $\nu_1 + 1$ nodes connected to it, and again these nodes have also triggered each $\nu_1 + 1$ other nodes. Therefore, we initially suppose that the triggered nodes, which are $\nu_1 + 2$ in number, are all known to the walker at the start of the simulation, and do not trigger again when later explored. Moreover, we assume that all links are new to the walker and have unitary weight. This initialization makes sure that in the initial stages of the simulation there are enough possible links between already known nodes. As we show in Sec. 3.6.3 where we test different initial graphs, the initialization procedure only affects thermalization times,

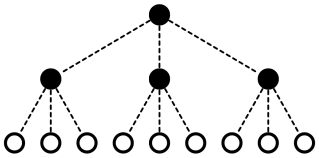


Figure 3.7: **Representation of the chosen initial network structure used in simulations of the ERRWT.** Although any initial network structure can be used for the ERRWT, in the simulations shown in Sec. 3.5.2 we consider a regular tree with branching parameter $\nu_1 + 1$ (equal to 3 in the figure) and 2 levels. This structure resembles the way new nodes are triggered during the exploration, so that the root and first layer (full nodes) are considered triggered and known, while the leaves (empty nodes) are considered new. All links are regarded as new (represented as dashed). The choice of this tree has been done to ensure that the triggering of new edges finds nodes that are already known by the random walker.

and becomes irrelevant asymptotically.

3.5.2 Numerical simulations

Balancing edge reinforcement and the node and edge triggering through the parameters ρ, ν_1 and ν_2, it is possible to control the pace of discovery of new nodes and edges, and consequently the exponents of the 1^{st}-order and the 2^{nd}-order Heaps' law associated to the sequences produced by the model. To systematically explore this, we simulate the ERRWT model with parameters $\rho = 10$, $\nu_1 = 0,\ 1,\ \ldots,\ 20$, and $\nu_2 = 0,\ 1,\ \ldots,\ 2\nu_1$, running 100 simulations for each set of parameters. Higher values of ν_2 have not been considered since they produce the same exponents as those for $\nu_2 = 2\nu_1$. Fig. 3.8(a) reports the increase in the number of 1^{st}-order and 2^{nd}-order novelties (continuous lines) for a specific set of parameters ($\rho = 10$, $\nu_1 = 10$, and $\nu_2 = 15$). The power-law fits (dashed lines) highlight that the Heaps' law is verified at higher-orders too, leading to an increase of the exponents values (from $\beta_1 = 0.56$ to $\beta_2 = 0.87$) as we increase the order.

The relationship between the different orders is explored in Fig. 3.8(b), where we show the scatter plot between the 1^{st}- and 2^{nd}-order Heaps' exponent. Each point refers to a different simulation, and we use the color to indicate the value of the used parameter ν_1 (see color bar). We notice that the ERRWT produces a wide range of exponents at both orders, which are no more trivially correlated as for previous models. This is even more clear when we look at Fig. 3.8(c), where Heaps' exponents are averaged across simulations for each set of parameters: each trajectory relates to a different value of ν_1, with ν_1 increasing from 1 to 20 from bottom left to top right of the panel. The color represents instead the variation of the parameter ν_2 from 0 to $2\nu_1$. For reference, we also flag with a red dot the pair of exponents related to the parameters used in Fig. 3.8(a).

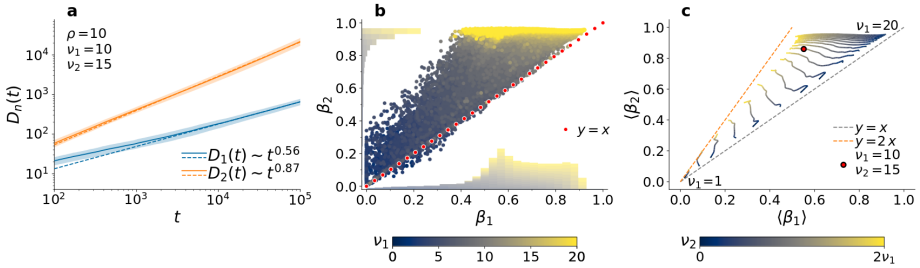


Figure 3.8: **Higher-order Heaps' exponents in the ERRWT model.** (**a**) Average number $D_n(t)$ of novelties of order n, with $n = 1$ and 2, as a function of the sequence length t for simulations of the ERRWT with parameters $\rho = 10$, $\nu_1 = 10$, $\nu_2 = 15$, and fit of the associated Heaps' laws (dashed lines), with estimated exponents shown in the legend. Shaded areas represent one standard deviation above and below the average. (**b**) Scatter plot between the (standard) Heaps' exponent β_1 and the 2$^{\text{nd}}$-order exponent β_2. Each point refers to a different simulation of the model, with colors representing the corresponding value of the parameter ν_1 ranging from 0 to 20 (see color bar), while $\rho = 10$ and $\nu_2 = 0, \ldots, 2\nu_1$. (**c**) Variation of the average n^{th}-order Heaps' exponents β_n, with $n = 1, 2$. Each trajectory refers to a different value of ν_1, increasing from 1 to 20 from bottom left to top right, with the color depending on the value of ν_2 (see color bar). The set of parameters used in (**a**) is here highlighted in with a red dot.

We can immediately notice how the 1$^{\text{st}}$- and 2$^{\text{nd}}$-order Heaps' exponents increase as ν_1 becomes larger. More interestingly, we can investigate the interplay with ν_2: given a single trajectory, by increasing ν_2 the difference between β_1 and β_2 becomes larger, and the point (β_1, β_2) moves away from the bisector, in a way that depends on the specific value of ν_1. In particular, for low values of ν_1, the trajectories are almost vertical, with only β_2 increasing. Instead, for higher values of ν_1, especially when $\nu_1 \geq \rho$, an increase of ν_2 produces a decrease of β_1, while the value of β_2, which is close to its upper bound value 1, does not change.

It is also possible to perform an analytical investigation of a simplified version of the ERRWT model, shown in Sec. 3.6.3, which leads to similar results. In particular, for such a model, we can prove that the values of the asymptotic Heaps' exponents β_1 and β_2 depend on the two ratios ν_1/ρ and ν_2/ρ. Moreover, we find that, for $\nu_1/\rho > 1$, the 2$^{\text{nd}}$-order Heaps' exponent is asymptotically equal to 1, while the 1$^{\text{st}}$-order one depends on ν_1/ν_2, in agreement with our numerical results. Finally, the exponents are asymptotically bounded by $\beta_1 \leq \beta_2 \leq 2\beta_1$, as also shown in the simulations in Fig. 3.8(c). This also explains why the exponents do not change when we increase ν_2 above $2\nu_1$.

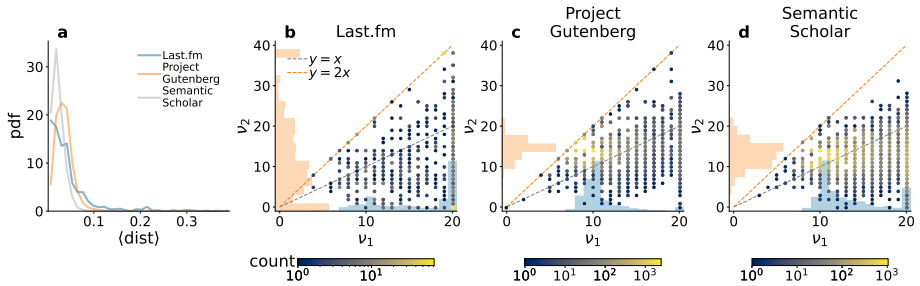


Figure 3.9: **Fitting the ERRWT model to real-world data sets.** (**a**) Distribution of the average distance between the pair of exponents $(\beta_1,\ \beta_2)$ of a real sequence and the pair $(\beta_1',\ \beta_2')$ obtained by the best fit of the ERRWT model. (**b-c**) Scatter plots of the best-fitted parameters ν_1 and ν_2 of the model across the sequences of the three data sets, i.e., Last.fm (**b**), Project Gutenberg (**c**), and Semantic Scholar (**d**). The color of a point refers to the number of sequences with that pair of parameters as corresponding best fit (see color bar).

3.5.3 Comparison with data

To show that the ERRWT model is able to reproduce the properties observed in real-world processes, we fit the model to the three data sets analyzed in Sec. 3.3, namely Last.fm, Project Gutenberg and Semantic Scholar. Given an empirical sequence and its pair of 1$^{\text{st}}$- and 2$^{\text{nd}}$-order Heaps exponents $(\beta_1,\ \beta_2)$, we compute the Euclidean distance between the pair $(\beta_1,\ \beta_2)$ and each of the pairs of exponents $(\beta_1',\ \beta_2')$ obtained by simulating the ERRWT model using the sets of parameters considered in Fig. 3.8. We then select the best model parameters by minimizing the average distance over 100 simulations for each set of parameters, and repeat the procedure for all the sequences of the three data sets.

Figure 3.9(a) shows the probability density distribution of the distances between the empirical sequences and the simulations of the best-performing ERRWT model. Notice how these distances are almost all below 0.1, that is below the uncertainty we expect on the values of the parameters. In fact, being ν_1, ν_2 integers and $\rho = 10$, the maximum precision we can gain on the estimate of the best parameters is approximately $1/\rho = 0.1$. The percentage of sequences with higher distance than this threshold is 7.67%, 0.73%, and 0.05% for Last.fm, Project Gutenberg, and Semantic Scholar, respectively.

The scatter plots of the best-fitted parameters ν_1 and ν_2 for the three data sets are shown in Fig. 3.9(b-d). The colors here indicate the number of empirical sequences which are best represented by each pair of parameters. We notice that most of the sequences of Last.fm are characterized by relatively large values of ν_1. Since ν_1 is related to the triggering of new nodes, this result indicates that the discovery of a new song exposes the user to a large variety of related songs, which were previously not accessible and can now be discovered. Conversely, the parameter ν_2, which refers to

the triggering of new edges between already existing items, takes values in a larger range, predominantly skewed towards the lower end. This suggests that, once a new association of two songs is established by a user, there is a high probability that the same association will be repeated over and over. Consequently, the user will preferably listen to songs in a similar order, instead of creating new associations. In the case of Project Gutenberg, most sequences have $\nu_2 > \nu_1$. This implies that writers tend to frequently generate new word associations, highlighting the incredible variety of expressions we can make combining a limited set of words. Finally, Semantic Scholar exhibits values of ν_1 and ν_2 similar to Project Gutenberg. However, some sequences of Semantic Scholar have a relatively high value of ν_1 with respect to ν_2. This is an indication that, when choosing words for titles, authors tend to use more original words, while the pace of creation of new word associations remains similar.

3.6 Analytical results

3.6.1 Analytical results for higher-order Heaps' laws in the UMT

The Urn Model with Triggering (UMT) features a triggering mechanism for the growth of the adjacent possible [8]. In particular, whenever a new color is drawn for the first time, $\nu+1$ new colors are triggered and added into the urn. Together with the reinforcement mechanism introduced in Polya's urn [24], the UMT manages to reproduce various features of innovation processes, including the Heaps' law. In particular, varying the parameters, the UMT produces different rates of discovery, which can be measured by the power-law exponent β_1 of the Heaps' law. As we have seen in Sec. 3.4, the UMT is also able to produce higher-order Heaps' laws, measuring the pace of discovery of combinations of more than one element. However, as shown in Fig. 3.5, the 2^{nd}-order Heaps' exponents obtained in simulations of the UMT are very close to the respective 1^{st}-order ones. In this section, we hence provide a complete analytical analysis of the higher-order Heaps' laws for the UMT, so as to have an analytical resulof the long-term behavior of the higher-order Heaps' laws

First-order Heaps' law

Following what already discussed in Sec. 2.2.4, the evolution of the number $D_1(t)$ of different colors that have appeared in the first t positions of the sequence $\mathcal{S}$ is ruled by the following master equation:

$$D_1(t+1) = D_1(t) + \mathbb{P}\left(\mathbb{N}^{(t+1)}\right) = D_1(t) + \frac{N_0 + vD_1(t)}{N_0 + \rho t + (v+1)D_1(t)}, \qquad (3.7)$$

where $\mathbb{N}^{(t+1)}$ is the event of drawing at time $(t+1)$ a ball of a color that has not been observed before. Its probability $\mathbb{P}\left(\mathbb{N}^{(t+1)}\right)$ can be expressed as the number of colors in the urn yet to be discovered, $N_0 + (\nu+1)D_1(t) - D_1(t)$, divided by the total number of balls available at time t in the urn. In the long time limit, Eq. (3.7) can be approximated by a differential equation, which leads to an analytical expression for $D_1(t)$ (see Sec. 2.2.4 for the analytical calculations):

$$\begin{cases} \frac{\mathrm{d}D_1(t)}{\mathrm{d}t} = \frac{N_0+vD_1(t)}{N_0+\rho t+(v+1)D_1(t)} \\ D_1(0) = 0 \end{cases} \implies D_1(t) \underset{t\to\infty}{\approx} \begin{cases} bt^{\beta_1} & \text{if } \nu < \rho, \\ b\frac{t}{\log t} & \text{if } \nu = \rho, \\ bt & \text{if } \nu > \rho, \end{cases} \tag{3.8}$$

where $\beta_1 = \nu/\rho$ and b is a constant depending on ν and ρ. In other words, in the sublinear case $\nu < \rho$, the Heaps' law is analytically verified, with asymptotic exponent $\beta_1 = \nu/\rho$ [8, 25, 26].

Second-order Heaps' law

In order to write down an equation similar to Eq. (3.8) for the number $D_2(t)$ of different pairs that have appeared in the sequence $\mathcal{S}_2$ of length t, i.e.,

$$\begin{cases} \frac{\mathrm{d}D_2(t)}{\mathrm{d}t} = \mathbb{P}(\text{“The } t\text{-th pair is new”}), \\ D_2(0) = 0, \end{cases} \tag{3.9}$$

we need to calculate the probability to observe a new pair. However, differently from Eq. (3.8), such a probability depends not only on the total number of balls and on the number of extracted colors, but also on the number of balls of each extracted color. Notice that the t-th pair (x_1, x_2) of $\mathcal{S}_2$ is composed by the color x_1 drawn at time t in $\mathcal{S}$ and the color x_2 drawn in the next time step. Hence, there are three separate events in which the t-th pair (x_1, x_2) is a novelty in $\mathcal{S}_2$: the event $\mathbb{A}$ in which x_1 is a novelty, i.e., it appears for the first time in the sequence $\mathcal{S}$ at time t; the event $\mathbb{B}$ in which x_1 is not a novelty but x_2 is a novelty; the event $\mathbb{C}$ in which both colors x_1 and x_2 are not novel, but the combination (x_1, x_2) appears for the first time. Consequently, the probability that the t-th pair is new is equal to the sum of the probabilities of such events. Using Eq. (3.8), for large values of t the probability of event $\mathbb{A}$ can be written as

$$\mathbb{P}(\mathbb{A}) = \mathbb{P}\left(\mathbb{N}^{(t)}\right) = \frac{\mathrm{d}D_1(t)}{\mathrm{d}t} \underset{t\to\infty}{\approx} b\beta_1 t^{\beta_1-1}. \tag{3.10}$$

Similarly, denoting with $\overline{\mathbb{N}^{(t)}}$ the opposite event of $\mathbb{N}^{(t)}$, the probability of event $\mathbb{B}$ reads

$$\mathbb{P}(\mathbb{B}) \approx \mathbb{P}\left(\overline{\mathbb{N}^{(t)}}\right)\mathbb{P}\left(\mathbb{N}^{(t+1)}\right) \approx \left(1-\frac{\mathrm{d}D_1(t)}{\mathrm{d}t}\right)\frac{\mathrm{d}D_1(t+1)}{\mathrm{d}t} \approx b\beta_1 t^{\beta_1-1} = \mathbb{P}(\mathbb{A}), \tag{3.11}$$

where we have disregarded infinitesimals of lower order. Thirdly, we can compute the probability of the event $\mathbb{C}$ by calculating the probability that each possible pair of old colors is a novelty in this time step. Since the number of old colors up to time t is $D_1(t)$, indicating with $\mathbb{C}_{i,j}^{(t)}$ the event in which i and j are two already extracted colors and their pair (i,j) is a novelty at time t in $\mathcal{S}_2$, we can write:

$$\begin{aligned}\mathbb{P}(\mathbb{C}) &\approx \mathbb{P}\left(\overline{\mathbb{N}^{(t)}}\right)\mathbb{P}\left(\overline{\mathbb{N}^{(t+1)}}\right)\mathbb{P}\left(\bigcup_{i,j=1}^{D_1(t)}\mathbb{C}_{i,j}^{(t)}\right)\\ &\approx \left(1-b\beta_1 t^{\beta_1-1}\right)^2\ \mathbb{P}\left(\bigcup_{i,j=1}^{bt^{\beta_1}}\mathbb{C}_{i,j}^{(t)}\right) \approx \sum_{i,j=1}^{bt^{\beta_1}}\mathbb{P}\left(\mathbb{C}_{i,j}^{(t)}\right).\end{aligned} \tag{3.12}$$

The last equality in Eq. (3.12) holds true because for any $(i_1,j_1) \neq (i_2,j_2)$ we have $\mathbb{C}_{i_1,j_1}^{(t)} \cap \mathbb{C}_{i_2,j_2}^{(t)} = \emptyset$, since only one pair can be extracted at each time step, and we have disregarded lower infinitesimals.

Let us now concentrate on computing the probability of $\mathbb{C}_{i,j}(t)$. Defining the event $\mathbb{E}_{ij}^{(\tau)}$ = "pair (i,j) appears (not necessarily for the first time) in the sequence at time τ", we can rewrite $\mathbb{C}_{i,j}^{(t)}$ as

$$\mathbb{C}_{i,j}^{(t)} = \overline{\mathbb{E}_{ij}^{(1)}} \cap \overline{\mathbb{E}_{ij}^{(2)}} \cap \cdots \cap \overline{\mathbb{E}_{ij}^{(t-1)}} \cap \mathbb{E}_{ij}^{(t)}, \tag{3.13}$$

where we denote with $\overline{\mathbb{E}_{ij}^{(\tau)}}$ the opposite event of $\mathbb{E}_{ij}^{(\tau)}$. We can hence compute its probability as

$$\begin{aligned}\mathbb{P}\left(\mathbb{C}_{i,j}^{(t)}\right) =& \mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(1)}} \cap \overline{\mathbb{E}_{ij}^{(2)}} \cap \cdots \cap \overline{\mathbb{E}_{ij}^{(t-1)}} \cap \mathbb{E}_{ij}^{(t)}\right)\\ =& \mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(1)}}\right)\mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(2)}} \mid \overline{\mathbb{E}_{ij}^{(1)}}\right)\cdots\mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(t-1)}} \mid \overline{\mathbb{E}_{ij}^{(1)}} \cap \cdots \cap \overline{\mathbb{E}_{ij}^{(t-2)}}\right)\\ & \mathbb{P}\left(\mathbb{E}_{ij}^{(t)} \mid \overline{\mathbb{E}_{ij}^{(1)}} \cap \cdots \cap \overline{\mathbb{E}_{ij}^{(t-1)}}\right).\end{aligned} \tag{3.14}$$

First, we notice that we can simplify the expressions in Eq. (3.14), since

$$\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \mid \overline{\mathbb{E}_{ij}^{(1)}} \cap \cdots \cap \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right) = \mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \mid \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right). \tag{3.15}$$

This equality in Eq. (3.15) holds true because, the probability of extracting the pair (i,j) at time τ can only be influenced by what has happened at time $(\tau-1)$, disregard-

ing all previous times.

Without loss of generality, let us index the colors in the urn in the same order they first appeared in the sequence, i.e., let us suppose that the i-th color has appeared at time t_i, with $t_{i+1} > t_i$, for $i = 1, 2, \ldots, D_1(t)$. Let us also suppose that the rate at which a new color appears is given exactly by the approximated solution given by Eq. (3.8). Then, it would be

$$i = D(t_i) \approx b t_i^{\beta_1} \implies t_i \approx \left(\frac{i}{b}\right)^{\frac{1}{\beta_1}}. \tag{3.16}$$

With Eq. (3.16) we are assuming that the behaviour of $D_1(t)$ at finite times can be approximated with the asymptotic one, and that colors appear deterministically at these expected moments. Even though strong, this assumption makes sense if we consider that, as it has been observed before, there is a good correspondence between this analytical solution and simulations at finite times. Moreover, we will confirm *a posteriori* the suitability of this assumption since, as we will see, there is correspondence between the analytical solution of $D_2(t)$ we obtain here and the results of model simulations.

Let us now define $n_i(t)$ as the number of times the color i has appeared before time t, supposing it has first appeared at time $t_i \leq t$. If $\mathbb{E}_i^{(t)}$ = "i appears at time t" (not necessarily for the first time), then we have that $\frac{dn_i}{\mathrm{d}t} = \mathbb{P}\left(\mathbb{E}_i^{(t)}\right)$. Thus, we can write:

$$\begin{aligned}
\begin{cases} \frac{dn_i(t)}{\mathrm{d}t} = \frac{\rho n_i(t)+1}{N_0+aD(t)+\rho t} \underset{t\to\infty}{\approx} \frac{n_i}{t}, \\ n_i(t_i) = 1, \end{cases} &\implies \begin{cases} n_i(t) \underset{t\to\infty}{\approx} \frac{t}{t_i} & \text{if } t \geq t_i, \\ n_i(t) = 0 & \text{if } t < t_i, \end{cases} \\
&\implies \begin{cases} \frac{dn_i(t)}{\mathrm{d}t} \underset{t\to\infty}{\approx} \frac{1}{t_i} = \left(\frac{b}{i}\right)^{\frac{1}{\beta_1}} & \text{if } t \geq t_i, \\ \frac{dn_i(t)}{\mathrm{d}t} = 0 & \text{if } t < t_i. \end{cases}
\end{aligned} \tag{3.17}$$

Let us observe that under these assumptions $dn_i/\mathrm{d}t$ is actually constant in time, depending just on t_i. Then, supposing that the number of balls $n_i(\tau),\ n_j(\tau)$ of the two colors in the urn follows exactly Eq. (3.9), we can calculate the probability of $\mathbb{E}_{ij}^{(\tau)}$ as

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)}\right) = \mathbb{P}\left(\mathbb{E}_i^{(\tau)}\right)\mathbb{P}\left(\mathbb{E}_j^{(\tau+1)}\right) &= \frac{dn_i(\tau)}{d\tau}\frac{dn_j(\tau+1)}{d\tau} \\
&\underset{t\to\infty}{\approx} \begin{cases} \frac{1}{t_i t_j} & \text{if } \tau \geq \max(t_i, t_j - 1), \\ 0 & \text{if } \tau < \max(t_i, t_j - 1). \end{cases}
\end{aligned} \tag{3.18}$$

Furthermore, if $\tau \geq \max(t_i, t_j - 1)$, we can write

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \cap \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right) = \mathbb{P}\left(\left[\left(\mathbb{E}_{ij}^{(\tau+1)} \cap \overline{\mathbb{E}_{ij}^{(\tau)}}\right) \cap \mathbb{E}_{j}^{(\tau)}\right] \cup \left[\left(\mathbb{E}_{ij}^{(\tau+1)} \cap \overline{\mathbb{E}_{ij}^{(\tau)}}\right) \cap \overline{\mathbb{E}_{j}^{(\tau)}}\right]\right) \\
&= \mathbb{P}\left(\left[\mathbb{E}_{i}^{(\tau)} \cap \mathbb{E}_{j}^{(\tau+1)} \cap \overline{\mathbb{E}_{i}^{(\tau-1)}} \cap \mathbb{E}_{j}^{(\tau)}\right] \cup \left[\mathbb{E}_{i}^{(\tau)} \cap \mathbb{E}_{j}^{(\tau+1)} \cap \overline{\mathbb{E}_{j}^{(\tau)}}\right]\right) \\
&= \mathbb{P}\left(\mathbb{E}_{i}^{(\tau)} \cap \mathbb{E}_{j}^{(\tau)}\right) \mathbb{P}\left(\mathbb{E}_{j}^{(\tau+1)}\right) \mathbb{P}\left(\overline{\mathbb{E}_{i}^{(\tau-1)}}\right) + \mathbb{P}\left(\mathbb{E}_{i}^{(\tau)} \cap \overline{\mathbb{E}_{j}^{(\tau)}}\right) \mathbb{P}\left(\mathbb{E}_{j}^{(\tau+1)}\right) \\
&= \delta(i,j) \frac{1}{t_i} \frac{1}{t_j} \left(1 - \frac{1}{t_i}\right) + (1 - \delta(i,j)) \frac{1}{t_i} \frac{1}{t_j}.
\end{aligned} \tag{3.19}$$

Therefore, we get

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \mid \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right) &= \frac{\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \cap \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right)}{\mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(\tau-1)}}\right)} = \frac{\delta(i,j) \frac{1}{t_i t_j} \left(1 - \frac{1}{t_i}\right) + \frac{1 - \delta(i,j)}{t_i t_j}}{1 - \frac{1}{t_i t_j}} \\
&= \frac{\delta(i,j) \left(1 - \frac{1}{t_i}\right) + (1 - \delta(i,j))}{t_i t_j - 1} \\
&= \delta(i,j) \frac{\left(1 - \frac{1}{t_i}\right)}{t_i t_j - 1} + (1 - \delta(i,j)) \frac{1}{t_i t_j - 1} \\
&= \delta(i,j) \frac{1}{t_i^2 + t_i} + (1 - \delta(i,j)) \frac{1}{t_i t_j - 1}.
\end{aligned} \tag{3.20}$$

In the following of this discussion, we make the following approximation:

$$\delta(i,j) \frac{1}{t_i^2 + t_i} + (1 - \delta(i,j)) \frac{1}{t_i t_j - 1} \approx \frac{1}{t_i t_j}. \tag{3.21}$$

Because of Eq. (3.18), the approximation in Eq. (3.21) implies in Eq. (3.20) that

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \mid \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right) &\approx \begin{cases} \frac{1}{t_i t_j} & \text{if } \tau \geq \max(t_i, t_j - 1) \\ 0 & \text{if } \tau < \max(t_i, t_j - 1) \end{cases} \\
&\implies \mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)} \mid \overline{\mathbb{E}_{ij}^{(\tau-1)}}\right) \approx \mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)}\right),
\end{aligned} \tag{3.22}$$

which is equivalent to assume that $\mathbb{E}_{ij}^{(\tau)}$ and $\overline{\mathbb{E}_{ij}^{(\tau-1)}}$ are statistically independent, i.e., that the extraction of a certain pair (i,j) at time τ is independent of its extraction at the previous time $(\tau - 1)$. Therefore, using Eq. (3.14), Eq. (3.15), Eq. (3.18) and Eq. (3.22), the probability of the event $\mathbb{C}_{ij}^{(t)}$ that the pair (i,j) is extracted at time t for

the first time can be approximated to

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{C}_{ij}^{(t)}\right) = \mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(1)}}\right)\mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(2)}}\right)\cdots\mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(t-1)}}\right)\mathbb{P}\left(\mathbb{E}_{ij}^{(t)}\right) \\
&= \prod_{\tau=1}^{t-1}\mathbb{P}\left(\overline{\mathbb{E}_{ij}^{(\tau)}}\right)\mathbb{P}\left(\mathbb{E}_{ij}^{(t)}\right) = \prod_{\tau=\max(t_i,t_j-1)}^{t-1}\left(1-\mathbb{P}\left(\mathbb{E}_{ij}^{(\tau)}\right)\right)\mathbb{P}\left(\mathbb{E}_{ij}^{(t)}\right) \\
&= \left(1-\frac{1}{t_i t_j}\right)^{t-\max(t_i,\,t_j-1)}\frac{1}{t_i t_j},
\end{aligned} \tag{3.23}$$

which can be used in Eq. (3.12) to obtain an approximated expression for the probability of event $\mathbb{C}$, i.e.,

$$\mathbb{P}\left(\mathbb{C}\right) \approx \sum_{i,j=1}^{bt^{\beta_1}}\mathbb{P}\left(\mathbb{C}_{ij}^{(t)}\right) \approx \sum_{i,j=1}^{bt^{\beta_1}}\left(1-\frac{1}{t_i t_j}\right)^{t-\max(t_i,\,t_j-1)}\frac{1}{t_i t_j}. \tag{3.24}$$

Summing up, by inserting Eq. (3.10), Eq. (3.11), and Eq. (3.24) into Eq. (3.9), we get the following differential equation for the number of new pairs in time in the UMT:

$$\frac{\mathrm{d}D_2}{\mathrm{d}t} \underset{t\to\infty}{\approx} 2b\beta_1 t^{\beta_1-1} + \underbrace{\sum_{i,j=1}^{bt^{\beta_1}}\left(1-\frac{1}{t_i t_j}\right)^{t-\max(t_i,\,t_j-1)}\frac{1}{t_i t_j}}_{\mathcal{C}(t)}. \tag{3.25}$$

In order to have an estimate of $\mathcal{C}(t)$, let us approximate the sum with the related integral:

$$\mathcal{C}(t) \underset{t\to\infty}{\approx} \int_1^{bt^{\beta_1}}\int_1^{bt^{\beta_1}}\left(1-\frac{1}{t_x t_y}\right)^{t-\max(t_x,\,t_y-1)}\frac{1}{t_x t_y}\mathrm{d}x\mathrm{d}y. \tag{3.26}$$

This way, using the change of variables $u = t_x = \left(\frac{x}{b}\right)^{\frac{1}{\beta_1}}$, $v = t_y = \left(\frac{y}{b}\right)^{\frac{1}{\beta_1}}$, we get

$$\mathcal{C}(t) \underset{t\to\infty}{\approx} (b\beta_1)^2\int_{\frac{1}{\rho-\nu}}^{t}\int_{\frac{1}{\rho-\nu}}^{t}\left(1-\frac{1}{uv}\right)^{t-\max(u,\,v-1)}\frac{\mathrm{d}u\mathrm{d}v}{(uv)^{2-\beta_1}}, \tag{3.27}$$

where we have substituted the initial value $\left(\frac{1}{b}\right)^{\frac{1}{\beta_1}} = \frac{1}{\rho-\nu}$, since $b = (\rho-\nu)^{\beta_1}$ when $\nu < \rho$, that is in this case [8]. Moreover, considering that u and v represent time variables, with $\tau \in (1, t)$, since between $t = 0$ and $t = 1$ there are no colors extracted yet, we can change the lower integral border to 1, i.e.,

$$\mathcal{C}(t) \underset{t\to\infty}{\approx} (b\beta_1)^2\int_1^{t}\int_1^{t}\left(1-\frac{1}{uv}\right)^{t-\max(u,\,v)}\frac{\mathrm{d}u\mathrm{d}v}{(uv)^{2-\beta_1}}, \tag{3.28}$$

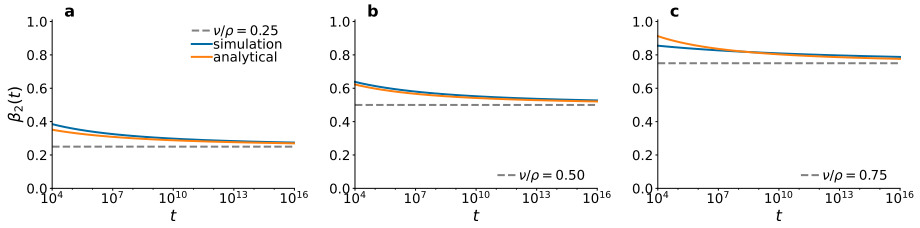


Figure 3.10: **2nd-order Heaps' exponent in the urn model with triggering.** Temporal evolution of the 2nd-order Heaps' exponents $\beta_2(t)$ of the urn model with triggering according to the simulations (continuous blue line) and the numerical integration of Eq. (3.25) (continuous orange line). Parameters are set to $\rho = 4$ and $\eta = 1$, while ν varies across panels: $\nu = 1$ (**a**), 2 (**b**), and 3 (**c**). Continuous lines are obtained by fitting $D_2(t)$ with a function $at^{\beta_1 + c/(d+\log(t))}$, as in Eq. (3.30). The expected 1st-order Heaps' exponent in each panel, respectively equal to $\beta_1 = \nu/\rho = 0.25, 0.5, 0.75$, is displayed as a dashed gray horizontal line.

where we have also simplified the exponent in the integrand, so that we can more easily calculate it as

$$\mathcal{C}(t) \underset{t\to\infty}{\approx} 2\,(b\beta_1)^2 \int_1^t \int_1^u \left(1 - \frac{1}{uv}\right)^{t-u} \frac{\mathrm{d}u\mathrm{d}v}{(uv)^{2-\beta_1}}. \tag{3.29}$$

We numerically solve the integral in Eq. (3.29) on specified points t_i using the command `NIntegrate` of *Mathematica* [141]. The points t_i have been chosen on a fine logarithmically spaced grid of $N = 1601$ points $1 = t_0 < t_1 < \cdots < t_N = 10^{16}$. By plugging the numerical approximation $\mathcal{C}(t_i)$ into Eq. (3.25), we also obtain a numerical approximation of $\mathrm{d}D_2/\mathrm{d}t$ in these points. We also obtain an analytical approximation of $\mathrm{d}D_2/\mathrm{d}t$ by fitting a function of the type $at^{b+c/(d+\log_2(t))}$ using `curve_fit` (in *Python*'s package `scipy`), where the minimization of the error has been done in logarithm scale. Finally, integrating Eq. (3.25) over t, we obtain a solution for $D_2(t)$. Again, we are not able to solve this integral analytically, so we solve it numerically using the analytical fit of $\mathrm{d}D_2/\mathrm{d}t$. In particular, we integrate using *Python*'s command `odeint` in the `scipy` package. We find that the numerical integration for $D_2(t)$ can also be fitted by a function of the type $at^{\beta_1 + c/(d+\log_2(t))}$.

To sum up, we have derived a solution of Eq. (3.25) of the type

$$D_2(t) \approx a\, t^{\beta_2}, \qquad \text{with } \beta_2 = \beta_1 + \frac{c}{d + \log(t)}, \tag{3.30}$$

where a, c, and d depend on the parameters ρ and ν. Fig. 3.10 shows that the analytical expression of β_2 we have found is in good agreement with the numerical simulations. From left to right, we consider parameters $\rho = 4$ and $\nu = 1, 2, 3$, and we run simu-

lations until $T = 10^7$. In each plot, continuous lines represent the 2nd-order Heaps' exponents of the power-law fits as a function of time t. The continuous blue line is obtained by fitting the best parameters a, c and d that minimize the error between the points $D_2(t)$ of the simulations with a function of the type $a\, t^{\beta_1 + \frac{c}{d+\log(t)}}$. The continuous orange line instead represents the result of our analytical approach in Eq. (3.30). The expected value of $\beta_1 = \nu/\rho$ is represented as a horizontal dashed gray line. Our results further confirm that 2nd-order Heaps' exponents differ from the 1st-order ones at finite times. However, they also highlight that in the UMT the difference between β_2 and β_1 slowly decays in time.

Higher-order Heaps' law

Finally, we point out that an analytical solution for higher-order Heaps' exponents can also be obtained by induction, with assumptions similar to those used for the 2nd-order one. For example, for the 3rd-order, we can repeat the same process as in Sec. 3.6.1 to compute the probabilities of obtaining a new triplet. In particular, supposing that

$$\frac{\mathrm{d}D_1(t)}{\mathrm{d}t} \approx a_1 t^{\beta_1 - 1}, \quad \frac{\mathrm{d}D_2(t)}{\mathrm{d}t} \approx a_2 t^{\beta_1 - 1 + \frac{c_2}{d_2 + \log_2(t)}}, \tag{3.31}$$

we can obtain a new triplet in the three following distinct cases.

($\mathbb{A}$): when at time $(t-1)$ a new pair is drawn, which happens with probability

$$\mathbb{P}(\mathbb{A}) = \frac{\mathrm{d}D_2(t-1)}{\mathrm{d}t} \approx a_2 t^{\beta_1 - 1 + \frac{c_2}{d_2 + \log_2(t)}}. \tag{3.32}$$

($\mathbb{B}$): when at time $(t-1)$ an old pair is drawn, and at time t a new color is drawn, which happens with probability

$$\begin{aligned} \mathbb{P}(\mathbb{B}) &= \left(1 - \frac{\mathrm{d}D_2(t-1)}{\mathrm{d}t}\right) \frac{\mathrm{d}D_1(t)}{\mathrm{d}t} \\ &\approx \left(1 - a_2 t^{\beta_1 - 1 + c_2/(d_2 + \log_2(t))}\right) a_1 t^{\beta_1 - 1} \approx a_1 t^{\beta_1 - 1}. \end{aligned} \tag{3.33}$$

($\mathbb{C}$): when at both times $(t-1)$ and t an old pair and an old color are extracted, but the corresponding triplet has never appeared in the sequence before. Following

the same steps of the 2nd-order case, we get the probability

$$\begin{aligned}\mathbb{P}(\mathbb{C}) &\approx \sum_{i,j,k=1}^{bt^{\beta_1}} \left(1-\frac{1}{t_i t_j t_k}\right)^{t-\max(t_i,t_j,t_k)} \frac{1}{t_i t_j t_k} \\ &\approx (a_1)^3 \int_1^t\int_1^t\int_1^t \left(1-\frac{1}{uvw}\right)^{t-\max(u,v-1,w-2)} \frac{\mathrm{d}u\mathrm{d}v\mathrm{d}w}{(uvw)^{2-\beta_1}} \\ &\approx 3!\,(a_1)^3 \int_1^t\int_1^u\int_1^v \left(1-\frac{1}{uvw}\right)^{t-u} \frac{\mathrm{d}u\mathrm{d}v\mathrm{d}w}{(uvw)^{2-\beta_1}},\end{aligned} \tag{3.34}$$

where $3! = 3 \cdot 2 \cdot 1$.

Then, summing up Eq. (3.32), Eq. (3.33) and Eq. (3.34), the probability to have a new triplet can be approximated as

$$\begin{aligned}\frac{\mathrm{d}D_3(t)}{\mathrm{d}t} \approx a_2 t^{\beta_1-1+\frac{c_2}{d_2+\log_2(t)}} + a_1 t^{\beta_1-1} \\ + 3!\,(a_1)^3 \int_1^t\int_1^u\int_1^v \left(1-\frac{1}{uvw}\right)^{t-u} \frac{\mathrm{d}u\mathrm{d}v\mathrm{d}w}{(uvw)^{2-\beta_1}}.\end{aligned} \tag{3.35}$$

In general, for the n^{th}-order Heaps' law, let us suppose by induction that all lower orders are known, i.e., for all orders $k = 1, \ldots, n-1$, with $n \geq 2$, we have

$$\frac{\mathrm{d}D_k(t)}{\mathrm{d}t} \approx a_k t^{\beta_1-1+\frac{c_k}{d_k+\log_2(t)}}, \quad D_k(t) = \tilde{a}_k t^{\beta_1+\frac{c_k}{d_k+\log_2(t)}}, \tag{3.36}$$

with $a, c, d > 0$. Then, following the same procedure used for the 3rd-order Heaps' law, the probability of extracting a new n-tuple is given by:

$$\begin{aligned}\frac{\mathrm{d}D_n(t)}{\mathrm{d}t} \approx \frac{\mathrm{d}D_{n-1}(t)}{\mathrm{d}t} + \frac{\mathrm{d}D_1(t)}{\mathrm{d}t} \\ + n!\,(a_1)^n \underbrace{\int_1^t\int_1^{u_1}\cdots\int_1^{u_{n-1}}}_{n \text{ integrals}} \left(1-\frac{1}{u_1\cdots u_n}\right)^{t-u_1} \frac{\mathrm{d}u_1\cdots\mathrm{d}u_n}{(u_1\cdots u_n)^{2-\beta_1}},\end{aligned} \tag{3.37}$$

which approximately gives

$$\frac{\mathrm{d}D_n(t)}{\mathrm{d}t} \approx a_n t^{\beta_1-1+\frac{c_n}{d_n+\log_2(t)}}, \quad D_n(t) = \tilde{a}_n t^{\beta_1+\frac{c_n}{d_n+\log_2(t)}}. \tag{3.38}$$

3.6.2 Comparison between analytical results and simulations of the UMT and the UMST

According to analytical results on the asymptotic Heaps' exponent found in the previous section, we have that the 1st-order Heaps' exponent β_1 is asymptotically ν/ρ. In this section, we check if this relation holds true at finite times in Fig. 3.11(**a**), where we show the scatter plots between ν/ρ and the fitted value of β_1 for simulations of the UMT with $\rho = 20$ and $\nu = 1, \ldots, 20$, run for $T = 10^5$ time steps. Each point refers to a different simulation, and we analyze 100 simulations for each set of parameters. We notice how the relationship holds true in most cases, although the fitted values are less than the theoretical ones, especially for high values of ν/ρ. We repeat this check also for higher-order Heaps' exponents in Fig. 3.11(**b-c**), finding that also in this case there is not so much difference between the theoretical value ν/ρ and the fitted β_2 and β_3, if only that the points in the plot are slightly higher than the bisector.

We repeat the same analysis for the Urn Model with Semantic Triggering (UMST), which introduces semantic triggering between the colors in the urn. In particular, two colors are considered semantically related if they have been triggered by the same color (siblings) or if one has triggered the other (parent and child). Then, whenever a

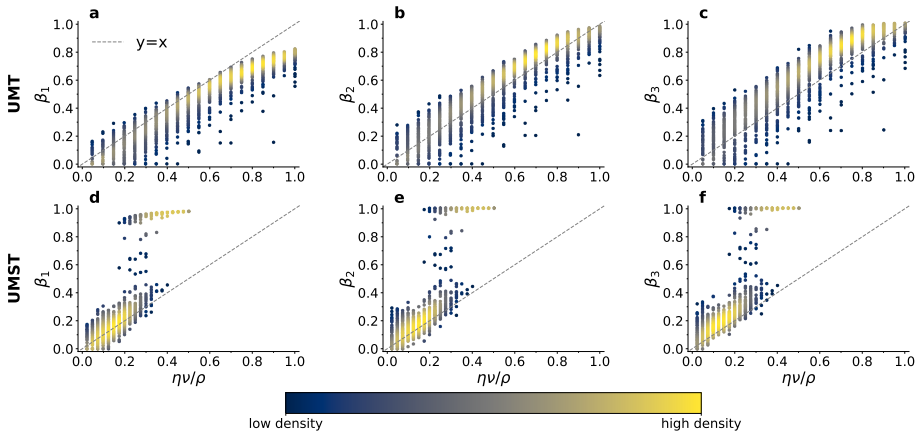


Figure 3.11: **Higher-order Heaps' exponents and their correlations with the expected asymptotic value in urn model simulations.** Scatter plots between the analytically expected lower bound $\eta\nu/\rho$ for the asymptotic 1st-order Heaps' exponent—the theoretical upper bound being $\min(1, \nu/\rho)$—and the Heaps' exponents β_1 (**a**), β_2 (**b**), β_3 (**c**), β_4 (**d**). Each point refers to a different simulation of 10^5 time steps, colored according to the density of points (see color bar). Each panel reports the value of the correlation coefficient r. The first row refers to the Urn Model with Triggering (UMT), with no semantic correlations ($\eta = 1$) and $\rho = 20$. The second row refers to the Urn Model with Semantic Triggering (UMST), with $\rho = 4$ and $\eta = 0.1$. Here, we show the results of 100 simulations for each set of parameters.

η	ρ	ν	$\beta_1 \approx \eta\nu/\rho$	$\beta_1 \approx 1$
0.1	4	1	100	0
0.1	4	2	100	0
0.1	4	3	100	0
0.1	4	4	100	0
0.1	4	5	100	0
0.1	4	6	100	0
0.1	4	7	99	1
0.1	4	8	91	9
0.1	4	9	80	20
0.1	4	10	51	49
0.1	4	11	44	56
0.1	4	12	21	79
0.1	4	13	12	88
0.1	4	14	0	100
0.1	4	15	0	100
0.1	4	16	1	99
0.1	4	17	0	100
0.1	4	18	0	100
0.1	4	19	0	100
0.1	4	20	0	100

Table 3.3: **Statistics on the number of simulations of urn models.** Number of simulations of the UMST ($\eta = 0.1$, $\rho = 4$, $\nu = 1, \ldots, 20$) that have an exponent approximately equal to the lower bound $\eta\nu/\rho$ or to the upper bound 1. For each set of parameters, 100 simulations have been launched.

new color needs to be extracted, a ball of a certain color in the UMST has a different weight depending on the semantic relationship with the previous color. If the two colors are related, then the ball has weight 1, otherwise it gets weight $\eta \leq 1$. Analytical results on the Heaps' law from the SI in Ref. [8] show that the asymptotic Heaps' exponent is found between $\eta\nu/\rho$ and $\min(1, \nu/\rho)$. We test this in Fig. 3.11(**d**), where we show the scatter plots between $\eta\nu/\rho$ and the fitted value of β_1 for simulations of the UMST with $\rho = 4$ and $\nu = 1, \ldots, 20$. We see that for low values of the $\eta\nu/\rho$ the value of β_1 corresponds to the theoretical lower bound. However, starting from about $\eta\nu/\rho = 0.2$ there start to be simulations in which the value of β_1 goes abruptly up to 1. Notice that for these values, we have that $\nu/\rho = 2$. Interestingly, up to $\eta\nu/\rho = 0.3$ and sometimes up to $\eta\nu/\rho = 0.4$, there are both simulations with Heaps' exponent $\beta_1 \approx \eta\nu/\rho$ and others with $\beta_1 \approx 1$, but almost none in between. After that, there remain only simulations with linear Heaps' law. We repeat the analysis for higher-order Heaps' exponents, finding the same behavior.

In Table 3.3 we also report the number of simulations with either of the two behaviors. Notice how the number of simulations with $\beta_1 \approx 1$ increases with higher

values of ν. This analysis shows the inadequacy of the UMST to reproduce the whole spectrum of paces of discovery. In fact, we are not able to obtain Heaps' exponents between 0.4 and 0.9 with $\eta = 0.1$. Moreover, if we knew that the Heaps' exponent lies in between these two bounds, simulations actually only produce exponents very close to these two bounds. The higher the theoretical value $\eta\nu/\rho$, the higher the chance of having a Heaps' exponent close to 1. A possible explanation of why this could happen lies on the way semantic triggering happens. In the UMST, indeed, when a color is drawn for the first time, $\nu+1$ balls of new colors are added to the urn, and they become semantically connected to the triggering color. Then, the probability to draw a ball of a color semantically close to the previous one is $1/\eta = 10$ times higher with respect to balls of other colors. This brings about two possible scenarios. On the one hand, if a small cluster of colors is highly reinforced in the beginning of the simulation, after one of them is drawn it is very likely that another of these colors is extracted in the next time step. On the other hand, if a new color is drawn, since it is highly probable to move to a semantic close color and almost all of them are new, if ν is high enough the next extracted color is also almost surely new. Then, once inside one of the two scenarios, it is very unlikely to break the loop, producing the two groups of Heaps' exponent we observe. This also explains why the likelihood of being in the linear case increases with ν, even though the two behaviors can coexist in the same set of parameters. Finally, this is also confirmed by simulations with higher number steps—we tested with 10^7 steps—, which show the same results, indicating that the behavior has already reached a stationary state.

3.6.3 Analytical details of ERRWT model

In this section we provide an analytical insight of the model proposed in Sec. 3.5, the Edge-Reinforced Random Walk with Triggering, or ERRWT. In particular, we refer to the definition of ERRWT given in Sec. 3.5.1, and try to build differential equations for the evolution of $D_1(t)$ and $D_2(t)$. From now on, we omit the explicit time dependence of the variables, e.g., $D_1 \equiv D_1(t)$, so that the mathematical expressions are easier to read. Moreover, in the following analysis we make an important simplification, that is, we do not consider an undirected update as defined the main text. Undirected update means that at any time a new link (i, j) is reinforced or triggered, the link (j, i) is updated as well; here we consider the directed version of the model, i.e., only the visited link (i, j) is updated.

We start from considering variables referring to the pace of discovery related to each node i, namely $D_{1i}^{in}, D_{1i}^{out}, D_{2i}^{in}, D_{2i}^{out}$, which represent respectively the number of times a new node is discovered *arriving* (in) in node i, and *leaving* (out) from node i, and the same for the number of times a new link is discovered arriving or leaving from node i. Notice that D_{1i}^{in} becomes 1 as soon as node i is visited for the first time.

These micro variables can be aggregated to obtain the macro variables D_1 and D_2, considering either in or out variables and summing over all the nodes:

$$D_1 = \sum_i D_{1i}^{in} = \sum_i D_{1i}^{out} \qquad D_2 = \sum_i D_{2i}^{in} = \sum_i D_{2i}^{out} \tag{3.39}$$

Let us now build differential equations for the evolution of such micro variables, which will be aggregated to obtain self-consistent equations for D_1 and D_2. Let us consider the probability of exploring a new node starting from node i, i.e., the probability that the variable D_{1i}^{out} increases by 1. On the one hand, the total weight of the links outgoing from node i is equal to

$$M_{0i} + \rho n_i + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in}, \tag{3.40}$$

where M_{0i} is the initial number of links connected with node i at time $t = 0$, and $n_i \equiv n_i(t)$ is the number of times node i has been visited up to time t. The other two terms refer to the new links triggered when arriving in node i. Indeed, when i is visited for the first time, $(\nu_1 + 1)$ links outgoing from i to new nodes are triggered. Moreover, whenever a link ending in i is traversed for the first time, $(\nu_2 + 1)$ new links from i to other explored nodes are triggered. On the other hand the total weight of links connecting i and never explored nodes is equal to

$$M_{0i} + (\nu_1 + 1)D_{1i}^{in} - D_{1i}^{out}, \tag{3.41}$$

i.e., the initial number of nodes connected to i yet to be discovered, plus the number of nodes triggered when discovering node i, minus the number of nodes already discovered starting from i. These considerations make possible to write that

$$\frac{dD_{1i}^{out}}{dt} = p(i,t)\frac{M_{0i} + (\nu_1 + 1)D_{1i}^{in} - D_{1i}^{out}}{M_{0i} + \rho n_i + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in}}, \tag{3.42}$$

where $p(i,t)$ is the probability of being on node i at time t, which is a needed condition for D_{1i}^{out} to evolve. Using the same argument, we can also write an equation for the evolution of D_{2i}^{out}:

$$\frac{dD_{2i}^{out}}{dt} = p(i,t)\frac{M_{0i} + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in} - D_{2i}^{out}}{M_{0i} + \rho n_i + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in}} \tag{3.43}$$

Notice that Eq. (3.42) and Eq. (3.43) cannot be obtained so easily in the undirected case. In fact, here we have implicitly assumed that any link in the adjacent possible that has never been visited before has weight 1. However, if the update is undirected, we may reinforce some link (j,i) never traversed before only because the walker might

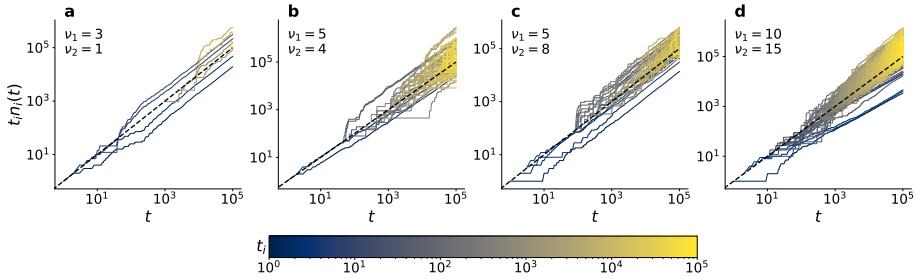


Figure 3.12: **Temporal evolution of the quantities $n_i(t)$.** In these figures we show the temporal evolution of $n_i(t)$, i.e., the number of times node i has been explored at time t, for many choices of the node i. In order to check the assumption $n_i(t) \sim t/t_i$, where t_i is the time when node i is discovered for the first time, we actually plotted $t_i n_i(t)$ vs t. We expect this quantity to go like t, which is represented by the dotted black line. As we can see from the four panels, the assumption is valid for a wide range of the parameters ν_1 and ν_2.

have visited the edge (i, j), making impossible to know the actual weight of never traversed links.

At this point we make another assumption in order to make the equations solvable. In particular, we assume a precise expression for the variable n_i. In fact, as we have seen in Sec. 3.6.1, in the UMT, at least in the sublinear regime, we have $n_i(t) \sim t/t_i$, where t_i is the first time item (node) i has been visited [8]. Exploiting the analogy between the UMT and the ERRWT model, we assume that $n_i(t)$ has the same behaviour. We also checked numerically the validity of this assumption. We have indeed measured the evolution of $n_i(t)$ in simulations of the ERRWT, showing that the assumption is reasonable for very different values of the parameters ν_1 and ν_2, as shown in Fig. 3.12. Further notice that $p(i, t) = dn_i(t-1)/dt \approx 1/t_i$, since the probability of being on i at time t is equal to the probability to move to node i in the previous time step. With all these elements we can rewrite Eq. (3.42) and Eq. (3.43) as

$$\frac{dD_{1i}^{out}}{dt} \approx \frac{1}{t_i} \frac{M_{0i} + (\nu_1 + 1)D_{1i}^{in} - D_{1i}^{out}}{M_{0i} + \rho\frac{t}{t_i} + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in}} \tag{3.44}$$

and

$$\frac{dD_{2i}^{out}}{dt} \approx \frac{1}{t_i} \frac{M_{0i} + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in} - D_{2i}^{out}}{M_{0i} + \rho\frac{t}{t_i} + (\nu_1 + 1)D_{1i}^{in} + (\nu_2 + 1)D_{2i}^{in}}. \tag{3.45}$$

The last step before aggregating the equations is to further simplify the denominator. First notice that D_{1i}^{in} is a variable which can only take values 0 or 1, since an arriving node can result to be new only once (this is not true for D_{1i}^{out}, which can be larger than 1). Therefore, we can neglect it with respect to the term with D_{2i}^{in}, because this can be larger than 1 and can go to infinity with time with a pace dependent on the

parameters as we will see later. Finally, we assume $D_{2i}^{in} \approx D_2/t_i$; this is a reasonable assumption given the fact that $n_i(t) \approx t/t_i$. Indeed, if a node i is visited with a frequency depending on the inverse of t_i, it is reasonable to assume that also the number of new links traversed arriving in node i occurs with the same frequency as well.

We can finally aggregate Eq. (3.44) summing over all nodes i, obtaining a self consistent equation for the evolution of D_1:

$$\begin{aligned}\frac{dD_1}{dt} &= \sum_{i=1}^{D_1} \frac{dD_{1i}^{out}}{dt} \approx \sum_i \frac{1}{t_i} \frac{M_{0i} + (\nu_1+1)D_{1i}^{in} - D_{1i}^{out}}{\rho\frac{t}{t_i} + (\nu_2+1)\frac{D_2}{t_i}} \\ &\approx \sum_i \frac{M_{0i} + (\nu_1+1)D_{1i}^{in} - D_{1i}^{out}}{\rho t + (\nu_2+1)D_2} \\ &= \frac{M_0 + (\nu_1+1)D_1 - D_1}{\rho t + (\nu_2+1)D_2} \approx \frac{\nu_1 D_1}{\rho t + (\nu_2+1)D_2},\end{aligned} \tag{3.46}$$

where in the last approximation we have disregarded M_0 in the numerator, considering that $D_1(t) \to \infty$ is the leading term in the numerator. Similarly for the 2[nd]-order Heaps' law, using Eq. (3.45) we can write

$$\begin{aligned}\frac{dD_2}{dt} &= \sum_{i=1}^{D_1} \frac{dD_{2i}^{out}}{dt} \approx \sum_{i=1}^{D_1} \frac{1}{t_i} \frac{M_{0i} + (\nu_1+1)D_{1i}^{in} + (\nu_2+1)D_{2i}^{in} - D_{2i}^{out}}{\rho\frac{t}{t_i} + (\nu_2+1)\frac{D_2}{t_i}} \\ &= \frac{M_0 + (\nu_1+1)D_1 + (\nu_2+1)D_2 - D_2}{\rho t + (\nu_2+1)D_2} \approx \frac{(\nu_1+1)D_1 + \nu_2 D_2}{\rho t + (\nu_2+1)D_2}.\end{aligned} \tag{3.47}$$

Notice that the initial structure of the network only enters in the equations through the constant $M_0 \equiv \sum_i M_{0i}$, and, as we have already pointed out, this term can be safely neglected with respect to the other variables. This means that the asymptotic behaviour of D_1 and D_2, and so the exponents β_1 and β_2, should not depend on the initial structure of the network. We have hence run simulations with different initial conditions and measured the exponents β_1 and β_2, checking that we obtain similar result in all cases. The results of this analysis is shown in Fig. 3.13. In particular, we consider regular trees with different number of levels, but same branching size. The idea is that we start from a node and trigger nodes, adding new levels of the tree. Therefore, the first initial network we consider is made by a root, considered triggered, connected to $\nu_1 + 1$ new nodes. The second one adds another level to the first one. Therefore, it is a regular tree with branching size $(\nu_1 + 1)$ and 2 levels. Here, the root and the first level are considered triggered, while the leaves are still new. This structure is the same used in the simulations of Sec. 3.5.2 (see also Fig. 3.7). Finally, the third one adds one more level, thus being much bigger than the previous ones. In the panels we show only two sets of parameters, but we find comparable results for other choices of the parameters. For both sets of parameters, we find that by increasing the number of levels (and hence the number of nodes and links) in the initial network,

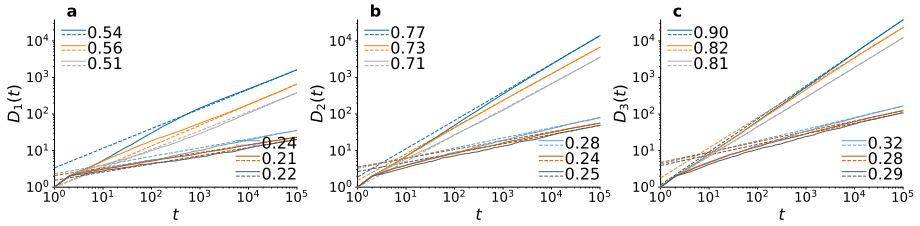


Figure 3.13: **Heaps' exponents for different choices of the initial conditions.** The three panels show the behaviour of $D_1(t)$, $D_2(t)$ and $D_3(t)$ versus t for three different initial conditions (i.e., initial structure of the network). In particular, we consider a regular tree with branching parameter $\nu_1 + 1$ and number of levels equal to 1 (gray lines), 2 (orange lines), 3 (blue lines). All nodes apart from the leaves are considered known (or discovered, or triggered) by the ERRWT at the beginning of the simulation. In each panel, the lines with higher Heaps' law (see top left legend), refer to simulations with $\rho = 10$, $\nu_1 = 8$, and $\nu_2 = 8$, while the other lines (see bottom right legend) with $\rho = 10$, $\nu_1 = 3$, and $\nu_2 = 1$. In the legend, the related extracted power-law exponent is reported. As we can see, the exponents measured in the three cases are similar across order, thus showing that the initial structure of the network is not relevant for the asymptotic behaviour of the ERRWT.

the higher-order Heaps' exponents slightly increase. Moreover, the bigger the network, the longer we see a transient time in which there is a much higher Heaps' exponent. For example, see the blue line in Fig. 3.13(**a**), where we can clearly find the initial higher slope. Nevertheless, notice that after this period, the pace of discovery, i.e., the exponent, seems to be similar across different initial conditions, thus showing that the initial structure of the network is not relevant for the asymptotic behaviour of the ERRWT.

Now, using Eq. (3.46) and Eq. (3.47), we are able to work out an analytical expression for the two exponents β_1 and β_2. Let us consider various cases. First, assuming a sublinear regime for D_2 (so that it can be neglected with respect to t), in the large time limit we can further simplify the equations and get

$$\frac{dD_1}{dt} \approx \frac{\nu_1 D_1}{\rho t}, \tag{3.48}$$

$$\frac{dD_2}{dt} \approx \frac{(\nu_1 + 1)D_1 + \nu_2 D_2}{\rho t}. \tag{3.49}$$

Solving both of these equations, we obtain an explicit expression for the two exponents β_1 and β_2; in the sublinear case we hence have asymptotically

$$\beta_1 = \frac{\nu_1}{\rho} \quad , \quad \beta_2 = \max\left(\frac{\nu_1}{\rho}, \frac{\nu_2}{\rho}\right) \tag{3.50}$$

From the expression of β_1 and β_2 in Eq. (3.50), we get that the sublinear regime holds only if $\nu_1 < \rho$ and $\nu_2 < \rho$.

Before moving on to the other regimes, let us notice that β_2 is constrained to be at most equal to $2\beta_1$. This is because if the number of nodes available to explore in the network is $O(N)$, then the number of available edges is $O(N^2)$. This means that D_2 can at most grow as the square of D_1 in the large time limit, imposing a constraint on the related exponents. Let us now consider the case in which D_2 grows linearly in time, but not D_1. Notice that this can happen only provided that $\nu_1 > \rho/2$; in fact, since β_2 is constrained to be smaller or equal than $2\beta_1$, then it would not be possible for β_2 to be equal to 1 if $\beta_1 = \nu_1/\rho < 1/2$. This regime can be obtained substituting a linear expression for $D_2 \sim at$ into Eq. (3.47). In this case, if we assume a sublinear behaviour for D_1, we can neglect the second term in the numerator, obtaining

$$\frac{dD_2}{dt} \approx \frac{\nu_2 at}{\rho t + (\nu_2+1)at} = \frac{\nu_2 a}{\rho + (\nu_2+1)a} = a \quad \Longrightarrow \quad a = \frac{\nu_2 - \rho}{(\nu_2+1)}, \tag{3.51}$$

thus showing that the condition for this regime to exist is $\nu_2 > \rho$, otherwise the coefficient a would be negative. Then, we can substitute $D_2 = \frac{\nu_2-\rho}{(\nu_2+1)}t$ into Eq. (3.46), to get the actual value of β_1:

$$\frac{dD_1}{dt} \approx \frac{\nu_1 D_1}{\rho t + (\nu_2 - \rho)t} = \frac{\nu_1 D_1}{\nu_2 t} \quad \Longrightarrow \quad \beta_1 = \frac{\nu_1}{\nu_2}. \tag{3.52}$$

Therefore, D_1 keeps growing sublinearly provided that $\nu_1 < \nu_2$. Notice that in this case there are no conditions on the value of ν_1, which can also be larger than ρ. Reminding also the network constraint $\beta_2 \leq 2\beta_1$, we have that this regime holds provided that $\beta_1 > 1/2$, which means $2\nu_1 > \nu_2$.

Finally, there is one last regime, in which both D_1 and D_2 are linear, i.e., with exponents $\beta_1 = \beta_2 = 1$. Substituting the two linear expressions $D_2 \sim at$ and $D_1 \sim bt$ in Eq. (3.46) and Eq. (3.47), we obtain the following system of equations

$$\begin{cases} \frac{dD_1}{dt} \approx \frac{\nu_1 b}{\rho + (\nu_2+1)a} = b \\ \frac{dD_2}{dt} \approx \frac{(\nu_1+1)b + \nu_2 a}{\rho + (\nu_2+1)a} = a \end{cases} \tag{3.53}$$

from which we can work out the values of the two coefficients:

$$a = \frac{\nu_1 - \rho}{(\nu_2+1)} \qquad b = \frac{(\nu_1-\rho)}{(\nu_2+1)}\frac{(\nu_1-\nu_2)}{(\nu_1+1)}, \tag{3.54}$$

which give the conditions $\nu_1 > \rho$ and $\nu_1 > \nu_2$ for this regime to hold. This comes out from the fact that, as we have seen before, whenever $\nu_2 > \nu_1$ we have a sublinear regime for D_1.

Summarizing the predicted exponents for the directed version of the model for any choice of the parameter ν_1, ν_2 and ρ, we have:

$$\begin{cases} \nu_2 < \rho,\ \nu_1 < \rho & \beta_1 = \frac{\nu_1}{\rho},\ \beta_2 = \min\left(\max\left(\frac{\nu_1}{\rho}, \frac{\nu_2}{\rho}\right), \frac{2\nu_1}{\rho}\right) \\ \nu_2 \geq \rho,\ \nu_1 \leq \frac{\rho}{2} & \beta_1 = \frac{\nu_1}{\rho},\ \beta_2 = \frac{2\nu_1}{\rho} \\ \nu_2 \geq \rho,\ \frac{\rho}{2} < \nu_1 < \nu_2 & \beta_1 = \frac{\nu_1}{\nu_2},\ \beta_2 = 1 \\ \nu_1 \geq \rho,\ \nu_1 \geq \nu_2 & \beta_1 = \beta_2 = 1 \end{cases} \tag{3.55}$$

The results above give us an analytical overview of a simplified version of the model, which can still provide the phenomenology we are interested in. In fact, with this analysis we still obtain a different behaviour for D_1 and D_2 with two different Heaps' exponents β_1 and β_2, which are controlled by the parameters ν_1 and ν_2, given ρ. In particular, when $\beta_2 < 1$, we have that β_1 is asymptotically ν_1/ρ, similarly to the UMT for which the Heaps' exponent is ν/ρ. In the same regime, β_2 depends instead on ν_2/ρ. Notice that if $\nu_2 < \nu_1$, then the 2$^{\text{nd}}$-order Heaps' exponent becomes equal to the 1$^{\text{st}}$-order one. Finally, when $\beta_2 = 1$, the value of β_1 depends on the relative value of ν_1 with respect to ν_2. For higher values of the triggering parameters, instead, Heaps' exponents at both orders are equal to 1.

3.7 Summary and conclusions

As we have seen in Chapter 2, the increasing wealth of data related to innovation processes has inspired various models of innovation, trying to uncover what are the mechanisms that drive such processes. Based either on extractions from urns or on random walks over complex networks, these models consider innovation as the exploration of a space of possibilities, where the elements represent ideas, concepts, artworks or other types of content that can be explored. The two key mechanisms of these models are the reinforcement of the elements explored and the triggering of new elements, or adjacent possibilities, whenever a discovery is made. Thanks to these mechanisms, such models reproduce the empirical Heaps' law observed in the empirical data, where the reinforcement pushes the explorer to exploit past discoveries, while the triggering of the adjacent possible pushes towards new ones.

However, there is more and more evidence that novelties can arise by combining existing elements [108, 109, 115, 117, 118]. In this chapter we have hence proposed the *higher-order Heaps' laws*, and their exponents, as a measure for the pace of new combinations realised in a system. In particular, we regard a novelty not only as the discovery of new items, but also as the first appearance of a new combination of different items. Notice how higher-order Heaps' laws differ from other measures for the pace of discovery that have been developed in the last years. For example, in Refs. [96,

142], the authors have used the number of all possible valid combinations that can be created using the elements so far acquired as a proxy of the level of innovation reached by the system. However, this does not take into account the actual number of novelties realised in the system, but rather their potential.

In Sec. 3.3 we have analyzed empirical data from different systems, discovering that higher-order Heaps' exponents can be used to distinguish users listening to music in Last.fm who feature a similar discovery rate of new songs and artists. The higher-order Heaps' exponent can indeed tell apart different ways to explore the same set of songs in terms of number of different consecutive pairs or higher-order structures explored. Analogously, we notice different patterns in texts of various nature by studying their Heaps' exponents at various orders: titles of peer-reviewed papers published in scientific journals show more creative juxtaposition of words with respect to the text of narrative books, encountering many more new n-grams, even if the total set of words used is similar in length. Overall, our analysis shows that, no matter the context, the space of possibilities grows in more complex ways than what was previously theorized. In fact, it does not depend solely on the balance between old items to exploit and new ones to explore, but also on the structure of their associations.

Then, in Sec. 3.4 we have checked if the existing models of discovery, from the urn model with triggering [8] to the edge-reinforced random walk [29], can also reproduce higher-order Heaps' exponents similar to the empirical data. On the one hand, these models are able to reproduce different behaviors in terms of 1^{st}-order Heaps' exponents. On the other hand, however, we find that they are not able to reproduce higher-order ones. As a matter of fact, they can only reproduce the same rate of discovery at all orders, therefore not considering the complex evolution of the space of possibilities and, in particular, the adjacent possible. In other words, this analysis manifests the need for a new generation of exploitation-exploration models based on the co-evolution of the network structure with the dynamical process of exploration.

In Sec. 3.5, we have thus proposed a new modelling framework, called the Edge-Reinforced Random Walk with Triggering, or ERRWT, which takes into account not only the exploration rate of new items, but also the predisposition to explore the same content in a more creative way. With this model, we imagine the process of innovation as a random walk over a growing complex network of ideas or other contents. In this model, we include the same reinforcement and triggering mechanisms of the previous models, but keeping into account the more general definition of novelty proposed in this chapter. Specifically, we assume that a novelty of a higher order also triggers new adjacent possibilities of the same order. For example, the exploration of a new link, also triggers the addition of new links in the space of possibilities. Overall, based on the reinforcement of links in a complex network and the triggering of new nodes and links whenever new parts of the adjacent possible space are explored, the mechanisms introduced give a new intuition of how the space of possibilities grows over time, shedding

light on how novel elements and combinations emerge along the innovation process.

We acknowledge there are multiple venues of improvement of the model we have proposed in this chapter. For example, the initial knowledge of the network to be explored could have an impact on the exploration process. Moreover, we have supposed that links start with unitary weight, but this can be an unrealistic assumption in certain contexts. Furthermore, we have assumed to trigger new links uniformly at random. This is a reasonable assumption when all elements and associations are intrinsically the same; in this case, the choice to explore a part or another of the space is left to the specific explorer, driven by stochastic events. Nevertheless, there are other cases in which there can be some preferential pathways in the space of possibilities. These can be attributed to the specific individual represented by the random walker, but can also be inherent in the space. For example, there can be an underlying structure that can be discovered through subsequent triggering of the adjacent possible. This could be implemented in our model by limiting the addition of new links to only those permitted by the underlying network, or adding more complex ways to trigger edges, e.g., using preferential attachment [37, 143]. Additionally, in this chapter we have not considered the presence of semantic correlations in the temporal sequence of visited items, which can be a consequence of the interplay between the network topology and a predisposition to move to items semantically close to the recent ones, reinforcing a clustered structure. All such limitations will be taken into account in Chapter 5, where we will model the exploration of the music space by expanding a space of possibilities through an underlying universal weighted network of all existing artists. Moreover, we will consider how semantic correlations in the sequence of elements explored affect the exploration of the content space.

Finally, so far we have considered innovation as an individual process of exploration of a complex space of possibilities. However, human progress is intrinsically a collective and social process. In fact, we learn the language and culture of our family when we are kids, we go to school to learn various disciplines, full of discoveries made along the history, and we keep discovering new things thanks to our peers. We meet new people as friends of friends, we listen to new songs or read new books suggested by our friends. Similarly, researchers bring their minds together to collaborate and make important discoveries, contributing to the growth of the collective knowledge [41]. Therefore, in Chapter 4 we take a step towards this direction, adding collaborative interactions between multiple individuals, each represented by an urn model with triggering. Specifically, we will develop another mechanism capable of expanding the space of possibilities in a system, which comes from the expansion of the adjacent possible in the social space. Then, in Chapter 5 we will propose a data-driven model of collective exploration of music, where we extend the exploration of a network of contents to a multi-agent context, including the adjacent possible expansion in the social space.

Chapter 4

The adjacent possible in the social space

4.1 Introduction and outline

Discoveries are essential milestones for the progress of our societies [95, 96, 144–152]. As we have seen in Chapter 2, different mathematical approaches have been recently proposed to model the dynamics of innovation [20, 30, 41, 68, 69, 107, 153–158]. Among these, of particular interest are those based on random processes with reinforcement [39, 159, 160], from basic Pòlya urns [24, 77] to more complicated urn or random walk models [8, 29], as discussed in Chapter 2.

Urns have been extensively used to study and model a variety of systems and processes, from evolutionary economics, voting and contagions [61, 161–163] to language and folksonomies [164, 165]. More recently, they have been employed to filter information [166] and grow social networks [30] (see Sec. 2.5.1). Interestingly, urns can also be used to model discovery processes, if opportunely combined with the concept of the *adjacent possible* (AP)—*the set of all those things which are one step away from what is already known* (Kauffman [13]). This formulation of the AP, which dates back to concepts previously introduced by Farmer, Langton and others [9–11], has been translated into the urn model with triggering (UMT), a particular process in which the space expands together with the discovery dynamics, and the appearance of a novelty opens up the possibilities of further discoveries [8, 25, 28, 106, 149] (see more details in Sec. 2.2.4). UMTs could successfully replicate the basic signatures of real-world discovery processes [8, 25–27], such as the Heaps' law [14, 15] and Zipf's law [55–57], often recurrent in complex systems [68, 167–172], as well as Taylor's law [82, 83]. Moreover, in Chapter 3 we have further extended our understanding and modelling of the AP in the content space by considering novel combinations of items as novelties.

The analysis of higher-order Heaps' laws, that is the pace of discovery of such combinations in a sequence of exploration, has revealed how complex the structure of this space is, being shaped by the exploration itself.

However, despite the fact that existing models can capture essential underlying mechanisms behind the discovery of novelties, little emphasis is given to the collective dynamics of exploration and to the benefits that social interactions could bring [173]. In fact, with the exception of Ref. [30], the modeled exploration dynamics refers to a single entity, representing, for example, the joint effort of researchers within a field [29]. Without taking into account the multiagent 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 [174–176], such as music listening, politics, voting, and language [177–179].

In this chapter, we propose a model of interacting discovery processes, named the *UrNet* model [1], where an explorer is associated to each of the nodes of a social network [32, 33, 36], and its dynamics is governed by an UMT. Hence, the local dynamics of each node accounts for the presence of an AP, more precisely the *adjacent possible in the content space*, as discussed in Sec. 2.2.4 [8]. Urns are then coupled through the links of a network so that each exploration process is also subjected to interactions with the processes of the neighboring nodes, and explorers can exploit opportunities (possible discoveries) coming from their social contacts, thus increasing their discovery possibilities in a cooperative way. In other words, 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 [35, 52]. Notice, however, that our setting crucially differs from the typical approach in which the network mediates, for example, the diffusion of innovations or social contagions [180, 181]. Here, 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.

This chapter is organized as follows. We start in Sec. 4.2 where the UrNet model is described, providing the analytical framework for the Heaps' laws in Sec. 4.3. Then, we explore the main features of the proposed dynamics on some small networks with extensive numerical simulations in Sec. 4.4, finding that the pace of innovation of an explorer strongly correlates with its position in the social network. In Sec. 4.5 we explore analytically the impact of the social network on the Heaps' laws of the agents, finding a general asymptotic approximation of the Heaps' laws enlightening the nu-

merical results previously discussed. Moreover, in Sec. 4.4.3 we analytically identify the key role played by node centrality measures on the discovery potential of the individuals. Finally, in Sec. 4.6 we discuss the obtained results and future perspectives.

4.2 UrNet: a model of coupled urns

Let us consider in the most general case 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 or agent, while the link (i, j) between two agents i and j denotes the existence of a directed social relation from individual i to j (such that i can benefit from j). The graph is fully 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.

In our UrNet model, each node i of the network represents an explorer, contributing to the collective process of discovery. Considering the ability of the urn model with triggering (UMT) to replicate Heaps' laws and to be analytically tractable (see Sec. 2.2.4 and Sec. 3.6.1), we equip each node i with an urn obeying to the rules of the UMT, describing the discovery process of the agent i [8]. We indicate the urn i at time t as $\mathcal{U}_i(t)$, 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 first initialized with M_0 balls of different colors, and its dynamics is characterized by two parameters, ρ_i and ν_i. As in the original UMT, the *reinforcement* parameter ρ_i accounts for the number of balls of the same color that are added to the urn i whenever a ball of a given color is extracted at time t. Furthermore, the *triggering* parameter ν_i controls the size of the *adjacent possible in the content space*, as $(\nu_i + 1)$ balls of new colors are added to the urn of node i whenever at time t a color is extracted for the first time [8]. In this abstract representation, the space of possibilities—made by all the colors—expands in time together with each discovery process, without relying on a predefined structure [25]. The discovery processes of different individuals are then coupled through the links of the network G, representing social interactions, from which the name of the model *"UrNet"*. Namely, at each time t, the individual i draws a ball from an enriched urn $\tilde{\mathcal{U}}_i(t)$, the so-called *social urn* of node i, composed by its own urn $\mathcal{U}_i(t)$ plus the additional balls present at time t in the urns of its neighbors, without their reinforcement. The latter represents the *adjacent possible in the social space*. Figure 4.1 illustrates the case of two nodes with a directed link. We thus have

$$\tilde{\mathcal{U}}_i(t) = \mathcal{U}_i(t) + \bigcup_{j \in \mathcal{N}} a_{ij} \mathcal{U}'_j(t) \tag{4.1}$$

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

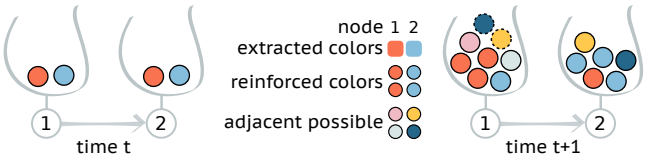


Figure 4.1: Illustration of the UrNet model of interacting urns in the case of a network with two nodes. Each node is equipped with an urn obeying the UMT with the same parameters $\rho = 2$, $\nu = 1$, and $M_0 = \nu + 1$. At the time t, the urns start with two balls, one red (R) and the other blue (B). Then, each node extracts a ball (1:R, 2:B), and therefore ρ additional balls of the same colors 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 colors are also added (adjacent possible). At $t + 1$, 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 neighbor (dashed borders).

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. Therefore, 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 urns.

Finally, in our model all initial colors and the other colors added by an individual i when triggered by a discovery will be taken from a single predefined set of discoverable balls of different colors. Notice that this set is shared by all the urns so that once a ball is drawn from an urn, it will not be available anymore to the others, except when enlarging the urn through the social adjacent possible (if they are connected).

4.3 The pace of discovery

As previous works have shown [8], 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 that is 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 at the individual level—of each sequence—as a function of time (sequence length), representing the pace of discovery.

We know, from standard results on the UMT analyzed in Sec. 2.2.4, that an isolated urn i follows a Heaps' law, i.e., a power law behavior $D_i(t) \sim t^{\beta_i}$ [15], $D_i(t) = |\mathcal{S}'_i(t)|$ being the number of different elements contained in the sequence $\mathcal{S}_i(t)$ up to time t. Thus, the Heaps' exponent β_i quantifies the speed at which the urn discovers new elements (by definition bounded by $0 < \beta_i \leq 1$).

Let us consider now a node i that interacts through the network. In general, since $D_i(t)$ increases by one whenever a color is extracted for the first time, we can write

$$D_i(t+1) = D_i(t) + P_i^{\text{new}}(t), \tag{4.2}$$

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,

$$P_i^{\text{new}}(t) = \mathbb{P}\left[D_i(t+1) = D_i(t) + 1 | D_i(t)\right], \tag{4.3}$$

which can be expressed as the fraction of discoverable balls over the total number of balls available to node i at time t. Thus, the master equation in Eq. (4.2) leads to an equation for the asymptotic Heaps' dynamics that in the continuous time limit reads

$$\frac{dD_i(t)}{dt} = P_i^{\text{new}}(t) = \frac{|\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}_i'(t)|}{\tilde{U}_i(t)}, \tag{4.4}$$

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).

In the most general case, where each node i is equipped with an UMT(ρ_i, ν_i), the equation for the Heaps' laws of each node $i \in \mathcal{N}$ in Eq. (4.4) can be explicitly written by accounting for all the neighbors that are part of the social urn of node i. This can be done by using the non-zero elements of $\boldsymbol{A}$, so that the number of balls $\tilde{U}_i(t)$ in the social urn of node i at time t reads

$$\begin{aligned} \tilde{U}_i(t) &= U_i(t) + \sum_{j \in \mathcal{N}} a_{ij} \big[M_0 + (\nu_j + 1) D_j(t)\big] \\ &= \rho_i t + \big[M_0 + (\nu_i + 1) D_i(t)\big] + \sum_{j \in \mathcal{N}} a_{ij} \big[M_0 + (\nu_j + 1) D_j(t)\big], \end{aligned} \tag{4.5}$$

where M_0 is the initial number of balls in each urn, or in its more compact form

$$\tilde{U}_i(t) = \rho_i t + \sum_{j \in \mathcal{N}} \big[a_{ij} + \delta_{ij}\big] \big[M_0 + (\nu_j + 1) D_j(t)\big] \tag{4.6}$$

where δ_{ij} stands for the Kronecker delta. Similarly, counting only the number of colors not yet discovered, the numerator in Eq. (4.4) can be written as

$$\begin{aligned} |\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}_i'(t)| &= \sum_{j \in \mathcal{N}} \big[a_{ij} + \delta_{ij}\big] \big[M_0 + (\nu_j + 1) D_j(t)\big] - D_i(t) \\ &= M_0 \sum_{j \in \mathcal{N}} \big(a_{ij} + \delta_{ij}\big) + \sum_{j \in \mathcal{N}} \big[a_{ij}(\nu_j + 1) + \delta_{ij} \nu_j\big] D_j(t). \end{aligned} \tag{4.7}$$

Finally, using Eq. (4.6) and Eq. (4.7), Eq. (4.4) for the asymptotic Heaps' dynamics becomes

$$\frac{dD_i(t)}{dt} = \frac{M_0 \sum_j (a_{ij} + \delta_{ij}) + \sum_j \left[a_{ij}(\nu_j + 1) + \delta_{ij}\nu_j\right] D_j(t)}{\rho_i t + \sum_j (a_{ij} + \delta_{ij}) \left[M_0 + (\nu_j + 1) D_j(t)\right]}. \tag{4.8}$$

Equation (4.8) forms a system of N coupled non-linear ordinary differential equations, with initial conditions $D_i(0) = 0 \ \ \forall i \in \mathcal{N}$, that can be numerically integrated for any network topology $\{a_{ij}\}$.

Notice that if a node i has an out-degree $\sum_j a_{ij} = 0$, its associated Eq. (4.8) reduces to the one of an isolated urn, for which $\tilde{\mathcal{U}}_i(t) = \mathcal{U}_i(t)$. Thus, its Heaps dynamics for $\nu < \rho$ follows $D_i(t) \sim t^{\nu/\rho}$ for $t \to \infty$ [8, 26] (see Sec. 2.2.4).

In Sec. 4.4 we will investigate the pace of discovery of the UrNet model through simulations and numerical results of the equations shown here. The explicit analytical solutions to the equations will instead be discussed in Sec. 4.5.

4.4 Numerical results

In this section, we investigate the behavior of the UrNet model we have introduced in Sec. 4.2. In particular, we simulate the model in the more simplified case where $\rho_i = \rho$ and $\nu_i = \nu$, focusing on the effect of the topology on the discovery dynamics. Moreover, we test the validity of our analytical formalism comparing the pace of discovery introduced in Sec. 4.3 against the simulations. To do this, we rely on small toy graphs to understand the basic mechanisms of the model and also on bigger empirical networks extracted from real-world data sets. Let us first give a brief overview of these data sets.

4.4.1 Description of the data sets

We consider four data sets of real-world networks representing different types of social interactions: the Zachary Karate Club (ZKC) network [182], a network of follower relationships among Twitter users [183], a co-authorship network in Network Science [184], and a collaboration network between jazz musicians [185]. The network of Twitter from the original data set (Ref. [183]) has been reduced by performing a random walk sampling.

Some basic properties of the networks are summarized in Table 4.1, like the total number of nodes N, the total number of links E, the average degree $\langle k \rangle$, and the maximum eigenvalue $\widehat{\mu}$ of the related adjacency matrix. Moreover, we have shown some properties related to the connectivity of the networks. In particular, we distinguished weakly-connected components (CCs) and strongly-connected components (SCCs), because they play an important role in the dynamics under investigation. Therefore we

showed the number of both CCs and SCCs, as well as the size of the respective largest one. As we can see, the networks we have chosen have all very different properties, either in size, average degree, and connection.

Data set	Label	Type	N	E	$\langle k \rangle$	$\widehat{\mu}$
ZKC	(a)	Undirected	34	78	4.6	6.7
Twitter	(b)	Directed	4968	26875	10.8	5.2
NetSci	(c)	Undirected	1589	2742	3.4	19.0
Jazz	(d)	Undirected	198	2742	27.7	40.08

Data set	Label	Type	N. CCs	N. SCCs	Size LCC	Size LSCC
ZKC	(a)	Undirected	1	1	34	34
Twitter	(b)	Directed	1	4164	4968	770
NetSci	(c)	Undirected	396	396	379	379
Jazz	(d)	Undirected	1	1	198	198

Table 4.1: Statistics and properties of the four real-world networks considered (cfr. Fig. 4.6): number of nodes N, number of edges E, average node degree k, maximum eigenvalue $\widehat{\mu}$, number of (weakly-) connected components (CCs), and number of strongly-connected components (SCCs), size of the largest (weakly-) connected component (LCC), and size of the largest strongly-connected component (LSCC).

4.4.2 Numerical simulations on simple graphs

We start exploring the behavior of the UrNet model on the famous Zachary Karate Club network (ZKC) [182]. This is a social undirected network of a university karate club made of 34 nodes and 78 edges studied by Wayne W. Zachary for a period of three years from 1970 to 1972 [182], often used in networks science as a test bed for community detection. In fact, during the study, because of a conflict between the instructor and the administrator the club split, thus forming two separate communities that can be predicted through community detection algorithms. In our modelling scheme, each node is equipped with an UMT($\rho = 6$, $\nu = 3$) with same parameters and initial condi-

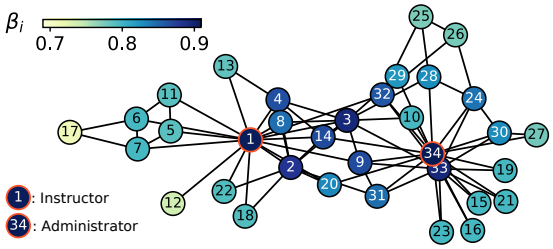


Figure 4.2: Discovery dynamics of the interacting urns on the Zachary Karate Club network, whose nodes are colored according to the resulting Heaps' exponent.

tions. 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 value of the Heaps' exponent of each node as $\beta_i = \beta_i(T)$, where $\beta_i(t) = \ln D_i(t)/\ln t$ and $T = 10^4$. Such measure can be considered as the slope in logarithmic scales of the line connecting the origin with the last point $(t, D(t))$. We avoid using more complicated fits, such as the one used in Chapter. 3 because of the high number of fits each simulation requires. Figure 4.2 shows the nodes of the ZKC colored according to the extracted Heaps' exponents. 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 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. Fig. 4.3(a-e) shows the temporal evolution of $D_i(t)$ for each node i of the networks displayed on the left. We report the simulated Heaps' laws (colored points), whose extracted exponents β_i are shown in the legend. In addition, to assess the validity of Eq. (4.8), we also plot the numerically integrated solutions (continuous black lines) obtained using the appropriate $\{a_{ij}\}$. It can be seen that the analytical formalism introduced perfectly captures the Heaps' laws, since lines are almost indistinguishable from (simulated) points. In particular, in Fig. 4.3(a) we observe the highest pace of discovery in the node with more outgoing links. However, the non-trivial behaviors observed in Fig. 4.3(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. 4.3(d) node 2 has two outgoing links, while the others have one link only. In contrast with what is observed in Fig. 4.3(a), here the highest pace of discovery is the one of node 1, whose social urn gets the benefits only of the urn of node 2. Moreover, in Figs. 4.3(c) and (d) a simple change of direction of link $4 \to 2$ translates into completely different dynamics. We further notice that in both Fig. 4.3(c) and (e) the presence of a cycle enhances the pace of discovery in a process of mutual exchange. However, while in Fig. 4.3(d) node 1 is linked to the cycle and captures the same behavior of those in the cycle, in Fig. 4.3(e) node 1 behaves as an individual urn.

We have also investigated whether the extracted β_i may depend on the maximum time T at which we have stopped the simulations. The curves reported in Fig. 4.3(f-j) as a function of time for time up to 10^8 clearly indicate that the systems, even for the small graphs considered, have not yet reached a stationary state. Thermalization times, that are typical of empirical trajectories of diffusion process [186], here seem to be strongly influenced by the topology of the network. This can be seen by comparing the two $\beta_1(t)$ of Fig. 4.3(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 persists at all finite times. In the next section we will

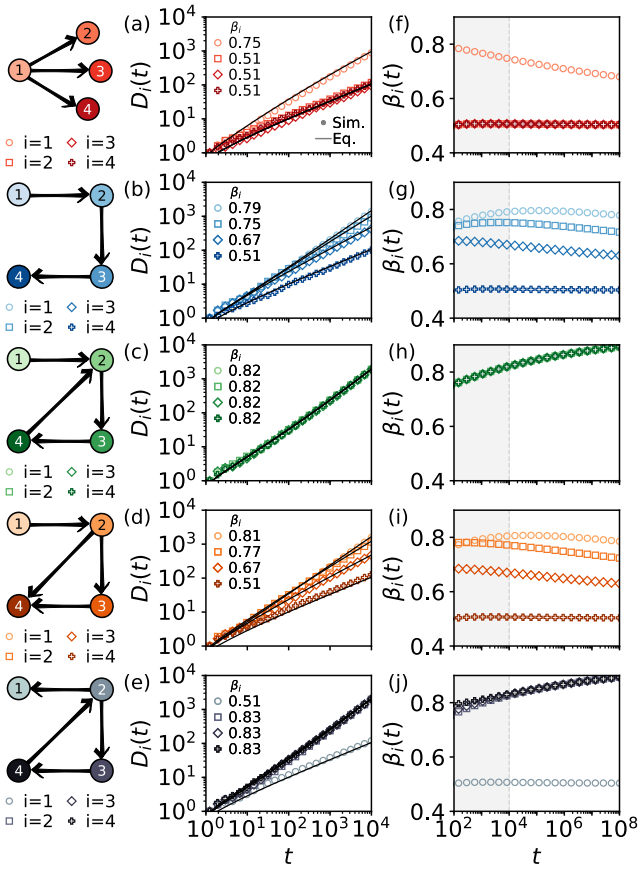


Figure 4.3: Discovery dynamics of the interacting urns on five directed toy graphs (different symbols correspond to different nodes). Each node is equipped with an UMT with the same parameters $\rho = 6$, and $\nu = 3$. (a-e) Temporal evolution of the number of discoveries $D_i(t)$ for each node i (associated Heaps' exponents β_i in the legend). The solutions of Eq. (4.8), shown as continuous black lines, are in perfect agreement with simulations. (f-j) Temporal behavior of the associated Heaps' exponents extracted at different times. The gray area up to $T = 10^4$ corresponds to the values of (a-e).

further investigate this characteristic behavior, ultimately proving its universality for all networks.

4.4.3 Node ranking persistence

In the previous sections we have developed and studied a networked model for the dynamics of discovery that introduces an heterogeneity in the paces of discovery, as it happens in real-world social networks. As we have seen in Fig. 4.3, the paces of discoveries, represented by the fitted Heaps' exponents, change in time, depending on the network topology and the model parameters. Nonetheless, the ranking of the nodes based on these fitted exponents always remains almost the same. This is even more

clear looking at Fig. 4.4, where we plot the fitted Heaps' exponents $\beta(T)$ at $T = 10^8$ as a function of the same exponents at $T = 10^4$ for four real-world networks. These networks are described in Sec. 4.4.1, and represent (a) the Zachary Karate Club network [182], (b) a network of follower relationships of Twitter [183], (c) a co-authorship network in network science [184] and (d) a collaboration network between jazz musicians [185]. All simulations in this section are performed with model parameters: $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$. In all cases, we get a Spearman's correlation of 1.00, meaning that even though the distribution of the fitted exponents change (plotted at the sides of each plot), the ranking is time-invariant, i.e. it does not depend on the particular T at which Heaps' exponents are fitted.

This rank persistence remains also in more extreme cases, as it can be seen in Fig. 4.4(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. Similarly, at the opposite case in Fig. 4.4(d), most of the exponents are very close to 1 (remember that these are limited between 0 and 1), but the Spearman's rank correlation remains equal to 1. All this is a strong indication that the various paces of discovery have to depend on some structural characteristics of the networks. For example, we have tested numerically the correlation between the eigenvector centralities and the measured Heaps' exponents at transient times. A lot of importance has been given in the past to eigenvector centrality, also known as the Bonacich centrality [187], which is related to the highest eigenvalue of the adjacency matrix of the graph. This centrality measure accounts for both local and global properties of the network, as it is not just dependent on the degree of the node, but also on the positioning of each node in the network [188]. Figure 4.5(a) shows the scatter plot and the Spearman's rank correlation of the eigenvector centralities and the fitted Heaps' exponents at time $T = 10^4$ for the ZKC network studied in Sec. 4.4, while in Fig. 4.5(b) its visualization with color-coded nodes can be seen (cfr. Fig. 4.2). The resulting Spearman's rank correlation higher than 0.98 persists changing the parameters in the simulations. We can hence conclude that the eigenvector centrality is an optimal proxy for the distribution of Heaps' exponents, at least in this case.

However, for other graphs this node measure does not have the same optimum correlation, for example in some directed graphs. As we will see analytically in Sec. 4.5.7, the correlation with the eigenvector centralities persists if the graph is strongly connected. Fortunately, a similar correlation can be found with a more general node centrality, in particular the α-centrality, first introduced in Ref. [189] to extend the eigenvector centrality to asymmetric graphs and widely used in network analysis [190, 191]. We have investigated numerically the correlation between the α-centrality and the pace of discovery of different nodes in real-world networks. For analytical reasons that will be clarified in Sec. 4.5.7, we have set α to $0.85/\widehat{\mu}$, where $\widehat{\lambda}$ is the maximum eigenvalue of the matrix $\boldsymbol{M} = \frac{\nu}{\rho}\boldsymbol{I} + \frac{\nu+1}{\rho}\boldsymbol{A}$, which can be numerically approximated. Fig. 4.6

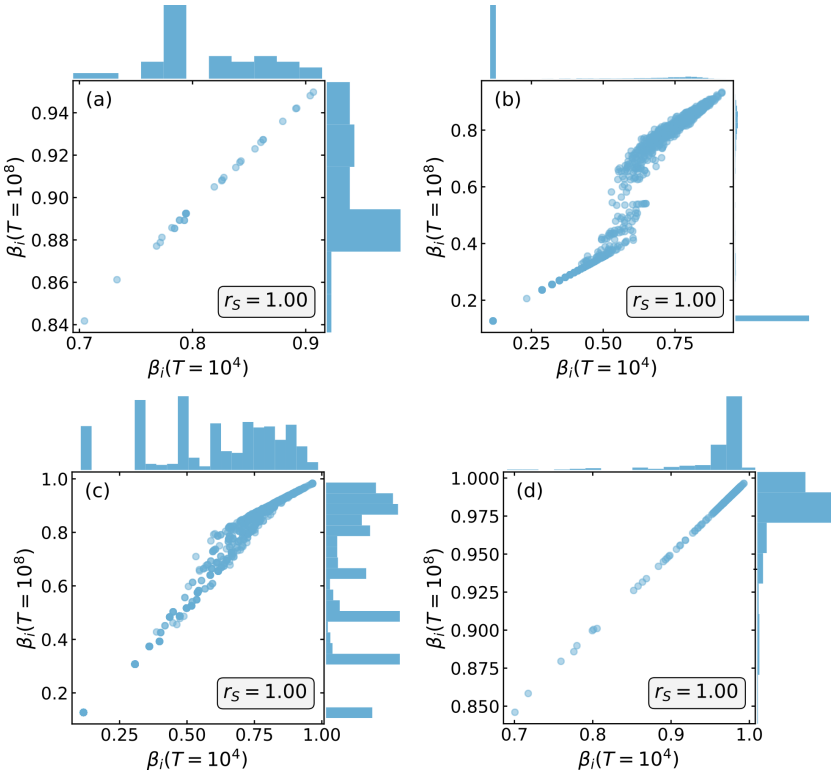


Figure 4.4: Scatter plot and Spearman's rank correlation coefficient r_S between fitted Heaps' exponents $\beta_i(T)$ at $T = 10^4$ and $T = 10^8$ associated to the $i = 1, \ldots, N$ nodes off the four empirical networks considered: (a) the Zachary Karate Club network [182], (b) a network of follower relationships of Twitter [183], (c) a co-authorship network in network science [184] and (d) a collaboration network between jazz musicians [185]. The parameters of the model are $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$.

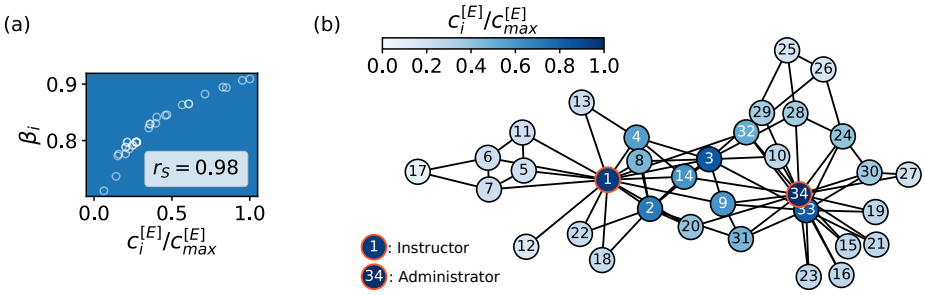


Figure 4.5: Correlation between discovery dynamics of the interacting urns and eigenvector centrality on the Zachary Karate Club network [182]. (a) Scatter plot and Spearman's rank correlation coefficients r_S between fitted Heaps' exponents $\beta_i(T = 10^4)$ and normalized eigenvector centrality $c_i^{[E]}/c_{\max}^{[E]}$ associated to the $i = 1, \ldots, N$ nodes of the network. (b) Nodes are colored according to the resulting normalized eigenvector centrality.

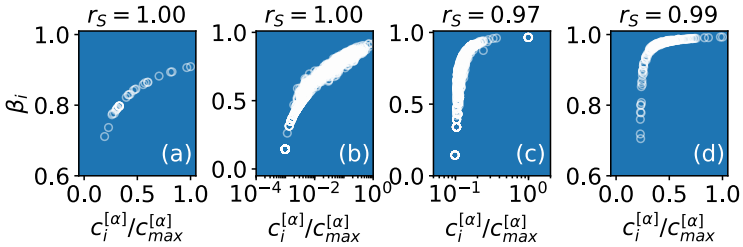


Figure 4.6: Scatter plot and Spearman's rank correlation coefficients r_S between fitted Heaps' exponents β_i and normalized α-centrality $c_i^{[\alpha]}/c_{\max}^{[\alpha]}$ associated to the $i = 1, \ldots, N$ nodes of four different real-world networks: (a) the Zachary Karate Club network [182], (b) a network of follower relationships of Twitter [183], (c) a co-authorship network in network science [184] and (d) a collaboration network between jazz musicians [185]. The parameters of the model are $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$.

shows the scatter plot of the number of discovered colors $D_i(T)$ and the normalized α-centrality $c_i^{[\alpha]}/c_{\max}^{[\alpha]}$ in the four empirical social networks analyzed. 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) is in agreement with our predictions. This result confirms that, together with the AP in the content space, it is crucial to take into account of an AP in the social space.

4.5 Analytical results

In this section we concentrate on finding an analytical solution to the Heaps' law set of equations in Sec. 4.3 [1]. As we have done in our numerical analysis in Sec. 4.4, here we consider the same parameters for each urn, so that $\rho_i = \rho$ (*reinforcement*) and $\nu_i = \nu$ (*triggering*) $\forall i = 1, \ldots, N$. We will be able to extract the asymptotic values of the Heaps' exponents and their dependence on the network topology. In order to do so, we first derive an analytical solution of Eq. (4.8) for $t \to \infty$ in some simple cases, in order to understand the underlying analytical mechanism with basic examples. In particular, we analyse a pair of nodes in Sec. 4.5.1, a chain in Sec. 4.5.2, a cycle in Sec. 4.5.3, and a clique in Sec. 4.5.4.

After this, in Sec. 4.5.5, we derive an explicit analytical solution for the asymptotic Heaps' dynamics of a generic graph. Furthermore, in Sec. 4.5.6, we apply this formalism to understand the dynamics taking place on the five toy graphs numerically examined in Fig. 4.3

Finally, in Sec. 4.5.7 we explain analytically why the node ranking provided by the Heaps' exponents persists at different times by looking deeper into the relation between the Heaps' exponents and the eigenvector and α-centralities.

4.5.1 Two coupled urns

The simplest example would be an isolated node, or equivalently, an urn on a node i for which the out-degree $\sum_j a_{ij}$ is null. In this case, the dynamics is the same of the Urn Model with Triggering (UMT) [8, 26], since the node does not have access to the balls of the neighbors, implying that its social urn will not be enriched, i.e., $\tilde{\mathcal{U}}_i(t) = \mathcal{U}_i(t)$. Therefore, the Heaps' law follows Eq. (2.20).

Let us hence consider the case of two coupled urns, that is a network with only two nodes connected by a directed edge $(1 \to 2)$, as already shown in Fig. 4.1. 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. As described in our general analytical framework in Eq. (4.2) and Eq. (4.4), the associated equations to determine the asymptotic Heaps' laws can be written expressing the probabilities $P_i^{\text{new}}(t)$ to draw a new ball as the fraction of discoverable balls over the total number of balls available to node i at time t, that is

$$\begin{cases} \dfrac{dD_1(t)}{dt} = \dfrac{|\tilde{\mathcal{U}}_1(t) \ominus \mathcal{S}_1'(t)|}{\tilde{U}_1(t)} & \text{(4.9a)} \\ \dfrac{dD_2(t)}{dt} = \dfrac{|\tilde{\mathcal{U}}_2(t) \ominus \mathcal{S}_2'(t)|}{\tilde{U}_2(t)} = \dfrac{U_2'(t) - D_2(t)}{U_2(t)}. & \text{(4.9b)} \end{cases}$$

Notice that the right-hand side of Eq. (4.9b) is simplified with respect to Eq. (4.9a), 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, for $\nu < \rho$ we have

$$D_2(t) \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{4.10}$$

Addressing the dynamics of node 1, the denominator $\tilde{U}_1(t)$ of Eq. (4.9a) can be expressed in terms of the contributions coming from the two urns at time t, which reads

$$\begin{aligned} \tilde{U}_1(t) &= \overbrace{M_0 + \rho t + (\nu+1)D_1(t)}^{U_1(t)} + \overbrace{M_0 + (\nu+1)D_2(t)}^{U_2'(t)} \\ &= 2M_0 + \rho t + (\nu+1)\big[D_1(t) + D_2(t)\big]. \end{aligned} \tag{4.11}$$

Similarly, the numerator of Eq. (4.9a), consisting of the number of balls present in the social urn of node 1 at time t which have not yet appeared in $\mathcal{S}_1(t)$, can be written as the total number of balls in the social urn of 1 at time t, without duplicates, minus the number of balls that do not represent a novelty anymore with respect to the sequence $\mathcal{S}_1(t)$, i.e.,

$$|\tilde{\mathcal{U}}_1(t) \ominus \mathcal{S}_1'(t)| = \tilde{U}_1(t) - \rho t - D_1(t). \tag{4.12}$$

Then, using Eq. (4.11) and Eq. (4.12), the final expression for Eq. (4.9a) reads

$$\frac{dD_1(t)}{dt} = \frac{2M_0 + \nu D_1(t) + (\nu+1)D_2(t)}{2M_0 + \rho t + (\nu+1)\big[D_1(t) + D_2(t)\big]}. \tag{4.13}$$

For large times ($t \gg M_0$) we can approximate Eq. (4.13) as

$$\frac{dD_1(t)}{dt} \approx \frac{\nu D_1(t) + (\nu+1)D_2(t)}{\rho t + (\nu+1)\big[D_1(t) + D_2(t)\big]}. \tag{4.14}$$

Let us assume now that the dynamics of node 2 relaxes before the one of node 1, so that we can solve Eq. (4.14) independently from Eq. (4.10). In addition, if we suppose that $\lim_{t\to\infty} D_i(t)/t = 0$, Eq. (4.14) can be approximated as

$$\frac{dD_1(t)}{dt} \approx \frac{\nu D_1(t)}{\rho t} + \frac{(\nu+1)D_2(t)}{\rho t}. \tag{4.15}$$

The related homogeneous equation has a similar solution of Eq. (4.10), i.e.,

$$\frac{d\overline{D}_1(t)}{dt} \approx \frac{\nu \overline{D}_1(t)}{\rho t} \implies \overline{D}_1(t) \sim (\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{4.16}$$

We now look for a solution in the family of functions $D_1(t) = \kappa(t)\overline{D}_1(t)$. Plugging this into Eq. (4.15), we obtain

$$\frac{d\kappa(t)}{dt}\overline{D}_1(t) + \kappa(t)\frac{d\overline{D}_1(t)}{dt} \approx \kappa(t)\frac{d\overline{D}_1(t)}{dt} + \frac{(\nu+1)D_2(t)}{\rho t}. \tag{4.17}$$

Thus, from Eq. (4.10) and Eq. (4.16) we get

$$\frac{d\kappa(t)}{dt} = \frac{\nu+1}{\rho t}\frac{D_2(t)}{\overline{D}_1(t)} \approx \frac{\nu+1}{\rho t}, \tag{4.18}$$

whose solution is

$$\kappa(t) \approx \frac{\nu+1}{\rho}\ln t. \tag{4.19}$$

The asymptotic solution ($t \to \infty$) of $D_1(t)$ is then approximated by

$$D_1(t) \sim \frac{\nu+1}{\rho}(\rho-\nu)^{\frac{\nu}{\rho}} \ln(t)\, t^{\frac{\nu}{\rho}}. \tag{4.20}$$

In conclusion, comparing the solutions in Eq. (4.10) and Eq. (4.20), the presence of an outgoing link effectively increases the number of novelties with respect to the dynamics of an isolated urn. However, as we have shown here, this increase is approximately only given by a multiplicative logarithmic factor, meaning that we can see a slight increase of the discovery rate at finite times, which practically disappears for larger

times. Further notice that the presence of this logarithmic term makes the functional form of $D_1(t)$ not follow a precise Heaps' law. Nevertheless, since the multiplicative factor is logarithmic, we could still locally approximate the function to a power law, where the logarithmic factor is incorporated in the coefficient of the power-law.

Moreover, let us point out that the dynamics of the system discussed here only applies to a pair of urns connected by a directed link. In the case of an undirected link, instead, we would get identical Heaps' laws for both nodes $i = 1, 2$, without logarithmic corrections, but with higher exponents. This can be seen as a cycle of two nodes, and will be explicitly discussed in Sec. 4.5.3, after solving the case of a directed chain of N nodes.

4.5.2 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 $(i \to i+1)$, with $i = 1, 2, \ldots, N-1$. This is the case considered in Fig. 4.3(b, g), 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:

$$\begin{cases} \dfrac{dD_1(t)}{dt} \approx \dfrac{\nu D_1(t) + (\nu+1)D_2(t)}{\rho t + (\nu+1)\big[D_1(t) + D_2(t)\big]} & \text{(4.21a)} \\ \qquad\vdots & \\ \dfrac{dD_{N-1}(t)}{dt} \approx \dfrac{\nu D_{N-1}(t) + (\nu+1)D_N(t)}{\rho t + (\nu+1)\big[D_{N-1}(t) + D_N(t)\big]} & \text{(4.21b)} \\ \dfrac{dD_N(t)}{dt} \approx \dfrac{\nu D_N(t)}{\rho t + (\nu+1)D_N(t)} & \text{(4.21c)} \end{cases}$$

Eq. (4.21) is a system of N equations, which can be solved starting from the last one and recursively substituting its solution into the equation above. Indeed, since node $i = N$ does not have any outgoing links, its related equation (Eq. (4.21c)) is independent and can be immediately solved, resulting in the known asymptotic solution:

$$D_N(t) \sim (\rho - \nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{4.22}$$

As done for the couple of nodes studied in Sec. 4.5.1, in Eq. (4.21b) we can consider $D_{N-1}(t)$ to be the only unknown variable. Then, following the same analytical steps, we obtain

$$D_{N-1}(t) \approx \frac{\nu+1}{\rho}(\rho - \nu)^{\frac{\nu}{\rho}} \ln(t) t^{\nu/\rho}. \tag{4.23}$$

The same reasoning can be iterated for each node i. Let us now prove by induction on i that the asymptotic solution is

$$D_i(t) = \frac{(\rho-\nu)^{\nu/\rho}}{(N-i)!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-i} t^{\nu/\rho}. \tag{4.24}$$

We have already proved that Eq. (4.24) 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 $i-1$ reads

$$\frac{dD_{i-1}(t)}{dt} \approx \frac{\nu D_{i-1}(t) + (\nu+1)D_i(t)}{\rho t + (\nu+1)\big[D_{i-1}(t) + D_i(t)\big]}. \tag{4.25}$$

For the induction hypothesis, in Eq. (4.25) the only unknown variable is $D_i(t)$. Therefore, we can consider the homogeneous associated equation

$$\frac{d\overline{D}_{i-1}(t)}{dt} \approx \frac{\nu\overline{D}_{i-1}(t)}{\rho t}, \tag{4.26}$$

whose solution is

$$\overline{D}_{i-1}(t) \approx (\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{4.27}$$

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. (4.25), leads to

$$\frac{d\kappa(t)}{dt}\overline{D}_{i-1}(t) + \cancel{\kappa(t)\frac{d\overline{D}_{i-1}(t)}{dt}} \approx \cancel{\kappa(t)\frac{d\overline{D}_{i-1}(t)}{dt}} + \frac{(\nu+1)D_i(t)}{\rho t}. \tag{4.28}$$

Thus, we get

$$\frac{d\kappa(t)}{dt} \approx \frac{\nu+1}{\rho t}\frac{D_i(t)}{\overline{D}_{i-1}(t)} \approx \frac{1}{(N-i)!}\frac{\nu+1}{\rho t}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-i}, \tag{4.29}$$

and therefore

$$\kappa(t) \approx \frac{1}{(N-(i-1))!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-(i-1)}. \tag{4.30}$$

Finally, after combining Eq. (4.27) and Eq. (4.30), we reach the solution for the dynamics of node $i-1$, that reads

$$D_{i-1}(t) \approx \frac{(\rho-\nu)^{\nu/\rho}}{(N-(i-1))!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{N-(i-1)} t^{\nu/\rho}, \tag{4.31}$$

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.

4.5.3 Cycle of N urns

Directed cycle

Let us consider a social network in the form of a 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 also in the asymptotic limit (without logarithms). Let us hence suppose that every node i is connected just to the following one with a directed link $(i \to i+1)$, with $i = 1, 2, \ldots, N$, 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

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)D_{i+1}(t)}{\rho t + (\nu+1)\big[D_i(t) + D_{i+1}(t)\big]}. \tag{4.32}$$

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. (4.32) becomes

$$\frac{dD_i(t)}{dt} \approx \frac{(2\nu+1)D_i(t)}{\rho t + 2(\nu+1)D_i(t)}, \tag{4.33}$$

that is equal to the equation of an individual urn [see Eq. (2.17)], with $\nu' = 2\nu + 1$. Therefore, if $\rho > \nu'$ we have the solution

$$D_i(t) \approx (\rho - 2\nu - 1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}. \tag{4.34}$$

Undirected cycle

Let us now consider a cycle 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 connected to two different nodes, $i-1$ and $i+1$, and the associated equations to be solved are

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)D_{i-1}(t) + (\nu+1)D_{i+1}(t)}{\rho t + (\nu+1)\big[D_i(t) + (\nu+1)D_{i-1}(t) + D_{i+1}(t)\big]}. \tag{4.35}$$

Again, for symmetry reasons, we can equivalently write Eq. (4.35) as

$$\frac{dD_i(t)}{dt} \approx \frac{(3\nu+2)D_i(t)}{\rho t + 3(\nu+1)D_i(t)}, \tag{4.36}$$

that is equal to the equation of an individual urn [see Eq. (2.17)], with $\nu'' = 3\nu + 2$. Therefore, if $\rho > \nu''$ we have the solution

$$D_i(t) \approx (\rho - 3\nu - 2)^{\frac{3\nu+2}{\rho}} t^{\frac{3\nu+2}{\rho}}. \tag{4.37}$$

Finally, notice that, in both directed and undirected version, the dynamics of each

node does not depend on the length of the cycle. Obviously, since all connections are mutual in an undirected network, the resulting paces of discovery are higher than those in the directed case.

4.5.4 Clique of N urns

Let us consider a N-clique, that is a fully connected network of N nodes, either directed or undirected. Being every node i connected to all other nodes, there is complete equivalence between all nodes, and the general equation for the growth of the number of novelties of node i reads

$$\frac{dD_i(t)}{dt} \approx \frac{\nu D_i(t) + (\nu+1)\sum_{j\neq i} D_j(t)}{\rho t + (\nu+1)\sum_{j=1}^{N} D_j(t)}. \tag{4.38}$$

For symmetry reasons, each urn follows the same dynamics, so that we can write Eq. (4.38) as

$$\frac{dD_i(t)}{dt} \approx \frac{[N(\nu+1)-1]D_i(t)}{\rho t + N(\nu+1)D_i(t)}, \tag{4.39}$$

that is equal to the equation for an individual urn [see Eq. (2.17)], with $\nu''' = N(\nu+1)-1$. Therefore, if $\nu''' < \rho$ we have the solution

$$D_i(t) \approx (\rho - N(\nu+1) - 1)^{\frac{N(\nu+1)-1}{\rho}} t^{\frac{N(\nu+1)-1}{\rho}}. \tag{4.40}$$

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.

4.5.5 The general solution

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 colors 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 colors present in the urn of node i and in the one of its neighbors (without their multiplicity) minus the number of colors 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 colors added to the urn j is exactly $(\nu+1)D_j(t)$, we can write:

$$\begin{aligned}\frac{dD_i(t)}{dt} = P_i^{\text{new}}(t) &= \frac{|\tilde{\mathcal{U}}_i(t) \ominus \mathcal{S}_i'(t)|}{\tilde{U}_i(t)} \\ &= \frac{M_0 + \nu D_i(t) + \sum_{j\neq i} a_{ij}\big[M_0 + (\nu+1)D_j(t)\big]}{\rho t + M_0 + (\nu+1)D_i(t) + \sum_{j\neq i} a_{ij}\big[M_0 + (\nu+1)D_j(t)\big]},\end{aligned} \tag{4.41}$$

or, equivalently:

$$\frac{dD_i(t)}{dt} = \frac{M_0 \sum_j (a_{ij}+\delta_{ij}) + \sum_j \big[\delta_{ij}\nu + a_{ij}(\nu+1)\big]D_j(t)}{\rho t + \sum_j (a_{ij}+\delta_{ij})\big[M_0 + (\nu+1)D_j(t)\big]}. \tag{4.42}$$

For $t \gg M_0$ we can disregard the presence of M_0 in Eq. (4.42). Moreover, 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 right hand side of Eq. (4.42) to ρt. After finding the approximated solution, we will estimate the set of parameters for which this approximation is valid for any given topology. Therefore, in the asymptotic limit and with a proper choice of the parameters, Eq. (4.42) can be rewritten as

$$\frac{dD_i(t)}{dt} \approx \frac{\sum_j \big[\delta_{ij}\nu + a_{ij}(\nu+1)\big]D_j(t)}{\rho t}, \tag{4.43}$$

which can be expressed in a more compact way as

$$\frac{d\vec{D}(t)}{dt} \approx \frac{1}{t}\left(\frac{\nu}{\rho}\boldsymbol{I} + \frac{\nu+1}{\rho}\boldsymbol{A}\right)\vec{D}(t) = \frac{1}{t}\frac{f(\boldsymbol{A})\vec{D}(t)}{t} = \frac{1}{t}\boldsymbol{M}\vec{D}(t), \tag{4.44}$$

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. (4.44) 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

$$\vec{D}(t) \approx \sum_{\ell=1}^{r}\sum_{p=0}^{m_\ell-1} \vec{c}_p \ln^p(t)\, t^{\lambda_\ell}, \tag{4.45}$$

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 behavior of the number of novelties $D_i(t)$ discovered by node i at time t is then governed by the leading term in Eq. (4.45), so that we can write

$$D_i(t) \underset{t\to\infty}{\approx} u_i \ln^{\widehat{p}(i)}(t)\, t^{\widehat{\lambda}(i)}. \tag{4.46}$$

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. For example, in the case of a chain as in Sec.4.5.2, the asymptotic solution is $D_i(t) \sim u_i \ln^{N-i}(t)\, t^{\nu/\rho}$, where $\widehat{\lambda}(i) = \nu/\rho$ and $\widehat{p}(i) = N - i$. In this example all the Heaps' 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 behavior also seen in Fig. 4.3(g). In order to better understand the solution of Eq. (4.46), in the next paragraphs we further study the general solution taking into consideration the strongly-connected components (SCCs) of the network. As we will see, nodes in the same SCC have the same exponents, while they may vary from SCC to SCC.

Strongly-connected network.

Let us first suppose that the graph $G(\mathcal{N}, \mathcal{E})$ is strongly connected. In this case the solution given by Eq. (4.46) simplifies. Indeed, for such a graph the corresponding adjacency matrix $\boldsymbol{A} = \{a_{ij}\}$ is irreducible [192]. Moreover, let us recall that for irreducible matrices the Perron–Frobenius theorem holds [193, 194], according to which there exists a positive eigenvalue $\widehat{\mu}$ greater or equal to all other eigenvalues in absolute value. 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 [187]. 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 [36]. Simple algebraic steps can prove that if μ is an eigenvalue for $\boldsymbol{A}$, then $\lambda = f(\mu)$ is an eigenvalue for $\boldsymbol{M} = f(\boldsymbol{A})$, where $f(x) = \frac{\nu}{\rho} + \frac{\nu+1}{\rho}x$. 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. (4.46) becomes

$$D_i(t) \underset{t\to\infty}{\approx} u_i\, t^{\widehat{\lambda}}, \tag{4.47}$$

meaning that all nodes have similar Heaps' laws, and the only difference is made by the coefficient, which is proportional to the eigenvector centrality. As we have seen in

the Sec. 4.4.2, these differences, more pronounced in transient times, will contribute to determine the fastest explorers in the network. In graphs such as the ZKC (see Fig. 4.2), the different values of u_i hence play a very important role. Most central nodes, as the instructor and the chief administrator, are the fastest explorers (highest β_i), even having the same asymptotic Heaps' exponent $\widehat{\lambda}$. On the contrary, in the case of cycles and cliques seen in Sec. 4.5.3 and Sec. 4.5.4, which are also strongly-connected graphs, nodes are all structurally equivalent (so $u_i = u \; \forall i$), and they have the same analytical form for $D_i(t)$.

Finally, from these calculations we can also deduce that the approximation used in Eq. (4.43) 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. (2.20), since in the original system in Eq. (4.41) we have $dD_i(t)/dt \leq 1$.

Non-strongly-connected network.

In the more general case in which a graph is not strongly connected, Eq. (4.46) still holds, and the same argument can be applied to each of the strongly connected components, i.e. maximal strongly-connected subgraphs, to recursively find the values of u_i, $\widehat{p}(i)$, and $\widehat{\lambda}(i)$. As we will see in Sec. 4.5.7, in such cases the role of the eigenvector centrality as a factor of nodes ranking according to the their pace of discovery is replaced by its natural extension to non-strongly-connected graphs, i.e., the α-centrality [189].

Let us hence construct an algorithm to determine the pace of discovery of each node, which will help us to better understand analytically why some nodes have a higher pace of discovery. Let us first partition the graph into its strongly-connected components (SCCs), which can be found in linear computational time, for example with a DFS-based algorithm [195]. Let all the SCCs be indexed as $C_1, \ldots, C_p$, 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 further assume that the number of SCCs is $p > 1$, because otherwise the graph would be strongly connected, which we have already discussed in the previous paragraph. Notice that for a weakly-connected graph we have more than one strongly-connected graph only if the graph is directed. 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. Permutating the indexes of the SCCs without loss of generality, 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:

$$D_i(t) \underset{t\to\infty}{\approx} \gamma_i^{(q)}\, t^{\widehat{\lambda}^{(q)}} \;\forall i \in C_q, 1 \leq q \leq p_1, \tag{4.48}$$

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 the SCCs of the second round, 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. (4.48) 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

$$D_i(t) \underset{t\to\infty}{\approx} \begin{cases} \gamma_i^{(q)}\, t^{\overline{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} > \tilde{\lambda}^{(q)} \\ \gamma_i^{(q)} \ln(t)\, t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)} \\ \gamma_i^{(q)}\, t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} < \tilde{\lambda}^{(q)} \end{cases} \quad \forall i \in C_q,\ p_1+1 \leq q \leq p_2, \tag{4.49}$$

meaning that the Heaps' exponent $\widehat{\lambda}^{(q)}$ for node i in C_q, $p_1 + 1 \leq q \leq p_2$, is

$$\widehat{\lambda}^{(q)} = \max(\overline{\lambda}^{(q)},\ \tilde{\lambda}^{(q)}), \tag{4.50}$$

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 Heaps' 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. (4.44), 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:

$$\frac{dD_i(t)}{dt} \approx \frac{1}{t}\left(\frac{\nu}{\rho}D_i(t) + \frac{\nu+1}{\rho}\sum_{j\in C_q} a_{ij}D_j(t) + \frac{\nu+1}{\rho}\sum_{j\notin C_q} a_{ij}D_j(t)\right), \quad (4.51)$$

$\forall i \in C_q$, where we have isolated the contributions coming from nodes outside C_q, which are known by induction. Considering the general asymptotic solution for each individual Heaps' law derived for a strongly-connected graph in Eq. (4.46), we can write explicitly the functions $D_j(t)$, $j \notin C_q$. We can hence write

$$\begin{aligned}\frac{\nu+1}{\rho}\sum_{j\notin C_q} a_{ij}D_j(t) &\approx \frac{\nu+1}{\rho}\sum_{j\notin C_q} a_{ij}u_j \ln^{\widehat{p}_j}(t)\, t^{\widehat{\lambda}_j} \\ &\underset{t\to\infty}{\approx} \tilde{u}_i \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}} = f_i(t),\end{aligned} \quad (4.52)$$

where we have used the fact that $\tilde{u}_i \ln^{\tilde{p}_i}(t)\, t^{\tilde{\lambda}_i}$ is the leading term of the expression $\sum_{j\notin C_q} a_{ij}u_j \ln^{\widehat{p}_j}(t)\, t^{\widehat{\lambda}_j}$ in the asymptotic regime. Then, using Eq. (4.52) 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. (4.51) in a compact form as

$$\frac{d\vec{D}^{(q)}(t)}{dt} \approx \frac{\boldsymbol{M}^{(q)}\vec{D}^{(q)}(t)}{t} + \frac{\vec{f}^{(q)}(t)}{t}. \quad (4.53)$$

The associated homogeneous system corresponds to the Heaps' dynamics of 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. (4.47), which is

$$\vec{\overline{D}}^{(q)}(t) \underset{t\to\infty}{\approx} \vec{\overline{u}}^{(q)}\, t^{\overline{\lambda}^{(q)}}, \quad (4.54)$$

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 its eigenvector centrality. Let us search a solution for Eq. (4.53) 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. (4.53) gives:

$$\begin{aligned}&\frac{d\vec{u}^{(q)}(t)}{dt} \circ \vec{\overline{D}}^{(q)}(t) + \cancel{\vec{u}^{(q)}(t) \circ \frac{d\left[\vec{\overline{D}}^{(q)}(t)\right]}{dt}} \\ &\qquad \approx \cancel{\vec{u}^{(q)}(t) \circ \frac{\boldsymbol{M}^{(q)}\vec{\overline{D}}^{(q)}(t)}{t}} + \frac{\vec{f}^{(q)}(t)}{t},\end{aligned} \quad (4.55)$$

where the cancellation is due to the general solution in Eq. (4.54) of the associated

homogeneous system. Therefore, recalling Eq. (4.52) and Eq. (4.54) we have

$$\frac{d\vec{u}^{(q)}(t)}{dt} \approx \vec{\tilde{u}}^{(q)} \circ \left(\vec{\overline{u}}^{(q)}\right)^{-1} \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}} = \vec{\gamma}\, \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}}, \tag{4.56}$$

or equivalently, considering the i-th components:

$$\frac{du_i(t)}{dt} \approx \tilde{u}_i^{(q)} \left[\left(\vec{\overline{u}}^{(q)}\right)^{-1}\right]_i \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}} = \gamma_i \frac{\ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}}{t^{\overline{\lambda}+1}}, \tag{4.57}$$

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

$$u_i(t) \approx \frac{\gamma_i}{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} + u_i \underset{t\to\infty}{\approx} u_i, \tag{4.58}$$

   which gives the solution:

$$D_i(t) \underset{t\to\infty}{\approx} u_i t^{\overline{\lambda}^{(q)}}. \tag{4.59}$$

2. Similarly, for $\overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)}$ we have

$$u_i(t) \approx \frac{\gamma_i}{\tilde{p}^{(q)} + 1} \ln^{\tilde{p}^{(q)}+1}(t) + u_i \underset{t\to\infty}{\approx} \frac{\gamma_i}{\tilde{p}^{(q)} + 1} \ln^{\tilde{p}^{(q)}+1}(t), \tag{4.60}$$

   which gives:

$$D_i(t) \underset{t\to\infty}{\approx} u_i \ln^{\tilde{p}^{(q)}+1}(t)\, t^{\tilde{\lambda}^{(q)}}. \tag{4.61}$$

3. Finally, if $\overline{\lambda}^{(q)} < \tilde{\lambda}^{(q)}$ we have

$$\begin{aligned} u_i(t) \approx \frac{\gamma_i}{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} + d_1 \underset{t\to\infty}{\approx} \\ \underset{t\to\infty}{\approx} \frac{\gamma_i}{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}} \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)} - \overline{\lambda}^{(q)}}, \end{aligned} \tag{4.62}$$

   hence the solution:

$$D_i(t) \underset{t\to\infty}{\approx} a_i \ln^{\tilde{p}^{(q)}}(t)\, t^{\tilde{\lambda}^{(q)}}. \tag{4.63}$$

To sum up, we have the following solutions:

$$D_i(t) \underset{t\to\infty}{\approx} \begin{cases} u_i t^{\overline{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} > \tilde{\lambda}^{(q)} \\ u_i \ln^{\tilde{p}^{(q)}+1}(t)\, t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} = \tilde{\lambda}^{(q)} \\ u_i \ln^{\tilde{p}^{(q)}}(t)\; t^{\tilde{\lambda}^{(q)}} & \text{if } \overline{\lambda}^{(q)} < \tilde{\lambda}^{(q)} \end{cases} \quad \forall i \in C_q, q > p_1, \tag{4.64}$$

Comparing this solution with the general one we gave in Eq. (4.46), 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 strongly-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 using Eq. (4.64). 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. (4.46). As an example, in the next section this algorithmic method is applied to the simple networks with $N = 4$ nodes studied numerically in Sec. 4.4.2.

4.5.6 Application to the five graphs in Fig. 4.3

As an application of the analytical results of the previous sections, we study here the same five networks reported in Fig. 4.3. 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 network. Explicit solutions are summarized in Table 4.2.

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 $1 \to 2$, $1 \to 3$, and $1 \to 4$ to all other nodes (see network representation in Table 4.2). Let us observe that the dynamics here is very similar to the case of a couple of urns with the only link $1 \to 2$. 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. (4.10), that is

$$D_2(t) \underset{t\to\infty}{\approx} D_3(t) \underset{t\to\infty}{\approx} D_4(t) \underset{t\to\infty}{\approx} (\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{4.65}$$

As for node 1, the differential equation for the Heaps' law is approximated by

$$\begin{aligned} \frac{dD_1(t)}{dt} &\approx \frac{\nu D_1(t)}{\rho t} + \frac{(\nu+1)\big(D_2(t)+D_3(t)+D_4(t)\big)}{\rho t} \\ &\approx \frac{\nu D_1(t)}{\rho t} + \frac{3(\nu+1)(\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}}{\rho t}. \end{aligned} \tag{4.66}$$

The resolution of Eq. (4.66) 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

$$D_1(t) \underset{t\to\infty}{\approx} 3\,\frac{\nu+1}{\rho}(\rho-\nu)^{\frac{\nu}{\rho}} \ln(t)\, t^{\frac{\nu}{\rho}}, \tag{4.67}$$

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 consider is a chain of 4 nodes, with links $1 \to 2$, $2 \to 3$, and $3 \to 4$. This network has already been studied in Sec. 4.5.2, and the solutions are:

$$D_i(t) \approx \frac{(\rho-\nu)^{\nu/\rho}}{(4-i)!}\left(\frac{\nu+1}{\rho}\ln(t)\right)^{4-i} t^{\nu/\rho}, \quad i = 1,2,3,4. \tag{4.68}$$

This analytical result shows us why node 1 has a higher pace of discovery than the other nodes, which is due to the presence of different powers of the logarithm. In the end, however, they all have the same asymptotic Heaps' exponent (the power-law 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 $2 \to 3$, $3 \to 4$, and $4 \to 2$, and another node 1 linked directly to node 2 ($1 \to 2$). 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 Sec. 4.5.3, the solution is given by Eq. (4.34) with N = 3, that is

$$D_i(t) \underset{t\to\infty}{\approx} (\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}, \quad i = 2,3,4. \tag{4.69}$$

Now let us consider the remaining SCC, namely node 1. Its equation is the same as Eq. (4.14) for the two coupled urns case in Sec. 4.5.1, with the only difference that here the solution of $D_2(t)$ has a higher exponent. Therefore, 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

$$\begin{aligned}\frac{d\kappa(t)}{dt} = \frac{\nu+1}{\rho t}\frac{D_2(t)}{\overline{D}_1(t)} &\approx \frac{\nu+1}{\rho t}\frac{(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}}{(\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}} \\ &= \frac{\nu+1}{\rho}\frac{(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{\nu+1}{\rho}-1}}{(\rho-\nu)^{\frac{\nu}{\rho}}},\end{aligned} \tag{4.70}$$

whose solution is

$$\kappa(t) \approx \frac{\nu+1}{\rho}\frac{(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{\nu+1}{\rho}}}{(\rho-\nu)^{\frac{\nu}{\rho}}}. \tag{4.71}$$

This gives the following asymptotic solution for node 1:

$$D_1(t) \approx \frac{\nu+1}{\rho}(\rho-2\nu-1)^{\frac{2\nu+1}{\rho}} t^{\frac{2\nu+1}{\rho}}. \tag{4.72}$$

We could have obtained the same result using the algorithm developed in Sec. 4.5.5. 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). This further shows the power of the algorithmic and analytical solution we have provided.

Graph (d)

In this case we consider the same network as the last graph we have just analyzed, but now we swap the direction of the link $4 \to 2$. Therefore, the cycle is broken (see network representation in Table 4.2), 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 apply the algorithm we have developed to showcase a full example of how it can be used.

Let us start from node 4, which has no outgoing links. This node is hence an individual urn, with the usual solution:

$$D_4(t) \underset{t\to\infty}{\approx} (\rho-\nu)^{\frac{\nu}{\rho}} t^{\frac{\nu}{\rho}}. \tag{4.73}$$

Let us move on to the next 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). This means that the actual solution for node 3 has that same exponent and a logarithmic factor. Indeed, the dynamics of node 3 is the same derived for the couple of urns in Sec. 4.5.1, which is

$$D_3(t) \underset{t\to\infty}{\approx} \frac{\nu+1}{\rho}(\rho-\nu)^{\frac{\nu}{\rho}} \ln(t)\, t^{\frac{\nu}{\rho}}. \tag{4.74}$$

Proceeding with node 2, we compare its power-law 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 $4 \to 2$. Thus, the solution for node 2 has another logarithmic factor and another constant multiplicative factor with respect to those of node 3. We hence get the solution

$$D_2(t) \underset{t\to\infty}{\approx} \left(\frac{\nu+1}{\rho}\right)^2 (\rho-\nu)^{\frac{\nu}{\rho}} \ln^2(t)\, t^{\frac{\nu}{\rho}}. \tag{4.75}$$

To complete, similarly for node 1 we obtain

$$D_1(t) \underset{t\to\infty}{\approx} \left(\frac{\nu+1}{\rho}\right)^3 (\rho-\nu)^{\frac{\nu}{\rho}}\, \ln^3(t)\, t^{\frac{\nu}{\rho}}. \tag{4.76}$$

We can hence see that the asymptotic solutions are equal to those of the chain in Sec. 4.5.2, and there are only some slight differences at finite times due to the presence of another link, which disappear at long enough times.

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 4.2). 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. (2.20). Then we have to solve the equations for the cycle, which in this case are given by

$$\begin{cases} \dfrac{dD_2(t)}{dt} \approx \dfrac{\nu D_2(t)}{\rho t} + \dfrac{(\nu+1)D_3(t)}{\rho t} + \dfrac{(\nu+1)D_1(t)}{\rho t} \\ \dfrac{dD_3(t)}{dt} \approx \dfrac{\nu D_3(t)}{\rho t} + \dfrac{(\nu+1)D_4(t)}{\rho t} \\ \dfrac{dD_4(t)}{dt} \approx \dfrac{\nu D_4(t)}{\rho t} + \dfrac{(\nu+1)D_2(t)}{\rho t}. \end{cases} \tag{4.77}$$

In this system, we can consider $D_1(t)$ known. Therefore, following the algorithm described in Sec. 4.5.5, we first solve this system without the external sources (i.e., node 1), in order to find the leading solution, and then compare these exponents with the Heaps' exponent of node 1. The solution of the associated homogeneous system is the same of a directed cycle as in Eq. (4.34), 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. Thus, 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 4.2.

4.5.7 Node ranking persistence

In this section, we study more in details the relation between the eigenvector centrality and α-centrality with the Heaps' law solutions found in the previous sections. Firstly, we concentrate on strongly-connected graphs, and we explain why the eigenvector centrality leads to the persistent ranking of the Heaps' exponents extracted in the simulations discussed in Sec. 4.4.3. Then, we consider generic networks, making use of the α-centrality. This analysis will ultimately make us understand that the rank persistence

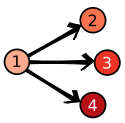
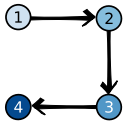
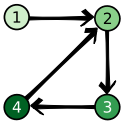
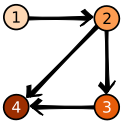
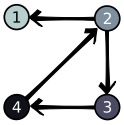
Graph	 (a)	 (b)	 (c)	 (d)	 (e)
$D_1(t)$	$u_1 \ln(t)\, t^{\frac{\nu}{\rho}}$	$u_1 \ln^3(t)\, t^{\frac{\nu}{\rho}}$	$u_1\, t^{\frac{2\nu+1}{\rho}}$	$u_1 \ln^3(t)\, t^{\frac{\nu}{\rho}}$	$u_1\, t^{\frac{\nu}{\rho}}$
$D_2(t)$	$u_2\, t^{\frac{\nu}{\rho}}$	$u_2 \ln^2(t)\, t^{\frac{\nu}{\rho}}$	$u_2\, t^{\frac{2\nu+1}{\rho}}$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

Table 4.2: Summary of the asymptotic Heaps' laws derived analytically for the 4 nodes composing the five networks reported in Fig. 4.3 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.

found in the Heaps' exponents is directly connected and intertwined with the network structure at a local and global level. Consequently, we will see how these centralities can be used as a valid predictor of the rank of the nodes in a social network where a cooperative discovery dynamics is in place.

Pace of discovery and eigenvector centrality

In Sec. 4.5.5 we have shown that in strongly-connected graphs each urn has the same asymptotic Heaps' exponent, and that the driving factor for each node is the associated asymptotic coefficient. As we saw when we derived the asymptotic expression of the Heaps' law for strongly-connected graphs in Eq. (4.46), the Heaps' exponent corresponds to the maximum eigenvalue $\widehat{\lambda}$ of the matrix $\boldsymbol{M} = \frac{\nu}{\rho}\boldsymbol{I} + \frac{\nu+1}{\rho}\boldsymbol{A}$, where $\boldsymbol{A}$ is the adjacency matrix. In particular, because of the Perron-Frobenius theorem [193, 194], we know that $\widehat{\lambda}$ is positive and simple, and the related eigenvector $\vec{u}$ has all positive entries. We also derived that the coefficients of the Heaps' laws are all multiples of this eigenvector. A lot of importance has been given in the past to this vector, from which we can derive the eigenvector centrality, also known as the Bonacich centrality [187]. As a definition, the eigenvector centrality $c_i^{(E)}$ of node i is the i-th coefficient of the normalized solution of the equation:

$$\boldsymbol{M}\vec{c}^{\,(E)} = \widehat{\lambda}\,\vec{c}^{\,(E)}, \tag{4.78}$$

where $\widehat{\lambda}$ is the highest positive eigenvalue [193]. This centrality measure accounts for both local and global properties of the network, as it is not just dependent on the degree of the node, but also on the positioning of each node in the network [188].

Our analytical investigation 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. For example, in Fig. 4.5(a) we have showed that the correlation between the eigenvector centralities and the measured Heaps' exponents at transient times for a simulation on the Zachary Karate Club network is higher than 0.98, and it persists changing the parameters in the simulations, even for sets of parameters in contrast with the approximations used in the analytical study, i.e. $\rho < \nu + (\nu + 1)\widehat{\mu}$. 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.

Pace of discovery and α-centrality

In this section we focus on generic directed graphs and the usage of the α-centrality as a proxy for the ranking of the nodes based on their pace of discovery in these more general cases. The α-centrality, widely used in network analysis [190, 191], has been first introduced in Ref. [189] to extend the eigenvector centrality to asymmetric graphs. The underlying idea is to tune the influence of the adjacency matrix structure with a parameter α to add exogenous sources to the centrality [36, 189]. Formally, it is defined as the vector $\vec{c}^{(\alpha)}$ such that

$$\vec{c}^{(\alpha)} = \alpha \boldsymbol{A} \vec{c}^{(\alpha)} + \vec{e}, \tag{4.79}$$

where $\vec{e}$ is an N-dimensional vector of ones. The matricial form of Eq. (4.79) reads

$$\vec{c}^{(\alpha)} = (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1} \vec{e} = \left(\sum_{k=0}^{\infty} \alpha^k \boldsymbol{A}^k \right) \vec{e}, \tag{4.80}$$

where $\boldsymbol{I}$ is the N-dimensional identity matrix. It has also been shown that this centrality is equivalent to the Katz-centrality [196] given by

$$\vec{c}^{(K)} = \left(\sum_{k=1}^{\infty} a^k \boldsymbol{A}^k \right) \vec{e}, \tag{4.81}$$

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 [189]. From Eq. (4.79) and (4.80), it is clear that the α-centrality can be both a local and global measure. In fact, for $\alpha \to 0^+$, the relative importance of the structure given by the adjacency matrix $\boldsymbol{A}$ decreases, in favor of the exogenous factor given by $\vec{e}$. With higher values of α, instead, the role of the exogenous part is damped.

For an undirected graph, the α-centrality becomes proportional to the eigenvec-

tor centrality when $\alpha \to (1/\widehat{\mu})^-$, where μ is the highest positive eigenvalue of the adjacency matrix. In fact, in this case all eigenvalues are real and the eigenvectors are orthogonal. Following Ref. [189], let $\{\mu_\ell\}$ and $\{\vec{u}_\ell\}$ be the (eventually multiple) eigenvalues and eigenvectors of the adjacency matrix $\boldsymbol{A}$, with $\widehat{\mu} = \mu_1 > \mu_\ell$ for $\ell \neq 1$. We can hence write $\boldsymbol{A} = \sum_{\ell=1}^{N} \mu_\ell \vec{u}_\ell \vec{u}_\ell^T$. Considering that $\boldsymbol{A}^k = \sum_{\ell=1}^{N} \mu_\ell^k \vec{u}_\ell \vec{u}_\ell^T$, from Eq. (4.80) we have

$$\begin{aligned} \vec{c}^{(\alpha)} &= \left(\sum_{k=0}^{\infty} \alpha^k \sum_{\ell=1}^{N} \mu_\ell^k \vec{u}_\ell \vec{u}_\ell^T \right) \vec{e} = \left(\sum_{\ell=1}^{N} \left(\sum_{k=0}^{\infty} \alpha^k \mu_\ell^k \right) \vec{u}_\ell \vec{u}_\ell^T \right) \vec{e} \\ &= \sum_{\ell=1}^{N} \frac{1}{1-\alpha\mu_\ell} \vec{u}_\ell \vec{u}_\ell^T \vec{e}. \end{aligned} \tag{4.82}$$

When $\alpha \to (1/\widehat{\mu})^-$, the factor relative to $\ell = 1$ in the last term of Eq. (4.82) becomes the leading term, thanks to the Perron-Frobenius theorem, so that we can write:

$$\lim_{\alpha \to (1/\widehat{\mu})^-} (1-\alpha_1)\vec{c}^{(\alpha)} = (\vec{u}_1^T \vec{e})\vec{u}_1 \propto \vec{u}_1 \propto \vec{c}^{(E)}, \tag{4.83}$$

where we have noted with $\vec{c}^{(E)}$ the eigenvector centrality.

Let us now generalize the analytical steps above to understand why the α-centrality correlates with the extracted Heaps' exponents for generic graphs, as showed numerically in Fig. 4.6. Let us suppose that the social network is a weakly-connected graph, since otherwise we can repeat the same argument for each weakly-connected component. As we have shown before, the asymptotic behavior of the Heaps' law for node i is of the type $u_i \ln^{\widehat{p}(i)}(t) t^{\widehat{\lambda}(i)}$. We have shown also that the values of $\widehat{p}(i)$ and $\widehat{\lambda}(i)$ for each strongly-connected component can be determined algorithmically. Here we will show that not only the α-centrality can account for the coefficient u_i like the eigenvector centrality, but also for the different values of $\widehat{p}(i)$ and $\widehat{\lambda}(i)$. Let us first concentrate on what happens when the highest eigenvalues appear with multiplicity higher than 1, for which the biggest difference in the dynamics is primarily given by $\widehat{p}(i)$. Therefore, let us suppose for now that all SCCs in the graph have the same Heaps' exponent $\widehat{\lambda}(i) = \widehat{\lambda}$, but different values of $\widehat{p}(i)$, and that in the leading terms the maximum value assumed by $\widehat{p}(i)$ is $\widehat{p}_{\max} < N$. This is the case for example of an open chain (already studied above), where $\widehat{p}(i) = 0, 1, \ldots, N-1$ for $i = N, N-1, \ldots, 1$ respectively, and $\widehat{p}_{\max} = N - 1$. Notice that, in this particular case, the adjacency matrix has only one eigenvalue $\widehat{\mu}$, related to the Heaps' exponent $\widehat{\lambda}$ through the relationship $\widehat{\lambda} = f(\widehat{\mu})$, with $f(x) = \frac{\nu}{\rho} + \frac{\nu+1}{\rho} x$. Therefore, the Jordan canonical form of the adjacency matrix

is

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1} = \boldsymbol{P}\begin{bmatrix} \widehat{\mu} & 1 & 0 & \cdots & 0 \\ 0 & \widehat{\mu} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \widehat{\mu} & 1 \\ 0 & \cdots & 0 & 0 & \widehat{\mu} \end{bmatrix}\boldsymbol{P}^{-1} = \boldsymbol{P}(\widehat{\mu}\boldsymbol{J_0} + \boldsymbol{J_1})\boldsymbol{P}^{-1}, \tag{4.84}$$

where $\boldsymbol{P} = \begin{bmatrix} \vec{u}_1 \mid \vec{u}_2 \mid \ldots \mid \vec{u}_N \end{bmatrix}$ has the generalised eigenvectors in each column, and $\boldsymbol{J_j}$ denotes the $N \times N$ matrix with ones only in the $(j+1)$-th upper diagonal and null everywhere else, with $\boldsymbol{J_0} = \boldsymbol{I}$. From Eq. (4.84) we can write

$$\boldsymbol{A}^k = \boldsymbol{P}(\widehat{\mu}\boldsymbol{J_0} + \boldsymbol{J_1})^k\boldsymbol{P}^{-1} = \boldsymbol{P}\left(\sum_{j=0}^{\min(N-1,\,k)} \binom{k}{j}\widehat{\mu}^{k-j}\boldsymbol{J_j}\right)\boldsymbol{P}^{-1}. \tag{4.85}$$

Hence, from Eq (4.80), similarly to what we have done in Eq. (4.82), we have

$$\begin{aligned}
\vec{c}^{\,(\alpha)} &= \left(\sum_{k=0}^{\infty}\alpha^k\boldsymbol{P}(\widehat{\mu}\boldsymbol{J_0} + \boldsymbol{J_1})^k\boldsymbol{P}^{-1}\right)\vec{e} \\
&= \left(\sum_{k=0}^{\infty}\alpha^k\boldsymbol{P}\left(\sum_{j=0}^{\min(N-1,\,k)}\binom{k}{j}\widehat{\mu}^{k-j}\boldsymbol{J_j}\right)\boldsymbol{P}^{-1}\right)\vec{e} = \\
&= \left(\sum_{j=0}^{N-1}\left(\sum_{k=j}^{\infty}\alpha^k\binom{k}{j}\widehat{\mu}^{k-j}\right)\boldsymbol{P}\boldsymbol{J_j}\boldsymbol{P}^{-1}\right)\vec{e} \\
&= \left(\sum_{j=0}^{N-1}\left(\sum_{k=0}^{\infty}\alpha^{k+j}\binom{k+j}{j}\widehat{\mu}^{k}\right)\boldsymbol{P}\boldsymbol{J_j}\boldsymbol{P}^{-1}\right)\vec{e} \\
&= \left(\sum_{j=0}^{N-1}\alpha^j\left(\sum_{k=0}^{\infty}\binom{k+j}{j}\alpha^k\widehat{\mu}^k\right)\sum_{\ell=1}^{N-j}\vec{u}_\ell\vec{u}_{\ell+j}^{\,T}\right)\vec{e} \\
&= \sum_{j=0}^{N-1}\sum_{\ell=1}^{N-j}\frac{\alpha^j}{(1-\alpha\widehat{\mu})^{j+1}}\vec{u}_\ell\vec{u}_{\ell+j}^{\,T}\vec{e} \\
&= \sum_{\ell=1}^{N}\left(\sum_{j=0}^{N-\ell-1}\frac{\alpha^j}{(1-\alpha\widehat{\mu})^{j+1}}\vec{u}_{\ell+j}^{\,T}\vec{e}\right)\vec{u}_\ell.
\end{aligned} \tag{4.86}$$

From Eq. (4.86) above, it is clear that the nodes ℓ for which $(\vec{u}_1)_\ell$ is positive have the greatest α-centrality when $\alpha \to (1/\widehat{\mu})^-$, since in the last term the contribute related to $l = 1$ and $j = 0$ tends to infinity for such nodes. Let us remind that these nodes are associated to the highest power in the logarithm $\widehat{p}(i) = \widehat{p}_{\max}$. Among these, as with the eigenvector centrality, nodes with higher coefficients, which correspond to the

eigenvector centralities in that SCC, have higher ranking. Then, the nodes who have zeroes in $\vec{u}_1$ but positive entries in $\vec{u}_2$ are next in the ranking, and so on. This confirms the fact that, when comparing nodes with same asymptotic Heaps' exponent, those with higher discovery rates, i.e. those with higher powers in the logarithm factor, have the highest α-centrality.

A similar approach to the one we used to derive the algorithmic solution of the Heaps'law in Sec. 4.5.5 for a generic graph can be used to treat the ranking in generic weakly-connected graphs. Let us divide the network into its SCCs. For each component C_q, we denote with $\mu^{(q)}$ the highest between the maximum eigenvalue the component would have if isolated and the maximum eigenvalue of the neighboring SCCs, following the same order used in the developed algorithm. In this setting, it is then possible to compute the α-centrality at $\alpha \to (1/\mu^{(q)})^-$, that might be different across SCCs. The final ranking is given by ordering the evaluated α-centralities starting from those with the highest $\mu^{(q)}$.

It is worth noticing that this method can be computationally not efficient, especially for big networks. For this reason, we test how reliable the α-centrality is if we choose a unique value of α close to $(1/\widehat{\mu})^-$ for all SCCs to be compared with the Heaps' exponents, regardless of the algorithmic procedure above. In Fig. 4.7, indeed, we investigate how the Spearman's rank correlation coefficient between the paces of discovery $\beta_i(10^4)$ and the α-centralities $c_i^{[\alpha]}$ changes as a function of α for 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 for any chosen value of α in this region. These results confirm what we have also found numerically in Fig. 4.6, where we have compared the extracted Heaps' exponents to the normalized α-centralities $c_i^{[\alpha]}/c_{\max}^{[\alpha]}$ in these networks, with $\alpha = 0.85/\widehat{\mu}$. The high values of the Spearman's rank correlations ($r_S \geq 0.97$ in all cases) found in both undirected [Fig. 4.6(a,c,d)] and directed networks [Fig. 4.6(b)] are in agreement with our analytical predictions. This confirms that, together with the adjacent possible (AP) in the content space, it is crucial to take into account of an AP in the social space, and that the α-centrality can be used as a predictor of the enhancement of the individual pace of discovery due to the social structure.

4.6 Summary and conclusions

In conclusion, in this chapter we have presented the UrNet model in which multiple processes of discovery are coupled over the nodes of a complex network, and analytical insights on the relations between structure and dynamics are possible. If in the previous chapters we have focused on the discovery process of an individual, here we have instead considered the dynamics of a group of individuals who explore and make

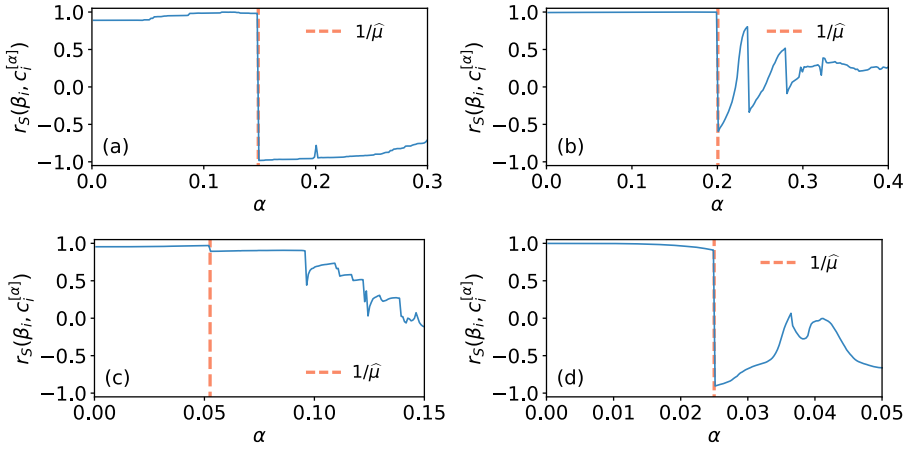


Figure 4.7: Spearman's rank correlation r_S between paces of discovery $\beta_i(10^4)$ and α-centrality $c_i^{[\alpha]}$ as a function of α for nodes $i = 1, \ldots, N$ belonging to four different real-world networks: (a) the Zachary Karate Club network [182], (b) a network of follower relationships of Twitter [183], (c) a co-authorship network in network science [184] and (d) a collaboration network between jazz musicians [185]. 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 $\rho = 10$, $\nu = 1$, $M_0 = \nu + 1$.

discoveries simultaneously. In particular, in the UrNet model, each explorer is associated to a node of a social network and equipped with an urn model with triggering controlling the individual discovery dynamics. Then, we couple the urns over the social network, expanding each urn through the adjacent possible in the social space. Thanks to this expansion, each urn is enriched with colors coming from their neighboring urns. Simulating the model on different structures, the results highlight that the structural—not just local—properties of the nodes can strongly affect their ability to make novelties. Let us remark that our UrNet model of socially-enriched urns is not just a simple extension of the UMT. What makes it novel and different is the very same idea of coupling together many urns over a complex social network, related to the concept of *"adjacent possible in the social space"* we have introduced. It is such a network coupling that spontaneously produces novel and heterogeneous behaviors, such as different exponents of the Heaps' law in a single system.

The work presented in this chapter represents only a first step toward the inclusion of structured interactions in discovery processes. Although the framework we have proposed is general, our results have focused on identical urns, so that the topological differences of the social network would be highlighted. Therefore, it would be interesting to study the relationship between differences in individual natural propensity to explore and their social structure. Moreover, urns can in fact result oversimplified models for the dynamics of individual explorers. As we have seen in Chapter 3, the

structural properties of the network of contents being explored can have a big impact on the discovery properties of the agents in the system. Therefore, in Chapter 5 we will consider an example where many interacting agents explore a complex network of music content, so as to investigate the relationship between adjacent possible in the content and social space in a real case study.

Another direct application of the UrNet could be related to team dynamics and optimization, in which more interdisciplinary and explorative individuals interact with others who are more specialist in a particular sector. To this end, one could extend the UrNet model by introducing social relationships unfolding across different network layers [197] or higher-order structures [198, 199], or through weighted social networks. For instance, we can see a team as a simplicial complex, in which each of the individuals shares their adjacent possible with the other teammates, while for all other pairwise connections only the discovered elements are available to the social contacts, similarly to what happens for scientific researchers reading papers suggested by their contacts. This study could hence lead to new insights on optimal team structures in discovery processes. This could in turn be exploited to obtain valuable insights on how to control and predict the impact of various constraints in the team structure on the emergence of novelties of the team, answering important questions from the perspective of funding and management choices. Due to the collaborative nature of the UrNet model, this could also be implemented in studies on efficient team structures in cooperative creative tasks [114, 200–203]. Results would be based on both collective and individual creativity-related measures, such as the discovery rate of the whole team, and the distribution of exploration rates of each team member.

Finally, let us list other possible extensions of the model, highlighting how general this framework is. As we have seen in Sec. 2.2.1, the reinforcement mechanism present in Polya's urns are very similar to the Barabási-Albert (BA) model of network generation [32]. In the recent past the BA model has been incrementally refined, starting from the Bianconi-Barabási model [143, 204] where each node is given a fitness, so that the usual "rich-get-richer" transforms into a "fit-get-rich" mechanism. It would hence be interesting to try and mimic a similar dynamics also on coupled urns, introducing different intrinsic weights, or fitness, to the various elements of the adjacent possible. This could be implemented in studies of substitutive systems [110] and social spreading processes [205–207], where the adoption of the new might trigger the abandoning of the old, thus leading to phenomena of waves of novelties [28, 97].

Chapter 5

Modelling the exploration of music in online platforms

5.1 Introduction and outline

In our everyday life, we are continuously exposed to novel ideas, new information, innovative cultural and technological products, and so on [107, 146, 208, 209]. Understanding the subtle balance between the exploration of new opportunities and the exploitation of what we already know is fundamental to unveil how we build our knowledge and set of skills [210–213]. Such a task is even more challenging if we consider that we live in a more and more interconnected society [32, 33, 36] and we do not explore the world alone. Indeed, we are constantly influenced by our peers, directly or indirectly [103, 176, 214, 215].

The recent increase in quantity and quality of digital traces has unlocked the possibility of tracking individual exploration trajectories in systems and processes as diverse as online music consumption [216–218], food purchases [219–221], Twitter post creation [222], and code development [28]. These opportunities have allowed, among other things, to identify the typical exploration patterns of individual users and to investigate the drivers that lead to a novelty, defined as the first time a user adopts or consumes a given content [8, 28, 95, 148]. We have indeed seen in Chapter 3 how the temporal analysis of the emergence of novelties, which can be extended to new combinations of a higher number of elements, can be used to characterize not only the pace of new discoveries in different systems, but also the growth of their inherent space of possibilities to make further discoveries.

Although social interactions play a crucial role in all these systems [1], as we have highlighted in Chapter 4, a thorough data-driven understanding of their impact on individual and collective exploration trajectories is still lacking. In this chapter, we try to

fill this gap by studying how social connections and communities affect the exploration patterns of different users of online (social) music platforms [2]. Our work is based on a unique data set that, differently from those used in previous studies [126, 223–228], contains information on both the whole listening histories and the social connections of a large and connected sample of users from the online music listening platform *Last.fm*. This platform is particularly suitable for our purposes, since it specifically encourages interactions among users and pushes them to explore new songs based on the rich metadata attached to each track [125]. As such, the data set we use represents the ideal testbed to *i*) characterize the exploration and discovery dynamics of each user, *ii*) measure the impact of social interactions, and *iii*) provide a structured view of the conceptual (or musical, in this case) space being explored by the agents [229–232]. A first analysis of the data set reveals that user exploration behaviors are very heterogeneous. We find that users with high discovery rates, the so-called explorers, tend to be connected with similar users in the social network, indicating the presence of homophily [233].

To get a better insight into the interplay between the topology of the social network and the different propensities of the users to explore, we introduce a multiagent model in which the agents simultaneously explore a space of contents. Our model, called *"ExploNet"* is based on the so-called *urn model with semantic triggering* (UMST) [8] discussed in Sec. 2.2.5, which is already capable of reproducing some statistical features of the empirical exploration trajectories, e.g., the Zipf's, Heaps' and Taylor's laws [25, 27–29]. UMSTs have been recently adapted both to model the evolution of a social network [30] (see Sec. 2.5.1) and to investigate peer effects in discovery dynamics, which we have done in Chapter 4. While the former model approach does not deal with content exploration [30], the latter lacks semantics in the space being explored [1]. Here, we explicitly consider a space endowed with a complex network of semantic relations between artists as given by similarity. As proposed in Chapter 3, we call this the *adjacent possible in the content space*. Multiple agents independently navigate this network with a reinforcement mechanism of the visited nodes, while also triggering the addition of new elements in the adjacent possible whenever a novelty appears. At the same time, agents interact with each other through an underlying social network, progressively enlarging their space of possibilities in what we have called the *adjacent possible in the social space* in Chapter 4. The collective nature of the dynamics allows us to correctly capture and reproduce the key empirical findings at a local and global level. At the local level, one observes an assortative arrangement of explorers, while at the global level, communities of people sharing similar music tastes emerge. The ensemble of these results offers valuable understandings regarding the mutual influence between the individual and collective experience of the new.

This chapter is organized as follows. In Sec. 5.2 we first crawl the Last.fm net-

work to obtain a smaller connected sample of active users. After analyzing the social network properties of this sample in Sec. 5.2, we download and investigate the complete listening history of such users, looking at their pace of discovery of new artists (Sec. 5.3) and at the presence of semantic correlations in the sequences (Sec. 5.2.4). We conclude our analysis of the social network in Sec. 5.2.5 by studying the interplay between the social network structure and both the individual rate of discovery and the overlap of tastes between users.

We use these empirical findings to create a data-driven agent-based model (ABM) in Sec. 5.3, leveraging on some mechanisms proposed in other models discussed in this thesis. We extract the content semantic structure from the empirical data in Sec. 5.3.1, and define the ABM in Sec. 5.3.2.

Subsequently, we run extensive simulations of the ABM, in which all agents follow the same evolutionary rules with the same parameters, finding, in Sec. 5.9, a reference set of parameters that reproduces the distribution of the pace of discovery and the semantic correlations similarly to the empirical data. We also find in Sec. 5.4.2 that the ABM correctly reproduces also the other empirical findings, for example regarding the impact of the social network on the pace of discovery and the overlapping tastes between friends. We further test the role of the social network by running and analyzing simulations in which, on the one hand, we switch off the interaction step in the model (in Sec. 5.4.3), and, on the other hand, we let the social interactions dynamically shape the social network structure (see Sec. 5.4.4). We conclude our analysis in Sec. 5.4.5 investigating how the results change if we run longer simulations, while in Sec. 5.4.6 we check the relationship between Heaps' and Zipf's exponent in the data and in the ABM simulations.

Finally, in Sec. 5.5 we summarize the results of this chapter, and we discuss further improvements of the ABM and future work.

5.2 The data set

Last.fm is an online digital music streaming platform born in 2002, famous for logging all listening activities, known as *scrobbles*, of its users [125]. We have crawled the *Last.fm* platform using its API, collecting all the listening sequences and social connections of a group of 4836 users. Such users have been found growing a breadth-first search sample from a random seed. We end up with 335 375 125 unique streaming events, complete of metadata and timestamps, with a total of 6 972 047 unique tracks authored by 958 732 artists. Let us hence start by going into more detail about how this data has been collected.

5.2.1 Data collection

The *Last.fm* data set presented in this chapter contains the complete listening history of a group of 4836 users and their social connections. The users have been selected randomly via a breadth-first search algorithm starting from a random seed node, using *Last.fm*'s API method `user.getFriends()`. More in detail, the initial seed is added to a growing list $\mathcal{L}$ of users to explore in a first-in-first-out manner. Therefore, at each step, the first user in $\mathcal{L}$ is included in the graph and removed from $\mathcal{L}$, while its neighboring users (or friends) that have not been sampled already are added at the end of $\mathcal{L}$. This process has been repeated until the graph has reached $10\,000$ users. Then, we have downloaded the complete history of streamed tracks of all users through the API `user.getrecenttracks()` endpoint. For each record, besides the timestamp, the API provides additional metadata, including the name and MusicBrainz Identifier [128] (MBID) of the track, artist, and album if present. Some of the users were not available to access correctly using the APIs above, or have not been very active in the platform. Therefore, we have filtered out those with less than $1\,000$ scrobbles. Finally, we have kept only the largest weakly connected component, resulting in a network of $4\,836$ users.

In total, the records span over almost 13 years, from August 2005 to February 2018, with $335\,375\,125$ unique streaming events, totaling $6\,972\,047$ tracks, $958\,732$ artists, and $1\,807\,150$ albums. We find that 31.8% of the set of tracks does not have an album, making up for 9.1% of all listening records in the data set. Considering the vast number of different tracks and the lack of consistency of albums, the analysis in this work has been carried out focusing only on the sequence of artists listened by the users, even though similar results can be obtained when considering the sequence of tracks instead.

Finally, let us remark that the data set we have collected is unique in nature and breadth. As a comparison, on the one hand, the music-listening histories data set presented in Ref. [227] consists of more than 27B logs from 583k users, for a total of 555k different artists and 46k logs per user on average. However, no social relations between the users are given, similarly to Refs. [126, 217]. On the other hand, in the data set shared in Ref. [234] the social relationships are present, but for each user there is only the tag assignments history, unfortunately much shorter (average length equal to approximately 98).

Our data set is available for download on figshare at `http://dx.doi.org/10.6084/m9.figshare.16652104`, while all the code to reproduce the results in this chapter is accessible at `https://github.com/gabriele-di-bona/Socially-enhanced-discovery-processes`.

5.2.2 Analysis of the social network

In Sec. 5.2.1 we have described how we have crawled the *Last.fm* platform growing a breadth-first search sub-graph from a random seed, ending up with a connected network made of 4836 users. In this section, we show that the sample we have collected is representative of the main statistics of the whole network.

Firstly, in Fig. 5.1, we see that this sample significantly reproduces the degree that users feature on the *Last.fm* platform. Quantitatively, we indeed have a Spearman rank correlation coefficient $r = 0.723$. ($p < 0.0001$) and a Pearson coefficient $r = 0.839$ ($p < 0.0001$) between the degree of a user in the sample network and the total number of followed users on the platform (the *original* degree). We further observe that the degree of a user in the sample is, on average, almost one tenth of the original degree on the platform. Therefore, in the sample we only maintain a smaller subset of the total connections of each sampled user. Nevertheless, the degree distribution has the same properties, since the original and sampled degrees are highly correlated. Notice also that the network is almost completely reciprocal (i.e., for each link $i \to j$, also the reciprocal link $j \to i$ exists), since 98.5% of existing links are reciprocal.

Moreover, in Fig. 5.2(A), we show a local snapshot of the users' social network $\mathcal{G}_S$, where nodes are colored according to their exploration propensity (i.e., their Heaps' exponent β_i defined in Sec. 5.2.3, the redder, the higher), their size is proportional to their betweenness centrality, and the link color intensity is proportional to the dynamical overlap (a measure of similarity between users based on their listening sequences, see Sec. 5.2.5 for the definition). Here we can clearly see at a glance how clusters of people with similar discovery rates are formed.

Furthermore, as shown in Fig. 5.2(B), we find that $\mathcal{G}_S$ features a scale-free out-degree distribution $P(k) \sim k^{-\mu}$, where k is the out-degree of a user, i.e., the number of users in the sample followed on *Last.fm*, and $\mu \approx 2.15$, with average out-degree

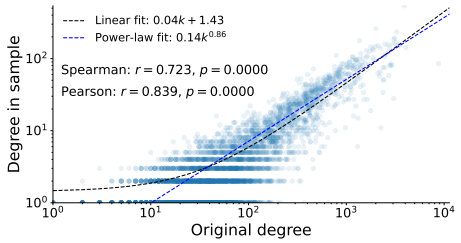


Figure 5.1: **Relationship between in-sample degree and number of followers in the data set.** We show the in-sample degree (y axis) versus the total number of followed users (x axis) on the *Last.fm* platform for all the users in the data set. We also plot the linear fit (black dashed line), $0.04k + 1.43$, and the power-law one (blue dashed line), $0.14k^{0.86}$. The values of Spearman and Pearson correlation coefficient are also displayed.

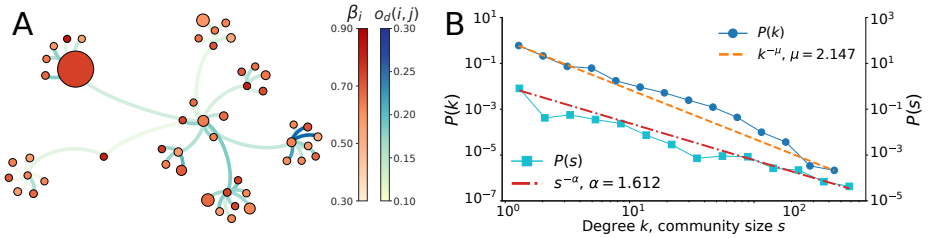


Figure 5.2: **Statistical analysis of the social network in the *Last.fm* data set. (A)** Sub-graph of the social network $\mathcal{G}_S$ of the analyzed sample. The node size is proportional to its betweenness centrality, while color intensity is proportional to its Heaps' exponent β (the redder, the higher). The color intensity of a link e_{ij} is proportional to the dynamical overlap $o_d(i,j)$ of the two nodes (the bluer, the higher). **(B)** Out-degree distribution $P(k) \sim k^{-\mu}$ (blue dots) and community size distribution $P(s) \sim s^{-\alpha}$ (cyan squares) of $\mathcal{G}_S$, both fitted as a power law with exponents $\mu \approx 2.147$ (orange dashed line) and $\alpha \approx 1.612$ (red dash-dot line).

$\langle k \rangle \approx 7.88$. The empirical distribution is plotted as a blue line with circle markers, while the power-law fit as a dashed orange line. $\mathcal{G}_S$ can be considered a typical small-world network [92], featuring, in fact, a small-world coefficient $\sigma \approx 1.35$ [235], with a relatively low characteristic path length (5.68) and high clustering coefficient (0.15).

Finally, the sample features a community structure [236]—i.e., nodes are arranged in tightly connected groups that are weakly linked between each other. Using the Louvain algorithm [237] we find 31 communities with more than ten users, with an average size of 150 users. In Fig. 5.2(B), we plot the community size distribution $P(s)$ (cyan line with squared markers), where the power-law fit (exponent $\alpha \approx 1.61$) is indicated as a red dash-dot line.

5.2.3 The pace of discovery

To quantify the exploration rate of new artists for the various users in the data set, we measure the Heaps' law exponent from the sequence $\mathcal{S}_i$ of artists listened by each user i in the sample, i.e., each node i of the social network $\mathcal{G}_S$. As we have seen in the previous chapters, the Heaps' law links the number $D(t)$ of distinct elements that are found in a sequence of t elements to a power-law behavior, namely $D(t) \sim t^{\beta}$, where $0 \leq \beta \leq 1$ is the *Heaps' exponent* [15]. The Heaps' law is also related to the Zipf's law [58, 81, 140], which states that the frequency f–rank R distribution of the elements in a sequence $\mathcal{S}$ decays as $f \sim R^{-1/\beta}$ for large ranks R [8, 57]. For a comparison between Heaps' and Zipf's laws for this system see Sec. 5.4.6.

In our analysis, the Heaps' exponent represents a natural proxy to measure the discovery rate of each user, which we hypothesize directly related to the propensity to search for new content during the discovery process. We extract the Heaps' exponent for each user by computing the slope in the log-log scale of the function $D(t)$ counting

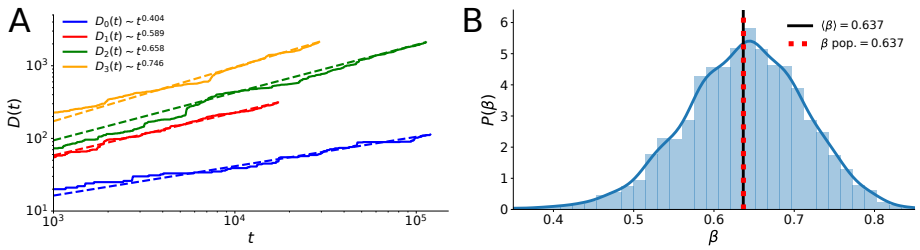


Figure 5.3: **The pace of discovery of new music by users of the *Last.fm* data set. (A)** Number $D(t)$ of distinct artists in the listening history of four random users in the data set as a function of the length t of the sequence (continuous lines). The Heaps' laws are also plotted (dashed lines), with their extracted exponents displayed in the legend. **(B)** Heaps' exponent distribution $P(\beta)$ for all the users in the data set. We also show the average Heaps' exponent $\bar{\beta} \approx 0.64$ (black line), and the global Heaps' exponent $\tilde{\beta} \approx 0.64$ measured on the global sequence $\mathcal{S}$ including all listening records in the whole data set (red dotted line).

the number of different tokens present in the sequence up to length t. In particular, if the sequence has length T, it is calculated as $\beta = \log(D(T))/\log(T)$. On average, for a power law behavior, this method gives a good approximation of the overall slope. Fig. 5.3(A) displays examples of Heaps' laws in four different users, highlighting how the Heaps' exponent is estimated from the listening sequences. Moreover, in Fig. 5.3(B), we observe how the Heaps' exponents β are heterogeneously distributed in the population. Some users are in fact more open to consume new music (higher β), while some others are more inclined to exploit already known tunes (lower β). We also note that a large fraction of the population features a discovery rate around the average $\bar{\beta} \approx 0.64$. Interestingly, this is the same value of the Heaps' exponent $\tilde{\beta} \approx 0.64$ found by measuring the Heaps' exponent on the whole global sequence $\mathcal{S}$, i.e., the single sequence obtained by merging, in temporal order, the records of all the users.

Notice that there are other methods to extract the Heaps' exponent. The problem arises when the power-law assumption is not completely verified, for example when the slope of the power law change over time. In this case, one might be interested in the slope of the tail specifically. Alternatively, given the points $(t, D(t))$, one can approximate $D(t)$ by fitting the best parameters of a function ax^{β_1} or $(1 + x/a)^{\beta_2}$. This way, though, there might be cases in which the fitted exponents β_1 or β_2 are higher than 1, which is not theoretically possible for the Heaps' law, since $D(t) \leq t$ is hardly constrained to be less than the linear function in time. The distributions of the Heaps' exponents according to these three different methods is shown in Fig. 5.4. As it can be seen, the number of cases for which β_1 or $\beta_2 > 1$ is significant in the two alternative methods. Since this problem does not arises with the method we have proposed, considering also that we are interested in the overall pace of discovery, we use the simpler approximation $\beta = \log D(t)/\log t$.

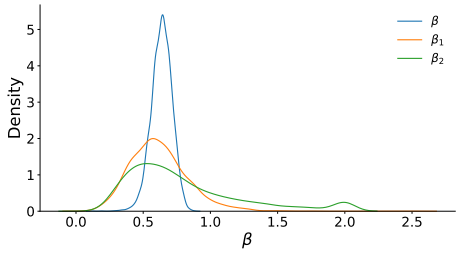


Figure 5.4: **Different estimation methods for the Heaps' exponents.** Comparison of the Heaps' exponent distribution on the sequences in the data set according to three different approximation methods. In particular, we extract $\beta = \log D(t)/\log t$, while β_1 and β_2 are found by fitting the functions ax^{β_1} and $(1 + x/a)^{\beta_2}$, respectively.

5.2.4 Semantic correlations

Another interesting observable to explore is the semantic correlations between the occurrences of songs from the same artist in the listening histories. It is indeed intriguing to quantify the extent to which the appearances of an artist in a chronologically ordered sequence of listening records are clustered, and therefore correlated. In simpler terms, we aim to determine whether listening to a particular artist increases the likelihood of listening to them again in the near future. To estimate these correlations, we compute, for each user i and each artist a listened by i, a modified Shannon entropy $S_i^a(f)$ introduced in Ref. [8], where f denotes the number of streams of artist a in the user's sequence $\mathcal{S}_i$. This entropy measures the extent of clustering among the events associated to the given semantic group, i.e., listening to songs of the same artist, with a larger cluster denoting stronger correlations among their occurrences.

In particular, let us define the number f_i^a of occurrences, or frequency, of artist a in the sequence $\mathcal{S}_i$ (whose length is T_i) of user i. We identify the sub-sequence $\mathcal{S}_i^a$ of $\mathcal{S}_i$ starting at the first occurrence of a, and we divide $\mathcal{S}_i^a$ in f_i^a parts of equal length. Computing the frequency $f_i^a(x)$ of a in each x^{th} interval, we obtain the normalized Shannon entropy of artist a in $\mathcal{S}_i$ as

$$S_i^a(f_i^a) = -\frac{1}{\log f_i^a} \sum_{x=1}^{f_i^a} \tilde{f}_i^a(x) \log \tilde{f}_i^a(x), \tag{5.1}$$

where $\tilde{f}_i^a(x) = f_i^a(x)/f_i^a$. When the occurrences are equally distributed in the f_i^a intervals, $S_i^a(f_i^a)$ hits its maximal value 1, whilst the entropy is at its minimum value $S_i^a(f_i^a) = 0$ when all the events are found in a single interval.

In Fig. 5.5 we plot the average entropy $S(f)$ as a function of the frequency f, where the average is made over all artists a and over all users i with frequency $f_i^a = f$ in $\mathcal{S}_i$.

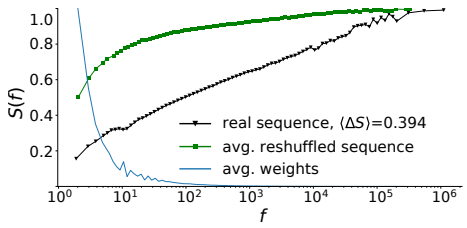


Figure 5.5: **Semantic correlations in the empirical sequences of exploration of users in the *Last.fm* data set.** Average normalized Shannon entropy $S(f)$ in the sequences of listened artists as a function of the artist frequency f (black line), compared to the average entropy $\tilde{S}(f)$ measured on the reshuffled sequences (green line). The weights for each frequency f (cyan line) are also shown, which are used to compute the weighted Shannon entropy difference $\langle \Delta S \rangle \approx 0.394$.

More precisely, we define the average Shannon entropy at a certain frequency f as

$$S(f) = \langle S_i^a(f_i^a) \rangle_{i,a|f_i^a=f}. \tag{5.2}$$

To test the statistical significance of this measure, we compare $S_i^a(f_i^a)$ with the randomized counterpart $\tilde{S}_i^a(f_i^a)$, i.e., the normalized Shannon entropy related to artist a listened by the user i computed after reshuffling the sequence $\mathcal{S}_i$. Therefore, we also plot the average normalized Shannon entropy $\tilde{S}(f)$ computed on the reshuffled sequence. The empirical data clearly shows a lower entropy (higher clustering) of the same-artist streaming events when compared to the reshuffled case.

To quantify the difference with the randomized sequence, we weight each frequency f with the global popularity w_f, namely the number of times we find an artist with frequency f in the different user sequences. Mathematically, we define w_f as the number of times $|\{i, a|f_i^a = f\}|$ an artist a appears in a sequence $\mathcal{S}_i$ with frequency f. The distribution of these weights is plotted in Fig. 5.5 as a blue line. Thus, we compute the weighted Shannon entropy difference $\langle \Delta S \rangle$, defined as

$$\langle \Delta S \rangle = \frac{\sum_f w_f \left(\tilde{\mathcal{S}}(f) - \mathcal{S}(f) \right)}{\sum_f w_f}. \tag{5.3}$$

Therefore, higher values of $\langle \Delta S \rangle$ are related to the presence of non-trivial semantic correlations in the process. In particular, in our data set, we obtain $\langle \Delta S \rangle \approx 0.394$.

Overall, the presence of significant semantic correlations implies that users tend to listen to music semantically close to the recent plays. Still, they also experience new content from time to time, according to the Heaps' law. To take into account the balance between semantic correlations and pace of discovery in the exploration process, both the weighted Shannon entropy difference $\langle \Delta S \rangle$ and the Heaps' exponent distribution $P(\beta)$ will be used in Sec. 5.3 to fit the simulations of the model to the data set.

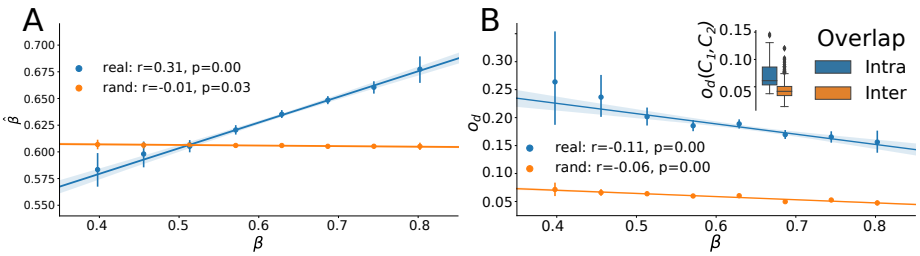


Figure 5.6: **The influence of the social network on the pace of discovery in the data set. (A)** Assortativity of the Heaps' exponents: the average exponent $\hat{\beta}$ measured over the neighbors of a user with exponent β is plotted as a function of β. Results from the original social network $\mathcal{G}_S$ (blue) are compared to those from the randomized network (orange), obtained using a configuration model. **(B)** Dynamical overlap o_d of each user with its neighbors as a function of the Heaps' exponent β of the user. Both the original (blue) and the randomized social network (orange) are shown. In the inset, the intra- (blue) and inter-community (orange) dynamical overlap distribution $o_d(C_1, C_2)$ are compared, where in the former $C_1 = C_2$, and in the latter $C_1 \neq C_2$.

5.2.5 The influence of the social network

We now shift the attention from the individual to the collective level. In particular, we focus on the relationship between the position of users in the social network and their respective exploration strategies. As we have qualitatively shown in Fig. 5.2(A), users tend to interact with people featuring a similar discovery rate and musical tastes. We quantitatively explore this assortativity in Fig. 5.6(A), where we observe a positive correlation between the Heaps' exponent β_i of a user i and the average exponent $\hat{\beta}_i = \langle \beta_j \rangle_{j \sim i}$ of its neighbors, featuring a Pearson correlation coefficient $r \approx 0.31$ ($p < 0.0001$). To test its significance, we measure the same correlation on a network obtained by randomly rewiring the edges of $\mathcal{G}_S$, obtaining in this case $r \approx -0.01$ ($p < 0.05$). This evidence is a clear sign of homophily based on the discovery rate β_i. In other words, explorers (exploiters), i.e., people with higher (lower) exponent β tend to form clusters with other explorers (exploiters).

The influence of the social network can also be measured by looking at the dynamical overlap $o_d(i, j)$ of a pair of users i and j, i.e., the fraction of common artists they listen to. This is calculated as

$$o_d(i, j) = \tilde{\mathbf{v}}_i \cdot \tilde{\mathbf{v}}_j, \tag{5.4}$$

where $\cdot$ is the scalar product and $\tilde{\mathbf{v}}_i$ is the vector of the normalized frequency distribution of artists listened by user i. If we average the dynamical overlap of a node i with its neighbors, that is

$$o_d(i) = \langle o_d(i, j) \rangle_{j \sim i}, \tag{5.5}$$

we can then compare it against its discovery rate β_i, as shown in Fig. 5.6(B). We find that the average dynamical overlap is much higher than the one calculated on the rewired network. This evidence represents another signature of homophily, i.e., friends share similar tastes. Moreover, we notice that $o_d(i)$ has a small negative correlation with the discovery rate β_i of the corresponding user (Pearson's $r \approx -0.11$ against $r \approx -0.06$ in the rewired network). This result reveals that explorers tend to interact slightly more with people sharing different musical tastes, thus enlarging the set of artists and genres they are exposed to. On the contrary, exploiters preferably surround themselves with people sharing similar tastes, limiting their chances to explore new content.

Moving away from the local scale, we test if the self-organization of users based on their tastes holds at the community level too. Therefore, for each pair of communities C_1, C_2 in the social network we compute the inter-community dynamical overlap as the average overlap of all possible pairs of individuals between C_1 and C_2, that is

$$o_d(C_1, C_2) = \langle o_d(i,j) \rangle_{i \in C_1, j \in C_2, i \neq j}. \tag{5.6}$$

When $C_1 = C_2 = C$, we obtain the intra-community overlap, defined as

$$o_d(C) = \langle o_d(i,j) \rangle_{i,j \in C, i \neq j}. \tag{5.7}$$

In the inset of Fig. 5.6(B) we compare the inter- and intra-community overlap distributions, considering only communities with at least ten users. We show that the average intra-community overlap (0.074) is significantly larger than the average inter-community one (0.039). In other words, users tend to create contacts within a community of people sharing similar musical tastes.

5.3 ExploNet: a model of collective exploration

In Sec. 5.2 we have analyzed different properties regarding the exploration of the space of artists in Last.fm. In particular, we have found three main results: *i)* users feature different propensities to explore new content, showing a heterogeneous distribution of discovery rate β; *ii)* individuals tend to interact and cluster with others that share a similar propensity to explore new content, highlighted by the positive assortativity of the Heaps' exponent β; *iii)* social connections are mainly established between groups of people sharing similar tastes, as proved by the higher dynamical overlap between friends and within communities. In this section, we develop an agent-based model (ABM) of content exploration and social interactions that reproduces these results. Before defining the model, which we refer to as *ExploNet*, let us measure the semantic structure of the content space to be explored from the empirical data, as we show in the

following paragraph.

5.3.1 Topology of the content space

The temporally-ordered sequences in our data set, containing the complete listening history of each user, have allowed us to characterize the discovery process of different people on Last.fm. Let us now use the same data to extract a semantic structure of the content space. Because of the scarcity of other meaningful musical features attached to the records in the data set, we quantify the similarity between artists from sequences of listening events. In doing so, we are hypothesizing the existence of an underlying semantic structure of the musical space that users navigate to consume content [29], as we have also shown in Chapter 3.

To this end, we construct a bootstrap-like statistics for the number of times, N_{ab}, a user listens to artist a and b sequentially (irrespectively of the order of a and b). In particular, we start from the reduced sequence $\tilde{\mathcal{S}}_i$ of each user i, i.e., the temporally-ordered sequence of artists listened by i without consecutive repetitions (multiple consecutive streams of one artist reduce to a single event). We then measure the number of times, $N_{a\to b}$, users listen to artist b after listening to a, with $a \neq b$. Since we aim at an undirected network of proximity, we define $N_{ab} = N_{a\to b} + N_{b\to a}$. Notice that in this step we filter out all artists with an overall number of appearances in the reduced sequences less than a threshold $\bar{f}$ to remove noise.

In order to have a statistically relevant measure of proximity, we repeat these steps $Q = 100$ times after shuffling all reduced sequences $\tilde{\mathcal{S}}_i$, counting for each reshuffle q the number N_{ab}^q of realization of the pair $a \to b$ or $b \to a$. Therefore, we compute the expectation $\overline{N}_{ab} = \langle N_{ab}^q \rangle_q$ and the standard deviation $\sigma_{ab} = \mathrm{std}(\mathrm{N_{ab}^q})$ of the count N_{ab}, under the assumption that there are no semantic relations in the sequences of streams. We hence define the proximity of two artists via the z-score $z_{ab} = (N_{ab} - \overline{N}_{ab})/\sigma_{ab}$. With this procedure, whose steps are illustrated in Fig. 5.7(A), we create a similarity network, drawing a link between two artists a and b whenever the related z-score z_{ab} is higher than a threshold $\bar{z}$.

Choosing $\bar{z} = 1$ and $\bar{f} = 100$ we obtain an undirected, weighted network $\mathcal{G}_C$ with 266 694 nodes and 17 765 819 edges, where we set the weight w_{ab} of a link to be $w_{ab} = min(z_{ab}, 100)$, directly related to the closeness of the two artists. A snapshot of such space of contents is shown in Fig. 5.7(B), where a snowball sample of the neighborhood of the artist *Pink Floyd* is displayed. In Fig. 5.7(C-D), we also show that both the z-score distribution and the degree distribution of the network encoding the space of content are stable with respect to the choice of different values of $\bar{z}$ and $\bar{f}$.

In summary, the procedure defined above transforms the universe of items in the data set—the artist IDs—into a weighted undirected network $\mathcal{G}_C$ where we can measure distances between nodes. This representation allows us to look at the sequence of

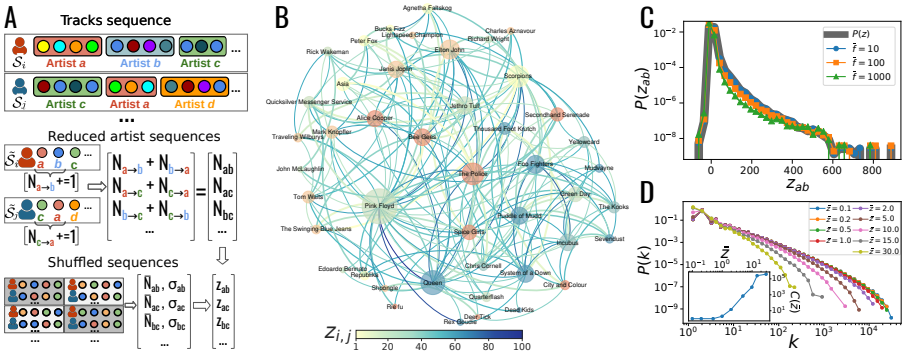


Figure 5.7: **From streams of songs to the creation of the content space. (A)** Illustration of the method. First, for each user i's sequence $\mathcal{S}_i$ we compress all consecutive appearances of an artist, say a, into a single occurrence of a, forming, in this way, the reduced sequence $\tilde{\mathcal{S}}_i$. Then, we count the number $N_{ab} = N_{a\to b} + N_{b\to a}$ of pairs of consecutive artists a and b in both orders in $\tilde{\mathcal{S}}_i$. We repeat the procedure $Q = 100$ times after reshuffling $\tilde{\mathcal{S}}_i$, calculating N^q_{ab} for each reshuffle q. We thus evaluate the z-score $z_{ab} = (N_{ab} - \bar{N}_{ab})/\sigma_{ab}$, where $\bar{N}_{ab}$ and σ_{ab} are respectively the average and standard deviation of all N^q_{ab}, for each pair of artists a and b. Artists with overall frequency less than $\bar{f}$ are disregarded, while we draw a link between two artists a and b if $z_{ab} > \bar{z}$. **(B)** Snowball-sampled snapshot of the neighborhood of *Pink Floyd* in the space of artists. Node sizes are proportional to their degree, while their color depends on the community of belonging. The color of the edges denotes their weight according to the z-scores, the bluer, the larger. **(C)** Comparison of the original z-score distribution (gray) with those obtained using different thresholds on the artist frequencies $\bar{f}$. **(D)** Degree distribution of the artists' network for different values of the threshold $\bar{z}$. The inset shows the number of connected components of the network as a function of the threshold $\bar{z}$.

listening records as the sequence of exploration steps of a network of artists. Moreover, knowing the semantic distance between artists, we can measure the propensity of users to explore items falling outside their comfort zones, as well as their willingness to accept recommendations (from others) not strictly meeting their current musical tastes. This mechanism will be a crucial ingredient of the model introduced below.

5.3.2 Model definition

To better understand the interplay between individual exploration and social interactions, let us develop an agent-based model (ABM), which we name *"ExploNet"*, capable of reproducing the empirical properties found in the Last.fm data set. We build upon the individual exploration-exploitation dynamics introduced in the *Urn Model with Semantic Triggering* (UMST) and analyzed in Sec. 2.2.5, drawing on concepts developed in Chapter 3 for the exploration of a content space and in Chapter 4 for the expansion of such space through social interactions.

In the UMST, the dynamics of an individual, or, from now onwards, an agent, is

modeled as random extractions of colored balls from an urn $\mathcal{U}$ to form a sequence of events $\mathcal{S}$ [8]. The urn $\mathcal{U}$ represents the *space of possibilities*, i.e., the set of possible choices the agent can make in the future. This space includes the so-called *actual space*, i.e., the subset of items already extracted by the agent and stored in a sequence $\mathcal{S}$. In addition to the actual space, the urn contains the so-called *adjacent possible space*, which consists of all those colors that are one step away from the actual space [9, 10, 13]. In the original UMST, the concept of proximity between colors is modeled through the definition of semantic relations between groups [8]. The idea is that the agent can realize only a subset of all the possibilities at any given time, preferentially extracting balls semantically related to the most recent ones. In particular, at every extraction, balls in $\mathcal{U}$ whose color is semantically related to the last drawn color keep their unitary weight. In contrast, all other balls get—temporarily—a weight $\eta \leq 1$ (with $\eta = 1$ we recover the classic urn model with triggering). Then, the agent draws a random ball with a probability proportional to these weights. The selected ball is put back in the urn with ρ additional copies of it (*reinforcement step*). Finally, if this ball has never appeared in $\mathcal{S}$, $\nu + 1$ brand-new balls are added to the urn (*triggering step*).

For the construction of the ExploNet model, we extend the UMST to account for the exploration of a shared conceptual space and peers' influence, taking inspiration from the edge reinforced random walk with triggering introduced in Chapter 3 for the former and from the interacting urns discussed in Chapter 4 for the latter. In particular, we allow N agents to independently explore a shared network of items $\mathcal{G}_C$ growing a personal space of possibilities, and also to interact with each other by exchanging information via a social network of contacts $\mathcal{G}_S$. Notice that the space of contents $\mathcal{G}_C$ to be explored can be, in general, a directed and weighted network, where link weights represent the strength of the semantic relation between pairs of items. Depending on the context, these items could be ideas, molecules, genomes, technological products, artists, etc. In the following analysis, we use the proximity network between artists we have created in Sec. 5.3.1. Regarding instead the social network $\mathcal{G}_S$, this is a directed and weighted graph too, whose links $i \to j$, of weight w^{ij}, may generally change over time. The weight of these links indicate the propensity of a node i to follow and copy the tastes of a friend j. To compare the model to the empirical data set analyzed in the previous sections, we consider the same $N = 4836$ agents as in the crawled *Last.fm* data set, using their social relationships for $\mathcal{G}_S$. Since we do not have information on how strong the relationships between users are, we assume to have unitary weights $w^{ij} = 1$ for each link $i \to j$ in the social network, zero otherwise. In Sec. 5.4.4, we will instead consider an initial complete graph, letting the agents evolve the weights of their links according to the interactions they have.

In this ABM representation, each agent i gradually increases his space of possibilities, that is, the set of all possible items that the agent can explore at any given time. We indicate this space as $\mathcal{G}_C^i$. The expansion of $\mathcal{G}_C^i$ happens in two dimensions. On

the one hand, $\mathcal{G}_C^i$ contains the set of nodes that have been discovered by i, that is the *actual space*, as well as the subset of items in $\mathcal{G}_C$ that are one step away from elements in i's actual space, i.e., the *adjacent possible in the content space*. On the other hand, $\mathcal{G}_C^i$ can be expanded through interactions with other agents, thus including an *adjacent possible in the social space*.

Finally, notice that we work in the system intrinsic time t, i.e., each listening event by any user increases the global time t to $t+1$. In particular, as summarized in Fig. 5.8, each time step is composed of two processes. An agent i, active at time t, *i*) independently explores its space of possibilities $\mathcal{G}_C^i$, and, *ii*) interacts with one of its neighbors in $\mathcal{G}_S$ to query for recommendations. The choice of the active agent i at time t can be done in different ways. In the following analysis, we opt for the same order found in temporally-ordered sequences of the empirical data set. Before explaining the two mentioned step in detail, let us go over how we initialize the space of possibilities of each user.

Initialization

At the beginning of the simulation, each agent i is placed on a node $x_i(t=0)$ of the conceptual space $\mathcal{G}_C^i$, so that the exploration of such space starts from these points. The problem of the initialization is related to whether to choose more widely distributed initial nodes, or, on the other extreme, the same node for every agent. It turns out that this choice has little impact on the outcome of the simulations, only delaying or reducing some effects due to interactions with peers initially too distant or too close. Since in this work we have considered the same agents as in the Last.fm data set crawled, we choose to initialize the agent's position with the first artist that the corresponding user has listened to in the empirical data.

This solution becomes natural when introducing an initial warm-up phase, in which the first steps of the model are not randomly chosen according to the rules of the model, but are instead the same steps made by the corresponding user in the empirical data. The impact of the length of the warm-up phase is shown in Sec. 5.4.4.

Finally, let us notice that another viable option for the initialization could be to choose the most frequent artists in the corresponding empirical listening history. In a way, this is similar to consider an initial warm-up phase, in which the initial space of possibilities contains a portion of the content explored by the corresponding user in the empirical data.

Exploration step

The exploration step of the ABM can be considered an adaptation of the UMST to the exploration of a universal space of content $\mathcal{G}_C$, which guides the expansion of the adjacent possible. Without loss of generality, let us suppose that agent i is active at

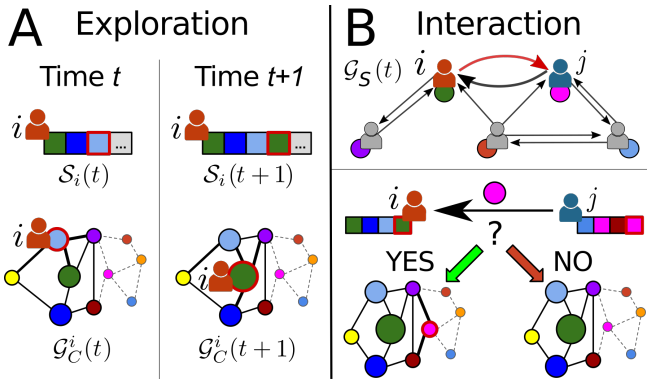


Figure 5.8: **Illustration of the ExploNet model. (A)** *Exploration step.* Agent i explores an item, represented by a colored node, at each time step. The node seen at time t is the cyan one, as highlighted in the sequence $\mathcal{S}_i(t)$ (top left). The space of possibilities $\mathcal{G}_C^i(t)$ of agent i at time t is represented as a weighted network (bottom left), where nodes that belong to the adjacent possible have a black continuous edge, while those connected through a gray dashed edge are not yet reachable to i. The bigger the node, the more it has been reinforced and the more likely it is to visit again. Here, we have highlighted the position of agent i on the cyan node. With probabilities depending on the weight of the nodes and on the semantic relation represented by the links, agent i moves to another color at time $t+1$. Notice that the outgoing links from the current position (cyan) are bigger, indicating a higher chance to move to these semantically related nodes. In this illustration, i chooses to move to the green node. Hence, its color is stored in $\mathcal{S}_i(t+1)$ (top right), i moves to the corresponding node in $\mathcal{G}_C^i$ and its weight is reinforced (bottom right). **(B)** *Interaction step.* The agents are connected to each other through a social network $\mathcal{G}_S(t)$ (top). We have represented the current position of each agent in their space of possibilities with a colored node underneath. After the exploration step, agent i randomly selects one of his friends to interact with, for example j, and their link is activated (red link). Therefore, agent j proposes its current pink node to i (middle). Depending on the distance of this node to the most reinforced nodes of i, the suggestion is accepted or refused. Notice how the addition of the pink node changes the space of possibilities $\mathcal{G}_C^i(t+1)$ of i for its next exploration step (bottom left). If the item is too far from i's tastes, the interaction does not lead to any change in $\mathcal{G}_C^i(t+1)$ (bottom right).

time t, and let $\mathcal{G}_C^i(t)$ be i's space of possibilities. We assume that $\mathcal{G}_C^i(t)$ is a weighted network growing in time, where the weight $w_a^i(t)$ of an item a in $\mathcal{G}_C^i(t)$ depends on the past history of i. Let us suppose that the current position of i on $\mathcal{G}_C^i(t)$ is $x_i(t)$. In the time step $t \to t+1$, i randomly moves to another node of his space of possibilities $\mathcal{G}_C^i(t)$, with probability depending directly on the weights $w_a^i(t)$ of all artists a in $\mathcal{G}_C^i(t)$ (*movement step*). Similarly to the UMST, we increase the probability to move to nodes semantically close to the last visited one. Therefore, we introduce a jump parameter $\eta \leq 1$ controlling the probability to move to neighbors of $x_i(t)$ versus jumping to other nodes. In particular, $2(\nu+1)$ nodes a are randomly sampled from the neighbors of $x_i(t)$ in $\mathcal{G}_C^i(t)$. Among these, we also allow $x_i(t)$ to be chosen again, recalling that a

user can consecutively listen to the songs of the same artist. Moreover, in order for the social recommendations to take effect, we further choose $2(\nu + 1)$ nodes surrounding (and including) the last node accepted via social interaction, if any recommendation has been accepted in the last interaction of i. Then, for the choice of the next node, on the one hand these sampled nodes keep their original weight $w_a^i(t)$; on the other hand all other nodes are considered with a reduced weight $\eta w_a^i(t)$. Notice that for $\eta = 1$ this process is equivalent to the UMT discussed in Sec. 2.2.4, since the agent can move to any node of his space of possibilities according only to their weight. On the contrary, for $\eta = 0$ this corresponds to a standard random walk with node reinforcement on the content space, i.e., there are no jumps to distant nodes.

After the choice of the next node $x = x_i(t + 1)$ is done, the item x is saved in the sequence $\mathcal{S}_i(t+1)$ of events of i. Moreover, the weight of x in $\mathcal{G}_C^i(t+1)$ is reinforced by $\rho > 0$, i.e., its weight becomes $w_x^i(t + 1) = w_x^i(t) + \rho$ (*reinforcement step*). Additionally, if it is the first time that i visits x, namely x has never appeared in $\mathcal{S}_i(t)$, then i expands its adjacent possible space by adding $\nu+1$ new items semantically close to x into $\mathcal{G}_C^i(t+1)$ (*triggering step*). In particular, $\nu+1$ new nodes randomly selected from the neighbors a of x in $\mathcal{G}_C$ are chosen with probability depending on the semantic relation w_{xa} between them. If the number of new neighbors—not already present in $\mathcal{G}_C^i(t)$—is less than $\nu + 1$, the search extends to nodes at a distance of two from x.

The mechanisms of reinforcement and triggering, which depend respectively on the parameters ρ and ν, account for the balance between listening again to artists already known by the agent (exploitation) and the expansion of the space of possibilities including new artists never heard by the user but semantically close (exploration). In particular, the ratio between ρ and ν sets the relative weight of exploitation and exploration. Moreover, the value of the jump parameter η influences how easily the agent can move to more distant items of its space of possibilities. Furthermore, as we will show later, the presence of social interactions represent a key ingredient to reproduce the assortativity of the exploration rates and the clustering of individuals with similar tastes.

Interaction step

After the exploration step, agent i randomly selects another agent among its neighbors in $\mathcal{G}_S$ with a probability distribution given by the weights w^{ij} of the links $i \to j$. Let us suppose that agent j is chosen. To mimic the dynamics of an online listening platform like *Last.fm*, i has the possibility to see what j is listening at the moment, i.e., what is the last token $y = x_j(t)$ explored by j, and, if interested, include it among his future possibilities in $\mathcal{G}_C^i(t + 1)$. We estimate i's potential interest in y with the semantic distance of y from i's core, i.e., the set of the first c items ranked by frequency found in the sequence $\mathcal{S}_i(t)$, which we refer to as c_i. Agent i actually adds y into $\mathcal{G}_C^i(t+1)$ with

probability $P = P(y|\mathcal{S}_i(t), c_i, \varepsilon)$, where ε is a noise factor that mimics i's imperfect capability of estimating the distance of items from the core c_i. In order to calculate the value of P, we first define the distance of a node from the core as

$$d_{c_i}(a) = \min_{b \in c_i} \{\text{dist}(a, b)\}, \tag{5.8}$$

where $\text{dist}(a, b)$ is the standard network distance between two nodes of the space $\mathcal{G}_C$. Then, we compute the following density function:

$$p(d|\mathcal{S}_i(t), c_i) = \frac{|\{a \in \mathcal{S}_i(t) : d_{c_i}(a) = d\}|}{|\mathcal{S}_i(t)|}. \tag{5.9}$$

Therefore, we can calculate P as

$$P = P(y|\mathcal{S}_i(t), c_i, \varepsilon) = \max\left(0, \min\left(1, \sum_{d=0}^{d_{c_i}(y)} p(d|\mathcal{S}_i(t), c_i) \pm \varepsilon\right)\right). \tag{5.10}$$

No significant differences in the analysed observables have been detected changing the size c of the core and the error ε. In the simulations shown in Sec. 5.4, we have fixed $c = 10$ and $\varepsilon = 0.1$.

Finally, the ExploNet model also allows the social network to co-evolve with the interactions. In this dynamical version, the weight of the link $i \to j$ is increased (decreased) by $+\Delta$ ($-\Delta$) if i accepts (rejects) the recommendation. In other words, we have $w^{ij}(t+1) = w^{ij}(t) + \Delta$ with probability P, or $w^{ij}(t+1) = w^{ij}(t) - \Delta$ with probability $(1 - P)$, with minimum weight equal to Δ. In the following, for the sake of simplicity, we focus on the case in which $\mathcal{G}_S$ does not change over time ($\Delta = 0$). We refer to Sec. 5.4.4 for the results on an evolving $\mathcal{G}_S$, starting from a fully connected network.

5.4 Results

In Sec. 5.2 we have investigated a data set containing the empirical sequences of listening activity of a group of users from the online plaform Last.fm, highlighting the heterogeneity in the pace of discovery of new content and the impact of social connections on the exploration process. We have hence built an ABM where multiple agents explore a shared network of artists, building their own space of possibilities, and interact with their friends. Therefore, in this section, we numerically investigate if this ABM is able to reproduce the key findings above. Firstly, in Sec. 5.4.1, we develop a procedure to select the parameters that reproduce a distribution of exploration rates similar to the empirical data. Secondly, we run multiple simulations to see if the ABM reproduces the findings of the data in Sec. 5.4.2. In order to understand the specific

role of social interactions, we also simulate the ABM without the interaction step in Sec. 5.4.3, while we let the social network dynamically evolve with the process in Sec. 5.4.4. Finally, we investigate finite-time effects on the simulations in Sec. 5.4.5 and the relationship with Zipf's exponents in Sec. 5.4.6.

5.4.1 Selection of parameters

Let us start our analysis of the ABM by developing a method to select a reference set of parameters that reproduce the main characteristics present in the data. Let us remember that, although in principle different parameters can be chosen for each agent, we consider the same set of parameters for all agents. We numerically simulate the ExploNet model using the fixed social network from Last.fm and the space of tokens given by the artists listened in the data set. Since the simulations are heavily computing-demanding, we explore the parameter space limiting ourselves to the first 10% of the original sequences in the data set, that is about 33.5 million total steps. In Sec. 5.4.5 we show how the reference set of parameters found in the limited sequence matches the empirical findings even when we let the simulation run for as many steps as the empirical number of streams. We explore the space of parameters varying the set of parameters (η, ρ, ν). In particular, we try the following values for the jump parameter η: 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, $\ldots$, 1; we also let the reinforcement parameter ρ span from 1 to 4 at steps of 0.1. Then, based on the results given in Ref. [8] on the limits of the Heaps' exponents in the UMST (see also Sec. 2.2.5), we consider integer values of the triggering parameter ν ranging from 0 to $\max(20,\, 1.5\rho/\eta)$. No significant differences have been observed changing the values of the core size c or the error ε in the probability to accept a recommendation. We hence fix them to $c = 10$ and $\varepsilon = 0.1$.

To select the reference set of parameters, we focus on just two basic footprints of discovery processes, namely the Heaps' exponent distribution, responsible for the balance between exploitation and exploration strategies, and the normalized Shannon Entropy of the distributions of artists in the listening sequences, which captures the presence of semantic correlations during music consumption. Since these two measures are not comparable, we first find the reference value of η by considering, for each simulated set of parameters $\mathbf{x}$, the normalized Shannon entropy weighted difference $\langle \Delta S^{(\mathbf{x})} \rangle = \sum_f w_f (\tilde{S}^{(\mathbf{x})}(f) - S^{(\mathbf{x})}(f)) / \sum_f w_f$ between the normalized Shannon entropy $S^{(\mathbf{x})}(f)$ calculated on the sequences generated by the simulation and the one $\tilde{S}^{(\mathbf{x})}(f)$ for the reshuffled ones (see Sec. 5.2.4 for the analysis of this measure on the data set). We group the various values of $\langle \Delta S^{(\mathbf{x})} \rangle$ by the parameter η of the corresponding simulations, that is 0.001, 0.01 0.05, 0.1, and > 0.1 (i.e., 0.2, 0.3, $\ldots$, 1), plotted in Fig. 5.9. We notice that only the groups of $\eta = 0.01$ and $\eta = 0.001$ range in an interval that includes the value of $\langle \Delta S \rangle$ calculated in the empiric data set in their first 10%. Therefore, we restrict our search only to simulations with $\eta \leq 0.01$.

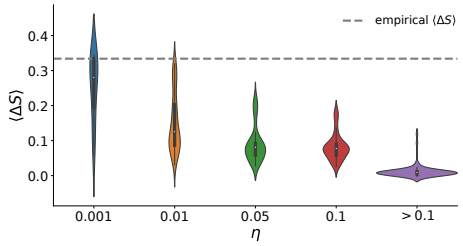


Figure 5.9: **Average Shannon entropy difference across simulations of the ExploNet model.** Violin plot showing the distributions of the Shannon Entropy weighted difference $\langle \Delta S \rangle$ with the randomized sequences, grouping the various simulations together according to their value of η. Notice that only the distributions relative to $\eta = 0.001$ and $\eta = 0.01$ overlap the value of $\langle \Delta S \rangle = 0.334$ of the empirical sequences stopped at 10%, i.e., the same length of the simulations.

	ρ	ν	η	KL	$\langle \Delta S \rangle$
1.	2.7	9	0.01	0.213398	0.239159
2.	3.0	10	0.01	0.216276	0.232123
3.	3.1	10	0.01	0.216367	0.283424
4.	2.0	7	0.01	0.218269	0.207541
5.	2.1	7	0.01	0.220671	0.243933
6.	2.9	10	0.01	0.222252	0.239109

Table 5.1: **Best parameters according to the comparison between model and data set.** The parameters ρ, ν and η related to the six best sets of parameters are shown together with their respective Kullback-Leibler divergence KL from the empirical data and normalized Shannon entropy weighted difference $\langle \Delta S \rangle$.

Then, for each simulated set of parameters $\mathbf{x}$, we calculate how much the Heaps' exponent distribution differs from the real one. To do this, we compute the Kullback-Leibler divergence $KL(\mathbf{x})$ between the empirical $P(\beta)$ (refer to Sec. 5.2.3) and the synthetic one. Finally, we select the reference set of parameters $\mathbf{x}$, among those with $\eta \leq 0.01$, such that $KL(\mathbf{x})$ is minimum.

The first six sets of parameters given by this process are shown in Table 5.1, together with the values of the normalized Shannon entropy weighted difference $\langle \Delta S(\mathbf{x}) \rangle$ and the Kullback-Leibler divergence $KL(\mathbf{x})$. We do not find particular differences between multiple simulations with the same quotient ρ/ν and same η. Notice indeed that all reference simulations have a quotient value between 3 and 3.5.

5.4.2 Analysis of the model

In this section we analyze simulations of the ABM developed in Sec. 5.3 with different values of the parameters, testing to which extent our model can reproduce the empirical observables investigated in Sec. 5.2. For computational reasons, we analyze

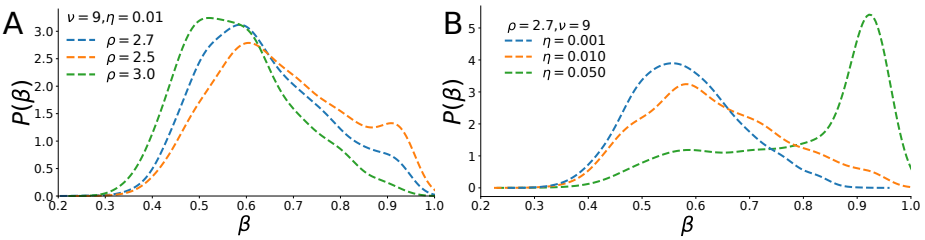


Figure 5.10: **Heaps' exponent distribution in ABM simulations with different sets of parameters. (A)** Heaps' exponent distribution $P(\beta)$ in the reference simulation ($\rho = 2.7$, $\nu = 9$, $\eta = 0.01$) (blue dashed line), compared to two cases with a lower ($\rho = 2.5$, orange line) and higher ($\rho = 3.0$, green line) value of ρ. **(B)** Heaps' exponent distribution $P(\beta)$ in the reference simulation (orange dashed line), compared to two cases with a lower ($\eta = 0.001$, blue line) and higher ($\eta = 0.05$, green line) value of η.

simulations with about 33.5 million time steps instead of the total 335 million present in the data set (see Sec. 5.4.5 for a comparison with the latter case). This way, we can inspect the influence of each parameter on the model outcomes, in relation to the empirical data. As discussed in Sec. 5.4.1, we select a reference set of parameters that feature the presence of semantic correlations similar to the empirical data and that minimize the Kullback-Leibler divergence between the empirical and synthetic Heaps' exponent distributions $P(\beta)$ in the population. Notice that we are not putting any constraints on the other empirical observables, such as assortativity or taste overlaps. The reference combination of parameters is given by $\rho = 2.7$, $\nu = 9$, and $\eta = 0.01$.

First, we check in Fig. 5.10(A,B) how the distribution $P(\beta)$ of Heaps' exponents changes, varying respectively only the reinforcement parameter ρ and the jump parameter η. We find that the average Heaps' exponent decreases when increasing the value of ρ or decreasing the value of η. In fact, the higher the reinforcement strength ρ is, the more likely it is to exploit already discovered items. Similarly, with lower values of η, the agent has higher chances to move only within semantically close tokens, thus usually ignoring more distant possibilities, even with large values of the triggering parameter ν. These plots also show that the ExploNet model can reproduce heterogeneous Heaps' exponents in the population, even if all agents share the same set of parameters and evolutionary rules. This evidence represents our first finding: collective exploration allows agents in our ABM to have heterogeneous propensities to discover new content.

Next, in Fig. 5.11, we show that with low values of η, the synthetic individual sequences of explored artists feature strong semantic correlations, as in the empirical case. In particular, the lower the parameter η, the lower the Shannon entropy $S(f)$ is, which implies that the various occurrences of the same artist are more closer to each other than randomly dispersed across the sequence. Hence, the evolution rule that promotes the extraction of items semantically related to the last extracted one correctly

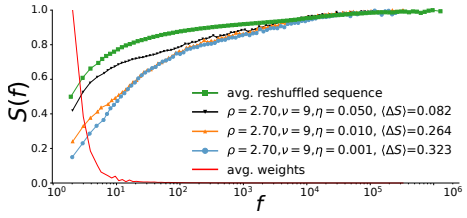


Figure 5.11: **Presence of semantic correlations in the sequences of ABM simulations.** Shannon entropy $S(f)$ of the occurrences of the artists in the user sequences in the same simulations showed in Fig. 5.10(B) as a function of the frequency of the artist f, compared to the Shannon entropy calculated on the reshuffled sequences of the reference case (green line). We also show the weights for each frequency f (red line) used to compute the difference $\langle \Delta S \rangle$.

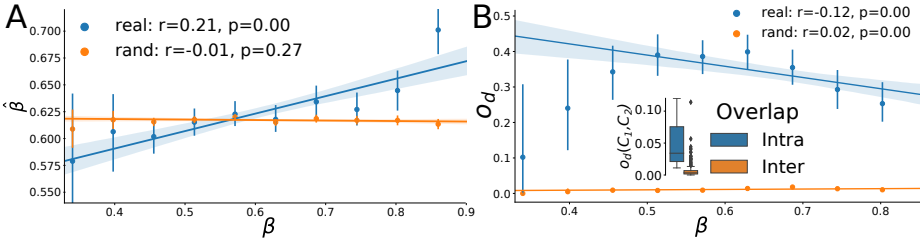


Figure 5.12: **Impact of the social network in the ABM. (A)** Scatter plot between the Heaps' exponents β and the average neighbors' Heaps' exponent $\hat{\beta}$, as found in the reference simulation (blue line) and when randomly rewiring the social network links using a configuration model (orange line). We observed the presence of assortativity between users through their exploration rates. **(B)** Scatter plot between the average dynamical overlap o_d of the neighbors' and the node's Heaps' exponent β, as found in the reference simulation (blue line) and when randomly rewiring the social network links using a configuration model (orange line). In the inset, we show the average dynamical overlap distribution between users in the same community (blue box) and between different communities (orange box).

generates highly correlated sequences of tokens, featuring low values of the Shannon entropy $S(f)$.

Moreover, in Fig. 5.12(A), we see that the model reproduces the assortative arrangement of explorers and exploiters in the network (Pearson correlation coefficient $r \approx 0.21$ and no correlation in the rewired case for the reference simulation). Taking into consideration the innovation rate β is heterogeneously distributed in the synthetic population and all agents follow the same evolutionary rules, this evidence is our second significant result: the social network's topology influences how individual exploration propensities are distributed. In other words, users feature an increased (decreased) exploration propensity when surrounded by agents with higher (lower) discovery rates.

Furthermore, in Fig. 5.12(B) we show that the ABM correctly reproduces the higher

dynamical overlap between the agent and his friends with respect to other random individuals. We also find a small negative correlation between the average dynamical overlap o_d of an agent with his neighbors and his discovery rate β (Pearson correlation coefficient $r \approx -0.12$, no correlation in the rewired case). Moreover, as shown in the inset of Fig. 5.12(B), the intra-community overlaps ($\langle o_d(C)\rangle_C \approx 0.057$) are also larger than the inter-communities ones ($\langle o_d(C_1, C_2)\rangle_{C_1 \neq C_2} \approx 0.004$). This last evidence is a third tangible effect of a shift from an individual to a collective exploration mechanism. The interaction dynamics between agents can increase their possibility to expand their space of possibilities by receiving suggestions from their neighbors, thus increasing the similarity of listened tokens between friends. Therefore, the different propensity for an agent to be an explorer (exploiter) depends not only on the neighbors featuring a similar pace of discovery, but also on the opportunity to be exposed to contents and tokens diverse (similar) from those already experienced. The model indeed pushes agents that explore different regions of the content space to easily accept suggestions outside their comfort zone, giving them a higher acceptance probability $P(x, \mathcal{S}_i, c_i, \varepsilon)$. In doing so, users' space of possibilities enlarges, and their pace of discovery increases, while the opposite happens to nodes that explore a limited portion of the contents space, leading to echo chambers. In Sec. 5.4.3, we further show that, when we switch off the interaction step of the model, the assortativity and the high overlaps within communities drop significantly.

5.4.3 Simulations with no interaction

Interaction is one of the important ingredients of the model we have proposed. The social neighborhood of a user indeed influences their exploration propensities, shaping their space of possibilities. In order to assess how much the topology has an influence in the phenomena we have observed, we have run some simulations with no interaction, and compared the results with the simulations with interaction with the same set of parameters. We find that the distribution of the Heaps' exponents is similar, with more heterogeneous results with lower values of η, as expected from the analytic results in [8], due to the presence of stochastic variations in the individual exploration process. However, as shown in Fig. 5.13(A), simulations with interaction have a higher Spearman's rank correlation r between the Heaps' exponents and the average one of their friends with respect to the simulations without interaction, also with a more significant p-value. This confirms that the exploration rates are affected by the social contacts in an assortative way when interaction is active: more explorative users tend to interact with peers more prone to explore new content.

Finally, unlike the simulations with interactions (see inset of Fig. 5.12(B)), in simulations without interactions the communities found with the Louvain algorithm on the social network have no dynamical influence on the process. Considering in fact the

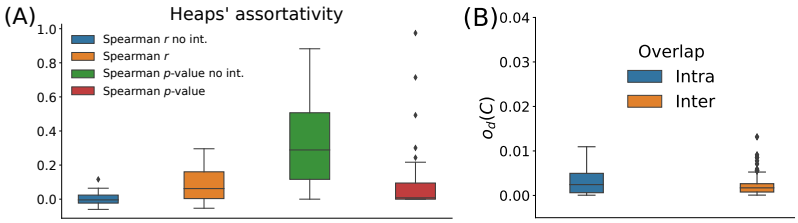


Figure 5.13: **Comparison of the key features in simulations without interaction. (A)** Comparison between Spearman's rank correlation r and related p-value between the Heaps' exponents β and the respective average $\hat{\beta}$ of the neighbor Heaps' exponents, without (blue and green) and with interaction (orange and red). Here the same sets of parameters have been used for both with and without interaction simulations. **(B)** Comparison between the average dynamical overlap distribution between users in the same community (blue) and between different communities (orange), calculated on the simulation with the reference set of parameters (cfr. inset of Fig. 5.12(B)), here without interaction.

simulation without interactions with the same reference set of parameters, the internal overlap of listening records with users of the same community is practically indistinguishable to the inter-community overlap, that is the overlap between users of different communities (see Fig. 5.13(B)). Moreover, they are much lower than the case with interaction.

5.4.4 Dynamical network simulations

So far we have considered a static social network $\mathcal{G}_S$ taken from the Last.fm data set, in which all weights are fixed to 1 throughout the simulation. In this section we study the case of dynamically evolving social networks, starting from an all-to-all network $\mathcal{G}_S$ with initial unitary weight $w_0 = 1$ to the edge between each pair of users. During the interaction step, the currently active agent i selects one of his neighbors j with probability $p(j) = w^{ij} / \sum_{k \sim i} w_{ik}$, and then sees what j has last listened to. In the dynamical version of the ABM, each time i interacts with j we let the respective edge weight evolve according to the law $w^{ij}(t+1) = w^{ij}(t) \pm \Delta$, with $\Delta = 0.1$ and with a sign + (-) when the interaction between i and j at time t is positive (negative), i.e., if the node suggested by j is added (or not) to i's space of possibilities $\mathcal{G}_C^i$. During the simulation, we clip each edge weight such that $0.1 \leq w^{ij}(t) \leq 10$ at every time step t. To better characterize the agents in the simulation, we let the first 1% of each individual's events to be equal to their original sequence as a warm-up.

We show in Fig. 5.14 the out-degree distribution $P(k)$ and the community size distribution $P(s)$ obtained letting the social network evolve from an initial all-to-all configuration for about 33.5 millions evolution steps, namely 10% of the total listening records in the data set. These distributions feature a scale-free behavior as in the em-

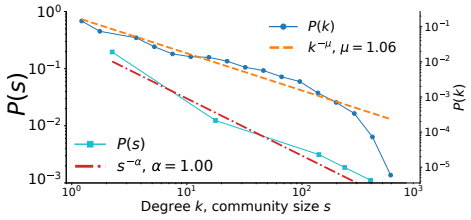


Figure 5.14: **Social network characteristics in simulations with dynamics on the social edges.** The out-degree distribution $P(k) \sim k^{-\mu}$ (blue circles) and the community-size distribution $P(s) \sim s^{-\alpha}$ (cyan squares) as found when letting the social network co-evolve during the model simulation, using the reference set of parameters. We also show the power-law fitting of the two, giving $\mu \approx 1.06$ (orange line) and $\alpha \approx 1.00$ (red line).

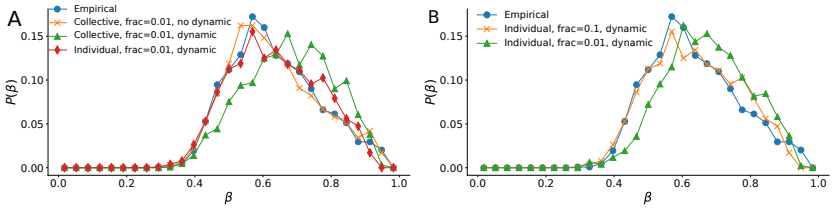


Figure 5.15: **Analysis of the pace of discovery of the model changing warm-up. (A)** Comparison of the Heaps' exponent distribution $P(\beta)$ in the empirical case (blue) and in the simulations with the reference set of parameters, using a collective warm-up for 1% of the events with (orange) and without (green) social dynamics on the edges and with an individual warm-up for 1% of the events with edge dynamics (red). **(B)** The same but limited to an individual warm-up of 10% of the events (orange) and 1% (green).

pirical case, and confirm that the ABM also accounts for the emergence of a complex, real-world-like topology based on the interactions between users alone.

Our results are robust with respect to the change of warm-up strategy and duration, as well as to the presence, or not, of a dynamical social network. We use either individual or collective warm-up for a certain fraction of events, meaning that the first fraction of, respectively, the agent's or the population events in the simulations corresponds exactly to the one in the empirical listening sequences. In Fig. 5.15(A) we show that the distribution of Heaps' exponents $P(\beta)$ found in the empirical case is still well reproduced if we change the kind of warm-up. Additionally, in Fig. 5.15(B) we highlight that the model reproduces the original distribution also varying the warm-up duration, keeping the other parameters fixed to the reference set of parameters.

Moreover, the results are robust with respect to the change of warm-up also when inspecting the degree distribution $P(k)$ and community size distribution $P(s)$, as can be seen in Fig. 5.16. Notably, in the simulations where we switch on the dynamics on the social edges, the model is able to reproduce both the degree and community size

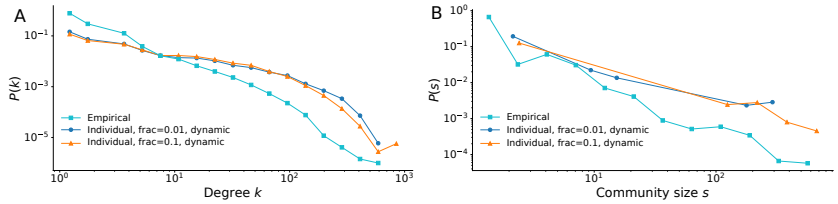


Figure 5.16: **Social network analysis of the model changing warm-up. (A)** Comparison of the out-degree distribution $P(k)$ in the empirical social network (cyan) with the one resulting from ABM simulations with dynamics on the social edges using an individual warm-up for 1% of the events (blue) and for the 10% of them (orange), using the reference set of parameters $\rho = 2.7$, $\nu = 9$, $\eta = 0.01$. **(B)** The same but focusing on the community-size distribution $P(s)$.

distribution, as well as the $P(\beta)$ distribution, highlighting the goodness of the chosen modeling framework to reproduce the empirical findings. Note that, in the simulations with the edge dynamics turned on, for the analysis of the social network in terms of degree and communities, we consider an unweighted directed social network in which we draw an edge between nodes i and j if the edge's weight $w^{ij} > \bar{w}$, where $\bar{w}$ is the largest weight cut-off at which we have a single weakly connected component in the resulting social network.

Remarkably, the model with the dynamics on the social edges reproduces the empirical dynamical overlap o_d and the assortativity between the Heaps' exponent β of an agent and the average exponent $\hat{\beta}$ of his neighbors. Furthermore, the assortativity between the focus node's β and the dynamical overlap o_d with his neighbors is robust with respect to the change of the warm-up kind and duration as well as to the presence of a dynamics on the social network's edge weights. As we show in Fig. 5.17(A-B) and Table 5.2, the individual warm-up with edge dynamics turned on results in a higher positive assortativity of β and $\hat{\beta}$ as well as in a stronger negative correlation between β and o_d, in line with the empirical findings. This is because the individual warm-up allows to better characterize the tastes of a single node, by imposing a larger fraction of its first exploration events. Moreover, we see in Fig. 5.17(C) that the edge dynamics, as expected, drives the system to communities with a higher internal overlap with respect to the non dynamical case. Notably, if we let the nodes to be characterized enough (individual warm-up with the first 10% of the node's events fixed to the original sequence, see Fig. 5.17(A)), the intra- and inter-communities overlap value approaches the empirical one (intra-community $\langle o_d(C)\rangle_C \approx 0.07$ inter-community $\langle o_d(C1, C2)\rangle_{C_1 \neq C2} \approx 0.025$).

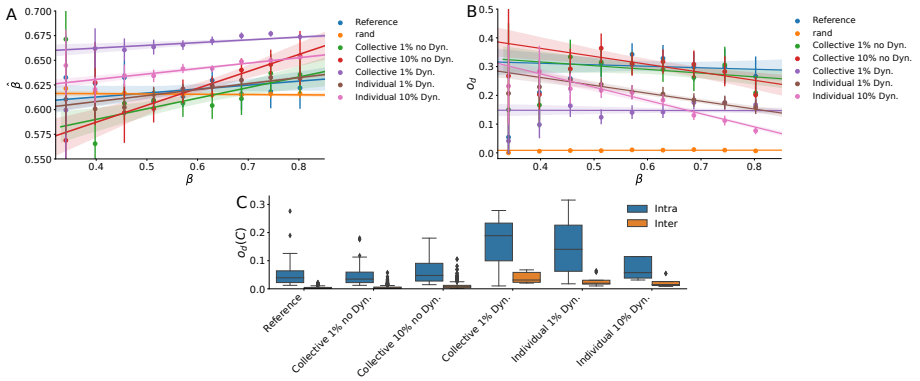


Figure 5.17: **Analysis of the influence of the social network on the pace of discovery changing warm-up in the model. (A)** The assortativity analysis between the Heaps' exponent β of a node and the average one $\hat{\beta}$ of its neighbors. Different warm-up and edge dynamics strategies are reported in the legend. The *Reference* case corresponds to the reference parameters set $\rho = 2.7$, $\nu = 9$, $\eta = 0.01$. The slope of the correlations found are reported in Table 5.2. **(B)** The same analysis but focusing on the assortativity of β with the dynamical overlap with neighbors o_d. **(C)** Comparison of the distribution $P(o_d(C_i, C_j))$ of the average dynamical overlap found within a community (blue box, intra) and between (orange box, inter) two distinct communities for different simulations settings (see x axis labels for details).

Warmup	Kind	Fraction	Edge dynamics	$r(\hat{\beta})$	$r(o_d)$
No, reference case	n.a	0	No	0.05	-0.018
No, reshuffled case	n.a	0	No	-0.05	0.003
Yes	Collective	1%	No	0.12	-0.048
Yes	Collective	10%	No	0.18	-0.095
Yes	Collective	1%	Yes	0.12	-0.02
Yes	Individual	1%	Yes	0.22	-0.19
Yes	Individual	10%	Yes	0.24	-0.36

Table 5.2: The correlation coefficients $r(\hat{\beta})$ of the assortativity between β and $\hat{\beta}$ and $r(o_d)$ between β and o_d for different warm-up strategies (Kind column), fraction of the warm-up, and with or without social edge dynamics.

5.4.5 Comparison with longer simulations

In the previous sections we have analyzed the simulations run on 10% of the total number of records in the data set. This choice has been done to better explore the space of parameters through a refined grid and find the reference set of parameters, as shown in Sec. 5.4.1. In this section we analyze the simulation run for a number of steps equal to the empirical number of streams. We hence consider the reference set of parameters ($\rho = 2.7,\ \nu = 9,\ \eta = 0.01$) found in the sequence limited to the first 10%, with the fixed parameters $c = 10$ and $\varepsilon = 0.1$, and we compare the long and

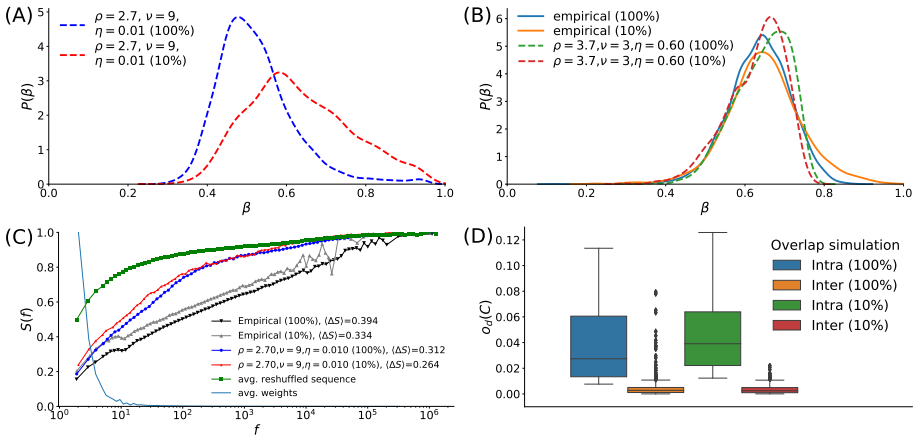


Figure 5.18: **Comparison of the key features in the whole sequences and in the first 10%. (A)** Heaps' exponent distribution of the simulation with reference set of parameters, with approximations of the exponents done after 10% (orange) and 100% (blue) of the sequences. **(B)** Heaps' exponent distribution of the simulation with optimal set of parameters constrained to $0.1 < \eta < 1$ at 10% (dashed green) and 100% (dashed red), compared to the empirical data at 10% (blue line) and 100% (orange line). **(C)** Shannon entropy distribution as a function of the frequency of the artists in the individual sequences of the simulation with reference parameters at 10% (red) and 100% (blue), compared to the empirical data at 10% (gray) and 100% (black line) and to the reshuffled sequences (green). **(D)** Distribution of dynamical overlaps between individuals in the same community (blue for 100%, green for 10%) or in different ones (orange for 100%, red for 10%) for the simulation with reference parameters at 100% and at 10%.

limited simulations. In Fig. 5.18(A) we show that the Heaps' exponent distributions of the long simulation is lower than the limited run. This is probably due to the very low value of η chosen in this simulation and the way we extract the Heaps' exponent. In fact, it seems that, in the simulations with very low values of η, the power-law exponent of the Heaps' laws for each individual decreases after a first part of high exploration. Moreover, the chosen approximation delays the capture of this decrease. This aspect is left for future improvements. As a matter of fact, the Heaps' distribution is more or less stable when comparing the whole sequence and only its 10% when considering either the empirical data set or simulations with $\eta \approx 1$, as shown in Fig. 5.18(B). In any case, even if quantitatively different, the heterogeneity of the individuals exploration rate shown in these simulations is still present also in longer runs.

In Fig. 5.18(C) we complete the analysis by comparing the semantic correlations of the simulation with the selected set of parameters and the empirical data set. We find a small decrease in the value of ΔS from the whole to the 10% of the sequences, both in the empirical data and in the simulations. Finally, there are not any significant differences in the effect of social interactions on the long or short run, as shown in Fig. 5.18(D) by checking the dynamical overlaps inside and outside the communities.

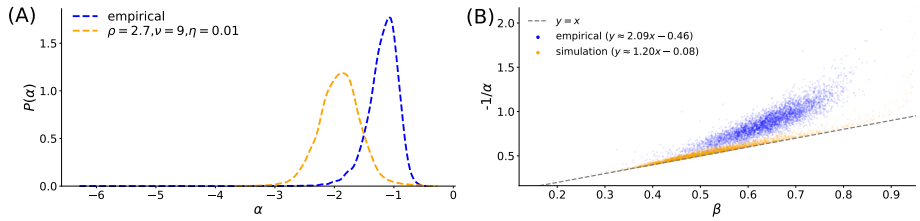


Figure 5.19: **Comparison between Heaps' and Zipf's exponents. (A)** Comparison of the Zipf's exponents distribution on the sequences in the data set (blue) and of the reference simulation (orange). **(B)** Scatter plot between the Heaps' exponents and the anti-reciprocal of the Zipf's exponents for the sequences in the data set (blue) and of the reference simulation (orange), with the respective linear regression displayed in legend.

5.4.6 Heaps' and Zipf's exponent correspondence

In this chapter we have discussed at plenty about the Heaps' law and its significance. However, also the Zipf's law has been of great importance for the analysis of exploration and innovation processes [8, 26, 57]. Decreasingly ordering the occurrence frequencies f of an element in the sequence of events, the Zipf's law can be expressed as $f(R) \sim R^{-\alpha}$, where R indicates the rank and α is called the *Zipf's exponent*. These two laws have been observed in various empirical systems, producing different values of Heaps' exponent β and Zipf's exponent α [8, 28, 165, 238]. Under mild assumptions they are asymptotically equivalent [58, 81, 140], being one the anti-reciprocal of the other. For these reasons, all of the models that have been recently proposed keep both of these laws into considerations [25].

We find that the Zipf's law is also present in the ABM we have proposed. In particular, in Fig. 5.19(A) we show the distributions of the Zipf's exponent for the sequences in the empirical data set (blue) and the reference simulation found in our analysis (orange). Moreover, in Fig. 5.19(B) we compare the Heaps' exponents β and the anti-reciprocal $-1/\alpha$ of the Zipf's exponents. As analytically found for the UMST [8], in the simulation we observe an almost linear correlation between the two exponents (Pearson $r = 0.94$, $p < 0.001$), although in the empirical case it is less strong (Pearson $r = 0.90$, $p < 0.001$), with a linear regression $-1/\alpha \approx 1.20\beta - 0.08$ in the simulation against $-1/\alpha \approx 2.09\beta - 0.46$ in the data.

5.5 Summary and conclusions

In this chapter we have presented an empirical study of online music consumption, and we have proposed a data-driven model of collective exploration, the *"ExploNet"* model, highlighting the central role of the social environment in shaping how we explore music and discover new content. Based on the concept that the space of possibilities is

expanded through the adjacent possible in both the content and social space, our model can indeed reproduce the considerable heterogeneity of individuals' exploration rates empirically observed and the relation between users' rate of discovery and their position in the social network. As it turns out, the opportunity to be connected to other individuals with a high propensity to explore new content makes an individual more likely to have a higher pace of discovery, on average. Moreover, since all the agents in the ExploNet follow the same evolutionary rules, the observed heterogeneity of the individual exploration rates can be explained by stochastic fluctuations (opportunities) and social influence (environment). It also gives insights into the emergence of communities composed of agents with similar tastes, as shown by the simulations in which the social network is shaped dynamically through random interactions.

Furthermore, our ExploNet model reproduces the semantic correlations found in the empirical sequences of music consumption by adequately modulating the individual propensity to select artists similar to their recent history or to move randomly across the whole space of possibilities. The possibility to move between similar artists is naturally accounted for in the underlying network of artists that can be explored, containing all possible artists observed in the empirical data. Such network has been obtained by analyzing the appearance of each pair of artists in the empirical sequences of listening records. As we have indeed shown in Chapter 3, the structure of the content space, and therefore the specific topology of the artist network, co-evolves along with the exploration process. We have hence used such structure as an underlying space that can be individually discovered and accessed by each agent in the model. In other words, we have embedded the expansion of the adjacent possible on an underlying universal network, containing the actual pathways that let the explorer move from one artist to another. Interestingly, even though the universal network of artists is fixed, the individual space of possibilities grows dynamically along with the exploration process of each individual, which defines the individual musical taste by reinforcing the explored parts of such space.

Finally, in our analysis of the ExploNet, we have assumed that all social links of a user in the social network of *Last.fm* have the same probability of being activated. However, in real life, one social link can be different from another. Nevertheless, in the absence of knowledge about the actual weights in the empirically studied network, our choice represents, in our view, a good educated guess. In this direction, in Sec. 5.4.4 we have developed and analyzed a dynamical version of the ExploNet, where the weight of the social links changes along the process. The simulation starts from an initial all-to-all social network, where every user can interact with any other one. Then, during the exploration process, the agents randomly interact with their friends, with probability proportional to their social link. Such weight is thus updated based on the outcome of the interaction. This interaction dynamics naturally creates a social network between the agents with topological properties similar to the empirical one. In particular, our

dynamical social network has a similar degree distribution and community structure, and features the same assortative mixing of users with similar musical taste and pace of discovery. Therefore, in this model social interactions drive the social structure development in an assortative way. However, we cannot conclude what is the precise causal mechanism at play. In fact, in our model users become friends with those users similar to themselves, and at the same time friends tend to induce each other to become more similar. One possible explanation is hence peer influence, where friends suggest music to their friends and therefore they actively influence their dynamics. Nevertheless, we cannot exclude that the same result could be obtained through simple homophily [178, 233, 239], i.e., considering that friends have a similar background and characteristics, making similar choices even if there is no direct influence. Furthermore, the same empirical findings could be partially explained by other confounding exogenous factors, such as the presence of a recommender system or other events. A more thorough analysis of the causal effects in this system, distinguishing between these various mechanisms, is left to future works with richer and more comprehensive empirical data, including precise individual behavioral data, users' demographics, geographic location, and other attributes [239].

Chapter 6

Conclusions and further work

6.1 Summary of contributions

In this thesis we have integrated empirical data analysis, various modelling techniques, and an interdisciplinary perspective to get a comprehensive understanding of innovation. In particular, we have emphasized the complex mechanisms that drive the growth of the space of possibilities, as well as the importance of social interactions, collaboration, and collective exploration in generating novel ideas and advancing the boundaries of what is possible.

We have started our exploratory journey into the dynamics of innovation in Chapter 2, where we have reviewed various mathematical models of innovation, mainly based on extractions from urns or random walks on networks. Such models manage to reproduce the Heaps' law, revealing two key mechanisms that influence the pace of discovery in these systems. On the one hand, the reinforcement of the elements explored makes these elements more likely to be explored again in the future. On the other hand, the appearance of novelties expands the space of potential discoveries that can be made in the future. Combining these two mechanisms, the discovery process can be seen as the exploration of a space of possibilities, made of the elements previously explored in the system, which are being reinforced during the process, and of all those elements that are one step away from the explored part of the space, or, in one expression, the adjacent possibilities. By properly balancing these two mechanisms, these models manage to reproduce various paces of discovery, from those in which the exploitation of past discoveries is prevalent to those in which new items are continuously explored.

Our first original contribution, in Chapter 3, has been the introduction of a more general definition of novelty. As a matter of fact, novelties can also arise from the combination or association of multiple elements for the first time. Therefore, we have

defined the *"n-th order Heaps' law"*, which quantifies the rate of discovery of novelties of order n in a sequence, i.e., combinations of n elements that appear for the first time consecutively. Thanks to the analysis of empirical sequences from different contexts, we have found that the higher-order Heaps' laws can distinguish sequences which show the same pace of discovery of single elements, as measured by the standard (1^{st}-order) Heaps law. This finding extends our understanding of how novelties emerge and how the space of possibilities grows, and emphasizes the importance of considering higher-order novelties in exploration processes. In fact, we have found that the existing models of innovation cannot effectively reproduce higher-order Heaps' laws, revealing that there are more complex mechanisms underlying the process of innovation.

We have hence proposed a novel modelling approach, based on the exploration of a complex network that co-evolves along with the exploration process. We can indeed see a 1^{st}-order or a 2^{nd}-order novelty as the first exploration of a node or a link, respectively, of a network, which represents the space of possibilities of the system. Therefore, the 1^{st}-order and 2^{nd}-order Heaps' laws can be exploited to characterize the pace of discovery of new nodes and links in this growing network. We then make use of this broader definition of novelty to adapt the two mechanisms of reinforcement and of expansion of adjacent possibilities present in the previous models. In particular, we propose a new model, called the *"Edge-Reinforced Random Walk with Triggering"* (ERRWT), in which we represent the explorer as a random walk over a growing network. On the one hand, at each time step the ERRWT moves from one node to a neighboring one, reinforcing the traversed edge and strengthening the association between the two nodes. On the other hand, whenever a node or a link is explored for the first time, new nodes or links between existing nodes are added to the network. Balancing these two mechanisms, the ERRWT is able to simulate the emergence of novelties at various orders, capturing the dynamic nature of innovation and the expansion of the adjacent possible in the content space. Moreover, thanks to this new approach, we have highlighted the importance of the networked structure of the space of possibilities, which grows while it is being explored.

Subsequently, in Chapter 4 we have unveiled the significant role played by social interactions and human connections in shaping an individual's propensity to generate novel ideas. Here, we have expanded the concept of the adjacent possible to incorporate a social dimension, which was missing in the previous modelling frameworks. We have proposed to model a group of individuals as an ensemble of interacting urns, coupled through the links of a social network. In particular, each urn, which is subject to the usual mechanisms of reinforcement and triggering of adjacent possibilities, is expanded with the set of opportunities coming from their social contacts. Depending on the structural properties of the network, the collaborative dynamics in our model, named the *"UrNet"* model, not only increases the average pace of discovery, but also creates

different behaviors in each agent. Simulating the model on different social structures, from small synthetic graphs to bigger real-world networks, we have indeed found that the structural—not just local—properties of the nodes can strongly affect their ability to make novelties. By integrating social interactions into the context of discovery models, we have demonstrated the crucial role of collaboration and social network properties in driving innovation and advancing the boundaries of what is possible. Moreover, we have revealed and characterized the impact of the expansion of the adjacent possible in the social space on the way a group of individuals can foster progress.

Finally, we have analyzed empirical sequences of music exploration in Chapter 5 to investigate the impact of peers on the individual pace of discovery of new music. Specifically, we have analysed a unique data set that contains information on both the whole listening histories and the social connections of a large and connected sample of users from the online music platform Last.fm. Our findings have shown that users with a high pace of discovery—measured through the Heaps' law—are more likely to be connected with other peers with a strong inclination for exploration. These results indicate how the pace of discovery of our friends, along with their musical taste, influences our own individual propensity to discover new music. We have thus leveraged the modelling framework created in the previous chapters to develop a data-driven agent-based model capable of reproducing the main empirical results and explain the effects of social interactions on individual music discovery. Our model, named the *"ExploNet"* model, combines the exploration of a growing space of possibilities—inspired by the ERRWT—with the expansion in the social space through social interactions—inspired by the UrNet. In this modelling scheme, each agent explores a universal network of artists, growing their personal space of possibilities through the mechanisms of reinforcement and triggering of the adjacent possible, while also interacting with other agents through means of a social network. By incorporating the expansion of the adjacent possible in both the content and social space, the ExploNet model captures the dynamics observed in the empirical data, explaining the observed heterogeneity in the discovery rates both in terms of stochastic fluctuations and social influence.

6.2 Further work

We hope that the findings of this thesis can contribute to shed light on the underlying rules controlling the emergence of novelties in innovation processes. Unveiling the hidden mechanisms behind the emergence of new ideas, understanding how novelties can trigger further ones, and explaining how they can effectively diffuse in a population is, indeed, not only interesting from a scientific point of view, but can also have a tangible societal and economic impact. In this thesis, we have hence proposed a new modelling framework capable to explain and reproduce the dynamics observed in various empir-

ical discovery processes, based on the exploration of a space of possibilities. Previous works have identified two important mechanisms driving these processes, namely the reinforcement of the elements explored in such space, and the expansion of the space through the triggering of new adjacent possibilities. Building on this, in our work, we have highlighted the role played by network structures. We have indeed found that the space of possibilities explored can be represented as a growing network, where new nodes and links are added while being explored. Moreover, we have highlighted the impact of social interactions, through the links of a social network, in influencing the emergence of novelties, adding a social dimension to the concept of adjacent possible. The findings presented in this thesis hence constitute a starting point to answer the following two questions:

1. How do groups of humans explore the seemingly infinite space of possibilities, leading to innovation and diffusion of new ideas, technologies, or cultural artworks?

2. How do group interactions influence our choices in these exploration processes?

Building on the findings of this thesis, a first natural step is to use a similar networked approach to characterize the evolution of innovation processes in other contexts, from science [3] and technology [240] to economics [241] and biology [242]. In order to have a richer representation of such processes, it could be worth investigating even more complex structures, accounting for combinations of more than two elements. To have a practical example, a patent can be considered as the combination of multiple technologies. Similarly, a scientific paper or a protein can be seen as the combination of various keywords or amino acids, respectively. Therefore, we can imagine the space of possibilities as a growing weighted hypergraph [199], where a new combination adds a new hyperedge, i.e., an edge made of more than two nodes. In this higher-order framework, a discovery would be represented by the first exploration of either a node or a group of already explored nodes (recombination). This higher-order structure would be especially important in all those cases in which the innovation process cannot be reduced to a sequence of single elements. Interestingly, in the analysis of this higher-order process one can integrate all the topological data analysis (TDA) tools [243, 244] if we consider simplicial complexes instead of hypergraphs, which can be obtained considering all sub-hyperedges of each hyperedge [245].

A similar higher-order direction can be used to answer the second question, i.e., how group interactions influence the innovation process. As a matter of fact, even if networks provide a powerful abstraction for complex systems representing the underlying set of pairwise interactions, much of the structure within social systems involves interactions that take place among more than two nodes at once. For example, a scientific paper is often the result of the collaborative effort of more than one or two authors. In this case, their resulting work cannot be solely attributed to any single author, nor

to the mere sum of their individual knowledge. Therefore, an interesting step forward from the findings of this thesis could be to study the effect of higher-order social interactions on innovation and exploration processes. Here, the social interactions could be modelled through a social hypernetwork, i.e., networks with hyperlinks constituting interactions between two or more individuals. Thanks to the UrNet and ExploNet models developed in this thesis, we believe that the adaptation to a higher-order social structure could shed light on the formation of optimal team structures and efficient collaboration networks. This line of research has the potential to contribute to the development of more comprehensive models of innovation dynamics, offering practical implications for team management and promoting innovation within complex organizations.

Bibliography

[1] I. Iacopini, G. Di Bona, et al. “Interacting discovery processes on complex networks”. In: *Physical Review Letters* 125.24 (2020), p. 248301.

[2] G. Di Bona, E. Ubaldi, et al. “Social interactions affect discovery processes”. In: *arXiv:2202.05099* (2022).

[3] G. Di Bona, A. Bracci, et al. “The decentralized evolution of decentralization across fields: from Governance to Blockchain”. In: *arXiv:2207.14260* (2022).

[4] G. Di Bona, L. Di Gaetano, V. Latora, and F. Coghi. “Maximal dispersion of adaptive random walks”. In: *Physical Review Research* 4 (4 Dec. 2022), p. L042051.

[5] C. D. B. Luft, I. Zioga, et al. “Social synchronization of brain activity increases during eye-contact”. In: *Communications biology* 5.1 (2022), pp. 1–15.

[6] G. Di Bona, A. Bellina, et al. “The dynamics of higher-order novelties”. In: *arXiv preprint arXiv:2307.06147* (2023).

[7] G. Di Bona and A. Giacobbe. “A Simple Theoretical Model for Lags and Asymmetries of Surface Temperature”. In: *Climate* 9.5 (2021), p. 78.

[8] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz. “The dynamics of correlated novelties”. In: *Sci. Rep.* 4 (2014), p. 5890.

[9] N. H. Packard. “Adaptation toward the edge of chaos”. In: *Dyn. Patterns Complex Syst.* 212 (1988), p. 293.

[10] C. Langton. *Computation at the edge of chaos: Phase transition and emergent computation*. Tech. rep. 1–3. 1990, pp. 12–37.

[11] C. Langton, C. Taylor, J. Farmer, and S. Rasmussen. *Artificial Life II*. Avalon Publishing, 2003. ISBN: 9780201525717.

[12] S. A. Kauffman et al. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.

[13] S. A. Kauffman. “Investigations: The Nature of Autonomous Agents and the Worlds They Mutually Create”. In: *SFI working papers*. Santa Fe Institute. 1996.

[14] G. Herdan. *Type-token Mathematics: A Textbook of Mathematical Linguistics*. Vol. 4. Mouton en company, 1960.

[15] H. S. Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.

[16] C. Darwin. *On the origin of species*. Routledge, 1859.

[17] J. G. March. "Exploration and exploitation in organizational learning". In: *Organization science* 2.1 (1991), pp. 71–87.

[18] C. S. Dweck. *Mindset: The new psychology of success*. Random house, 2006.

[19] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[20] S. Thurner, P. Klimek, and R. Hanel. "Schumpeterian economic dynamics as a quantifiable model of evolution". In: *New J. Phys.* 12.7 (2010), p. 075029.

[21] J. A. Schumpeter et al. *Business cycles*. Vol. 1. McGraw-Hill New York, 1939.

[22] J. A. Schumpeter. *Capitalism, socialism and democracy*. routledge, 2013.

[23] F. Eggenberger and G. Pólya. "Über die Statistik verketteter Vorgänge". In: *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik* 3.4 (1923), pp. 279–289.

[24] G. Pólya. "Sur quelques points de la théorie des probabilités". fr. In: *Annales de l'institut Henri Poincaré* 1.2 (1930), pp. 117–161.

[25] V. Loreto, V. D. Servedio, S. H. Strogatz, and F. Tria. "Dynamics on expanding spaces: modeling the emergence of novelties". In: *Creativity and Universality in Language*. Ed. by M. Degli Esposti, E. G. Altmann, and F. Pachet. Springer, 2016, pp. 59–83. ISBN: 978-3-319-24403-7.

[26] F. Tria, V. Loreto, and V. Servedio. "Zipf's, Heaps' and Taylor's Laws are Determined by the Expansion into the Adjacent Possible". In: *Entropy* 20.10 (Sept. 2018), p. 752. ISSN: 1099-4300.

[27] F. Tria, I. Crimaldi, G. Aletti, and V. D. P. Servedio. "Taylor's Law in Innovation Processes". In: *Entropy* 22.5 (2020). ISSN: 1099-4300. DOI: 10.3390/e22050573.

[28] B. Monechi, Ã. Ruiz-Serrano, F. Tria, and V. Loreto. "Waves of novelties in the expansion into the adjacent possible". In: *PLoS One* 12.6 (2017), e0179303.

[29] I. Iacopini, S. Milojević, and V. Latora. "Network dynamics of innovation processes". In: *Phys. Rev. Lett.* 120 (2018), p. 048301.

[30] E. Ubaldi, R. Burioni, V. Loreto, and F. Tria. "Emergence and evolution of social networks through exploration of the Adjacent Possible space". In: *Commun. Phys.* 4.1 (2021), pp. 1–12.

[31] G. D. Marzo, F. Pandolfelli, and V. D. Servedio. “Modeling innovation in the cryptocurrency ecosystem”. In: *Scientific Reports* 12.1 (2022), p. 12942.

[32] R. Albert and A.-L. Barabási. “Statistical mechanics of complex networks”. In: *Rev. Mod. Phys.* 74.1 (2002), p. 47.

[33] M. E. Newman. “The structure and function of complex networks”. In: *SIAM Rev.* 45.2 (2003), pp. 167–256.

[34] S. Boccaletti, V. Latora, et al. “Complex networks: Structure and dynamics”. In: *Physics reports* 424.4-5 (2006), pp. 175–308.

[35] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[36] V. Latora, V. Nicosia, and G. Russo. *Complex networks: principles, methods and applications*. Cambridge University Press, 2017.

[37] A.-L. Barabási and R. Albert. “Emergence of scaling in random networks”. In: *science* 286.5439 (1999), pp. 509–512.

[38] L. Lovász et al. “Random walks on graphs: A survey”. In: *Combinatorics, Paul erdos is eighty* 2.1 (1993), pp. 1–46.

[39] R. Pemantle et al. “A survey of random processes with reinforcement”. In: *Probab. Surv* 4.0 (2007), pp. 1–79.

[40] N. Masuda, M. A. Porter, and R. Lambiotte. “Random walks and diffusion on networks”. In: *Physics reports* 716 (2017), pp. 1–58.

[41] C. Cattuto, A. Barrat, et al. “Collective dynamics of social annotation”. In: *Proc. Natl. Acad. Sci. U.S.A.* 106.26 (2009), pp. 10511–10515.

[42] B. Monechi, P. Gravino, et al. “Significance and popularity in music production”. In: *Royal Society open science* 4.7 (2017), p. 170433.

[43] H. F. de Arruda, F. N. Silva, L. d. F. Costa, and D. R. Amancio. “Knowledge acquisition: A Complex networks approach”. In: *Information Sciences* 421 (2017), pp. 154–166.

[44] P. Allegrini, P. Grigolini, and L. Palatella. “Intermittency and scale-free networks: a dynamical model for human language complexity”. In: *Chaos, Solitons & Fractals* 20.1 (2004), pp. 95–105.

[45] T. Jia, D. Wang, and B. K. Szymanski. “Quantifying patterns of research-interest evolution”. In: *Nature Human Behaviour* 1.4 (2017), p. 0078.

[46] R. Axelrod and W. D. Hamilton. “The evolution of cooperation”. In: *science* 211.4489 (1981), pp. 1390–1396.

[47] M. Nowak and R. Highfield. *Supercooperators: Altruism, evolution, and why we need each other to succeed*. Simon and Schuster, 2011.

[48] A. Civilini, O. Sadekar, et al. “Explosive cooperation in social dilemmas on higher-order networks”. In: *arXiv preprint arXiv:2303.11475* (2023).

[49] J. C. Mitchell. “Social networks”. In: *Annual review of anthropology* 3.1 (1974), pp. 279–299.

[50] M. E. Newman and J. Park. “Why social networks are different from other types of networks”. In: *Physical review E* 68.3 (2003), p. 036122.

[51] S. P. Borgatti, M. G. Everett, and J. C. Johnson. *Analyzing social networks*. Sage, 2018.

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

[53] P. Bak and K. Sneppen. “Punctuated equilibrium and criticality in a simple model of evolution”. In: *Physical review letters* 71.24 (1993), p. 4083.

[54] J.-B. Estoup. *Gammes sténographiques*. Institut sténographique de France, 1916.

[55] G. K. Zipf. “Relative frequency as a determinant of phonetic change”. In: *Harvard studies in classical philology* 40 (1929), pp. 1–95.

[56] G. K. Zipf. *The Psychobiology of Language*. 1935.

[57] G. K. Zipf. “Human behavior and the principle of least effort. Cambridge, (Mass.): Addison-Wesley, 1949, pp. 573”. In: *Journal of Clinical Psychology* 6.3 (1949), pp. 306–306. DOI: 10.1002/1097-4679(195007)6:3<306::AID-JCLP2270060331>3.0.CO;2-7.

[58] L. Lü, Z.-K. Zhang, and T. Zhou. “Zipf’s law leads to Heaps’ law: Analyzing their relation in finite-size systems”. In: *PloS one* 5.12 (2010), e14139.

[59] G. De Marzo, F. S. Labini, and L. Pietronero. “Zipf’s law for cosmic structures: How large are the greatest structures in the universe?” In: *Astronomy & Astrophysics* 651 (2021), A114.

[60] G. De Marzo, A. Gabrielli, A. Zaccaria, and L. Pietronero. “Dynamical approach to Zipf’s law”. In: *Physical Review Research* 3.1 (2021), p. 013084.

[61] M. V. Simkin and V. P. Roychowdhury. “Re-inventing willis”. In: *Phys. Rep.* 502.1 (2011), pp. 1–35.

[62] Z. Wang, N. R. Clark, and A. Ma’ayan. “Dynamics of the discovery process of protein-protein interactions from low content studies”. In: *BMC Systems Biology* 9.1 (2015), pp. 1–10.

[63] V. Mori, B. J. Smith, B. Suki, and J. H. Bates. “Linking physiological biomarkers of ventilator-induced lung injury to a rich-get-richer mechanism of injury progression”. In: *Annals of biomedical engineering* 47 (2019), pp. 638–645.

[64] Á. Corral and Á. González. “Power law size distributions in geoscience revisited”. In: *Earth and Space Science* 6.5 (2019), pp. 673–697.

[65] A. Puglisi, A. Baronchelli, and V. Loreto. “Cultural route to the emergence of linguistic categories”. In: *Proceedings of the National Academy of Sciences* 105.23 (2008), pp. 7936–7940.

[66] J. Brundidge and R. E. Rice. “Political engagement online: Do the information rich get richer and the like-minded more similar?” In: *Routledge handbook of Internet politics*. Routledge, 2008, pp. 144–156.

[67] A. G. Collins and M. J. Frank. “How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis”. In: *European Journal of Neuroscience* 35.7 (2012), pp. 1024–1035.

[68] M. M. Dankulov, R. Melnik, and B. Tadić. “The dynamics of meaningful social interactions and the emergence of collective knowledge”. In: *Sci. Rep.* 5 (2015), p. 12197.

[69] F. Saracco, R. Di Clemente, A. Gabrielli, and L. Pietronero. “From innovation to diversification: a simple competitive model”. In: *PLoS One* 10.11 (2015), e0140420.

[70] E. James, M. G. Gaskell, A. Weighall, and L. Henderson. “Consolidation of vocabulary during sleep: The rich get richer?” In: *Neuroscience & Biobehavioral Reviews* 77 (2017), pp. 1–13.

[71] C. Cheng, H.-y. Wang, L. Sigerson, and C.-l. Chau. “Do the socially rich get richer? A nuanced perspective on social network site use and online social capital accrual.” In: *Psychological Bulletin* 145.7 (2019), p. 734.

[72] D. J. D. S. Price. “Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front.” In: *Science* 149.3683 (1965), pp. 510–515.

[73] G. U. Yule. “II.—A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S”. In: *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character* 213.402-410 (1925), pp. 21–87.

[74] J. C. Willis, H. De Vries, et al. *Age and area: a study in geographical distribution and origin of species*. CUP Archive, 1922.

[75] J. C. Willis and G. U. Yule. “Some statistics of evolution and geographical distribution in plants and animals, and their significance.” In: *Nature* 109.2728 (1922), pp. 177–179.

[76] H. A. Simon. “On a class of skew distribution functions”. In: *Biometrika* 42.3/4 (1955), pp. 425–440.

[77] F. M. Hoppe. “Pólya-like urns and the Ewens’ sampling formula”. In: *J. Math. Biol.* 20.1 (1984), pp. 91–94.

[78] W. J. Ewens. “The sampling theory of selectively neutral alleles”. In: *Theoretical population biology* 3.1 (1972), pp. 87–112.

[79] R. A. Fisher. *The genetical theory of natural selection.* Oxford Clarendon Press, 1930, p. 302.

[80] S. Wright. “Evolution in Mendelian populations”. In: *Genetics* 16.2 (1931), p. 97.

[81] M. Á. Serrano, A. Flammini, and F. Menczer. “Modeling statistical properties of written text”. In: *PloS one* 4.4 (2009), e5372.

[82] L. R. Taylor. “Aggregation, variance and the mean”. In: *Nature* 189.4766 (1961), pp. 732–735.

[83] Z. Eisler, I. Bartos, and J. Kertész. “Fluctuation scaling in complex systems: Taylor’s law and beyond”. In: *Advances in Physics* 57.1 (2008), pp. 89–142.

[84] J. Gómez-Gardeñes and V. Latora. “Entropy rate of diffusion processes on complex networks”. In: *Physical Review E* 78.6 (2008), p. 065102.

[85] E. Agliari, R. Burioni, and G. Uguzzoni. “The true reinforced random walk with bias”. In: *New Journal of Physics* 14.6 (2012), p. 063027.

[86] D. Boyer and C. Solis-Salas. “Random walks with preferential relocations to places visited in the past and their application to biology”. In: *Physical review letters* 112.24 (2014), p. 240601.

[87] J. Choi, J.-I. Sohn, K.-I. Goh, and I.-M. Kim. “Modeling the mobility with memory”. In: *EPL (Europhysics Letters)* 99.5 (2012), p. 50001.

[88] M. Szell, R. Sinatra, et al. “Understanding mobility in a social petri dish”. In: *Scientific reports* 2 (2012), p. 457.

[89] D. Denteneer, F. d. Hollander, and E. Verbitskiy. “Dynamics & Stochastics: Festschrift in honor of MS Keane”. In: *arXiv preprint math/0608289* (2006).

[90] J. G. Foster, P. Grassberger, and M. Paczuski. “Reinforced walks in two and three dimensions”. In: *New Journal of Physics* 11.2 (2009), p. 023009.

[91] D. Coppersmith and P. Diaconis. “Random walk with reinforcement”. In: *Unpublished manuscript* 19873 (1987).

[92] D. J. Watts and S. H. Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[93] M. E. Newman and D. J. Watts. “Scaling and percolation in the small-world network model”. In: *Physical review E* 60.6 (1999), p. 7332.

[94] S. Milojević. “How are academic age, productivity and collaboration related to citing behavior of researchers?” In: *PloS one* 7.11 (2012), e49176.

[95] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans. “Choosing experiments to accelerate collective discovery”. In: *Proc. Natl. Acad. Sci. U.S.A.* 112.47 (2015), pp. 14569–14574.

[96] T. Fink, M. Reeves, R. Palma, and R. Farr. “Serendipity and strategy in rapid innovation”. In: *Nat. Commun.* 8.1 (2017), p. 2002.

[97] K. Abbas, M. Shang, et al. “Popularity and novelty dynamics in evolving networks”. In: *Scientific reports* 8.1 (2018), pp. 1–10.

[98] G. C. Rodi, V. Loreto, and F. Tria. “Search strategies of Wikipedia readers”. In: *PloS one* 12.2 (2017), e0170746.

[99] S. Sreenivasan. “Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords”. In: *Scientific reports* 3 (2013), p. 2758.

[100] R. Adolphs. “The social brain: neural basis of social knowledge”. In: *Annual review of psychology* 60 (2009), pp. 693–716.

[101] A. Senju and M. H. Johnson. “The eye contact effect: mechanisms and development”. In: *Trends in cognitive sciences* 13.3 (2009), pp. 127–134.

[102] M. North. *Novelty: A history of the new*. University of Chicago Press, 2013.

[103] I. Iacopini and V. Latora. “On the dual nature of adoption processes in complex networks”. In: *Front. Phys.* 9 (2021), p. 109.

[104] G. De Marzo, A. Gabrielli, A. Zaccaria, and L. Pietronero. “Dynamical approach to Zipf’s law”. In: *Physical Review Research* 3.1 (2021), p. 013084.

[105] G. Armano and M. A. Javarone. “The beneficial role of mobility for the emergence of innovation”. In: *Scientific reports* 7.1 (2017), pp. 1–8.

[106] P. Gravino, B. Monechi, et al. “Crossing the horizon: exploring the adjacent possible in a cultural system”. In: *Proceedings of the Seventh International Conference on Computational Creativity*. 2016.

[107] J. McNerney, J. D. Farmer, S. Redner, and J. E. Trancik. “Role of design complexity in technology improvement”. In: *Proc. Natl. Acad. Sci. U.S.A.* 108.22 (2011), pp. 9008–9013.

[108] M. Abbasiharofteh, D. F. Kogler, B. Lengyel, et al. “Atypical combination of technologies in regional co-inventor networks”. In: *Papers in Evolutionary Economic Geography (PEEG)* 20 (2020).

[109] B. Lambert, G. Kontonatsios, et al. “The pace of modern culture”. In: *Nature Human Behaviour* 4.4 (2020), pp. 352–360.

[110] C. Jin, C. Song, et al. “Emergence of scaling in complex substitutive systems”. In: *Nat. Hum. Behav.* 3.8 (2019), pp. 837–846.

[111] A. M. Leroi, B. Lambert, et al. “On revolutions”. In: *Palgrave Communications* 6.1 (2020), pp. 1–11.

[112] E. Market and F. N. Papavasiliou. “V (D) J recombination and the evolution of the adaptive immune system”. In: *PLoS biology* 1.1 (2003), e16.

[113] J. M. Jones and M. Gellert. “The taming of a transposon: V (D) J recombination and the immune system”. In: *Immunological reviews* 200.1 (2004), pp. 233–248.

[114] S. Fortunato, C. T. Bergstrom, et al. “Science of science”. In: *Science* 359.6379 (2018), eaao0185.

[115] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. “Atypical combinations and scientific impact”. In: *Science* 342.6157 (2013), pp. 468–472.

[116] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini. “Defining and identifying sleeping beauties in science”. In: *Proceedings of the National Academy of Sciences* 112.24 (2015), pp. 7426–7431.

[117] J. Wang, R. Veugelers, and P. Stephan. “Bias against novelty in science: A cautionary tale for users of bibliometric indicators”. In: *Research Policy* 46.8 (2017), pp. 1416–1436.

[118] M. Fontana, M. Iori, F. Montobbio, and R. Sinatra. “New and atypical combinations: An assessment of novelty and interdisciplinarity”. In: *Research Policy* 49.7 (2020), p. 104063.

[119] L. Wu, D. Wang, and J. A. Evans. “Large teams develop and small teams disrupt science and technology”. In: *Nature* 566.7744 (2019), pp. 378–382.

[120] U. Alvarez-Rodriguez, F. Battiston, et al. “Evolutionary dynamics of higher-order interactions in social networks”. In: *Nature Human Behaviour* 5.5 (2021), pp. 586–595.

[121] R. Sinatra, D. Condorelli, and V. Latora. “Networks of motifs from sequences of symbols”. In: *Physical review letters* 105.17 (2010), p. 178702.

[122] L. Q. Ha, P. Hanna, J. Ming, and F. J. Smith. “Extending Zipf’s law to n-grams for large corpora”. In: *Artificial Intelligence Review* 32 (2009), pp. 101–113.

[123] J. Ryland Williams, P. R. Lessard, et al. “Zipf’s law holds for phrases, not words”. In: *Scientific reports* 5.1 (2015), p. 12209.

[124] P. Virtanen, R. Gommers, et al. “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3 (2020), pp. 261–272.

[125] Last.fm. *description page*. https://www.last.fm/about. Accessed: January 2021.

[126] O. Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010.

[127] Last.fm. *Music Recommendation Datasets for Research. Last.fm Dataset - 1K users*. http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html. Accessed: January 2021. 2009.

[128] MusicBrainz. *Home page*. https://musicbrainz.org/. Accessed: June 2022. 2022.

[129] M. Gerlach and F. Font-Clos. “A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics”. In: *Entropy* 22.1 (2020), p. 126.

[130] M. F. Porter. *Snowball: A language for stemming algorithms*. 2001.

[131] M. F. Porter. “An algorithm for suffix stripping”. In: *Program* (1980).

[132] C. D. Paice. “Another Stemmer”. In: *SIGIR Forum* 24.3 (Nov. 1990), pp. 56–61. ISSN: 0163-5840. DOI: 10.1145/101306.101310.

[133] W. Ammar, D. Groeneveld, et al. “Construction of the Literature Graph in Semantic Scholar”. In: *NAACL*. 2018.

[134] Allen Institute for AI. *Semantic Scholar's paper Field of Study classifier*. https://github.com/allenai/s2_fos. Accessed June 8, 2022. 2022.

[135] G. Di Bona, I. Iacopini, A. Petralia, and V. Latora. *Code used to download, process and analyse the data and the models at higher orders*. https://github.com/gabriele-di-bona/higher-order-heaps-laws. 2023.

[136] G. Csányi and B. Szendrői. “Structure of a large social network”. In: *Physical Review E* 69.3 (2004), p. 036131.

[137] W. Glänzel. “Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation”. In: *Journal of Informetrics* 1.1 (2007), pp. 92–102.

[138] S. Milojević. “Modes of collaboration in modern science: Beyond power laws and preferential attachment”. In: *Journal of the american society for Information science and technology* 61.7 (2010), pp. 1410–1423.

[139] S. Milojević. “Power law distributions in information science: Making the case for logarithmic binning”. In: *Journal of the American Society for Information Science and Technology* 61.12 (2010), pp. 2417–2425.

[140] M. Gerlach and E. G. Altmann. “Stochastic model for the vocabulary growth in natural languages”. In: *Physical Review X* 3.2 (2013), p. 021006.

[141] W. R. Inc. *Mathematica, Version 12*. Champaign, IL, 2022.

[142] T. Fink and M. Reeves. “How much can we influence the rate of innovation?” In: *Science Advances* 5.1 (2019), eaat6107.

[143] G. Bianconi and A.-L. Barabási. “Competition and multiscaling in evolving networks”. In: *EPL (Europhysics Letters)* 54.4 (2001), p. 436.

[144] J. Drews. “Drug discovery: a historical perspective”. In: *Science* 287.5460 (2000), pp. 1960–1964.

[145] D. Park, J. Nam, and J. Park. “Novelty and influence of creative works, and quantifying patterns of advances based on probabilistic references networks”. In: *EPJ Data Science* 9.1 (2020), p. 2.

[146] B. Hofstra, V. V. Kulkarni, et al. “The Diversity–Innovation Paradox in Science”. In: *Proc. Natl. Acad. Sci. U.S.A.* 117.17 (2020), pp. 9284–9291.

[147] D. H. Erwin and D. C. Krakauer. “Insights into innovation”. In: *Science* 304.5674 (2004), pp. 1117–1119.

[148] F. Wu and B. A. Huberman. “Novelty and collective attention”. In: *Proc. Natl. Acad. Sci. U.S.A.* 104.45 (2007), pp. 17599–17601.

[149] V. Sood, M. Mathieu, et al. “Interacting branching process as a simple model of innovation”. In: *Phys. Rev. Lett.* 105.17 (2010), p. 178701.

[150] M. Perc. “Self-organization of progress across the century of physics”. In: *Sci. Rep.* 3.1 (2013), p. 1720.

[151] A. T. Barron, J. Huang, R. L. Spang, and S. DeDeo. “Individuals, institutions, and innovation in the debates of the French Revolution”. In: *Proc. Natl. Acad. Sci. U.S.A.* 115.18 (2018), pp. 4607–4612.

[152] M. Coccia. “Why do nations produce science advances and new technology?” In: *Technol. in Soc.* 59 (2019), pp. 1–9.

[153] M. Coccia. “The theory of technological parasitism for the measurement of the evolution of technology and technological forecasting”. In: *Technol. Forecast. Soc. Change* 141 (2019), pp. 289–304.

[154] A. Pichler, F. Lafond, and J. D. Farmer. “Technological interdependencies predict innovation dynamics”. In: *arXiv:2003.00580* (2020).

[155] M. Andjelković, B. Tadić, et al. “Topology of innovation spaces in the knowledge networks emerging through questions-and-answers”. In: *PLoS One* 11.5 (2016).

[156] B. Tadić, M. M. Dankulov, and R. Melnik. “Mechanisms of self-organized criticality in social processes of knowledge creation”. In: *Phys. Rev. E* 96.3 (2017), p. 032307.

[157] T. M. A. Fink and M. Reeves. “How much can we influence the rate of innovation?” In: *Sci. Adv.* 5.1 (2019).

[158] T. M. A. Fink and A. Teimouri. “The mathematical structure of innovation”. In: *arXiv:1912.03281* (2019).

[159] M. Launay and V. Limic. “Generalized interacting urn models”. In: *arXiv: 1207.5635* (2012).

[160] G. Aletti, I. Crimaldi, A. Ghiglietti, et al. “Interacting reinforced stochastic processes: Statistical inference based on the weighted empirical means”. In: *Bernoulli* 26.2 (2020), pp. 1098–1138.

[161] M. Hayhoe, F. Alajaji, and B. Gharesifard. “A Polya urn-based model for epidemics on networks”. In: *Proceedings of the American Control Conference (ACC), 2017*. IEEE. 2017, pp. 358–363.

[162] M. Hayhoe, F. Alajaji, and B. Gharesifard. “Curing with the Network Polya Contagion Model”. In: *Proceedings of the 2018 Annual American Control Conference (ACC)*. IEEE. 2018, pp. 2644–2650.

[163] S. Berg. “Paradox of voting under an urn model: The effect of homogeneity”. In: *Public Choice* 47.2 (1985), pp. 377–387.

[164] T. Gong, L. Shuai, M. Tamariz, and G. Jäger. “Studying language change using price equation and Pólya-urn dynamics”. In: *PLoS One* 7.3 (2012), e33171.

[165] C. Cattuto, V. Loreto, and L. Pietronero. “Semiotic dynamics and collaborative tagging”. In: *Proc. Natl. Acad. Sci. U.S.A.* 104.5 (2007), pp. 1461–1464.

[166] R. Marcaccioli and G. Livan. “A Pólya urn approach to information filtering in complex networks”. In: *Nat. Commun.* 10.1 (2019), p. 745.

[167] F. Font-Clos, G. Boleda, and A. Corral. “A scaling law beyond Zipf’s law and its relation to Heaps’ law”. In: *New J. Phys.* 15.9 (2013), p. 093033.

[168] M. Perc. “The Matthew effect in empirical data”. In: *J. R. Soc. Interface* 11.98 (2014), p. 20140378.

[169] A. Mastrototaro. “A mathematical model for the emergence of innovations”. PhD thesis. Politecnico di Torino, 2018.

[170] A. Mazzolini, A. Colliva, M. Caselle, and M. Osella. “Heaps’ law, statistics of shared components, and temporal patterns from a sample-space-reducing process”. In: *Phys. Rev. E* 98.5 (2018), p. 052139.

[171] A. Mazzolini, M. Gherardi, et al. “Statistics of shared components in complex component systems”. In: *Phys. Rev. X* 8.2 (2018), p. 021023.

[172] A. Mazzolini, J. Grilli, et al. “Zipf and Heaps laws from dependency structures in component systems”. In: *Phys. Rev. E* 98.1 (2018), p. 012315.

[173] T. Carletti, A. Guarino, A. Guazzini, F. Stefanelli, et al. “Problem Solving: When Groups Perform Better Than Teammates”. In: *Journal of Artificial Societies and Social Simulation* 23.3 (2020), pp. 1–4.

[174] M. J. Salganik, P. S. Dodds, and D. J. Watts. “Experimental study of inequality and unpredictability in an artificial cultural market”. In: *Science* 311.5762 (2006), pp. 854–856.

[175] R. Pálovics and A. A. Benczúr. “Temporal influence over the Last. fm social network”. In: *Soc. Network Anal. Mining* 5.1 (2015), p. 4.

[176] J. Ternovski and T. Yasseri. “Social complex contagion in music listenership: A natural experiment with 1.3 million participants”. In: *Soc. Network* 61 (2020), pp. 144–152.

[177] P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The people’s choice.* Duell, Sloan & Pearce, 1944.

[178] R. M. Bond, C. J. Fariss, et al. “A 61-million-person experiment in social influence and political mobilization”. In: *Nature* 489.7415 (2012), p. 295.

[179] J. Bryden, S. P. Wright, and V. A. Jansen. “How humans transmit language: horizontal transmission matches word frequencies among peers on Twitter”. In: *J. R. Soc. Interface* 15.139 (2018), p. 20170738.

[180] E. Rogers. *Diffusion of Innovations, 4th Edition*. Free Press, 2010.

[181] D. Centola. *How Behavior Spreads: The Science of Complex Contagions*. Princeton Analytical Sociology Series. Princeton University Press, 2018.

[182] W. W. Zachary. “An information flow model for conflict and fission in small groups”. In: *J. Anthropol. Res.* 33.4 (1977), pp. 452–473.

[183] M. De Choudhury, Y.-R. Lin, et al. “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.

[184] M. E. J. Newman. “Finding community structure in networks using the eigenvectors of matrices”. In: *Phys. Rev. E* 74.3 (2006), p. 036104.

[185] P. M. Gleiser and L. Danon. “Community structure in jazz”. In: *Adv. Complex Syst.* 6.04 (2003), pp. 565–573.

[186] G. Dosi, A. Moneta, and E. Stepanova. “Dynamic increasing returns and innovation diffusion: bringing Polya Urn processes to the empirical data”. In: *Ind. Innovation* 26.4 (2019), pp. 461–478.

[187] P. Bonacich. “Factoring and weighting approaches to status scores and clique identification”. In: *J. Math. Sociol.* 2.1 (1972), pp. 113–120.

[188] L. Lü, D. Chen, et al. “Vital nodes identification in complex networks”. In: *Phys. Rep.* 650 (2016), pp. 1–63.

[189] P. Bonacich and P. Lloyd. “Eigenvector-like measures of centrality for asymmetric relations”. In: *Soc. Netw* 23.3 (2001), pp. 191–201.

[190] K. Ide, A. Namatame, et al. “A new centrality measure for probabilistic diffusion in network”. In: *Adv. Comput. Sci.* 3.5 (2014), pp. 115–121.

[191] R. Ghosh and K. Lerman. “Rethinking Centrality: The Role of Dynamical Processes in Social Network Analysis”. In: *Discrete Continuous Dyn. Syst. Ser. B* 19 (Sept. 2012).

[192] A. Berman and R. J. Plemmons. “Nonnegative Matrices”. In: *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979. Chap. 2, pp. 26–62.

[193] O. Perron. “Über Matrizen”. In: *Math. Ann.* 64 (1907), pp. 248–263.

[194] G. Frobenius. “Über Matrizen aus nicht negativen Elementen”. In: *S. B. Deutsch. Akad. Wiss. Berlin. Math-Nat. Kl.* 1912, pp. 456–477.

[195] R. Tarjan. “Depth-first search and linear graph algorithms”. In: *SIAM J. Comput.* 1.2 (1972), pp. 146–160.

[196] L. Katz. “A new status index derived from sociometric analysis”. In: *Psychometrika* 18.1 (1953), pp. 39–43.

[197] S. Boccaletti, G. Bianconi, et al. “The structure and dynamics of multilayer networks”. In: *Phys. Rep.* 544.1 (2014), pp. 1–122.

[198] I. Iacopini, G. Petri, A. Barrat, and V. Latora. “Simplicial models of social contagion”. In: *Nat. Commun.* 10.1 (2019), p. 2485.

[199] F. Battiston, G. Cencetti, et al. “Networks beyond pairwise interactions: structure and dynamics”. In: *Phys. Rep.* 874 (2020), pp. 1–92.

[200] A. Schecter, A. Pilny, et al. “Step by step: Capturing the dynamics of work team process through relational event sequences”. In: *J. Organ. Behav.* 39.9 (2018), pp. 1163–1181.

[201] V. S. Torrisi, S. Manfredi, I. Iacopini, V. Latora, et al. “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), University of Strathclyde, Glasgow, 2019*. 2019.

[202] B. Monechi, G. Pullano, and V. Loreto. “Efficient team structures in an open-ended cooperative creativity experiment”. In: *Proc. Natl. Acad. Sci. U.S.A.* 116.44 (2019), pp. 22088–22093.

[203] R. Sinatra, D. Wang, et al. “Quantifying the evolution of individual scientific impact”. In: *Science* 354.6312 (2016), aaf5239.

[204] G. Bianconi and A.-L. Barabási. “Bose-Einstein condensation in complex networks”. In: *Physical review letters* 86.24 (2001), p. 5632.

[205] J. P. Gleeson, J. A. Ward, K. P. O’Sullivan, and W. T. Lee. “Competition-induced criticality in a model of meme popularity”. In: *Phys. Rev. Lett.* 112.4 (2014), p. 048701.

[206] J. P. Gleeson, K. P. O’Sullivan, R. A. Baños, and Y. Moreno. “Effects of network structure, competition and memory time on social spreading phenomena”. In: *Phys. Rev. X* 6.2 (2016), p. 021019.

[207] J. D O’Brien, I. K. Dassios, and J. P. Gleeson. “Spreading of memes on multiplex networks”. In: *New J. Phys.* 21.2 (2019), p. 025001.

[208] J. Ziman. *Technological innovation as an evolutionary process*. Cambridge University Press, 2003.

[209] W. B. Arthur. *The nature of technology: What it is and how it evolves*. Simon and Schuster, 2009.

[210] P. Érdi, K. Makovi, et al. “Prediction of emerging technologies based on analysis of the US patent citation network”. In: *Scientometrics* 95.1 (2013), pp. 225–242.

[211] J. Kim and C. L. Magee. *Dynamic patterns of knowledge flows across technological domains: empirical results and link prediction*. 2017.

[212] A. Tacchella, A. Napoletano, and L. Pietronero. “The language of innovation”. In: *PloS one* 15.4 (2020), e0230107.

[213] D. Zhou, D. M. Lydon-Staley, P. Zurn, and D. S. Bassett. “The growth and form of knowledge networks by kinesthetic curiosity”. In: *Curr. Opin. Behav. Sci.* 35 (2020), pp. 125–134.

[214] N. O. Hodas and K. Lerman. “The simple rules of social contagion”. In: *Sci. Rep.* 4.1 (2014), pp. 1–7.

[215] D. Centola. *How Behavior Spreads: The Science of Complex Contagions*. Princeton Analytical Sociology Series. Princeton University Press, 2020. ISBN: 9780691202426.

[216] E. Zangerle, M. Pichl, W. Gassler, and G. Specht. "# nowplaying music dataset: Extracting listening behavior from twitter". In: *Proceedings of the first international workshop on internet-scale multimedia management*. 2014, pp. 21–26.

[217] B. Brost, R. Mehrotra, and T. Jehan. "The music streaming sessions dataset". In: *The World Wide Web Conference*. May 2019, pp. 2594–2600.

[218] M. Schedl, C. Bauer, et al. "Listener Modeling and Context-Aware Music Recommendation Based on Country Archetypes". In: *Front. Artif. Intell.* 3 (2021), p. 108. ISSN: 2624-8212.

[219] R. Di Clemente, M. Luengo-Oroz, et al. "Sequences of purchases in credit card data reveal lifestyles in urban populations". In: *Nat. Commun.* 9.1 (2018), pp. 1–8.

[220] L. M. Aiello, R. Schifanella, D. Quercia, and L. Del Prete. "Large-scale and high-resolution analysis of food purchases and health outcomes". In: *EPJ Data Science* 8.1 (2019), pp. 1–22.

[221] E. Schulz, R. Bhui, et al. "Structured, uncertainty-driven exploration in real-world consumer choice". In: *Proceedings of the National Academy of Sciences* 116.28 (2019), pp. 13903–13908.

[222] L. Weng and F. Menczer. "Topicality and impact in social media: diverse messages, focused messengers". In: *PloS One* 10.2 (2015), e0118410.

[223] C. Ivan, B. Peter, and K. Tsvi. "The yahoo! music dataset and kdd-cup'11". In: *Last.fm web 2.0 dataset. In 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems*. RecSys 2011, Chicago, I. 2011.

[224] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer. "The yahoo! music dataset and kdd-cup'11". In: *Proceedings of KDD Cup 2011*. PMLR. 2012, pp. 3–18.

[225] B. McFee, T. Bertin-Mahieux, D. P. Ellis, and G. R. Lanckriet. "The million song dataset challenge". In: *Proceedings of the 21st International Conference on World Wide Web*. 2012, pp. 909–916.

[226] D. Hauger, M. Schedl, A. Košir, and M. Tkalcic. "The million musical tweets dataset: What can we learn from microblogs". In: *Proc. ISMIR*. 2013, pp. 189–194.

[227] G. Vigliensoni and I. Fujinaga. *The Music Listening Histories Dataset.* 2017.

[228] M. Schedl. "The lfm-1b dataset for music retrieval and recommendation". In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 2016, pp. 103–110.

[229] R. M. MacCallum, M. Mauch, A. Burt, and A. M. Leroi. “Evolution of music by public choice”. In: *Proceedings of the National Academy of Sciences* 109.30 (2012), pp. 12081–12086.

[230] C. S. Siew, D. U. Wulff, N. M. Beckage, and Y. N. Kenett. “Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics”. In: *Complexity* 2019 (2019).

[231] C. W. Lynn and D. S. Bassett. “How humans learn and represent networks”. In: *Proc. Natl. Acad. Sci. U.S.A.* 117.47 (2020), pp. 29407–29415.

[232] D. M. Lydon-Staley, D. Zhou, et al. “Hunters, busybodies and the knowledge network building associated with deprivation curiosity”. In: *Nat. Hum. Behav.* 5.3 (2021), pp. 327–336.

[233] M. McPherson, L. Smith-Lovin, and J. M. Cook. “Birds of a feather: Homophily in social networks”. In: *Annu. Rev. Sociol.* 27.1 (2001), pp. 415–444.

[234] I. Cantador, P. Brusilovsky, and T. Kuflik. “2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)”. In: *Proceedings of the 5th ACM conference on Recommender systems*. RecSys 2011. Chicago, IL, USA: ACM, 2011.

[235] M. D. Humphries and K. Gurney. “Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence”. In: *PLoS One* 3.4 (2008), e0002051.

[236] S. Fortunato. “Community detection in graphs”. In: *Phys. Rep.* 486.3-5 (2010), pp. 75–174.

[237] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[238] C. Cattuto, A. Baldassarri, V. D. Servedio, and V. Loreto. “Vocabulary growth in collaborative tagging systems”. In: *arXiv preprint arXiv: 0704.3316* (2007).

[239] S. Aral, L. Muchnik, and A. Sundararajan. “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks”. In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21544–21549.

[240] A. Patelli, G. Cimini, E. Pugliese, and A. Gabrielli. “The scientific influence of nations on global scientific and technological development”. In: *Journal of Informetrics* 11.4 (2017), pp. 1229–1237.

[241] G. P. Swann. *The economics of innovation: an introduction*. Edward Elgar Publishing, 2014.

[242] F. Richter, B. Haegeman, R. S. Etienne, and E. C. Wit. “Introducing a general class of species diversification models for phylogenetic trees”. In: *Statistica Neerlandica* 74.3 (2020), pp. 261–274.

[243] A. R. Benson, R. Abebe, et al. “Simplicial closure and higher-order link prediction”. In: *Proceedings of the National Academy of Sciences* 115.48 (2018), E11221–E11230.

[244] A. E. Sizemore, E. A. Karuza, C. Giusti, and D. S. Bassett. “Knowledge gaps in the early growth of semantic feature networks”. In: *Nature human behaviour* 2.9 (2018), pp. 682–692.

[245] F. Baccini, F. Geraci, and G. Bianconi. “Weighted simplicial complexes and their representation power of higher-order network data and topology”. In: *Phys. Rev. E* 106 (3 2022), p. 034319.