universität
wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation/ Title of the Doctoral Thesis

„Statistical mechanics based approach to understanding information flow processes in complex systems“

verfasst von / submitted by

Dipl.Ing. Niklas Reisz, BSc, BSc (WU)

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796 605 411
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Physik
Betreut von / Supervisor:	Univ.-Prof. Mag. DDr. Stefan Thurner Univ.-Prof. Mag. Dr. Christoph Dellago

Statement of the thesis advisors

I hereby confirm that Niklas Reisz has contributed substantially to the following publications and patent, which are part of this thesis

[1] Niklas Reisz, Vito D. P. Servedio, Vittorio Loreto, William Schueller, Márcia R. Ferreira, and Stefan Thurner. *Loss of sustainability in scientific work*, New Journal of Physics, 24(5):053041, 2022.

[2] Márcia R. Ferreira, Niklas Reisz, William Schueller, Vito D. P. Servedio, Stefan Thurner, and Vittorio Loreto. *Quantifying Exaptation in Scientific Evolution*, pages 55–68. Springer International Publishing, Cham, 2020.

[3] Vito D. P. Servedio, Márcia R. Ferreira, Niklas Reisz, Rodrigo Costas, and Stefan Thurner. *Scale-free growth in regional scientific capacity building explains long-term scientific dominance*, Chaos, Solitons & Fractals, 167:113020, 2023.

[4] Niklas Reisz, Vito D. P. Servedio, and Stefan Thurner. *To what extent homophily and influencer networks explain song popularity*, arXiv:2211.15164, 2022.

[5] Vittorio Loreto, Vito D. P. Servedio, Niklas Reisz, and Stefan Thurner. *System, method, and computer-readable medium for tracking information exchange*, US Patent App. 16/375,017, 2020.

as well as to the projects that are presented in the appendix.

Prof. Stefan Thurner

Prof. Christoph Dellago

Abstract

It has been argued that information has become the most valuable resource in the world today and is crucial to the advancement of both science and innovation. People exchange information in many ways, including personal conversations, emails, scientific publications, and many more, creating a dynamic, directed multilayer network. Information spreads along the links of this network from person to person in anomalous diffusion processes. In this thesis, we study different aspects of these diffusion processes and present the outcomes in four publications and a patent. Our study has led to two different directions of research, which are structured in two parts.

In the first part, we study citation networks where existing knowledge is built upon in a combinatorial and evolutionary process over long time periods. Understanding the mechanism of how these networks arrange and rearrange over time allows one to estimate the change of active knowledge in a scientific field. This part of the thesis consists of two papers and a book chapter. In the first paper [1], we study how scientific knowledge is created and forgotten over time. This can be traced to the dynamic network of references between scientific works of entire fields. There, we find that the probability of attracting citations for individual publications is mainly defined by two power laws, preferential attachment and a sharp decline over time. While most papers are forgotten very quickly, we identify those that have the potential to become milestones. We further predict the citation landscape of 2050, estimating that 95% of the papers that will be cited at that time have yet to be written.

Out of the large number of forgotten papers, a tiny fraction is rediscovered in a different field than the one it was initially published in, in events called *exaptations*. In a book chapter [2], we focus on these rare exaptation events and attempt to formalize them for the first time. We present an entropy-based method to detect the state of any exaptation that allows us to distinguish them from regular papers at an early stage.

In a second paper [3], we study the diffusion process of knowledge on a geographic basis by looking at the movement of researchers between institutions. We want to explain the nature of this diffusion process by identifying sources, sinks, and flows of particular scientific expertise over time. We find that while the scientific capabilities in specific fields vary significantly between regions, the process of scientific knowledge accumulation is remarkably similar between regions and is best described by preferential attachment. Consequently, early birds tend to dominate. However, we find that no critical mass is required, i.e., any number of pioneering scientists can start a flourishing research field. Moreover, given enough resources, any region might outgrow its competitors and establish itself as the leader in a field.

In the second part, we study information diffusion in social networks on short timescales. Drawing parallels to our study on citation networks, we show how our findings can be extended to the domain of music, where information is not interpreted as knowledge but rather as knowing about novelties such as new songs [4]. Using a unique dataset we collected, we study the co-evolution of behavioral patterns, like music preference, and underlying social network structures, in particular, the friendship network. We find that on short timescales, information about new songs is exchanged along the edges of friendship networks in a diffusion process. These microscopic interactions lead to a macroscopic increase of similarity between connected users, called *homophily*. We show how homophily can be leveraged to significantly enhance state-of-the-art machine learning models' predictions of song popularity. Furthermore, we demonstrate how it relates to preferential attachment processes.

As part of this thesis, we designed and prototyped a new system for tracking multilayer social inter-

action networks. The system records information on different channels such as personal discussions, talks, group meetings, emails, and collaborative online writing tools. It has been patented in the US and Japan [5] and was presented at public outreach events at the *Palazzo delle Esposizioni* in Rome, as well as the *Maker Faire 2022* in Rome. The fine-grained data that is made available by this tool can be used to study and optimize organizational information flow.

In this work, we show how principles of diffusion in networks and simulations of microscopic interacting particles that have been studied in physics for the last decades can be applied to various domains of social dynamics, touching the fields of *science of science*, hit song science, and organizational communication. In this sense, this work can be seen as a physics-inspired, interdisciplinary exploration of information flow processes. Our findings increase our theoretical understanding of attachment processes across different domains through their contribution of microscopic generative models on the one hand and insights into flow processes in networks on the other hand. In addition to that, they contribute accurate predictions that allow for better policy-making and decision-making in science, the music industry, and organizational communication.

Kurzfassung

Zahlreiche Argumente deuten darauf hin, dass Information zur weltweit wertvollsten Ressource aufgestiegen ist, und für den Fortschritt von Wissenschaft und Innovation eine entscheidende Rolle spielt. Personen tauschen Information untereinander auf vielfältige Weise aus, zum Beispiel durch persönliche Gespräche, E-Mails, wissenschaftliche Veröffentlichungen und vieles mehr, wodurch ein dynamisches, gerichtetes, mehrschichtiges Netzwerk entsteht. Information breitet sich entlang der Verbindungen dieses Netzwerks von Person zu Person in anomalen Diffusionsprozessen aus. In dieser Arbeit untersuchen wir verschiedene Aspekte dieser Ausbreitungsprozesse und stellen die Ergebnisse in vier wissenschaftlichen Veröffentlichungen und einem Patent vor. Unsere Studie führte zu zwei verschiedenen Forschungsrichtungen, die im Folgenden in zwei Teile gegliedert sind.

Im ersten Teil, untersuchen wir Zitiernetzwerke, auf denen bestehendes Wissen in einem kombinatorischen und evolutionären Prozess über lange Zeiträume hinweg aufgebaut wird. Wenn man den Mechanismus versteht, wie sich diese Netzwerke im Laufe der Zeit anordnen und umgestalten, kann man den Wandel des aktiven Wissens in einem wissenschaftlichen Feld abschätzen. Dieser Teil der Arbeit besteht aus zwei Veröffentlichungen und einem Buchkapitel. In der ersten Publikation [1] untersuchen wir, wie Wissen im Laufe der Zeit entsteht und vergessen wird. Dies lässt sich anhand des dynamischen Netzwerks von Zitaten zwischen wissenschaftlichen Arbeiten ganzer Fachgebiete nachvollziehen. Dabei stellen wir fest, dass die Wahrscheinlichkeit, dass einzelne Publikationen zitiert werden, hauptsächlich durch zwei Potenzgesetze bestimmt wird: eine bevorzugte Bindung (Englisch *preferential attachment*) und ein starker Rückgang mit zunehmendem Alter der Publikation. Während die meisten Arbeiten sehr schnell in Vergessenheit geraten, identifizieren wir diejenigen, die das Potenzial haben, zu Meilensteinen zu werden. Darüber hinaus prognostizieren wir die wissenschaftliche Zitierlandschaft des Jahres 2050 und schätzen, dass 95% der Arbeiten, die zu diesem Zeitpunkt zitiert werden, zum heutigen Zeitpunkt noch nicht geschrieben wurden.

Unter der großen Zahl vergessener Arbeiten wird ein winziger Teil in einem anderen Feld als dem, in dem sie ursprünglich veröffentlicht wurden, wiederentdeckt, was als *Exaptation* bezeichnet wird. In einem Buchkapitel [2] konzentrieren wir uns auf diese seltenen Exaptationsereignisse und versuchen, sie zum ersten Mal zu formalisieren. Wir stellen eine auf Entropie basierende Methode vor, um den Zustand einer Exaptation zu erkennen, die es uns ermöglicht, sie in einem frühen Stadium von gewöhnlichen Veröffentlichungen zu unterscheiden.

In einer zweiten Publikation [3] untersuchen wir den Prozess der Wissensausbreitung auf geografischer Basis, indem wir die Bewegungen der WissenschaftlerInnen zwischen Institutionen untersuchen. Wir wollen diesen Ausbreitungsprozess erklären, indem wir Quellen, Senken und Ströme von besonderem wissenschaftlichen Fachwissen im Laufe der Zeit ermitteln. Dabei stellen wir fest, dass die wissenschaftliche Expertise in bestimmten Fachgebieten zwar von Region zu Region sehr unterschiedlich sind, der Prozess der wissenschaftlichen Wissensakkumulation jedoch bemerkenswert ähnlich ist und sich am besten durch eine bevorzugte Bindung beschreiben lässt. Folglich haben die ersten in einem Feld einen deutlichen Startvorteil. Wir stellen jedoch fest, dass keine kritische Masse erforderlich ist und dass jede Region, wenn sie über genügend Ressourcen verfügt, ihre Konkurrenten überholen und eine führende Position auf einem Gebiet einnehmen kann.

Im zweiten Teil untersuchen wir die Informationsausbreitung in sozialen Netzwerken auf kurzen Zeitskalen. Hier ziehen wir Parallelen zu unserer Studie über Zitiernetzwerke, und zeigen wie unsere Ergeb-

nisse auf den Bereich der Musik ausgeweitet werden können, wo Informationen nicht als wissenschaftliches Wissen, sondern als Wissen über Neuheiten wie neue Lieder interpretiert werden können [4]. Anhand eines einzigartigen Datensatzes, den wir gesammelt haben, untersuchen wir die Koevolution von Verhaltensmustern, wie beispielsweise der Musikpräferenz, und den zugrunde liegenden sozialen Netzwerkstrukturen, insbesondere dem Freundschaftsnetzwerk. Wir stellen fest, dass auf kurzen Zeitskalen Informationen über neue Lieder entlang der Kanten von Freundschaftsnetzwerken in einem Diffusionsprozess ausgetauscht werden. Diese mikroskopischen Interaktionen führen zu einer makroskopischen Zunahme der Ähnlichkeit zwischen verbundenen Nutzern, die als *Homophilie* bezeichnet wird. Wir zeigen, wie Homophilie genutzt werden kann, um die Vorhersagen moderner Machine-Learning Modelle zur Popularität von Songs erheblich zu verbessern. Darüber hinaus zeigen wir, wie sie mit bevorzugter Bindung zusammenhängt.
Im Rahmen dieser Arbeit haben wir ein neues System zur Aufzeichnung mehrschichtiger sozialer Interaktionsnetzwerke entwickelt und prototypisiert. Das System zeichnet Informationen auf verschiedenen Kanäle wie zum Beispiel in Diskussionen, Gesprächen, Gruppentreffen, E-Mails und kollaborativen Online-Schreibwerkzeugen auf. Es wurde in den USA und Japan patentiert [5] und auf öffentlichen Veranstaltungen im *Palazzo delle Esposizioni* in Rom sowie auf der *Maker Faire 2022* in Rom vorgestellt. Die hochaufgelösten Daten, die dieses Tool zur Verfügung stellt, können zur Untersuchung und Optimierung des organisatorischen Informationsflusses verwendet werden.
In dieser Arbeit zeigen wir, wie Prinzipien der Diffusion in Netzwerken und Simulationen mikroskopischer interagierender Teilchen, die in den letzten Jahrzehnten in der Physik erforscht wurden, auf verschiedene Bereiche der sozialen Dynamik angewandt werden können, die die Bereiche Wissenschaft der Wissenschaft, Hit-Song-Wissenschaft und Organisationskommunikation betreffen. In diesem Sinne kann diese Arbeit als eine von der Physik inspirierte, interdisziplinäre Erforschung von Informationsflussprozessen betrachtet werden. Unsere Ergebnisse erweitern unser theoretisches Verständnis von Bindungsprozessen in verschiedenen Bereichen, indem sie einerseits mikroskopische generative Modelle und andererseits Einblicke in Diffusionsprozesse in Netzwerken liefern. Darüber hinaus tragen sie zu genauen Vorhersagen bei, die eine bessere politische Gestaltung und Entscheidungsfindung in der Wissenschaft, der Musikindustrie und der Organisationskommunikation ermöglichen.

Acknowledgements

First and foremost, I would like to thank my thesis supervisor, Stefan Thurner, for his guidance and supervision along the way. Thank you for giving me the opportunity to do my Ph.D. at the Complexity Science Hub, a truly inspiring and exciting institution. And thank you for sharing your experience and deep knowledge in the field to help me figure out the solution to my problems when I was stuck, and for always challenging and motivating me to acquire the skills and the knowledge that I required on the way.

I would also like to thank my supervisor, Christoph Dellago, for his support and guidance and for welcoming my research at the Institute for Computational and Soft Matter Physics at the University of Vienna.

I am thankful for the great people I have worked with in the Innovation Dynamics group. I want to express my gratitude toward my colleague, group leader, and friend, Vito, for his supervision, endless patience, and tremendous support, for introducing me to the field and teaching me about complexity science and programming, for countless discussions and hundreds of whiteboards filled with equations and cartoons, for the freedom and flexibility I had while working with him as my group leader, for his unmatched expert knowledge and teachings in Italian cuisine, and for the fun we had working and traveling together. Many thanks to my colleagues and friends, Márcia for her knowledge and experience in the field of science of science, William for challenging and supporting me in improving my coding, Simone for his sharp observations and comments, and all of them for the great time we had together.

Thank you also to Vittorio Loreto and his team at Sony Computer Science Labs Paris, for the exciting projects and social experiments we worked on together, and for the inspiring insights into the cross-section of science and industry.

I am truly happy to have met so many great and bright people at the Complexity Science Hub, the University of Vienna, Sony Computer Science Labs, and the Centro Ricerche Enrico Fermi, the names of which would go beyond the scope of this page.

Finally, I want to express my sincere gratitude to my family and my partner. Thank you for always being there for me and for supporting me in finding my own way.

Contents

List of Figures

Preface

This thesis deals with information diffusion processes in networks, and how their statistical footprints and critical behavior can be understood using statistical physics-based theoretical and numerical modeling. It is structured as follows:
In Part I, an overview of the most significant theoretical ideas is offered together with a summary of the current empirical situation. Additionally, the subjects of the papers and patent in Part II are introduced, along with how they relate to information flow processes.
Part II includes three scientific articles, one book chapter, and one patent, that were written and published during the course of the author's Ph.D. studies. Each chapter introduces a different aspect of information flow processes. At the start of each chapter, the relevant publications are listed including the status of publication and the journal. This is followed by the original or pre-print version of the publication.
First, chapter 3 introduces a paper and a book chapter that investigate information diffusion in scientific citation networks. In the paper, we study how scientific knowledge is created and forgotten over time. We introduce a microscopic generative model for how likely scientific publications are cited with regard to their past popularity and age. We distinguish between average and exceptional publications and test the model's capability for forecasting future citation landscapes. The book chapter extends this model by dealing with a certain class of exceptional and rare papers called *exaptations*. In an exaptation event, previously found insights in one field are rediscovered and repurposed in a different field. We formalize these exaptation events and introduce an entropy-based method of assessing the current state of an exaptation.
Next, chapter 4 presents a paper on the diffusion process of knowledge on a geographic basis. Supplementary to chapter 3, here we are investigating knowledge diffusion through the movement of scientists between institutions. We introduce a microscopic generative model for the growth of scientific capacities of geographical regions and use it to compare regions with regard to their attractiveness for scientists.
In chapter 5, we transfer our findings to the domain of music. While on the surface, science and music seem quite different, the underlying microscopic interactions are remarkably similar. Here, information can be interpreted as knowing about novelties rather than scientific discoveries. In our paper, we investigate the diffusion of this information along the links of friendship networks. We study, how the microscopic interactions that drive the diffusion, lead to network topologies that enable cascading propagation of information. Additionally, we demonstrate how these findings can be translated into predictive models with potential industrial applications.
In chapter 6, we present a patent that, at the time of writing, is pending in the USA and has been accepted in Japan. The patent describes our experimental work, in which we developed a software tool for the purpose of recording multilayer social interaction networks. These networks are generated by day-to-day knowledge flows in organizations or at events such as conferences or exhibitions. With our tool, we aim to make microscopic interactions visible and open them up to subsequent study.
Part III gives a summary of the main results of this thesis. It discusses the relevance of the findings and gives an outlook on future work. Some of the results or application cases, that are unfinished or went unpublished yet, are discussed in the appendix.

Part I

Introduction

Chapter 1

Quantifying information flow networks

In this chapter, we shall discuss the basic concepts and methods needed to study information flow processes. Before we do, we need to define what we mean by *information* and *information flow*. Information and knowledge have several different definitions. In the context of this thesis, we will use the definition that we give in the following, which is firmly based on [6]. According to [6], information is structured, formatted data. It is inert because it doesn't enable us to act on it unless interpreted first. If interpreted by a human receiver, information becomes knowledge and enables us to act on it. Hence knowledge is related to cognitive capabilities. The implications of this definition become apparent when we look at an example [6]. The information in Elliot Lieb's *Density functionals for coulomb systems* [7] is straightforward to replicate. One simply has to copy the digital file of the paper, for example, by putting it on a USB stick. Replicating the knowledge, i.e., having another person (e.g., a student) understand its contents, is quite challenging. Consequently, we speak of *information flow* whenever structured, formatted data is transmitted and of *knowledge transfers* whenever a human receiver interprets the transmitted data.

In this thesis, we shall address the transfer of knowledge. However, knowledge depends on the interpretation of information by the receiver. As our approach is purely quantitative and grounded in physics, we do not study any psychological or cognitive aspects of knowledge transfer, hence on a methodological level, we are dealing with information flow processes rather than knowledge transfers. Whenever it is reasonable to do so, we take these information flows as a proxy for knowledge transfers. To give an example, it is reasonable to assume that if an email is sent to a colleague, that colleague will likely read and, to some extent, understand the content of that email. Hence the information flow of an email exchange can be taken as a proxy for a knowledge transfer. By making this connection between information flow and knowledge transfer, whenever it is reasonable to do so, our findings can be related to knowledge transfers. In other words, we study information diffusion to argue about knowledge transfers.

A particular set of concepts and tools is needed to study information flow. This becomes apparent when we look at recent developments. Over the last decades, technological progress has added more and more communication channels. Information is exchanged through conversations, written books, or letters and via phone, emails, websites, smartphone applications, and many more. Suppose one tries to understand these exchanges of knowledge quantitatively. In that case, one quickly finds characteristics common to complex systems: specific interactions, co-existence of multiple interactions of similar strength, and co-evolution of states and interactions. In the following, we give a structured overview of the main concepts and methods of complex systems that can be used to deal with this situation and that are relevant in the context of this thesis.

1.1 Complex networks

In the following section, which is based on the book by Thurner et al. [8], we give a brief definition of a complex network and an overview of the main concepts behind it and their relation to traditional physics.
In a one-sentence summary, Thurner et al. define complex systems as, "*Complex systems are co-evolving multilayer networks*", hence a complex network is a co-evolving multilayer network. In this sentence, multiple facts about complex systems are summarized. Let us take a closer look at them.
First, on the microscopic level, complex systems consist of many elements characterized by their states. These elements are not necessarily restricted to physical particles but instead take the form of generalized particles. Anything that can be described by states and interactions, can be an element of a complex system.
In physics, interactions consist of the four fundamental forces mitigated by the exchange of gauge bosons. Typically, given a specific problem, the strength of one or two of these forces is multiple orders of magnitude larger than the others. For example, to calculate the path of planet Earth around the sun, we usually do not consider strong, weak, or electromagnetic interactions in our calculation. Instead, one focuses only on gravitation. In complex systems, interactions are neither limited to the four fundamental forces nor are they limited to the exchange of gauge bosons. Instead, we are dealing with *generalized interactions*. Superpositions of multiple different types of interactions of similar strength are common. In addition to that, interactions might be specific, meaning that not every element interacts with every other element. A handy concept to describe such specific interactions are networks. Commonly, networks are denoted with a matrix M with matrix elements M_{ij} where i and j are the interacting elements, called nodes, and the matrix entries, called links, denote the strength of the interactions between each pair of elements. When there are multiple different types of interaction acting simultaneously, each type of interaction can be described in a separate layer α of the network by adding an index to our matrix M^{α}_{ij}. Networks with multiple layers are called multilayer networks.
In non-complex systems, typically, interactions do not change over time. For example, gravity can change the state of an object by changing its momentum and position through its pull. However, the changing mass of an object does not change the value of the gravitational constant or the fact that the Newtonian gravitational force is proportional to the square of the distance. In complex systems, we are facing a different situation. Here, the change of states caused by interactions can, in turn, induce change in the interactions. This leads to a co-evolution of states and interactions that comes with two implications. First, complex systems are often chaotic and heavily depend on the initial conditions. Second, the system's phase space evolves with it and might be open-ended, leading to a vast or infinite number of macro states that cannot be simply inferred from the properties of the single elements. Hence, self-organizing microscopic processes between individual elements can lead to complex macroscopic outcomes. This phenomenon is called *emergence*. In our multilayer network, we can hint at the change of interactions by adding a time dependence to our matrix $M^{\alpha}_{ij}(t)$. We will discuss the co-evolution of states and interaction in more detail in section 1.4.
At this point, we arrived at our definition from earlier: "*Complex systems are co-evolving multilayer networks*". By complex, we do not mean complicated (though, in practice, some degree of correlation between the two is not uncommon). Instead, we mean the specific set of characteristics of a system that we discussed here. To give an example of a complex system, let us put the definition in the context of information flow processes. Information, interpreted as knowledge, is transferred between people (the nodes of our network) when they interact with each other. Not every person interacts with every other person, making the interactions specific (described by the links of our network). There are many ways of exchanging knowledge orally, through books, publications, emails, and many more. Each of them can be mapped to a layer in the network, leading to a multilayer network. This network is not static. The exchange of knowledge between nodes might, at any point, lead to a change in interactions as well. People might get encouraged to communicate with someone previously unknown to them. Or they might choose not to interact anymore with a specific node after receiving a particular piece of information from them. They might even use their knowledge to create a new means of transferring

knowledge, as was the case with the invention of the internet, leading to a whole new layer in our network. In summary, we can state that information flow occurs on a complex co-evolving multilayer network.

1.2 Diffusion processes in networks

In this section, which is based on the books by Thurner et al. [8] and Newman [9], we discuss the central concepts of diffusion on networks.
In classical physics, diffusion is the flow of any kind of matter from regions of high concentration to regions of low concentration due to the random motion of individual particles. It is described by Fick's law [10] as follows,

$$J = -D\frac{d\phi}{dx} \tag{1.1}$$

where J is the diffusion flux, D is the diffusion constant, ϕ is the concentration, and x is the position. Analogous to classical diffusion, diffusion on networks can be modeled with randomly moving particles, i.e., random walkers [11]. In each time step, the walkers jump from one node to another along the network links. Different diffusion processes can be modeled by different *decision rules* for the random walkers, i.e., rules for the walkers to decide which one out of multiple possible links to follow. These decision rules can be formalized in update equations. A general form of an update equation is given by,

$$x_i(t+1) = \sum_j n(j \to i)x_j(t) + \beta_i \tag{1.2}$$

where $x_j(t)$ is the number of random walkers present at node j at time t, $x_i(t+1)$ is the number of random walker at node i in the next discrete time step $t+1$, $n(j \to i)$ is a transition element and β is a constant birth rate of random walkers [8]. The specific version of update equation 1.2 that is closest to physical diffusion is the *Laplacian diffusion*. In the case of Laplacian diffusion on an unweighted and undirected network, the birth rate equals zero $\beta_i = 0$. The transition elements are given by $n(j \to i) = \frac{A_{ij}}{k_j}$, where A_{ij} is the adjacency matrix that is one if there is a link between nodes i and j and zero otherwise, and k_j is the degree of node j, i.e., the number of links that connect to node j. Hence, the update equation for Laplacian diffusion is given by,

$$x_i(t+1) = \sum_j \frac{A_{ij}}{k_j}x_j(t) \tag{1.3}$$

where the total number of random walkers is conserved [8]. Typically, one is interested in the stationary distribution x^* of random walkers, i.e., the distribution that does not change from one time step to the next $x^*(t+1) = x^*(t)$. In vector notation, we can write,

$$x^* = AD^{-1}x^* \tag{1.4}$$

The solution is the eigenvector of the matrix AD^{-1} with eigenvalue 1. In general, the eigenvector entries are not identical and reveal a property of each node that can aid the analysis of the network [8, 12]. For random walk-based diffusion models, there exists a range of different variations, such as the Eigenvector centrality [13], Katz prestige [14], or Google's famous PageRank [15].
It is also possible to introduce diffusion to networks without the concept of random walkers. Fick's law for networks can be written as,

$$\frac{dx_i}{dt} = D^{diff}\sum_j A_{ij}(x_j - x_i) \tag{1.5}$$

where D^{diff} is the diffusion constant [9]. In vector notation, we get,

$$\frac{dx}{dt} + D^{diff}\Lambda x = 0 \tag{1.6}$$

with the Laplacian $\Lambda = D - A$, where A is the adjacency matrix and D is a diagonal matrix $D_{ii} = \sum_j A_{ij}$. We can solve this equation by writing x as a linear combination of its eigenvectors, where we can see that the term with the smallest eigenvalue $\lambda_N = 0$ dominates. In analogy to classic diffusion, the result is a uniform distribution of walkers across all nodes.

The two examples of Laplacian diffusion provide an introduction and some intuition for diffusion models on networks. Many real-world systems are characterized by the diffusion of particles, walkers, or information packages on networks [16, 17]. Important applications include the internet [18], social networks [19–22], or technological diffusion [23]. For most of these real-world systems, specific problems like the ones that are the topic of this dissertation can not be solved with Laplacian diffusion alone. We need more detailed models for these problems, and often we rely on numerical solutions. The next section gives an overview of some of the relevant methods and models in our context.

1.3 Microscopic diffusion models and numerical methods

Complex networks are dynamic by definition. Often, they grow or rearrange over time. In this thesis, we discuss multiple examples of growing networks, the citation network of scientific literature, discussed in chapter 3, the mobility network of scientists, discussed in chapter 4, and the listening network of the music streaming platform *last.fm*, discussed in 5. When we want to understand these complex networks' macroscopic growth dynamics, the underlying microscopic interactions often reveal valuable insights. In this context of growing networks, the *attachment kernel* is a useful concept. The attachment kernel is a non-normalized weight for each node that depends on the properties of that node. When a new link, l, is formed, it attaches with probability, p_i, to node, i, where p_i is proportional to the attachment kernel, f_i, [24],

$$P(l \text{ connects to } i) \propto f_i \tag{1.7}$$

The attachment kernel is updated as the node's properties change over time. In [25], the author suggests a method to estimate an attachment kernel from data, which we will briefly discuss here. If we assume that the probability for a new link, l, that enters the system at time t, to attach to a node, i, depends on some property, x, of that node, then the probability, $P(x,t)$, for the link to attach to *any* node with property x, is given by

$$P(x,t) = f(x,t)\frac{n(t)}{N(t)} \tag{1.8}$$

where $f(x,t)$ is the attachment kernel, $n(t)$ is the number of nodes that have property x at time t, and $N(t)$ is the total number of nodes at time t [1]. The attachment kernel can then be estimated by building a histogram for different values of x, where each new link contributes with a weight of $\frac{N(t)}{n(t)}$. If the attachment probability does not depend on property x, the attachment kernel is constant $f(x,t) = 1$. Otherwise, one can fit a function to the histogram to determine the functional shape of the attachment kernel.

The particular case where the attachment kernel is proportional to the degree of the node is called preferential attachment and is highly relevant in many real-world systems. Preferential attachment is self-reinforcing and is sometimes also called *rich-get-richer* or the *Matthew effect* [8, 26] because it leads to a highly skewed degree distribution where a small number of nodes, called *hubs*, have a much larger than average number of connections. This can be understood from one of the most famous and fundamental network growth models, called the Barabási-Albert model [27], that can be constructed using preferential attachment only. In the following, we briefly overview the model and the resulting degree distribution.

In the Barabási-Albert model, a new node is added at each time step, t. The node attaches to precisely one other node, i, where the node is chosen at random from all existing nodes with a probability, $\Pi(k_i)$,

proportional to the degree of the node, k_i, as in the following

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{1.9}$$

where the sum in the denominator is the sum over the degrees of all other nodes that exist at time t [8]. If we consider, that the rate at which a node, i, acquires new links over time is $\frac{d}{dt}k_i = \Pi(k_i)$, and $\sum_j k_j = 2t$ because we are adding a link at every time step and each link connects two nodes, then we get $\frac{dk_i}{k_i} = \frac{dt}{2t}$ with the solution

$$k_i(t) = \left(\frac{t}{t_i}\right)^{1/2} \tag{1.10}$$

where t_i is the time at which node i entered the system [8]. If we now look for the probability $P(k_i(t) < k)$ that the degree of node i is smaller than k, we see that this is equal to the probability that the node was added after time $t_i = \frac{t}{k^2}$. For the probability, we get

$$P(k_i(t) < k) = P\left(t_i > \frac{t}{k^2}\right) = 1 - \frac{t}{k^2(t + m_0)} \tag{1.11}$$

where for large t, we can neglect m_0. After differentiating with respect to k, we end up with the probability density function for the degree

$$\rho(k) \sim k^{-3}. \tag{1.12}$$

Hence, the network's degree distribution emerges as a power law with exponent $\gamma = 3$ from the preferential attachment. Other values for the exponent γ can be obtained if the preferential process is proportional $\Pi(k_i) \sim k_i^\alpha$ with an $\alpha \neq 1$ [28, 29]. Networks with a power law degree distribution are called *scale-free networks* [27]. Consequently, in these networks, there exist *hubs*, which are extremely well-connected nodes that typically play a central role in the network's structure. Many real-world networks, such as the world wide web, are scale-free networks characterized by the existence of hubs [30]. In this thesis, we discuss scale-free networks in which knowledge is exchanged.
If the attachment kernel of a complex network is known, it can be used to build a microscopic generative model and numerical simulations. From these, one can forecast the future state of the network. In publication [1] of this thesis, we demonstrate this on a scientific citation network.

1.4 Evolutionary processes

As part of this thesis, we study the evolution of scientific knowledge. While evolution originates from the field of biology and is best known there, the concept has been transferred successfully to other domains like economics [31–33], technology [34, 35], and science [36–40] the latter two being crucial in our context. When new scientific discoveries are published, they build on existing knowledge. Previous works are cited, as they are combined into something new. When a new publication is added to the body of scientific knowledge, the system grows and changes, as other people's research is impacted by it. In this sense, scientific progress is an evolutionary process.
A general way of formalizing evolutionary processes is by using complex evolutionary networks. In section 1.1, we stated that complex networks are co-evolving, meaning that the states of the elements that compose the nodes of the network are changing over time due to the interactions, and the interactions between the nodes themselves are changing due to the changes in the states. We indicated this co-evolution by adding a dependence on time to our interaction matrix $M_{ij}^\alpha(t)$. A more specific but still quite general way of stating this is through the following two equations

$$\begin{aligned} \frac{d}{dt}\sigma_i(t) &\sim F(M_{ij}^\alpha(t), \sigma_j(t)) \\ \frac{d}{dt}M_{ij}^\alpha(t) &\sim G(M_{ij}^\alpha(t), \sigma_j(t)) \end{aligned} \tag{1.13}$$

where M_{ij}^{α} is the multi-layer interaction matrix between nodes i and j, σ_j is the state of node j, and G and F are some functions that depend on the interactions and the states [8]. These two equations relate to a three-step algorithmic process that is described in [8]. In the first step, a new entity (e.g., species, product, or scientific discovery) is created and placed in a given environment. In the second step, which relates to the first equation, the new entity interacts with its new environment, causing it to either get destroyed or survive and proliferate. Here, the state of the new entity is influenced by the interaction with the existing entities. If the entity survives, there is the third step. The new entity is included in the environment in the third step, which relates to the second equation. Consequently, the environment changes and existing entities relax toward a new equilibrium. The existing environment is updated and presents a new boundary for newly arriving entities.

Built around this three-step algorithmic process, there are several models. Examples of well-known evolutionary models include the replicator equation [41, 42], the NK [43, 44] and NKCS [45] models, the Bak-Sneppen model [46] and the Jain -Krishna model [47]. One model that we want to mention in particular in the context of this thesis is the co-evolving combinatorial critical model (CCC) [31] because it highlights the characteristics of complex systems.

The CCC model has, as its name suggests, three main characteristics [8]. First, it is co-evolving, meaning that the space of entities and the space of interactions influence each other and evolve together. Second, new entities in the model emerge spontaneously through the combination of existing entities, hence it is combinatorial. And third, the model captures the critical phenomena and the emergence of power laws typically observed in evolutionary systems. In the CCC model, existing entities, σ_j and σ_k, can interact with each other through a random productive interaction tensor, M_{ijk}^{+} at time, t, to create a new entity, σ_i, in the next timestep, $t+1$. This can be formally expressed as

$$\sigma_i(t+1) = \sum_{j,k} M_{ijk}^{+} \sigma_j(t) \sigma_k(t) \tag{1.14}$$

Analogously, elements can interact and destroy an existing element through a random destructive interaction tensor, M_{ijk}^{-}. In addition to that, any element can be spontaneously created or destroyed with a small probability, p. Together, these simple rules lead to metastable periods, where the mix of entities stays relatively constant, which are then punctuated by phases of disruptive change. Power laws emerge for the distribution functions of the size of the change in diversity, the lifespan of the metastable phases, and the lifespan of the phases of disruptive change [8]. It is also worth mentioning that the CCC model is closely related to spin models and can be formulated as such in an infinitely large tensor space where a mean-field solution can be obtained [48].

An exciting aspect of the CCC model is that it is open-ended. While most evolutionary models deal with the production and destruction of a fixed set of entities, the CCC model allows for creating new entities in each time step. Consequently, the phase space of the system evolves. The concept of dynamically evolving phase spaces is not unique to the CCC model. It was first introduced by Kauffman [49], who coined the term *adjacent possible*. The adjacent possible is the part of a (potentially infinite) phase space that can be reached in the next time step. This concept is quite helpful in the context of biological, technological, or scientific evolution, where a limited number of potential new species, technologies, or ideas exist for the immediate next time step. Still, an infinite number of these entities might be possible when time approaches infinity.

In science or technology, we are interested in exploring the adjacent possible, as it tells us which new ideas or technologies could be created at the next time step. In recent works, a particularly intuitive model for this exploration process was created using Polya urns [50–52]. These models use a generalized version of Polya urns to model entities as colored balls in an urn. The adjacent possible is, at each time, given by the set of all colors of balls present in the urn. It can be expanded by adding balls of new colors to the urn. Previously known balls can be reinforced when drawn by adding copies of the same color to the urn. This simple model can reproduce both the rate at which novelties occur as well as the distribution of attention each entity receives in a way that matches records from multiple real-world systems.

Initially, for the subjects of this thesis, we intended to rely on the CCC model and generalized Polya urns for modeling information flow processes. However, due to unexpected promising opportunities arising in the domain of scientific citation networks, we shifted our focus toward diffusion-based models. There, we were presented with the opportunity to pioneer quantitative work in exaptation, a particular aspect of evolution that received very little attention despite its importance. In the following, we shall briefly define exaptation and relate it to the concepts discussed here.
Exaptation, as initially defined in the context of biology, is the name for features that now enhance fitness but were not built by natural selection for their current role [53]. A typical example is feathers on the wings of birds used for flying. Feathers are said to have evolved originally for thermal insulation. However, they proved to provide a survival advantage for the feathered species by enabling it to jump further, glide, or later on fly due to the increased air resistance provided and their low weight [53]. In the context of scientific evolution, an exaptation might be a publication that was originally published in one field but re-discovered in a different field after and hence used for a different purpose than originally intended. Significant exaptations have the potential to substantially contribute to a field, build bridges between fields, and trigger cascades of further discoveries, similar to the phases of rapid change described by the CCC model. However, in contrast to the CCC model, exaptation is not combinatorial. Instead, an existing entity spontaneously changes by itself. In that sense, it can be related to the small probability p for an entity to get created or destroyed spontaneously that we discussed in the context of the CCC model above. In chapter 3 of this thesis, we present a book chapter that goes deeper into the definition of exaptation and its relevance in the context of scientific evolution and presents an entropy-based method of identifying possible candidates for exaptation at an early stage.

1.5 Complex social networks

In this last section of the chapter on information flow networks, we want to discuss social networks. By *social network*, we are referring to networks of interactions that happen between people in our society [54]. These networks might include many different kinds of interactions, such as the spread of diseases, the spread of opinions, innovation, or knowledge transfers.
The idea of applying concepts from physics to gain insights into societal problems is not new. The earliest mention of this line of thought might have been when Thomas Hobbes visited Galileo Galilei in Florence, and Hobbes began to frame the idea of explaining the "physical phenomena" of society through the laws of motion [55]. A few centuries later, in the 19th century, August Compte coined the term *social physics*, defining it as a science of social phenomena, considered in the same light as astronomical or physical phenomena [56]. At the heart of social physics is an analogy that Ludwig Boltzmann, inspired by the works of Henry Thomas Buckle, drew:

> *The molecules are like so many individuals, having the most various states of motion, and the properties of gases only remain unaltered because the number of these molecules which on the average have a given state of motion is constant* [57]

Hence, already at the birth of statistical mechanics, its close relation to the social domain was remarked. Since then, social physics has evolved as a field of study, where physicists contribute concepts and methods to understand the fundamentals of social systems [58, 59].
While it has been argued that describing human behavior with equations or algorithms inevitably leads to over-simplifications such as the concept of the *Homo economicus*, the author of [58] argues that similar to the Ising spin model, in many cases, the details of the model just don't matter. While specific individual actions are and remain unpredictable, the collective behavior is often quite robust, allowing for a statistical description of social interactions [58, 60]. As long as one is aware of the assumptions and limitations of any model, the model can prove an invaluable tool for understanding and improving the underlying social system.

In the field of complex systems, social dynamics are modeled as complex networks that are composed of people and their social interactions [61]. These networks have been studied by a number of physicists in the past, and the direct application of statistical physics-based methods has led to advancements in many social areas [59]. Important examples include urban environments and transportation systems [62–64], innovation dynamics and the modeling of knowledge flows [36, 50, 51, 65–67], the analysis of diseases and the modeling of disease spreading, that became even more relevant in lights of recent events [68–72], and the modeling of opinion dynamics that gives insights into collective behavior related to decision making such as in the context of elections [73–76].
One concept related to social networks, that is especially relevant in our context is the concept of social triads. Social triads are social groups of three people and their interactions with each other [77]. In undirected networks, each pair of people in a triad can be either linked or not linked, resulting in four different possible types of triads (zero to three pairs being linked). In directed networks, there are 16 unique types of triads, because each pair of people can be linked either in one direction, in the opposite direction, in both directions, or not at all [77]. Social triads can be used to understand in which constellations people in a social system are interacting with each other. By counting the frequency of each type of triad and comparing it to the frequencies of a suitably randomized network, one might find types of triads that are significantly over- or underrepresented. One pattern that is frequently observed is that two persons that interact with the same third person, tend to also interact with each other. To give an example, people who have a friend in common, typically have an increased probability of becoming friends over randomly picked pairs of people [78]. This pattern is called triadic closure, and it has been predicted in [79]. It has been suggested that triadic closure is in part responsible also for what is called homophily [80]. Homophily reflects the idea that social ties form predominantly between people with similar features. Homophily in social networks has been found with regard to multiple attributes, such as obesity [81], performance in school [82], and musical preference [83].

With the concepts introduced in this section, we are prepared to deal with information flow processes that happen on underlying social networks. In the last part of this section, we want to give a brief overview of the social networks that we are studying in this thesis. In chapter 5 we present our publication on information diffusion in the domain of music. There, we are dealing with two social networks that we built based on a unique dataset we collected from *last.fm*, the friendship network and the influencer network. The friendship network is an undirected, unweighted network where nodes are user accounts on *last.fm*, and links indicate a bidirectional friendship between two users. The influencer network, I_{ij}, on the other hand, is a directed and weighted network where nodes represent user accounts and a link, I_{ij}, indicates the tendency of user j to influence user i into listening to the same music as user j. A friendship between users i and j is a requirement for influence, hence the influencer network is a sub-network of the friendship network. In chapter 5, we will show how the topology of these two networks relates to a diffusion of information that enables the cascade-like spreading of new songs. We will further show, how this information can be leveraged to improve state-of-the-art machine-learning models for hit-song prediction.
In chapter 6 we present a system to record information diffusion in a social context. The system monitors the exchange of knowledge in an organization on different channels, such as through group meetings, talks, discussions, exchanges on chat applications such as *Slack*, and through emails or collaborative writing tools. The recorded information can then be mapped onto a directed and weighted multilayer network, M_{ij}^{α}. In this network, nodes are not necessarily restricted to being people, but might also represent *generalized events*, such as meetings, lectures, chat groups, publications, or collaborative projects. Interactions, M_{ij}^{α}, indicate the contribution of knowledge from node j to node i, where α corresponds to the channel on which information was transmitted (e.g., in person, via email, or via chat). The time-resolved interactions can be annotated with details about the exchange to provide more context. What results from this, is a detailed data set that might be analyzed to gain deep insights into the way knowledge diffuses in organizations. In chapter 6, we present the patent application that describes the system used to record information diffusion in organizations. In the appendix, we present three use cases where the application has been deployed, and data has been

collected and analyzed as part of social experiments.

With these concepts and methods, we are now well-equipped to deal with information flow processes in networks. In the next chapter, we will introduce the topics of this thesis, explain their background, and place them in the context of prior work.

Chapter 2

Network information diffusion in the modern age

Information plays a central role in our lives. In every human interaction, whenever we talk to someone, every time we read something or listen to a recording, information is exchanged. If a human receiver understands information, it is turned into knowledge. Over the last few thousand years, we, as humans, have found increasingly sophisticated methods for preserving and transmitting our knowledge. It can be argued that this knowledge has given rise to our society and distinguishes us from our ancestors.

Knowledge's high importance is reflected in how our society has been centered more and more around it. The extent to which knowledge impacts our society has led to the term *knowledge society* [84], coined by economist Peter Drucker. In the early 90s, Drucker claimed that knowledge was assuming more importance than ever before in history and has surpassed both capital and labor in importance to the wealth of nations [84]. The development of new technologies has become fundamental to our economy and most of its actors [6]. Global research and development spending has reached a record high of 1.7 trillion USD [85], and its output drives economic success [6]. The trend of knowledge becoming increasingly important will likely continue. Society is transforming, with knowledge at the core of that transformation [86].

At the same time, the global transfer of knowledge is getting more and more complex. There is a rapid growth of the global research community [87], and increasing global exchange of knowledge through new channels such as online social media like *Twitter*, *Facebook*, and *LinkedIn*, chat and voice chat services like *Skype*, *Zoom*, *Slack* and countless websites and web services. Globally, more knowledge is produced, and faster [6, 88]. In this context, concerns about information overload are voiced across many domains [89–91]. Their message is that there is so much knowledge being produced we do not know anymore what to focus our attention on. More scientific papers are being produced than one could read in a lifetime, and scientific fields have become so numerous and profound that no one can hope to keep up to date with even a tiny fraction of them. Technologies that enable products like modern computers or cars have become so complicated that no single person can understand them in all their details. This issue of information overload raises the question if the current system of knowledge transfer is sustainable in the long run.

Due to the crucial role of knowledge in modern society, and the issues arising connected to its increasing complexity, there is an urgent need for a deeper understanding of knowledge flows. Understanding knowledge flows can help us pinpoint problems, identify elements or actors critical in facilitating knowledge flow and make predictions that can aid our planning. A deep understanding is elemental for the improvement or redesign of existing systems in both scientific publishing as well as in organizations and companies to make sure they have the robustness and flexibility that is needed in times of transformation. Information flow needs to be optimized so that the right knowledge is delivered to those who need it. Otherwise, information overload is bound to happen.

If we want to understand knowledge flows, we need to consider their interdisciplinary nature and broad scope. The transfer of knowledge happens on many different channels, such as scientific publications, conferences, or books on long-time scales, as well as discussions, emails, or chat applications on short-time scales. Knowledge is transmitted on different scales, ranging from short distances in local exchanges within a research unit to global exchanges between regions or countries. It is also transmitted across multiple domains, where it can come in the form of a new technology, a new idea, a business practice, or just any information that the receiver understands.
We need data when we want to study knowledge flows in a quantitative, predictive, and experimentally testable way. Through the rise of online tools and websites that collect user data and online databases like *Scopus*, the *Web of Science*, and *Dimensions*, more and more data has become available. This data provides a ground for the quantitative study of information flow and allows us to understand and forecast knowledge flows.
In this thesis, we want to leverage big data sets that have recently become available or that we collected to address the apparent need for a deeper understanding of knowledge flows. The big questions motivated by this need and that we want to address here are: Where is knowledge created, and how is it transferred? What will the future scientific publishing landscape look like? How do geographical regions accumulate and exchange knowledge? And how do the underlying social networks influence knowledge transfer in a society?
To answer these questions, we have to consider the interdisciplinary nature of information flows and the fact that they stretch over extended time and length scales across multiple channels. Hence our approach, while strongly founded in Physics, is interdisciplinary and touches the fields of scientometrics and computer science.

In this thesis, we study information diffusion in networks in five contributions:

- In a *New Journal of Physics* publication, we study citation networks that represent knowledge diffusion on long timescales through the medium of scientific publications.
- In a book chapter, we show how the information embedded in citation networks can be used to uncover a rare type of publication called *exaptation*.
- In a *Chaos, Solitons & Fractals* publication, we study how knowledge diffusion through the mobility of scientists leads to the scale-free growth of regional scientific capacities.
- In an *Arxiv* pre-print, we study how cascade-like information spreading in the domain of music is enabled by the topology of the underlying social network and how this finding can be used to improve song popularity predictions.
- In a patent that was accepted in Japan and is pending in the USA, we present a novel system for recording information flows on short timescales that can be used to understand and optimize knowledge diffusion in organizations.

In the following sections, we shall introduce the topics of these five contributions and discuss the background and prior work.

2.1 Scientific citation networks

In the first section, we introduce the traditional way of transferring scientific knowledge that we are familiar with, scientific publishing. Since the inception of the first scientific journals *Philosophical Transactions of the Royal Society* and *Journal des sçavans* in 1665 [92], scientific progress has been captured in papers and systematically made available in journals, where researchers can access the knowledge and keep up to date on the latest developments. During these three-and-a-half centuries of scientific publishing that have passed since the first inception of scientific journals, there has been

an enormous, exponential growth of scientific output [93]. The total number of journals has increased from 2 to more than 30.000, with more than seven million publications published per year in recent years [93].
With the invention and global adoption of the internet, the possibility of tracking information about new publications online has emerged. Large-scale datasets such as *Scopus*, *Dimensions*, and the *Web of Science* have become available and allow for detailed study of the body of scientific literature. A common way of dealing with the large size of available publication records is to abstract away the details and to only look at the metadata of papers and their references. What is left is a set of elements representing publications linked to each other via their references. This is called a *citation network*. A citation network, M_{ij}, encodes the information about which paper j cites paper i. If paper j cites paper i, or in other words, paper i is in the reference list of paper j, then $M_{ij} = 1$, otherwise $M_{ij} = 0$. Due to each paper being associated with a publication date and citations always only pointing to the past, the citation network is a directed acyclic graph. Based on the citation network, two other citation-based networks can be computed. First, the co-citation network, $C_{ij} = MM^T$, encodes the information of how many papers have both paper i and j in their reference list [94]. For instance, $C_{ij} = 3$ means that three papers cite both, i and j. Second, the bibliographic coupling network, $B_{ij} = M^T M$, states how many references papers i and j share [95]. For instance, $B_{ij} = 2$ means that two papers are both in the reference list of paper i and paper j. If one also considers information about the authors of the publications, one can construct a co-authorship network. The co-authorship network, A_{ij}, states how many papers authors i and j have authored together [96]. For example, $A_{ij} = 2$ means that author i and author j have published two papers together.
Citation networks can be seen as a proxy for how established knowledge is combined into new knowledge when a new publication cites established ones. Based on the composition of references and the publication times of the individual works, one can study the dynamics of scientific publishing. Several influential works in the fields of Scientometrics and Science of Science have done this in the past. They have contributed valuable insights concerning, for instance, the increasing interdisciplinarity of natural sciences [67, 97, 98], the distribution of research funding [99], or the community structure of researchers [37, 100, 101].
In addition to these, there are multiple works concerned with the task of forecasting the success of individual papers [38, 102–104]. However, what is missing there, are predictions concerned with the system as a whole on longer timescales. These are needed to answer important questions such as: How will the citation landscape of the future look like? And which papers have the potential to become immortal and why?
In the first contribution of this thesis, [1], we closed this research gap. We study the citation network of the American Physical Society (APS) and contribute a microscopic generative model for how papers are cited. The model is based on two competing mechanisms, preferential attachment, and aging. Preferential attachment means that the more attention a particular paper has received in the past, the more likely it is to continue to receive attention. We explain this mechanism in more detail in chapter 1.3. Aging, on the other hand, means that the older a paper is, the less likely it is to be cited. In combination, we get a model that lets us investigate why most papers are only cited for a short time and then quickly forgotten. At the same time, some exceptional papers have managed to retain a high level of popularity over decades. It also allows us to forecast the citation landscape of the next decades. The full study is presented in chapter 3.

In our first contribution, we model the bulk of quickly forgotten papers and the exceptional ones that become famous and stay popular for a long time. Out of this large number of forgotten papers, some are rediscovered in a rare event called exaptation. We covered the concept of exaptation in chapter 1.4. There we defined exaptation in the context of scientific publications as a publication published initially in one field but re-discovered in a different field after.
Exaptations might happen for different reasons. In some cases, a field might not have any use for a particular discovery at one point in time but might develop a need for it later. It might also be that a catalyst is missing - the piece of technology that allows the original discovery to be transferred and

applied in a different field. On the other hand, it might happen that due to information overload, a field is simply not aware of all the published work in other fields and might miss a solution to one of its problems that might exist in a different field. In that sense, an exaptation is often a particular case of *sleeping beauty*, a publication that goes unnoticed for a long time, after which it receives a sudden increase of attention [105].
Exaptations are often relevant because they can act as bridges in connecting fields in an interdisciplinary effort. As such, they have a first-mover advantage that gives them the potential to trigger cascades of new discoveries [106]. Despite their relevance, there is very little quantitative work on exaptations in the context of science. To close this gap, our second contribution [2] is dedicated to these rare but impactful exaptation events. There, we show how citation networks can be used to identify exaptations at an early stage. The basic recipe for this is easy to follow. By using appropriate clustering algorithms, such as the Leiden algorithm [107], we can cluster the citation network and use these clusters to assign each paper in a body of literature to a scientific field. If we then track which fields are citing a particular paper over time, we might find a sudden change of fields in the case of an exaptation. In chapter 3, we explain how this basic idea can be leveraged to design an entropy-based indicator for exaptations.
Lastly, we want to mention that exaptations are not to be confused with *serendipities*. In serendipities, a researcher would look for the solution to one problem but unintentionally find the solution to another problem. A famous example in physics is the serendipitous discovery of cosmic background radiation [108]. While both in serendipities and exaptations, the output is unexpected, for serendipity, the relevance is known at the time of publishing. For that reason, they produce very different patterns in the citation network.

2.2 Scientific mobility

Supplementary to scientific publications, traveling is an even older and more established way of transferring knowledge. Already in ancient Greece, philosophers traveled to other countries to exchange knowledge and get inspired [109]. Nowadays, it is common for scientists to travel to different research institutes. We travel for a few days to attend conferences or give talks, a few weeks or months for research visits or exchanges, or a few years to move from one position to another. University contracts are often deliberately kept short to keep a constant influx of new knowledge and perspective.
Over the last decades, the mobility of scientists has slightly increased [110]. In absolute numbers, though, the number of traveling scientists has increased tremendously, primarily due to the strong growth in the number of people working as scientists [110]. Scientists that move to a different location transfer knowledge to that location and establish links between their origin and target regions, along which knowledge is exchanged and collaborations are formed [111, 112]. Hence, mobile scientists moving around the world are one of the drivers in the diffusion of knowledge.
In the context of this thesis, scientific mobility is a central mechanism for knowledge diffusion over long timespans (typically multiple years) and long distances (international or intercontinental). Not only is the knowledge and expertise of the moving scientist transferred to their target region, but connections to the origin region are often established, as the scientist is still in contact with their former institution [111]. These connections open up new channels through which knowledge can flow, effectively reshaping the network of knowledge flows for the involved actors by building bridges between clustered members of institutions.
The transfer of knowledge facilitated by the mobility of scientists is not only relevant in the context of science but also has a significant impact on the economy [113]. Traditionally, some locations have established themselves as the leaders in their field and are acting as magnets for attracting new talent, such as the *Silicon Valley* in the USA has established itself as the leading region in the field of semiconductors [114]. Despite its scientific and economic importance, the mechanisms behind scientific mobility are poorly understood. It is evident that the global distribution of scientific capacities is quite imbalanced, but the conditions under which a region can reach comparatively high levels of scientific

capacity are unknown. As one of the preconditions, the requirement for a *critical mass* of researchers has often been assumed [115, 116].
This lack of understanding opens up some essential questions regarding scientific mobility that we want to address in this thesis. How do regions attract scientists? How can a region establish itself as a scientific leader, and is a critical mass of researcher required for it? And how can other regions catch up on long-established leaders or even overthrow them? We address these questions in a publication that we present in chapter 4, where we build a microscopic generative model that lets us investigate the dynamics of scientific mobility. The model is based on publication metadata collected in the *Dimensions* database. There, we use the affiliations of authors of papers and the geographical location of these affiliations to track the movement of scientists and the scientific output of each region. We explain how regions grow their scientific capacities, show how newcomers can catch up on established leaders, and clarify the question of critical mass.

2.3 Knowledge diffusion in social networks

Knowledge can take on many forms and shapes. It can come in the form of scientific discoveries, where knowing a scientific discovery might translate into having read and understood the corresponding paper. Outside of the scientific context, it might mean that a person has experienced something. For instance, that person has read a new book, watched a new movie, or listened to a new song. In that context, the concept of novelties is relevant. Innovation differs from novelties because it is new to everyone, while novelties are new only to those who have never seen them before [50]. Hence, novelties are specific to a person. For example, a person's favorite song is not a novelty to them because they have listened to it many times, but it might be a novelty to you if you have never heard it before.
Novelties play an essential role in how we explore new knowledge. A new book might get us interested in a new topic and motivate us to explore more in that direction, and a new song might inspire us and trigger a search for more music of the same artist or genre [50]. When we experience something new of high quality, we might also want to share it with the people around us, for instance, by recommending them to read a particular book or listen to a particular song.
In the article we like to introduce in this section, we study knowledge flows related to music. Historically, the connection between music and knowledge is strong. Prior to literacy, knowledge was transmitted through songs and stories over millennia [117, 118]. In our context, knowledge related to songs is about novelties, i.e., knowing a new song. This does neither require nor exclude that the song's lyrics carry knowledge or that acoustic elements reflect the culture or traditions of the musician. Information diffusion in music is driven significantly by people recommending songs to their friends. Hence it strongly depends on the topology of the underlying social network. Before we detail the knowledge flows in music, we want to motivate the transition from scientific knowledge to music at this point in this thesis. At first glance, it might seem like songs and scientific publications do not have much in common. However, if we look at the microscopic mechanisms that drive the spread and adoption of both, we find striking parallels. In chapter 2.1, we introduced our model for scientific publications, where the probability of a paper getting cited is influenced by preferential attachment and aging. In music, we find a similar mechanism, and the reasons for this are pretty intuitive. If we think of a particular song, the more popular it is, the more likely it will be played on the radio or a TV show, found in a recommended playlist, or introduced to someone by their friends, which in turn increases its popularity. This pattern is a clear indicator of preferential attachment. On the other hand, most songs lose their appeal after some time, and listeners want to listen to something new. Hit songs on the radio rarely stay in the charts for long. Typically, they are replaced by new songs. Some old songs that we used to listen to many years ago might even feel awkward to us or others now. Hence there is a strong aging effect as well. However, some timeless pieces of music – evergreens – such as the works of *Elvis Presley* or *the Beatles*, stay famous for decades and are in concept related to scientific milestones. In summary, we could reasonably expect to find preferential attachment, aging, and exceptional works in music, in the same way we did in scientific publishing.

These apparent parallels between publications and songs are reflected in a striking similarity on the microscopic level that encouraged us to extend the scope of our work and study information flow processes in the domain of music, where we expected a high potential for promising new insights. Studying knowledge transfer in music is highly relevant because music suffers from the same problem as scientific publishing – information overload.
The total music production has long surpassed our capacity to listen to everything produced. Every day, more than 130 days' worth of new tracks are uploaded to *Spotify* [119, 120], making it virtually impossible to listen to more than a small fraction of new songs. Due to this enormous production, there is fierce competition for attention, where most artists struggle to survive, while a few hit artists receive billions of listenings [121]. From the publishers' perspective, identifying the potential hit songs to focus their marketing attention on is essential to their business. For this reason, machine learning models for forecasting song popularity have important commercial applications and have received substantial interest over the last few years [122].
An aspect that has received little attention in these forecasting models in the past is the social interactions of music listeners. This is even though it has been shown that social networks substantially influence music listening behavior [123–126]. In the work we present in chapter 5, we want to put our focus on this social aspect and show how it can be leveraged to improve song popularity predictions. For this, we collected a unique data set from the *last.fm* social music listening platform. On *last.fm*, users can listen to music and recommend songs to their friends on the platform. From this dataset, we can reconstruct the two networks that we introduced in chapter 1.5. In the friendship network, nodes are users, and undirected links represent friendships between users. In the influencer network, nodes are users, and directed links represent a tendency of the origin user to influence the target user into listening to the same songs as the origin user. We use these two networks to show how knowledge about new songs flows through social networks in cascades of recommendations between friends. On the modeling side, we show how these cascades contribute to the preferential attachment mechanism, which determines how likely a song is listened to. On the application level, we leverage our insights into social influence to build an improved machine-learning model for predicting song popularity with potential commercial applications.
The publication on this topic is presented in chapter 5. A part of this work that deals with the microscopic mechanisms found in scientific citation networks and music that did not make it into the final publication is presented in appendix A. The data we collected for this study is proprietary to *last.fm* and may not be published. However, we have published an open-source software tool that allows downloading the data through the official *last.fm* API under `https://github.com/NiklasRz/lastfm-data-downloader`.

2.4 Knowledge diffusion on short timescales

In this last section, we want to highlight the puzzle piece to understanding knowledge flows that has been missing in the context of this thesis: knowledge diffusion on short timescales. Most pieces of work that contain knowledge, be it scientific publications, songs, books, or websites, are neither made in isolation nor in a single day. Typically, much knowledge is exchanged beforehand, which goes into the making of the piece of work. This might include activities such as reading other works, discussing with colleagues, listening to talks, getting feedback, brainstorming together, and many more. These frequent exchanges of knowledge with our social contacts on very short timescales are fundamental in the creative process of creating new knowledge. At some level, both research institutes and companies are aware of that and are trying to create an environment that fosters this exchange of knowledge [127]. Nevertheless, knowledge diffusion in organizations (here, with organizations, we mean anything from research institutes to universities, companies, NGOs, or governmental units) is still often treated as a black box. Management or members of the organization might introduce new measures to the organization, such for example the introduction of weekly town hall meetings. As a result, an improvement in knowledge diffusion might be noticed sometime after. However, the mechanism behind the diffusion

is often not well understood, and immediate feedback is lacking. We argue that the main reason for this is the lack of fine-grained data. If we cannot observe when, between whom, and on what channels, and what specific piece of knowledge diffuses in an organization, then we cannot study and understand the underlying processes.
Initially, as part of this thesis, we intended to study software collaborations on the open-source software community *GitHub*. There, we hoped to get fine-grained and time-resolved insights into the knowledge flow in collaborations. However, even the fine-grained data available on *GitHub* only provided a partial picture as a lot of the communication in *GitHub* went through other channels, like chats, emails, apps, or other websites. While this lack of information prevented us from gathering and publishing insights in this domain, it inspired us to explore a different, more promising avenue. Instead of working with partial data, we decided to put our focus on what was missing in this picture: capturing daily knowledge flows in organizations on all channels.
We built a software system in collaboration with Sony Computer Science Labs Paris called *KOUZAN* to achieve this. The system is designed to record microscopic knowledge exchanges on short timescales across multiple channels. It works both as a mobile phone application that can be used to record knowledge flows in discussions, meetings, lectures, or talks and as an interface to automatically record knowledge exchanged on other applications such as chat apps like *Slack*, voice chat apps like *Skype*, or collaborative tools like online text editors. It is built as a flexible REST-API, an interface for other programs with standardized input and output. Any meeting, project, lecture, document, or other forms of knowledge exchange is mapped to a generalized event, where the contributions of each person are recorded as generalized interactions. In addition, the system can be connected to a blockchain to ensure the information can not be tampered with retrospectively. The application's output is what we called a *co-evolving multilayer network* in chapter 1.1. It represents dynamic generalized interactions on multiple channels between generalized agents.
The application adds the missing piece required to analyze knowledge flows in organizations. With it, one can map all knowledge flows and visualize them. On the scientific side, this allows us to study and understand where knowledge is created in organizations and how it is transferred. On the application side, it allows for an analysis of knowledge flows that might reveal bottlenecks, such as a lack of knowledge exchange between certain groups within the organization, or redundancies, such as multiple members of the organization working on the same issue. It also allows us to identify essential actors, such as people facilitating knowledge exchange between groups. Often, informal channels of knowledge exchange are invisible to the management and only discovered when they go missing, for example, because an employee that was central in facilitating knowledge exchange over informal channels is retiring. In addition to this, the system also enables us to do experiments. To give an example, an organization might decide to start hosting weekly talks to improve knowledge diffusion. In this context, the system can be used to analyze how knowledge flows change due to the introduction of the weekly talks. A few impressions of the user interface of the *KOUZAN* application are shown in figure 2.1.
In chapter 6, we present the patent filed by SONY in relation to that system. During the work on this thesis at the Complexity Science Hub, the system was deployed in four different use cases, at the Complexity Science Hub Vienna, at Sony Computer Science Labs Paris, at an exhibition *Incertezza* at the *Palazzo delle Esposizioni* in Rome, and the *Maker Faire 2022* in Rome. The author of this thesis has co-supervised a master's thesis about optimizing knowledge flows in exhibitions that is based on the data collected by the system. In the main body of this thesis, we present the patent as the main contribution, while the use cases are presented in appendix A.

With this, we now have an overview of all the topics covered in this thesis. In the following part, the individual contributions are presented.

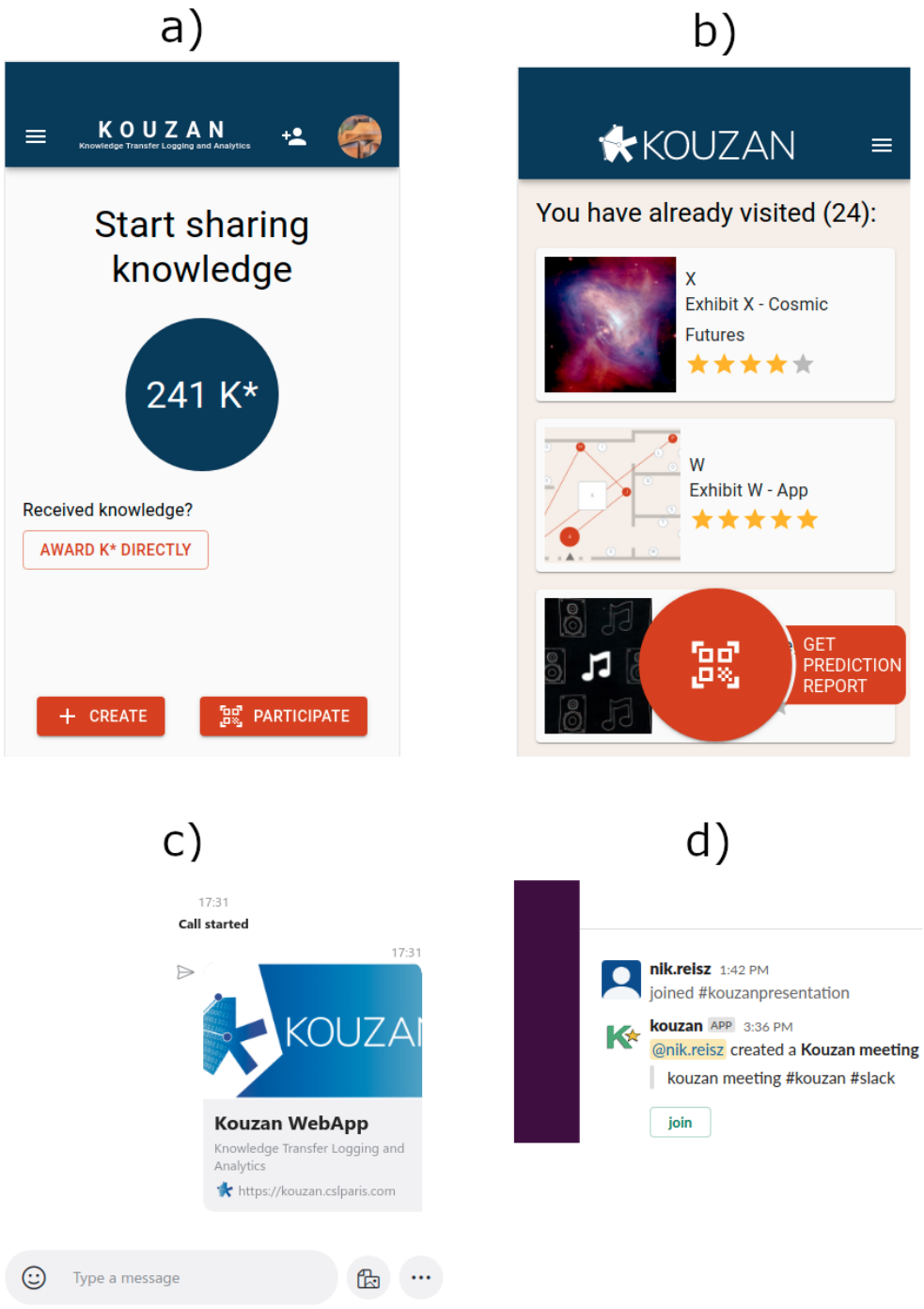


Figure 2.1: The *KOUZAN* application for recording knowledge flows. a) the home screen of the *KOUZAN* mobile application, where users can record knowledge exchanges in personal meetings, discussions, talks, or lectures. b) an example home screen of the *KOUZAN EXPO* application, where users can track knowledge flows during exhibitions and interact with and rate exhibits. c) link to a *KOUZAN* meeting on the *Skype* voice chatting platform. d) interaction with the *KOUZAN* API for the *Slack* collaborative work environment.

Part II

Publications

Chapter 3

Knowledge diffusion in science

Loss of sustainability in scientific work

Niklas Reisz, Vito D. P. Servedio, Vittorio Loreto, William Schueller, Márcia R. Ferreira and Stefan Thurner

Published in: New Journal of Physics, Volume 24, May 2022

Quantifying exaptation in scientific evolution

Márcia R. Ferreira, Niklas Reisz, William Schueller, Vito D. P. Servedio, Stefan Thurner and Vittorio Loreto

Published in: La Porta, C., Zapperi, S., Pilotti, L. (eds) Understanding Innovation Through Exaptation. The Frontiers Collection. Springer, Cham.

Preprint version: arXiv:2002.08144

IOP Publishing New J. Phys. 24 (2022) 053041 https://doi.org/10.1088/1367-2630/ac6ca1

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
15 February 2022

REVISED
18 April 2022

ACCEPTED FOR PUBLICATION
4 May 2022

PUBLISHED
20 May 2022

PAPER

Loss of sustainability in scientific work

Niklas Reisz[1], Vito D P Servedio[1], Vittorio Loreto[1,2,3], William Schueller[1], Márcia R Ferreira[1] and Stefan Thurner[1,4,5,*]

[1] Complexity Science Hub Vienna, Josefstädter Strasse 39, A-1080 Vienna, Austria
[2] Sony Computer Science Lab, 6, Rue Amyot, 75005, Paris, France
[3] Sapienza Unversity of Rome, Physics Dept, Piazzale A. Moro 2, 00185, Rome, Italy
[4] Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Austria
[5] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 85701, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: stefan.thurner@meduniwien.ac.at

Keywords: scaling laws, science of science, citation networks, preferential attachment, complex networks

Supplementary material for this article is available online

Abstract

For decades the number of scientific publications has been rapidly increasing, effectively out-dating knowledge at a tremendous rate. Only few scientific milestones remain relevant and continuously attract citations. Here we quantify how long scientific work remains being utilized, how long it takes before today's work is forgotten, and how milestone papers differ from those forgotten. To answer these questions, we study the complete temporal citation network of all American Physical Society journals. We quantify the probability of attracting citations for individual publications based on age and the number of citations they have received in the past. We capture both aspects, the forgetting and the tendency to cite already popular works, in a microscopic generative model for the dynamics of scientific citation networks. We find that the probability of citing a specific paper declines with age as a power law with an exponent of $\alpha \sim -1.4$. Whenever a paper in its early years can be characterized by a scaling exponent above a critical value, α_c, the paper is likely to become 'ever-lasting'. We validate the model with out-of-sample predictions, with an accuracy of up to 90% (area under the curve ~ 0.9). The model also allows us to estimate an expected citation landscape of the future, predicting that 95% of papers cited in 2050 have yet to be published. The exponential growth of articles, combined with a power-law type of forgetting and papers receiving fewer and fewer citations on average, suggests a worrying tendency toward information overload and raises concerns about scientific publishing's long-term sustainability.

1. Introduction

When asked by a student, 'Dr Einstein, are not these the same questions as last year's (physics) final exam?', he replied: 'yes; but this year the answers are different'. To Einstein, it was very clear that his field was evolving at a rapid pace and that new scientific discoveries outdated historical knowledge. This observation might be even more relevant today. Since Einstein's days, the global scientific output, measured in the number of publications per year, has grown exponentially. The number of articles published every year in all of physics increased a thousandfold from 200 papers in 1900 to about 200 000 in 2010 [1, 2]. The total scientific output is about ten times higher than in physics [3].

Scientific work usually builds on pre-existing knowledge, which is acknowledged through citations in publications. Only a small fraction of this knowledge is actively accessed and actually cited [4]. We find that less than half of all papers published in American Physics Society (APS) journals received more than one citation during the last ten years. Most works become outdated quickly, are never looked at, are forgotten and never mentioned again. This is especially true for the increasing number of non-reproducible papers

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

that do not contribute knowledge sustainably [5, 6]. On the other hand, there are scientific milestone papers—typically exceptional works—that continue to attract citations, even after decades [7]. What emerges from the process of knowledge creation, selection, and forgetting is a 'living' or active body of knowledge that is actively used—ideally—for pushing the scientific Frontier forward. The active knowledge is a tiny subset of all the existing knowledge stored in papers. Its extent can be estimated using citation data.

Large-scale publication and citation data has become increasingly available in recent years. It opens up the possibility to reconstruct the history of science (production and reception) on the granular basis of individual publications. In particular, it becomes possible to trace citations in so-called citation networks.

There are different types of citation networks. In its simplest form, a citation network, M_{ij}, captures which paper j cites paper i. If paper j cites paper i then $M_{ij} = 1$, otherwise $M_{ij} = 0$. These networks should not be confused with other popular citation based networks such as the co-citation network or the bibliographic coupling network. The co-citation network, $C_{ij} = MM^{\mathrm{T}}$, contains the information of how often papers i and j appear *together* in the reference list of another paper [8]. $C_{ij} = 2$ means that there exist two papers that cite both, i and j. The bibliographic coupling network, $B_{ij} = M^{\mathrm{T}}M$, on the other hand, states how many references papers i and j have in common [9]. Every paper, i, is uniquely associated with a publication date, t_i, which makes it possible to bring in the time component of citations.

There are immediate lessons that have been learned from citation networks that are important in our context. Early contributions date back to the 1960s [10] where citation networks were pioneered by investigating links between scientific publications. Along with other insights, most strikingly it was noted, that the percentage of papers receiving n citations in any year approximately scales as a power law in citations n. It follows that most papers receive a small number of citations, while a few papers in the 'fat tail' of the distribution receive many. This observation has been confirmed [11, 12]. In follow-up work, an explanation was offered in terms of 'cumulative advantage' [13], stating that the probability for a paper to get cited is proportional to the number of citations it accumulated up to that point in time. This process is also called *preferential attachment* [14]. It has been studied in depth and confirmed by a growing number of authors, both in a theoretical context and in the context of citation networks [15–21].

While preferential attachment is a plausible mechanism for understanding the structure of citation networks, recent studies suggest that preferential attachment alone does not result in accurate predictions of several features of citation networks [22, 23]. Other factors need to be taken into account. Timing of publication has been found to be a contributing factor and a first-mover advantage was proposed along with preferential attachment [24]. If a work is among the first in an emerging field, then there is not much competition, and it may receive more citations. It was noted that papers that receive citations unusually quickly tend to continue to receive more citations than usual, even though they might not be the most cited papers at the time [25].

Time not only plays a role in the success of a paper, but also in its 'decline'. In this context, the notion of 'half-life of literature' has been introduced some time ago [26]. It is defined as the time during which the second half of all papers that are still being cited were published. This time is rather short, because most papers cite the recent past [26]. This fact was also observed in [7, 10]. In contrast to the half-life of nuclei, the 'half-life of literature' is not a constant, but depends on the overall growth of the citation network. For that reason, citation networks must always be considered against the exponential growth of the number of publications. In physics, this growth is characterized by a doubling time of about 11.8 years [2]. In [2, 27] it is noted that the number of publications per author (productivity) has practically not increased during the last century while the average number of authors per paper has increased [2]. The observed growth in the number of publications can therefore be largely attributed to the growth in the overall number of scientists and research funding. From the existence of a half-life of literature and the growth of the network follows, immediately, that the distribution of the age of citations (time between the publication of the citing and the cited paper) is highly skewed towards more recent papers. This general trend also persists if one accounts for the exponential growth of the number of articles published each year [7]. For this reason it is apparent, that there is a strong preference of papers to cite more recent works, an effect that is often referred to as *aging* [28, 29].

Preferential attachment in combination with aging are the two mechanisms that form the basis for a number of mathematical models that reasonably explain citation networks and their dynamics. In [30], an evolving network model with preferential attachment and a power law aging of nodes is considered. It demonstrates a strong first mover advantage. Another model with preferential attachment and power-law aging [31] shows that if aging is included in the model, the associated degree distribution of the resulting network deviates from a strict power law. They confirm their findings on a Hollywood actor network. In [22] an author-paper network is simulated, where authors can read, write, and cite papers, and search for references in a recursive manner. Aging is included by fitting a Weibull distribution to the time-dependent attachment probability of papers. This model explains the deviations from strict power laws in the degree

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

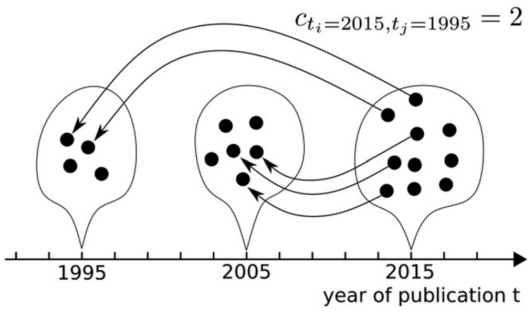


Figure 1. Number of citations, c_{t_i,t_j}, originating from papers published in year t_i and citing papers published in year t_j.

distribution and is supported by empirical data from articles of the Proceedings of the National Academy of Sciences (PNAS). A stochastic model of citation dynamics that builds on [22] combines preferential attachment and aging with a recursive search and copy mechanism [32]. The main result is a non-linearity in the network growth that allows for individual papers to achieve diverging citation trajectories. The rate at which both, patents and scientific articles, acquire citations is studied in [33, 34]. There, this rate is modeled with power law preferential attachment and exponential aging. Strikingly, in this model, the aging component is completely independent of the preferential attachment. Yet another model for patent citation networks assumes the probability to cite a patent to follow linear preferential attachment, as well as power law forgetting with an exponential cutoff [35]. It presents a method to determine this probability from empirical data and shows that the scaling behavior can be explained by the combination of preferential attachment and aging. In all the mentioned models, preferential attachment and aging have an impact on the fate of a paper. Some models allow for short-term performance predictions (citations) [25, 32, 34]; some also do predictions on longer horizons [28] of individual papers.

What has been studied in much less detail is the dynamics of scientific work being forgotten. For how long is scientific work relevant? How long will it take before today's work is forgotten? Which works have the potential to become milestones, and under what conditions does this happen? Can these conditions be identified by seeing papers not in isolation, but as a part of the citation network? What has been missing in this context so far is a model that focuses on the forgetting aspect in the context of a growing citation network. We want to add that missing piece to understand the dynamics of forgetting in order to better estimate the future state of the system as a whole. From that, we can then understand how certain papers stay relevant for decades, while most papers are forgotten.

2. Method

To understand the forgetting process more quantitatively, we introduce a variable to represent the total number of citations originating from papers published in year i that cite papers published in year j,

$$c_{t_i t_j} \equiv \sum_{k,l} \begin{cases} M_{kl}, & \text{if publication year of } k = t_j \\ & \text{and publication year of } l = t_i, \\ 0 & \text{else} \end{cases} \tag{1}$$

This is shown in figure 1, where each dot represents a paper published in a particular year and each arrow represents a citation from one paper to another.

In figure 2(a) the values of $c_{t_i=2015,t_j}$ (citations per year from papers published in 2015) are shown (blue) for all American Physical Society journals that include about 600 000 papers published between 1893 and 2015 and around 5000 000 citations. Next, we look at a hypothetical situation, where all papers have the same probability of being cited, independent of their age. We call the corresponding variable $\bar{c}_{t_i t_j}$ (red dashed line). It is given by $\bar{c}_{t_i t_j} = \frac{N_{t_j}}{N} \sum_{t_j} c_{t_i t_j}$, where N_{t_j} is the number of papers published in year t_j, N is the total number of papers published in all years $N = \sum_{t_j} N_{t_j}$ and the sum represents the total number of references of papers published in year t_i. We can see that $\bar{c}_{t_i t_j}$ is directly proportional to the number of papers published in every year and can be used as a normalization factor, to account for the growing number of publications. By dividing $c_{t_i=2015,t_j}/\bar{c}_{t_i=2015,t_j}$, we get a measure that is independent of the number

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

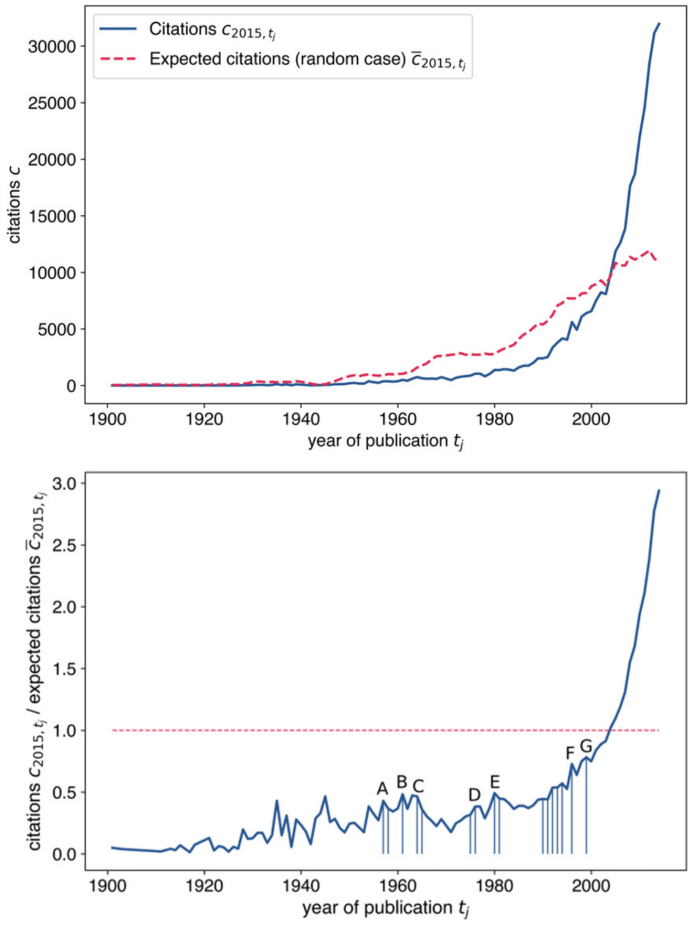


Figure 2. (a) Number of citations per year from papers published in 2015 (blue solid line) and expected values of the random case, where citations are assigned to publications randomly with equal probability (red dashed line). The red line is proportional to the number of papers published each year. (b) 'Forgetting curve' or citation preference of papers published in 2015. Values >1 indicate a relative preference with respect to equal probability citations, values <1 indicate an under-representation. Vertical lines mark the publishing years of milestone scientific papers such as: (A) 'theory of superconductivity' (Bardeen, Cooper, Schrieffer); (B) 'effects of configuration interaction on intensities and phase shifts' (Fano); (C) 'inhomogeneous electron gas' (Hohenberg, Kohn); (D) 'special points for Brillouin-zone integrations' (Monkhorst, Pack); (E) 'ground state of the electron gas by a stochastic method' (Ceperley, Alder); (F) 'generalized gradient approximation made simple' (Perdew, Burke, Ernzerhof) & 'efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set' (Kresse, Furthmüller); (G) 'from ultrasoft pseudopotentials to the projector augmented-wave method' (Kresse, Joubert).

of papers published each year. We call this resulting curve the 'forgetting curve' of the APS, shown in figure 2(b). Note a general trend of a strong preference towards citing papers that are not older than ten years. In a randomized scenario, where new publications would cite other publications with equal probability, the forgetting curve would fluctuate around a constant value of 1. For simulation results on $\bar{c}_{t_i t_j}$, see SI 1 (https://stacks.iop.org/NJP/24/053041/mmedia). Note that spikes in the forgetting curve coincide with the publishing years of milestone papers. These heavily cited papers stand out from the bulk of papers and cause the corresponding year to be over-represented. We find that in years, where a milestone paper was published, these papers, while making up only 0.03% of the papers published, attract on average 14% of all citations to papers published in that year. For more details, see SI 2.

The forgetting curve can be used to infer the age structure of the literature new knowledge is built on. We explain the curve with the help of a simple model of citation dynamics that takes into account in different ways the bulk of publications and the exceptional works.

To model the dynamical process of citing literature, we devise a simple generative model. We consider the case where a new paper i is published at the time t_i and ask how likely will the paper i cite the paper j?

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

To answer it, we measure the function that modulates the probability of attracting citations for individual publications. This function is based on the age of the referenced paper j, $t = t_i - t_j$, and the number of citations the paper j received in the past, up to time t_i. We call this function the *attachment kernel*, $f(k, t)$. Assuming that the underlying mechanisms of how we choose our references do not change over time, the dependence of t_i and t_j of the attachment kernel can be condensed into one variable t. To measure the attachment kernel, we adapt a method used in [19] and apply it to the 120 years of empirical citation data of the APS; see SI 3. The resulting kernel can be approximated well by a linear function in k and a power law decay in time,

$$f(k, t) = (k + 1)(t + 1)^{-\alpha}. \tag{2}$$

The exponent α cannot be completely decoupled from the number of citations k, as it shows a slight dependency $\alpha = \alpha(k)$. However, we show that for the bulk of papers, α can be well approximated as a constant from the empirical data by its mean value over all observations. For the mean, we find a value of $\overline{\alpha} = 1.42$. We confirm the linear preferential attachment in k found in other works [21]. Here, the offset by 1 represents the finite chance of a publication currently without citations to get cited for the first time. The temporal term represents forgetting.

To this point, the model covers the bulk of average papers. However, as noted before, there are milestone works that are clearly exceptional. These works are causing visible spikes in the forgetting curve figure 2(b) and thus follow a different attachment kernel. If we assume, that the underlying preferential attachment mechanism is the same also for milestones, we can determine their attachment kernel by measuring the value of the exponent α_m of any individual milestone paper m. For our analysis, we define milestones as the top 30 most cited papers in the APS, excluding review papers. For each of these papers, we measure the exponent α_m based on the waiting times between events, where these papers receive citations. For details, see SI 4. The attachment kernel for milestones has the form:

$$f(k, t, \alpha_m) = (k + 1)(t + 1)^{-\alpha_m}, \tag{3}$$

where α_m can be different for every individual milestone, m. With these ingredients, we can now define the generative model as follows. An exponentially increasing number of papers is published over time. Each of these papers, i, cites multiple other papers j that were published before. The probability of j to be cited by i is proportional to the attachment kernel $f(k, t)$, where k is the number of citations that paper j has at that point in time and t is the time difference in years between the publishing times of paper i and paper j. To normalize the probability, $f(k, t)$ is calculated for all average papers and $f(k, t, \alpha)$ for all milestone papers at each point in time. Here, we define milestones as the top 30 most cited papers in the APS, excluding review papers. Following this model, one can generate a simulated citation network.

3. Results

3.1. Null model validation

To test the model, we generate a *randomized* citation network that serves as a null model. To generate it, we take the publishing dates and out-degrees of papers from the APS, but omit the information about which paper cites which. The process of citing is replaced by a random process, where probabilities for individual papers to receive a citation are proportional to their respective attachment kernel, $f(k, t)$. We include the publishing dates and individual attachment kernels $f(k, t, \alpha_m)$ for milestones in this model. We then assign the references paper by paper, ordered in time, until we have reconstructed the full citation network. For details, see SI 5. On the reconstructed network, we measure the forgetting curve. In figure 3 the resulting forgetting curve for the null model (red dashed line) is compared to the empirical forgetting curve (blue solid line) that was introduced in figure 2(b).

The general features of the null-model are in good agreement with the empirical values, especially for the more recent decades. To assess how far the null model deviates from the empirical data, we calculate the mean absolute percentage error (MAPE) on the forgetting curves. This value is calculated by taking the absolute value of the relative deviation between null model and empirical data for each data point. The resulting values are then averaged to arrive at the MAPE. When comparing the null model to empirical data, the MAPE is below 6% for the last 30 years. It increases for the earlier years, up to values of 35% if the whole 120 years of the APS, including early years with low numbers of publications and high volatility, are taken into account. The low MAPE for recent years implies that the null model accurately reproduces the forgetting curve of the real APS. Note that also some spikes caused by milestone papers are present in the null model. The magnitude of these spikes is captured well for years after the 1990s. For earlier years, low publication numbers are resulting in higher volatility. Nevertheless, the impact of the largest milestones is reflected moderately well, with exception of the year 1935. In this year, a major milestone of physics was

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

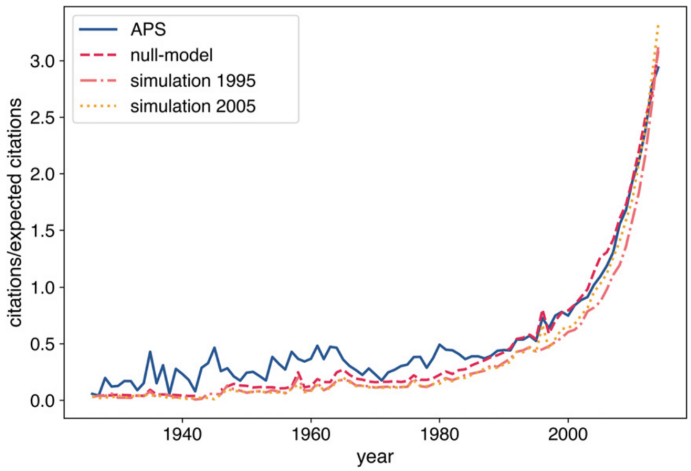


Figure 3. Forgetting curve for the APS citation network for the empirical data (blue), the null model based on attachment kernel $f(k, t)$ (red) and the out-of-sample predictions for the years 1995–2015 (orange) and 2005–2015 (yellow), based on empirical data from 1893 to 1995 and 1893 to 2005 respectively.

published by Einstein, Podolsky and Rosen, which remained almost unnoticed for decades, serving as a prime example of a 'sleeping beauty' [36]. This unusual citation pattern leads to an underestimation of citations in the null model. For more details, see SI 6. Overall, the similarity of the null model and the empirical data implies that the attachment kernel was measured accurately and captures the essential underlying citation dynamics.

3.2. Out-of-sample prediction

Next, we estimate the predictive power of the model by asking if the model can be used to simulate plausible future citation networks. To test the predictive power, we split our empirical data in two parts, separated by age, and try to predict the second part using data only from the first part. The quality of the prediction can then be assessed by comparing the prediction results with the empirical data. The first part consists of all papers published until 1995, the second part that we predict are all papers published between 1995 and 2015. We assume (realistically) that the number of references per paper stays more or less constant and that the number of publications published per year grows exponentially with a rate $\gamma = 0.0587$ corresponding to a doubling time of 11.8 years as determined in [2]. For our prediction, we use the attachment kernel that we measured on the full APS citation network using data until 1995 only. We simulate the occurrence of new milestone papers with a Poisson process with exponentially distributed waiting times with an estimated rate λ from the empirical data. The simulation works as for the null model, but instead of taking the publishing dates and out-degrees from empirical data, we assume an exponentially increasing stream of publications with a constant number of references. We bootstrap the empirical citation network until 1995 and simulate the next 20 years for the forgetting curve of 2015. We then repeat the process on a shorter time period of ten years, between 2005 and 2015.

For the predicted citation network, we compute the forgetting curve and compare it to the empirical one from the APS, see figure 3. The orange line shows the forgetting curve of the simulation. It yields a reasonable approximation to the empirical forgetting curve (blue line). The MAPE for the last 30 years is around 21%. This means, that the last three decades of the forgetting curve we predicted for 2015 based on data up to the year 1995 on average deviates less than 21% from the empirical forgetting curve. The second prediction, for the years between 2005 and 2015, is shown in yellow. Here, the MAPE for the last 30 years is around 17%. If all earlier years, back to the 19th century, are included, the error increases up to values of 50%. Note that the absence of a spike in the year 1996 in the orange curve is due to the simulation only using publication data until 1995.

3.3. Predicting milestones

We next test the predictive power of the model for milestone papers by asking how predictive the exponent α_m is for a paper becoming a milestone in the future. We first define a set of *potential milestones* as all those papers that had at least 200 citations by the end of 2005, a condition that is matched for 814 papers in the

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

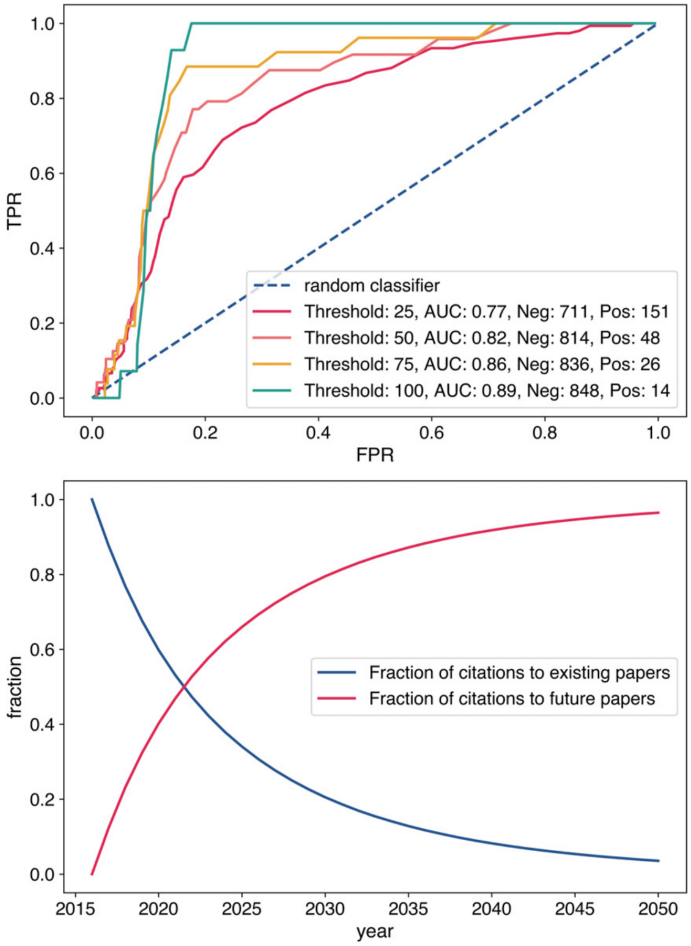


Figure 4. (a) ROC curve (true positive rate vs false positive rate) for predicting the success of potential milestone papers, m, after ten years, based on their exponent, early in their existence, α_m. We show the ROC curves for different thresholds. These thresholds are defined as the minimum number of citations, that a paper must receive in one year, ten years after we measured its exponent α_m. The legend shows the values for the AUC as well as for the number of positives (POS) and negatives (NEG) in our data set for each threshold. (b) What we will cite in the future? For every year on the x-axis we consider the references of papers published in that year. The blue curve represents the fraction of references to papers published up to—and including—2015. The red curve is the fraction of references to papers, published after 2015. Roughly 20% of citations from papers published in 2030 are going to papers published until 2015. 96% of the papers cited 30 years from now are not yet written.

data set. Next, *true milestones* are defined as those papers that receive at least 50 citations in the year, ten years after they received their first 200 citations. At this citation rate they would be ranked among the top cited papers within a few decades. For each potential milestone, m, we measure the exponent, α_m, based on the first 200 citations. If this exponent is small, the forgetting can be over-compensated by the preferential attachment mechanism. Thus, if the measured exponent is below a threshold, $\alpha_m \leqslant \alpha_c$, we predict that the paper will receive more than 50 citations in one year after ten years.

The predictive power of this binary classifier is seen in the receiver operating characteristic (ROC) curve in figure 4(a), where we find an area under the curve (AUC) of up to 0.89. The high value for the AUC suggests that α_m is a good predictor of the future success of a paper, with high specificity and sensitivity. While our definition of *true milestones* was somewhat arbitrary, we confirm that a wide range of thresholds results in reasonably high AUC values and the choice of threshold thus does not significantly impact our results. Our definition of *potential milestones* is based on the requirement for a high number of data points in order to get a low-variance measurement of the exponent α_m. For details, see SI 4. Note that on the left side of the ROC curve in figure 4(a) there is a convex section that indicates, that there are some papers that initially have a very small α_m exponent but fail to stay popular in the long run and thus are erroneously

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

classified as milestones in our model. We hypothesize that the convex section is caused by papers with a sudden but brief success—'flash in the pan' papers.

For the prediction, we estimate an individual constant α_m based on the first 200 citations of each potential milestone. If instead of taking the first 200 citations, these exponents α_m are derived from several hundreds of citations, a closer look reveals that many of the exponents come with a positive slope as a function of the number of citations k. This indicates that for milestone papers, the attachment kernel cannot be factorized as $f(k, t, \alpha_m) = g(k)h(t, \alpha_m)$. For that reason, if one is interested in predicting citation trajectories of individual papers with more than a few 100 citations, α_m should not be taken as a constant but rather as a function of k. We find, that this function is best approximated by a linear function $\alpha_m(k) = \alpha_{m_0} + \beta k$, which can be estimated from the data in an analogous way. For details, see SI 4.

3.4. Predicting future citation landscapes

After confirming that the model accurately predicts the general shape of the forgetting curve, as well as the success of individual papers, we use it to predict what citations today's research will get. We simulate the future citation network as in the situation of the out-of-sample prediction. We bootstrap it on all empirical data until the end of 2015 and then simulate the next 35 years into the future, based on a continued exponential increase of papers and on the attachment kernel measured until 2015. We simulate the network by introducing paper by paper, until we arrive at the possible citation network for the year 2050. Each simulation is performed ten times and the results are averaged.

From this simulated citation network for the years 2016 to 2050, we estimate how quickly today's work will be forgotten. For each of these years, we define the fraction of citations that go to publications published before—and including—2015, which we call *existing papers*. The fraction of citations that go to publications published after 2015 we call *future papers*. Figure 4(b) shows that the fraction of citations to *existing papers* decreases to about 30% within ten years into the future and reaches a value of 3.6% in 2050. If we shift the curve by six years to estimate what will happen to papers published in 2021, we conclude that more than 95% of the papers cited in 2050 are not yet written.

3.5. Empirical citation trends

The empirical data of the APS reveals additional trends that are important in our context. First, the average number of citations that papers published in a certain year receive, has been decreasing since the 1960s. And second, larger and larger fractions of our references consist of comparably old and highly cited milestone papers, while the fraction of milestones among the newly published papers has been steadily declining. For details, see SI 7.

4. Discussion

In this paper, we focused on understanding how scientific work is being utilized or forgotten. In empirical citation networks, we observe two main mechanisms: a linear preferential attachment and a power-law temporal forgetting. We capture these empirical findings in a generative model in which newly published papers cite older papers with a probability proportional to an attachment kernel $f(k, t)$. The kernel reflects two main factors: the cited paper's academic age and the number of citations it already received. This result implies that an article's probability of receiving a citation grows linearly with the citations already received and decays with its publication age as a power-law, characterised by an exponent α.

While there is a strong preference towards citing recent papers, some milestone works –potentially of exceptional impact—continue to attract citations even after decades. These singular milestones, m, are characterized by an exponent $\alpha_m < 0.85$. In comparison, the exponent for the bulk of non-exceptionally successful papers is about $\alpha \sim 1.4$. Based on α_m, one can predict whether a paper will continue to be successful beyond the age of ten years. The corresponding ROC analysis yields an AUC of 0.83, confirming that the exponent is a good predictor indeed.

The model reproduces the highly skewed distributions of the number of citations and the age of references observed in other works. In particular, the forgetting curve that results from the model shows a strong tendency to cite recent papers [10, 26], while older papers fade progressively in oblivion [7, 10, 26]. The spikes observed in the forgetting curve highlight the impact of milestone papers that are remembered for a long time [7, 32].

The linear preferential attachment mechanism in our model confirms the linear dependence on the number of citations, originally found in [21]. The second mechanism we find is power-law-like forgetting. While forgetting, or aging, is a mechanism found in most models of citation dynamics, its detailed functional description remained controversial [22, 29–32, 34, 35]. In [22] an author-paper network yields an attachment kernel with an aging component that follows a Weibull distribution. [32, 34] both investigate

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

the citation rate of individual papers and find exponential aging, while [35] studies a patent citation network and finds power-law aging with an exponential cutoff in *intrinsic time*, where each new patent represents one timestep [30, 31], both find power-law aging in intrinsic time. [30] applies it to the context of citation networks while [31] applies it to a Hollywood-actor network. In [37], the attachment kernel of online media content is described by a model based on preferential attachment, modulated by an exponential forgetting, for which exponents are experimentally determined from data. Our model finds a power-law-like forgetting mechanism with an exponent larger than one for the bulk of average papers. For milestone scientific papers, we find values below one. Power-law-like forgetting implies that papers are forgotten relatively quickly and only have a limited time to attract attention. However, power laws also indicate that old papers keep a chance to get cited. This result enables a few notable publications that attracted much attention early on to continue attracting citations for a long time.

There are several limitations to our model. In particular, it relies on the assumption that the way scientific literature evolves will not change fundamentally. This includes the assumption that the scientific output continues to grow exponentially, an assumption that might not hold forever [38]. This observation limits the prediction of the number of citations a paper will receive in the future, simply because the total number of citations is directly proportional to the number of articles published. However, our results only partially depend on this assumption, as the general shape of the forgetting curve is also recovered well for sub-exponential growth and even for linear growth. On the other hand, our model also assumes no fundamental changes in how citations are chosen. This assumption might not hold if, for instance, disruptive new technologies are introduced and adopted globally, as had been the case with the internet.

To forecast citation trajectories of individual highly cited papers, our model provides a simple indicator based only on the citation network. For higher accuracy, additional parameters, such as the popularity of the authors, the size of the field, or the novelty of the publication, should be taken into account [2, 24, 25, 28, 39–43]. However, even taking these factors into account, there remains uncertainty due to unpredictable events such as a highly relevant discovery that went unnoticed for a while, termed 'sleeping beauty' [44] or the rediscovery of a paper outside its original domain, called exaptation [45].

Restricting our model to only age and number of citations makes it easily transferable to other domains. The possibility that the life-cycle of scientific literature might be different in physics than for other fields like biology, medicine, or philosophy, and even in non-scientific domains such as the music and entertainment industry, is an interesting research direction.

5. Conclusion

Our model allows for projecting the current citation ecosystem into the future and investigating its sustainability. Through the empirical forgetting kernel, and under the assumption that the number of papers will keep growing exponentially, it is possible to simulate possible future citation networks. For instance, one learns that 95% of today's (and yesterday's) work will be outdated before the end of the next three decades. It is this finding, and the fact that papers are receiving fewer and fewer citations on average and have decreasing chances of becoming a milestone, that has substantial implications both for individuals and for scientific policies. Partially, the observed dynamics might be attributed to an increasing fragmentation in the system, driven by the emergence of sub-fields. The fragmentation becomes evident, if one looks at the increase in the number of journals in the APS over the last decades. In this fragmented system, papers might become isolated inside of sub-fields and reach smaller scientific audiences on average. On the other hand, a much more worrisome interpretation of our findings hints towards an increasing quantity-over-quality dynamic, where increasingly short lived publications receive little attention and have diminishing chances of becoming works of great impact.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare no conflict of interest.

IOP Publishing New J. Phys. 24 (2022) 053041 N Reisz et al

Author contributions

NR, ST, VS, VL conceptionalized the work and designed the model. NR, MF and WS did the data preparation. NR and WS did the implementation. NR did the analysis. NR and ST produced a first draft, all contributed to the final paper.

Funding

This work was supported by the Austrian Research Promotion Agency FFG (Grant Nos. 882184, 873927).

ORCID iDs

Niklas Reisz https://orcid.org/0000-0002-5568-4572

References

[1] Sinatra R, Deville P, Szell M, Wang D and Barabási A-L 2015 A century of physics *Nat. Phys.* **11** 791–6
[2] Martin T, Ball B, Karrer B and Newman M E J 2013 Coauthorship and citation patterns in the physical review *Phys. Rev.* E **88** 012814
[3] White K National Science Foundation publications output: U.S. trends and international comparisons 2021 https://ncses.nsf.gov/pubs/nsb20206/publication-output-by-region-country-or-economy (accessed 09 February 2021)
[4] Meho L I 2007 The rise and rise of citation analysis *Phys. World* **20** 32–6
[5] Loken E and Gelman A 2017 Measurement error and the replication crisis *Science* **355** 584–5
[6] Ioannidis J P A 2005 Why most published research findings are false *PLoS Med.* **2** e124
[7] Redner S 2005 Citation statistics from 110 years of physical review *Phys. Today* **58** 49–54
[8] Small H 1973 Co-citation in the scientific literature: a new measure of the relationship between two documents *J. Am. Soc. Inf. Sci.* **24** 265–9
[9] Martyn J 1964 Bibliographic coupling *J. Doc.* **20** 236
[10] de Solla Price D J 1965 Networks of scientific papers *Science* **149** 510–5
[11] Seglen P O 1992 The skewness of science *J. Am. Soc. Inf. Sci.* **43** 628–38
[12] Redner S 1998 How popular is your paper? An empirical study of the citation distribution *Eur. Phys. J.* B **4** 131–4
[13] Price D D S 1976 A general theory of bibliometric and other cumulative advantage processes *J. Am. Soc. Inf. Sci.* **27** 292–306
[14] Barabási A-L and Albert R 1999 Emergence of scaling in random networks *Science* **286** 509–12
[15] Krapivsky P L and Redner S 2001 Organization of growing random networks *Phys. Rev.* E **63** 066123
[16] Dorogovtsev S N and Mendes J F F 2002 Evolution of networks *Adv. Phys.* **51** 1079–187
[17] Albert R and Barabási A-L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47–97
[18] Newman M E J 2003 The structure and function of complex networks *SIAM Rev.* **45** 167–256
[19] Newman M E 2001 Clustering and preferential attachment in growing networks *Phys. Rev.* E **64** 0104209
[20] Capocci A *et al* 2006 Preferential attachment in the growth of social networks: the internet encyclopedia wikipedia *Phys. Rev.* E **74** 036116
[21] Jeong H, Néda Z and Barabási A L 2003 Measuring preferential attachment in evolving networks *Europhys. Lett.* **61** 567–72
[22] Börner K, Maru J T and Goldstone R L 2004 The simultaneous evolution of author and paper networks *Proc. Natl Acad. Sci. USA* **101** 5266–73
[23] Lehmann S, Jackson A D and Lautrup B 2005 Life, death and preferential attachment *Europhys. Lett.* **69** 298–303
[24] Newman M E J 2009 The first-mover advantage in scientific publication *Europhys. Lett.* **86** 68001
[25] Newman M E J 2014 Prediction of highly cited papers *Europhys. Lett.* **105** 28002
[26] Burton R E and Kebler R W 1960 The 'half-life' of some scientific and technical literatures *Am. Doc.* **11** 18–22
[27] Fanelli D and Larivière V 2016 Researchers' individual publication rate has not increased in a century *PLoS One* **11** e0149504
[28] Wang D, Song C and Barabási A-L 2013 Quantifying long-term scientific impact *Science* **342** 127–32
[29] Yin Y and Wang D 2017 The time dimension of science: connecting the past to the future *J. Inf.* **11** 608–21
[30] Dorogovtsev S N and Mendes J F F 2000 Evolution of networks with aging of sites *Phys. Rev.* E **62** 1842–5
[31] Safdari H *et al* 2016 Fractional dynamics of network growth constrained by aging node interactions *PLoS One* **11** 1–13
[32] Golosovsky M and Solomon S 2017 Growing complex network of citations of scientific papers: modeling and measurements *Phys. Rev.* E **95** 1–19
[33] Higham K W, Governale M, Jaffe A B and Zülicke U 2017 Fame and obsolescence: disentangling growth and aging dynamics of patent citations *Phys. Rev.* E **95** 1–7
[34] Higham K W, Governale M, Jaffe A B and Zülicke U 2017 Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields *J. Inf.* **11** 1190–200
[35] Valverde S, Solé R V, Bedau M A and Packard N 2007 Topology and evolution of technology innovation networks *Phys. Rev.* E **76** 1–7
[36] Ke Q, Ferrara E, Radicchi F and Flammini A 2015 Defining and identifying sleeping beauties in science *Proc. Natl Acad. Sci. USA* **112** 7426–31
[37] Lefortier D, Ostroumova L and Samosvat E 2012 Evolution of the media web https://doi.org/10.48550/ARXIV.1209.4523
[38] Brown J H *et al* 2011 Energetic limits to economic growth *BioScience* **61** 19–26
[39] Thurner S, Liu W, Klimek P and Cheong S A 2020 The role of mainstreamness and interdisciplinarity for the relevance of scientific papers *PLoS One* **15** e0230325
[40] Klimek P, Jovanovic A S, Egloff R and Schneider R 2016 Successful fish go with the flow: citation impact prediction based on centrality measures for term-document networks *Scientometrics* **107** 1265–82

[41] Acuna D E, Allesina S and Kording K P 2012 Predicting scientific success *Nature* **489** 201–2
[42] Shen H, Wang D, Song C and Barabási A-L 2014 Modeling and predicting popularity dynamics via reinforced Poisson processes *Proc. of the 28th AAAI Conf. on Artificial Intelligence, AAAI'14* (AAAI Press) pp 291–7
[43] Sinatra R, Wang D, Deville P, Song C and Barabasi A-L 2016 Quantifying the evolution of individual scientific impact *Science* **354** aaf5239
[44] van Raan A F J 2004 Sleeping beauties in science *Scientometrics* **59** 467–72
[45] Ferreira M R, Reisz N, Schueller W, Servedio V D P, Thurner S and Loreto V 2020 *Quantifying Exaptation in Scientific Evolution* (Cham: Springer) pp 55–68

Supplementary information

Text S1: Forgetting curve in the random case

In the method section, we mention that the forgetting curve would fluctuate around a value of 1 if each paper had the same probability of getting cited. The result of a simulation implementing this simple random model is shown in figure 5. The curve results from averaging the results of 20 independent simulation runs.

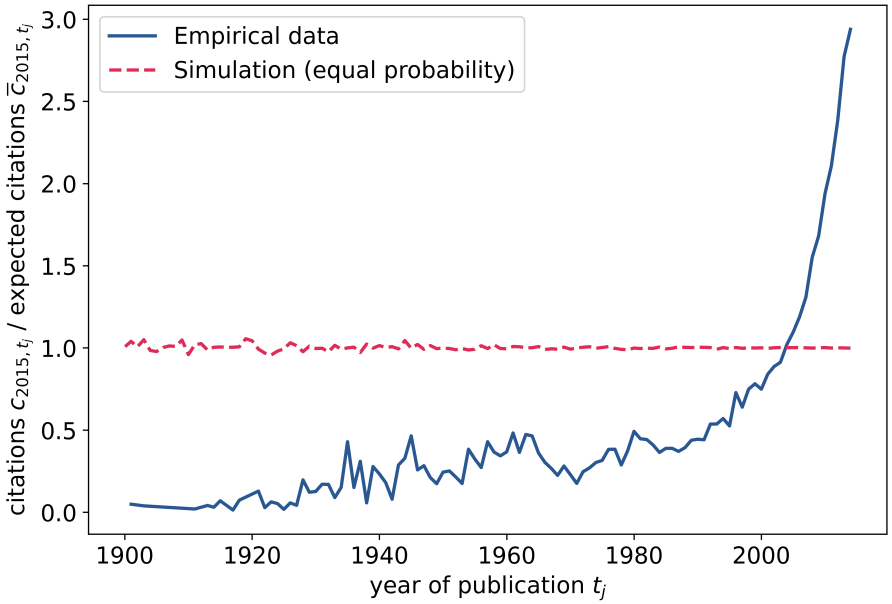


Figure 5. Forgetting curve for the APS citation network for the empirical data (blue), and the simple simulation where every paper is equally probable to get cited (red, dashed).

Text S2: Milestone significance

In the forgetting curve in Fig. 2 (b) we observed spikes coinciding with the publishing years of the top 30 papers. SI Fig. 6 shows how these spikes develop over the years.

We want to show that these spikes can be caused by works of exceptional quality, here called milestones. To do so, we calculate the relative share cit_{ms} of citations that go to milestone papers, compared to citations that go to all papers. The calculation is simple: For each year, we count the number of citations to milestone papers published in that year, divided by the total number of citations to all papers published in that year.

We then average this value over all the years when milestone papers were published. We find, that in the years when milestone papers were published, they attract around 6% of all citations. If we repeat the analysis but consider only citations from papers published in 2015, the last year of our data set, we find that in the years when milestone papers were published, these milestones attract around 14% of all citations of papers published in 2015. Since the forgetting curve is directly proportional to the number of citations, these values are indeed high enough to cause a visible spike in the forgetting curve.

Text S3: Measuring the attachment kernel

To measure the attachment kernel, $f(k,t)$, we modify the method originally used by Newman in co-authorship networks[19] to two dimensions. We then apply it to the 120 years of empirical citation data of the American Physical Society. In this method, we go through the papers of the APS, one by one, ordered in time, and build a histogram from which we can read off the attachment kernel. The steps included are as follows. First, we go through the APS citation network $M_{i,j}$ node by node and link by link, starting from the first paper ever published and ending on the most recent paper. At each point in time, the probability of a paper j citing a specific paper i with age, t, and k citations, is given by:

$$P(k,t) = f(k,t)\frac{n(t_j)}{N(t_j)}, \tag{4}$$

where $n(t_j)$ is the number of papers present at publishing time t_j of paper j that have the same exact values of k and t as paper i, and $N(t_j)$ is the total number of papers present at that time. Then the attachment kernel $f(k,t)$ can be estimated

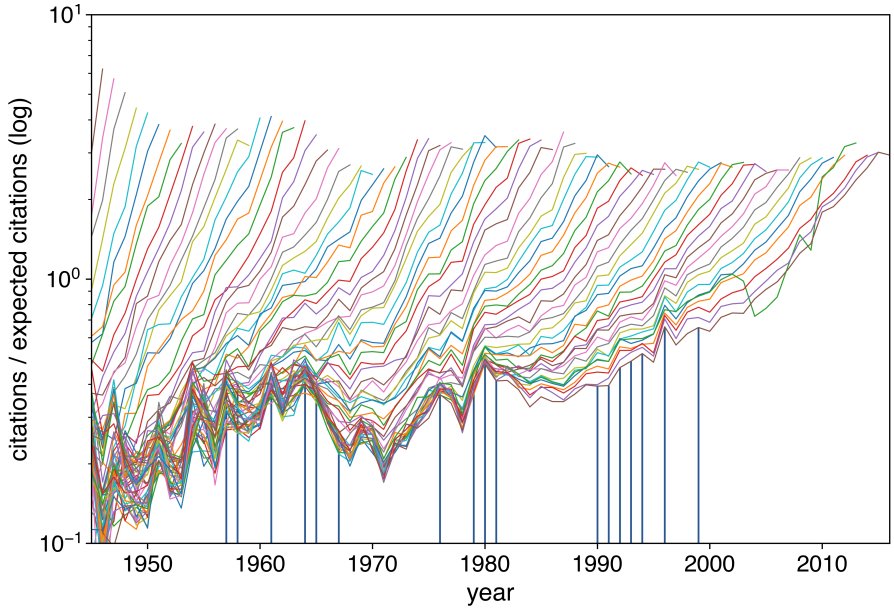


Figure 6. The forgetting-curves of the APS. Each line represents the forgetting-curve for papers published in the year after the right-most point of the curve (e.g. if the curve ends in 2014, it is the forgetting curve of papers published in 2015). Years where milestone papers were published are marked with vertical lines. In these years, one can observe the formation of spikes due to the adoption of the milestone paper by following the vertical line from the top (earlier years) to the bottom (recent years).

from a histogram where each contribution is weighted with the inverse probability $\frac{N(t_j)}{n(t_j)}$. A sketch of this method is shown in Fig. 7. From the resulting histogram, we can extract slices for fixed values of k or t. These let us fit the functions in k and t respectively. If the slices only differ in a factor, then the kernel is of the form $f(k,t) = g(k)h(t)$. We find, that this is true in good approximation for the bulk of average publications.

The result is a two-dimensional histogram that approximates the probability kernel, $f(k,t)$. Note, however, that many combinations of (k,t) are underrepresented in the history of the APS and thus also in the histogram. For example, it is evident that there are very few papers that have a high number of citations already in the first few years. Thus, the histogram is only representative in areas where sufficient data is available. For this reason, it is not practicable to directly fit a function to the histogram.

By fixing k or t to specific values, slices in t or k can be extracted from the surface. We slice the landscape for all integer values of k and t and fit functions to the slices. If the probability kernel is factorizable, $f(k,t) = g(k)h(t)$, then the slices would only differ by a factor. We fit a linear function $g(k) = k$ and a power $h(t) = t^{-\alpha}$. From this, we calculate an average exponent as the weighted average over all exponents, where the weights correspond to the number of times a particular value of k was observed in the history of the APS.

Text S4: Milestone attachment kernel

In the introduction, we discussed that some old publications receive exceptionally high numbers of citations. We call these the milestone papers. In the context of a citation network, we define them through their citations and take the top 30 most cited papers of all time that are not review papers. The choice is arbitrary, but the results hold for different values as long as the papers are receiving numerous citations.

The milestone papers are exceptions, they stand out from the bulk. As such, their number is low, and the statistics are not sufficient to measure their attachment kernel with the method described in SI 3. Instead, we propose a slightly different method, described below.

To figure out an estimator for the probability of being cited, we draw a comparison to an urn model. Suppose we have an urn with different colors of balls. We are only interested in one special color. There are multiple copies of that color in the urn. In each run, we draw with replacement until we draw a ball of our color. If we do many runs, and on average it takes n draws until a ball of our color is drawn, then the most plausible estimator for the probability of drawing our color would be $\frac{1}{n}$.

In addition to that, we introduce the concept of intrinsic time for a citation network. Each citation event, where any paper is cited, represents one discrete time step in intrinsic time. For a milestone paper m, the probability of being cited at any intrinsic

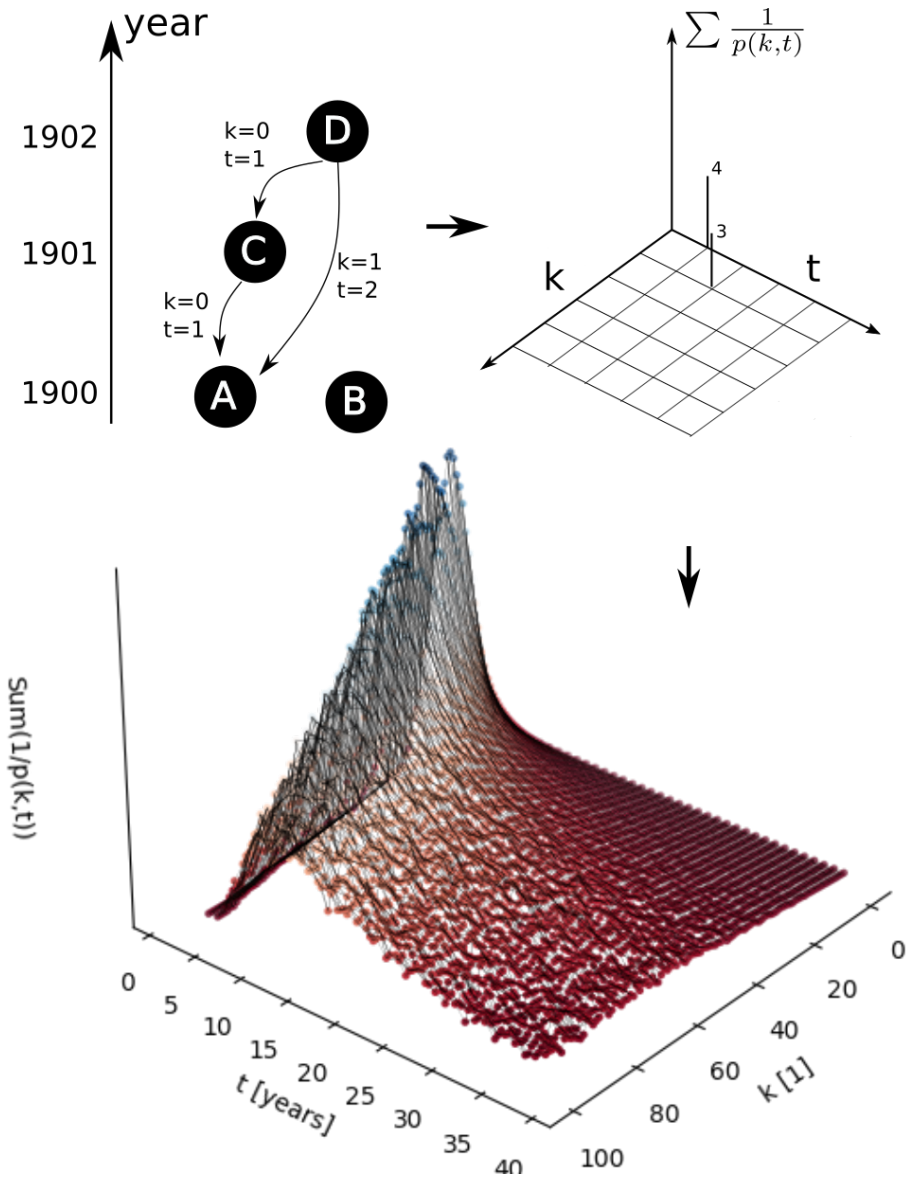


Figure 7. Approximating the kernel $f(k,t)$ of the APS citation network.
a) The citation network is rebuilt paper by paper. Each back dot is a paper, each link is a citation. The papers are published in alphabetical order. For reach citation, the values of k, t and the state of the network are used to build the histogram.
b) The histogram is generated as the sum of the inverse probabilities when reconstructing the citation network.
c) The histogram for the entire network. This landscape in k and t represents the shape of the probability kernel $f(k,t)$. The shape is only representative for areas where sufficient data is available.

time step τ is given by:

$$p_m = \frac{f(k_{m,\tau}, t_m, \alpha_m)}{\sum_i f(k_{i,\tau}, t_i, \alpha_i)}, \tag{5}$$

where f is the probability kernel of the form:

$$f_i = f(k_{i,\tau}, t_i, \alpha_i) = (k_{i,\tau} + 1)(t_i + 1)^{-\alpha_i}, \tag{6}$$

and $k_{i,\tau}$ is the in-degree of paper i at intrinsic time-step τ, t_i is the age of paper i in years and α_i is the forgetting-exponent of paper i.

The most likely estimator for the number of citations C it takes until paper m is cited is $\frac{1}{p_m}$. We can estimate:

$$C \approx \frac{1}{p_m} = \frac{\sum_i f_{i,\tau}}{(k_{m,\tau}+1)(t_m+1)^{-\alpha_m}} \tag{7}$$

leading to

$$\alpha_m \approx \log\left(\frac{C(k_{m,\tau}+1)}{\sum_i f_{i,\tau}}\right)\frac{1}{\log(t_{m,\tau}+1)}\,. \tag{8}$$

We know $k_{m,\tau}$ and t_m. C is the number of citation events between two events where the milestone paper m is cited. The α_i are unknown, but they only appear in the sum. Since we have a good approximation for them, supposedly, the sum should be approximated correctly. Thus, we can estimate α_m directly. For each time the milestone paper m is cited, we get a different C and $k_{m,\tau}$ is increased by 1, so we collect a set of values for α_m

$$\overline{\alpha}_m = \log\left(\frac{C(k_m+1)}{\sum_i f_{i,\tau}}\right)\frac{1}{\log(t_{m,\tau}+1)}\,. \tag{9}$$

These values can be averaged (before applying the logarithm) in order to determine an average exponent. However, if we apply a linear fit to the set of values for α, we notice a general slightly increasing trend. This suggests that in our approximation for the kernel, $f(k,t) = (k+1)(t+1)^{-\alpha}$, the functions of k and t cannot be separated completely for milestone papers. Instead, there is a small dependency on k in α. A more fitting approximation for milestone papers is thus given by $f(k,t) = (k+1)(t+1)^{-(a+bk)}$ i.e., the exponent α is a linear function of k with a very small and positive slope.

To validate the method, we construct the null model explained in SI 5. We simulate the full system and then repeat the measurement of the exponent in the simulated system. Since the exponent in the attachment kernel of the simulation are configuration parameters, we have perfect knowledge of them. Thus, we can compare the values we measure with the ones we used as an input. If the measurement method is valid, we would expect these values to be the same, apart from some noise. Indeed, we measure an average absolute deviation of $\Delta\alpha = 0.1$. If binning is used to reduce the noise, the average absolute deviation can be reduced to $\Delta\alpha = 0.01$. The result for one example paper is shown in fig. 8.

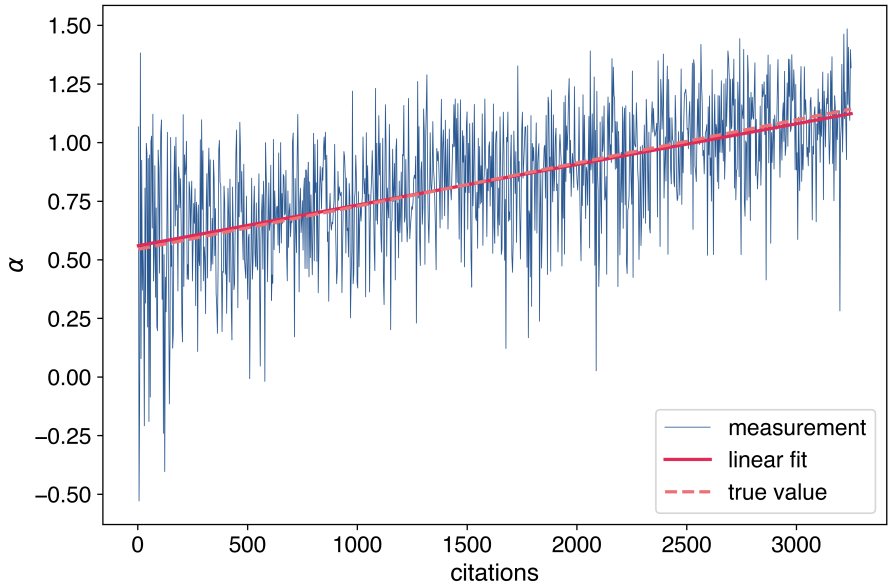


Figure 8. Exponent α of $(k+1)(t+1)^{-\alpha}$ vs in-degree k for a representative milestone paper, measured on the simulated null model. Blue: individual measurements of α, Red: linear fit on the individual measurements, Orange, dashed: true value (simulation input)

Two additional effects might have an impact on the measured value and thus need to be ruled out. First, in equation 9 the sum $\sum_i f_{i,\tau}$ changes after each citation event (each intrinsic time step τ). For the calculation, we take the value at the end of the interval, i.e., the value of τ at the time when the milestone is cited. This is technically inaccurate, but the inaccuracy is

very small. We measure this both at the beginning and the end of the interval, confirming that the sum changes only by an insignificant amount. Thus, this effect can be neglected. Second, each paper cites n other papers, but each of the cited papers can only appear once in the reference list. Thus, the probabilities for getting cited by one paper change depending on what else was cited by the same paper. Also, this effect can be neglected since on average the number of references per paper is very small compared to the total number of citable papers.

Text S5: Null model

We construct a null-model of the APS citation dynamics. In this model, we keep all publishing dates and out-degrees (number of references) of real papers from the empirical data, but we attach each outgoing link to a random paper that was published before based on the probability kernel $f(k,t)$. For milestone papers we use the exponent that we determined from the real data, for the bulk of average papers we take the average exponent. A sketch of the null-model is shown in SI Fig. 9.

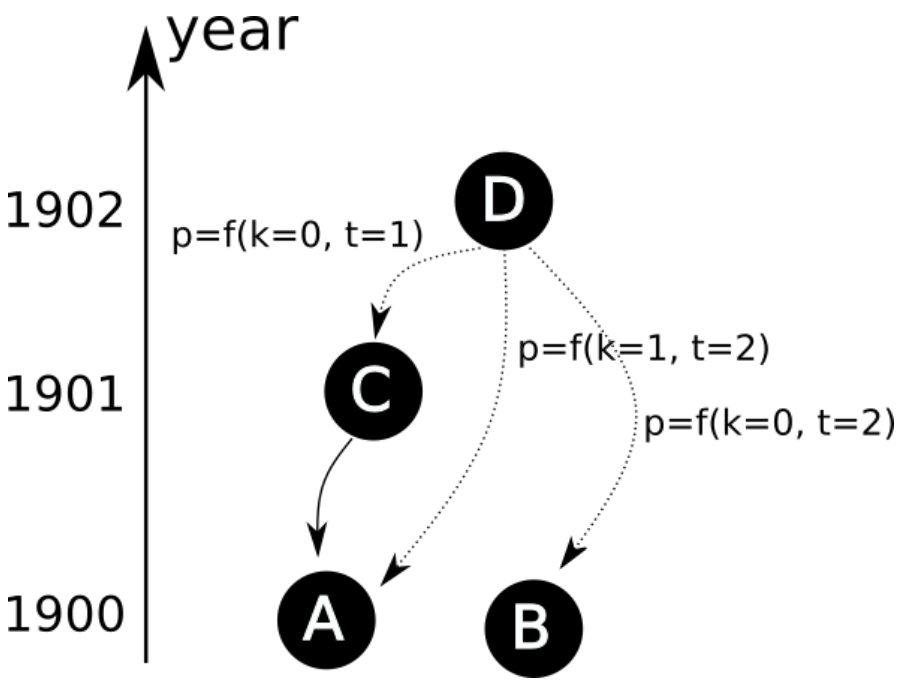


Figure 9. Null model of the APS citation dynamics. Papers are published in the same order, at the same time and with the same out-degree (number of references) as the papers of the APS citation network. Each paper randomly cites papers that were published before with a probability proportional to the probability kernel, $f(k,t)$.

Text S6: Milestone simulation performance

Figure 10 shows a comparison of forgetting curves and allows the estimation of the simulation's performance in terms of milestones. Three forgetting curves are plotted there. The first (blue, solid) is the forgetting curve of the empirical APS. The second curve (red, dashed) is the forgetting curve of the empirical APS, with all reference to milestones removed. By comparing the first two curves, one can understand how milestones affect the entire forgetting curve. The third curve represents our null-model simulation. One can see, that in almost all cases where there is a difference between the curves with and without milestones, the simulated curve has a spike of comparable magnitude.

There is one apparent exception in the year 1935, that is not captured well in the simulation. It concerns the famous paper titled "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" by Einstein, Podolsky and Rosen. Curiously, this paper received less than 10 citations from papers published in the APS in the first 15 years after being published, presumably in part due to the war. It wasn't until the 1990s that this paper gained traction and began to be widely quoted. As a result, it is known as a prime example of a "sleeping beauty." This late discovery and peculiar citation pattern are not adequately reflected by a time-independent forgetting exponent, and as a result, the citations of this publication are underestimated in the simulation.

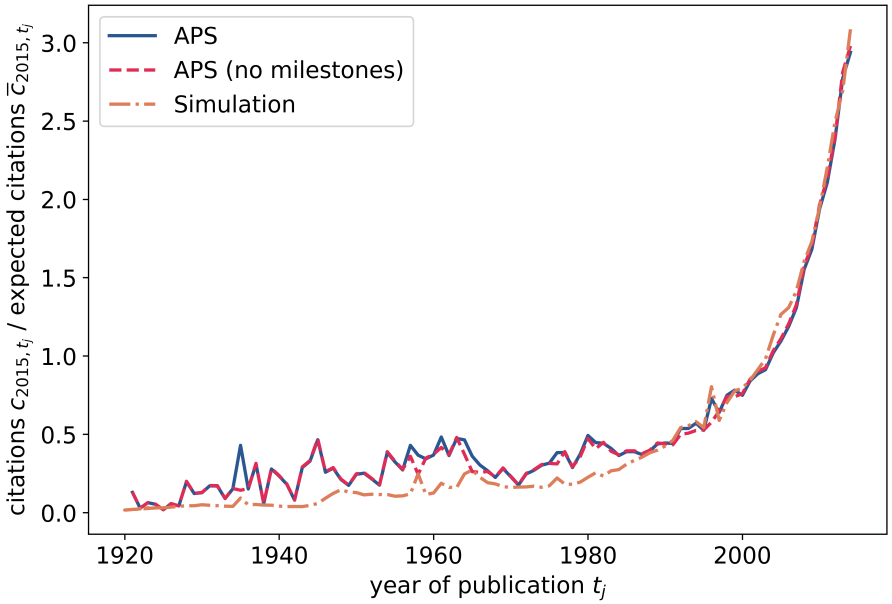


Figure 10. A comparison between forgetting curves. Forgetting curve of the empirical APS (blue, solid), forgetting curve of the empirical APS where citations to milestones were removed (red, dashed) and null-model simulation (orange, dashed and dotted). The apparent outlier in 1935 is the famous paper by Einstein, Podolsky, Rosen, and a prime example of a "sleeping beauty".

Text S7: Citation trends

To reveal general citation trends in the APS, we investigate the average number of citations that a paper published in a certain year receives, see fig. 11. Note that the steep decline during the last decade is due to the papers not having had enough time to accumulate all their citations yet.

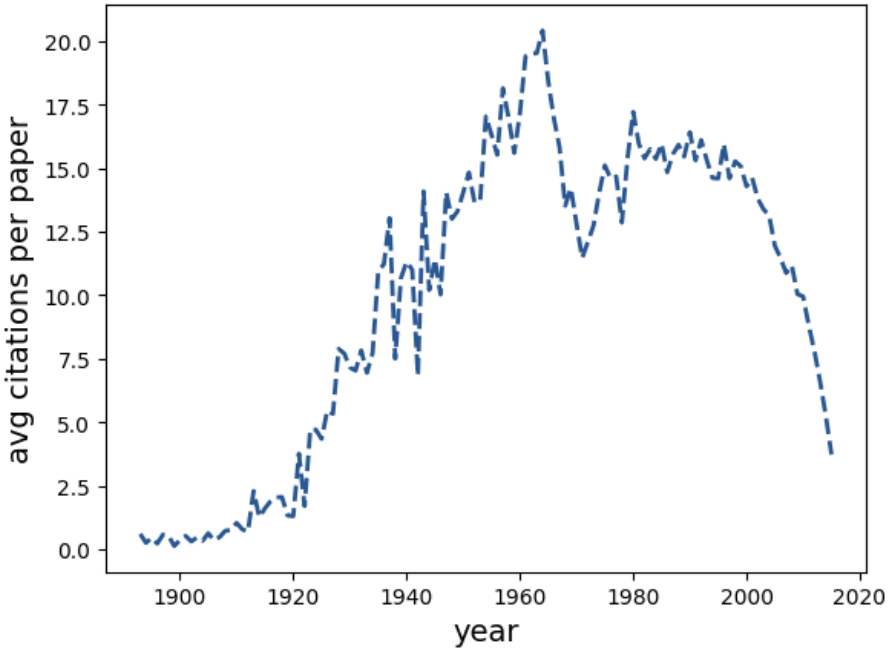


Figure 11. Average number of citations that papers published in each year between 1893 and 2015 received. Note that the steep decline during the last decade is due to the papers not having had enough time to accumulate all their citations yet.

To show that an increasing number of citations go to milestone papers, we plot the fraction of references of all papers published in a certain year, that reference highly cited papers (papers with ≥ 1000 citations) over the total number of references. This is shown in fig. 12.

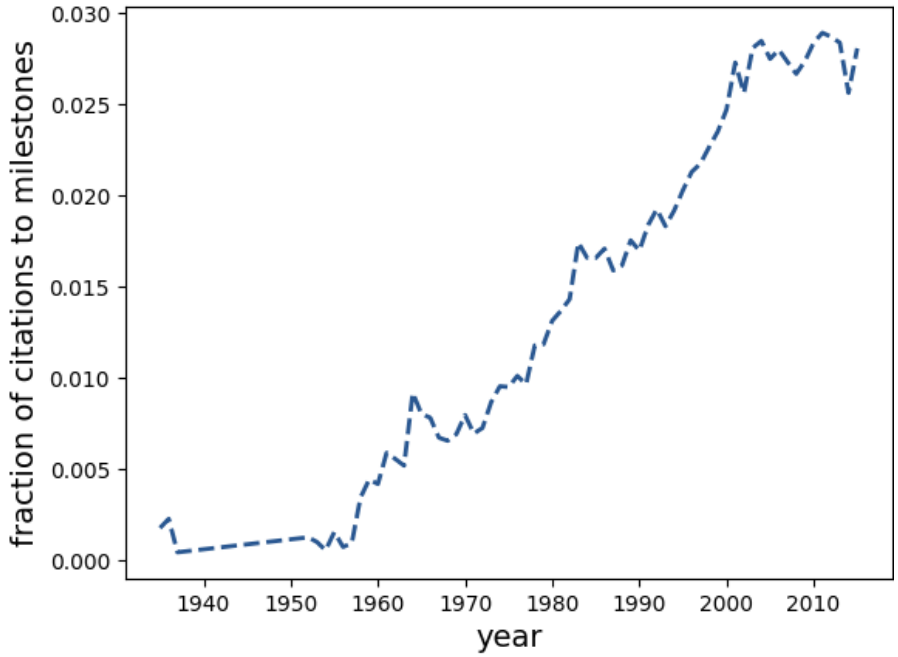


Figure 12. Fraction of references to highly cited papers (≥ 1000 citations) over references to all papers for each year between 1893 and 2015.

In fig. 13 we show the number of highly cited papers (papers with ≥ 1000 citations) published in a certain year over the total number of papers published in that year.

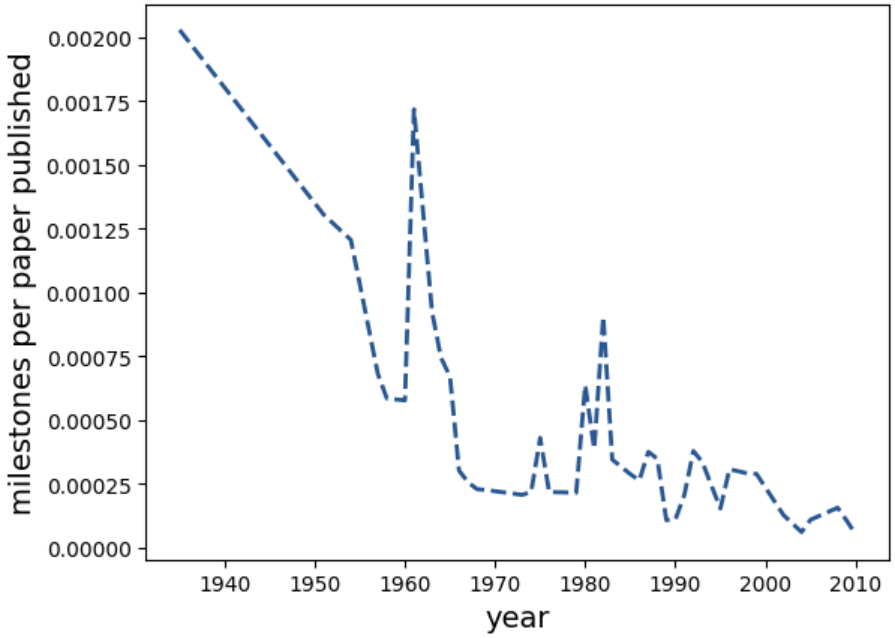


Figure 13. Fraction of highly cited papers published ($\geq$ 1000 citations) over total number of papers published for each year between 1893 and 2015.

Quantifying exaptation in scientific evolution

Márcia R. Ferreira[1, 2], Niklas Reisz[1], William Schueller[1], Vito D.P. Servedio[1], Stefan Thurner[1, 3, 4, 5], and Vittorio Loreto[1, 6, 7]

[1]Complexity Science Hub Vienna, Josefstädter Str. 39, 1080 Vienna, Austria
[2]TU Wien (Vienna University of Technology), 1040 Vienna, Austria
[3]Section for Science of Complex Systems, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, A-1090 Vienna, Austria
[4]International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria
[5]Santa Fe Institute, Santa Fe, NM 85701
[6]Sony Computer Science Lab, Paris, 6, Rue Amyot, 75005 Paris, France
[7]Sapienza University of Rome, P.le Aldo Moro 2, 00185, Rome, Italy

February 20, 2020

Abstract

Rediscovering a new function for something can be just as important as the discovery itself. In 1982, Stephen Jay Gould and Elisabeth Vrba named this phenomenon *Exaptation* to describe a radical shift in the function of a specific trait during biological evolution. While exaptation is thought to be a fundamental mechanism for generating adaptive innovations, diversity, and sophisticated features, relatively little effort has been made to quantify exaptation outside the topic of biological evolution. We think that this concept provides a useful framework for characterising the emergence of innovations in science. This article explores the notion that exaptation arises from the usage of scientific ideas in domains other than the area that they were originally applied to. In particular, we adopt a normalised entropy and an inverse participation ratio as observables that reveal and quantify the concept of exaptation. We identify distinctive patterns of exaptation and expose specific examples of papers that display those patterns. Our approach represents a first step towards the quantification of exaptation phenomena in the context of scientific evolution.[1]

Keywords: Exaptation, Evolution, Innovation, Scientometrics, Science of Novelty

1 Introduction

The discovery of a new function or application for a given object or concept can be just as important as the discovery of something for the first time. This phenomenon is known as *exaptation*, and the related verb is *to co-opt*. It characterises the process of acquiring new functions for which a trait, which originally evolved to solve one problem, is co-opted to

[1]This article is a preprint version of results of a study published in an upcoming Springer-Nature: The Frontiers Collection book entitled: "Understanding innovation through exaptation" edited by Stefano Zapperi and Caterina La Porta.

solve a new problem [21]. The definition is similar to the concept of *preadaptation* [7]. However, since this term may suggest teleology, Vrba and Gould urged scholars to replace that term by exaptation. The idea of exaptation was also proposed to distinguish the concept from *adaptation* [13]. While exaptations are traits that have applications that deviate from their original purpose [21, 25], adaptations have been shaped by natural selection for their current use [7, 13]. One canonical example of exaptation in biology is the evolution of feathers. It is often argued that feathers were not originally developed for flight, but emerged in the reptilian ancestors of today's birds for thermal regulation [21]. Other examples include the ability of a metabolic reaction network to survive on different food sources which can allow adoption of alternative substrates [5]. Exaptation events seem to be important to give birth to adaptive innovations, diversity and, more generally, to complex traits [5].

Gould and Vrba propose exaptations have adaptive and non-adaptive origins: *preaptation* and *nonaptation*. Preaptation refers to characteristics or traits that are adapted and 'selected' for one evolutionary purpose (adaptations). These are later co-opted to serve another purpose leading to an increase in the fitness of the co-opted trait. Thus, preaptations are adaptations that have undergone a significant change in function [21, 34]. Another source of exaptation is nonaptation. Nonaptation refers to traits that emerge through a process of trial and error that generates lots of 'leftover' features (e.g., DNA, molecules, cells) [21]. Nonaptation concerns the effective use of co-opted leftover traits to serve a particular function, but whose origin cannot be ascribed to the process of 'natural selection' [21]. Thus, nonaptations are byproducts of the evolution of some other trait [13] that do not add to the fitness of the co-opted trait [34, 21]. In summary, adaptations, preaptations, and nonaptations are essential processes that drive the evolution of living matter, cells, humans, organisms, and biological ecosystems. They allow us to understand the adaptive and non-adaptive origins of biological novelty.

Recently, the notion of exaptation has been applied to the study of technological change [8, 15]. One set of studies have focused on the development of technological speciation narratives [15, 32, 3] and niche construction theory [14]. Other studies focused on the adoption of technology [40], its commercial application [42], and its economic impact [14], but not on the origin of those inventions [18]. Small-scale studies of technological exaptation abound in the management and innovation literature (e.g., [47, 14, 10, 41]). These studies point to the role of chance such as serendipitous discovery of a new function [3]. Serendipity (accidental circumstances leading to fortunate findings) and exaptation (a shift in the function of something) are intricately related by the fact that accidental discoveries that contribute to the redeployment of a component in a different context lead to a shift in the function of that component. A related stream of research has attempted to model the dynamics of invention mathematically by analysing specific knowledge spaces [51, 35, 49, 28, 27, 23, 46, 43, 24, 20]. The aim of these studies is not to explain why some entities produce more innovations than others (productivity), or what influences the ability of these entities to produce them, but how knowledge evolves in a mechanistic view. These studies have provided evidence for the existence of innovation bursts in national economies [49, 27], the rediscovery of publications leading to the emergence of new scientific fields [50, 53], and the novel combination of components as a source of everyday novelties [51]. Few attempts focused on explicitly quantifying exaptation, one notable exception being [1], where it is estimated that about 40% of pharmaceutical drugs have started as something else.

Similarly, many scientific discoveries find applications that are not foreseen from the outset. The scientific context is particularly relevant for the study of exaptation since it

encompasses a variety of human activities where knowledge is frequently rediscovered and re-used. In scientific evolution processes, different disciplines may come together, "to tell one coherent interlocking story" [56], or a field may subdivide into new disciplines. Both may form the basis on which concepts can further recombine. The recent use of statistical physics to examine people's behaviour in crowds, traffic, or stock markets is an example of co-opting theories and techniques to the social sciences. Research on laser technologies [8], pharmaceuticals [1], and fibre optics [10] keep expanding their scope of application in very diverse fields. This kind of repurposing may enhance (though not necessarily) the fitness of entities [21]. Exaptation in the context of science thus refers to a diversification logic, where publications build on an existing knowledge base and succeed in entering other fields by creating new scientific niches or sub-fields. We thus interpret exaptation in science as how research insights from publications in one field are co-opted (i.e., cited) by publications from different scientific domains.

To arrive at a formal understanding of exaptation in science and technology we start by briefly reviewing related perspectives on the origins of innovations.[2] We then attempt to detect the fingerprints of exaptation by using publication data indexed in the APS (American Physical Society) database, which contains over 450,000 articles in the field of physics. We use the direct citation network to construct clusters of publications that represent sub-fields of physics. Citations have been considered as a proxy for academic relevance, with citation-based indicators offering approximate information about the scientific impact of publications [45]. We focus on seminal publications that initially appear in a given domain and later receive acknowledgements and new functions from other domains. The idea of function is critical here because it reflects knowledge-use of specific research outcomes in different scientific contexts. By investigating the "citation paths" of individual publications across different domains, we quantify both their overall importance, as well as their impact on the structure of the field of physics as a whole.

The paper is structured as follows. In Section 2, we introduce the concept of exaptation as a theoretical concept and clarify the scientific context in which we will use it in this paper. In Section 3, we describe the data and the empirical approach. In Section 4, we present results. Finally, in Section 5, we conclude the paper and summarise our results and proposal in this area of research.

Exaptation in science and technology

A growing number of scholars in the area of innovation theory propose exaptation as the ultimate source of novelty. They argue that exaptation can explain the emergence of markets, technologies, and technical functions [14, 10, 1, 37]. In their view, exaptation leads to *technological speciation* or the creation of *technological niches* [2]. In technological speciation processes, new technology develops from the effective transfer of existing knowledge to a new situation, where the transferred knowledge is interpreted and exploited in new ways [12]. Famous cases that illustrate this pattern include Corning Inc., a company that used its long-standing experience on glass engineering to deliver ground-breaking fibre-optic work that has transformed the telecommunications landscape [56], or the microwave, which was discovered by chance through the repurposing of parts of a radar system [41]. Often, a distinction is made between radical and incremental innovations. While radical innovations transform the technological landscape, incremental innovations are minor improvements in existing technologies [16]. Exaptation has been associated with radical

[2]It is not our aim to provide an in-depth discussion and definition of the concepts of novelty, innovation, or invention, which can be found elsewhere [17, 4]; these terms will be used interchangeably throughout the paper.

innovations leading to the creation of new technological niches [3]. Most empirical studies of exaptation in those contexts have been limited to small-scale or narratives of specific technologies.

In the context of science, contributions to the study of invention have used scientific publications, and metadata, such as author affiliation, organisation, location, and citation linkages to assess novel research outputs. Citation network analysis has been used extensively. Uzzi et al. [52] used co-citation linkages from publications in various fields to differentiate between typical and atypical co-citations. Atypical co-citations are papers that are rarely cited together. They found that high-impact publications were usually not those that had the most atypical or novel combination of ideas, nor those that used typical combinations of ideas, but papers that cite a mix of new and conventional ideas. This result implies that, while originality is a crucial feature of high-impact science, the building blocks for new ideas are often embodied in existing knowledge [52]. Further, papers with high novelty as measured by atypical combinations tend to be less cited at the start, but are more likely to be cited after several years after publication [44]. Other studies, which highlight the role of recombining ideas in driving innovation, suggest that older, seminal works are more likely to inspire ground-breaking science [29].

The combination of different theories, fields, and tools is also central to interdisciplinary research [54]. Interdisciplinarity is likely to lead to more 'innovative' research [50], which is associated with higher levels of citation impact [31]. Furthermore, the citation influence of papers is enhanced by the thematic distance (i.e., cognitively different fields) from the articles they cite [26]. Yet, such outputs often face more resistance than mainstream publications (i.e., publications drawing mainly on the knowledge of a single field), thus requiring more time to get recognised by the wider scientific community [50]. This idea relates to the "first-mover" advantage where mediocre papers will often receive more citations early on, than a later excellent one [38]. There is, however, conflicting evidence that interdisciplinary research obtains higher citation rates at the level of journals in natural and medical sciences [33], research departments [39], and in the field of biomedicine [30]. This evidence shows that the relationship between novel research - as defined by interdisciplinary combinations - and impact depends on the characteristics of the fields and type of analysis involved [30].

2 Methodology

To track scientific progress, we use the APS dataset [19], which includes publications in the leading physics journals since 1893. Following [6], we apply the Louvain algorithm to design an alternative classification scheme that clusters the set of all publications (articles and reviews) in the APS between 1893 and 2017 into research areas. The method is based on first determining pairs of publications that cite one another, and second, on clustering publications into a research area. The procedure uses a directed citation network where nodes consist of publications and links consist of citations between publications. Each publication belongs to a unique research area. We disregard the direction of the links in the network and exclude publications without citations.

Fig. 1 shows the distribution of cluster sizes. For several clusters the number of publications is very small. For practical reasons we excluded clusters that have less than 10 publications. This method allows us to examine the influence of publications that belong to a specific field on other papers pertaining to different fields. This information is essential to determine whether a paper has been co-opted by papers in another field. This also allows

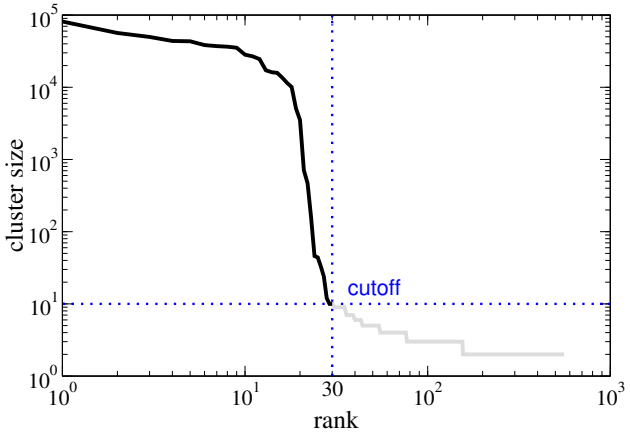


Figure 1: **Cluster size vs rank.** The number of APS articles in a cluster as a function of the rank of the clusters. Clusters were detected by the Louvain algorithm applied to the citation network. We cluster all articles appeared in APS from 1893 to 2017. In our analysis, we remove the articles belonging to clusters with less than 10 articles. These removed clusters constitute a few percent of all articles.

us to trace the bibliographic properties of the citing publications.

Quantifying exaptation

To quantify the idea of functionality, we define a quantity that we call *forward normalised entropy* (FNE), S_i, of a generic article a_i. In Fig. 2, we consider the articles citing a_i (red circle), each belonging to a cluster, C_n, here we use all papers published in APS until 2017. Denoting by A_i the set of articles citing a_i and by C_n the clusters in the system, we define $p_{i,n} = |C_n \cap A_i|/|A_i|$. In other words, $p_{i,n}$ is the fraction of articles citing the reference article a_i, that belong to cluster C_n. We define the normalised forward entropy, S_i, as:

$$S_i = -\frac{1}{\log N_i} \sum_{n=1}^{N_i} p_{i,n} \log p_{i,n} \ , \tag{1}$$

where N_i is the number of clusters for which $p_{i,n} > 0$. We call S_i a forward entropy because it is computed based on articles published after a_i and citing a_i.

It gives information on how heterogeneous the composition of the citing articles is in terms of cluster composition. If an article is cited by articles belonging to only one cluster, i.e., it belongs to a well-defined scientific field, $S_i = 0$. If a paper has $S_i = 1$, then its citations are uniformly distributed among different areas. Therefore, the forward entropy can be thought of as a score for interdisciplinary impact.

To estimate the effective number of clusters from which the paper a_i received citations, we consider the Inverse Participation Ratio (IPR) I_i of article a_i:

$$I_i = \left(\sum_{n, p_{i,n}>0} (p_{i,n})^{-2} \right)^{-1} \ . \tag{2}$$

Its value approximates the number of effective clusters citing a_i. We calculate the forward entropy and IPR for every year, by considering only those articles published in a given year,

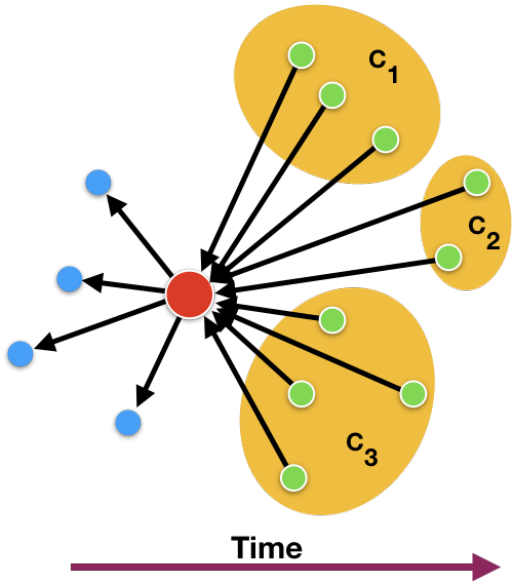


Figure 2: **Definition of normalised forward entropy cartoon.** Circles represent published articles and arrows stand for citations. Only those articles and links are displayed, which are relevant for the entropy definition. The article a_i denoted with the large red circle cites four papers (blue circles at its left side) and has been cited by nine papers (green circles on its right side). These nine articles belong to three different clusters C_n depicted as orange ovals. Denoting with p_n the fraction of articles of cluster n in the whole citation pool of the red article, we have $p_1 = 1/3$, $p_2 = 2/9$ and $p_3 = 4/9$. According to Eq. (1), the normalised forward entropy is then $S_i = (-1/3 \log 1/3 - 2/9 \log 2/9 - 4/9 \log 4/9)/\log 3 \approx 0.966$. The IPR for a_i is $I_i \simeq 2.79$.

and citing a_i. In this way, we can follow the trajectory of an article over time, keeping track of the scientific areas it belonged to.

3 Results

We considered all publications in the APS database until 2017 and selected the top 200 most cited ones. The reasons for using highly cited publications are both conceptual and pragmatic. First, we assume that exaptation results from the significant adoption or acknowledgement of a paper. This implies that a co-opted publication should have a high citation impact. Second, a large number of citations enhances the statistical significance of the results. To identify distinctive patterns of exaptation, we looked at the yearly number of citations vs. the forward normalised entropy (FNE), S_i, and the IPR, I_i, for all articles. We present a few examples with a well-defined signature of exaptation.

Before that, let us clarify how this pattern should ideally look like. A good candidate article for exaptation, say a_{ex}, ideally belongs to a well-established field and is disciplinary in nature. If the paper initially received citations only by papers in the same scientific sub-field, then its FNE score, S_i, would be zero. We hypothesise that at some point in time, the number of citations to a_{ex} starts increasing and possibly remains in the same scientific sub-field. At a later stage, the article may start receiving citations by papers from

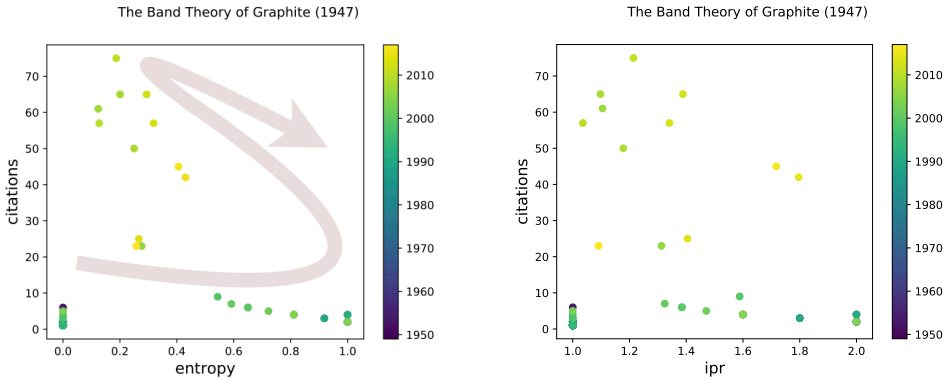


Figure 3: **A typical pattern for exaptation of an important paper.** Left: yearly number of citations vs. the forward normalised entropy of the paper of Ref.[55]. Each circle represents a different year, and the time-evolution is marked with a colour code, from dark tones (older) to light ones (newer). The ideal dynamics of citations and entropy over time for an exapted paper is also depicted by a thick arrow. The paper starts in one field ($S_i = 0$), then is exapted in another field leading to an increase of S_i; then, citations grow since the paper gets more popular while S_i decreases, since the new field prevails; eventually, the number of citations decreases as an ageing effect while S_i increases again, since the paper is rediscovered in the older field. Right: yearly number of citations vs. IPR of [55].

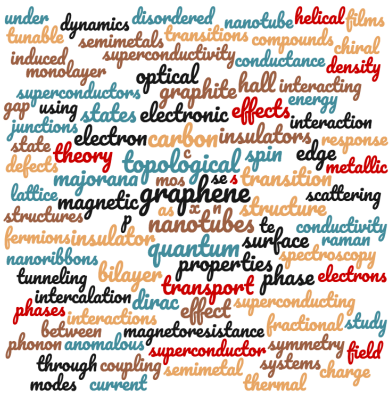
other scientific sub-fields, causing S_i to increase. The very fact that papers are co-opted in another scientific sub-domain also brings more citations to a_{ex}. If the new citing domain is highly productive, i.e., with many published papers, then S_i may eventually decline, as most of the citations will now come from the new citing field, overshadowing the original one. Eventually, while the citation rate of the paper will start to decrease as a natural consequence of ageing, its FNE, on the contrary, may increase slightly as other fields may become interested in the article and grow in relative importance.

The paper *The band theory of graphite*, published in 1947 [55] seems to follow the hypothesised pattern of exaptation. Before taking a closer look, let us first dig into the graphene background. Graphite is a material made up of carbon atoms arranged in parallel hexagonal layers. Like the diamond, it is an allotropic form of carbon. While graphite is a conductor at room temperature, the diamond is an insulator. To understand why the carbon atoms with different crystal geometries are conducting or insulating it is necessary to understand how electrons behave once an electrical potential is applied. One of the great success of quantum mechanics was the possibility to understand these phenomena. The electronic structure of simple materials became computable right after the formalism of quantum mechanics was established. The geometry of graphite suggests that its electronic properties can be incrementally determined by first analysing a single layer of carbon - what is today known as *graphene* - and then by considering the mutual interaction between layers. As Wallace wrote in his manuscript [55]:

> Since the spacing of the lattice planes of graphite is large (3.37Å) compared with the hexagonal spacing in the layer (1.42Å), a first approximation in the treatment of graphite may be obtained by neglecting the interactions between planes, and supposing that conduction takes place only in layers.

Wallace did not consider the possibility that a hexagonal layer of carbon atoms could

Word cloud for the graphene cluster



Word cloud for the ESP cluster

Figure 4: **Word clouds of the two communities citing Wallace's paper.** Word clouds [57] were generated by merging all the titles belonging to the main two clusters citing the graphite/graphene article. Stop words, numbers and punctuation marks were removed. Font size is proportional to the logarithm of the number of occurrences of words. The composition of the main cluster (graphene) with 613 articles citing Wallace's graphite article [55] is shown on the left; the second important cluster (electronic structure properties) with 33 articles is on the right.

exist by itself and used it as a mere tool to solve the "more complicated" problem for the three-dimensional graphite. As recently pointed out in a review essay on graphene [9]:

> It was P. R. Wallace, who in 1946 wrote the first papers on the band structure of graphene and showed the unusual semi-metallic behaviour in this material (Wallace, 1947). At that point in time, the thought of a purely 2D structure was a mere fantasy and Wallace's studies of graphene served him as a starting point to study graphite, a very important material for nuclear reactors in the post-World War II era.

Let us see now why the seminal paper of Wallace on graphene can be considered an example of exaptation according to the pattern highlighted above. Figure 3, shows two panels of the yearly number of citations vs. the forward normalised entropy (FNE) S_i (left), and the IPR, I_i (right). The paper starts with a small number of citations and a very small FNE (bottom-left corner of the plot in the left panel). After that, the pattern follows a very peculiar "S-shaped" trajectory, mirroring the following three situations: (i) the adoption by another field leading to an increase of S_i, (ii) the growth in the number of citations while S_i decreases, (iii) the decrease of the number of citations while S_i increases again.

This suggests that Wallace's article that was meant to belong to the field of electronic structure properties (ESP), was later co-opted in the graphene research field. According to the cluster analysis, the ESP field contains approximately 12,000 articles and the much more recent graphene area contains about 5,000 articles. Despite the larger number of articles in ESP, Wallace's paper has been cited mostly by the graphene community (613 times), and much less from the ESP (33 times). The forward normalised entropy of this paper follows the exaptation pattern we assumed. Its IPR pattern (Fig. 3 (right panel)) demonstrates how it switched from the ESP to the graphene community. To visualise the differences between the two ESP and graphene communities, in Fig. 4 we show two word-clouds from the articles citing Wallace's paper: from the graphene community (left) and the ESP one (right).

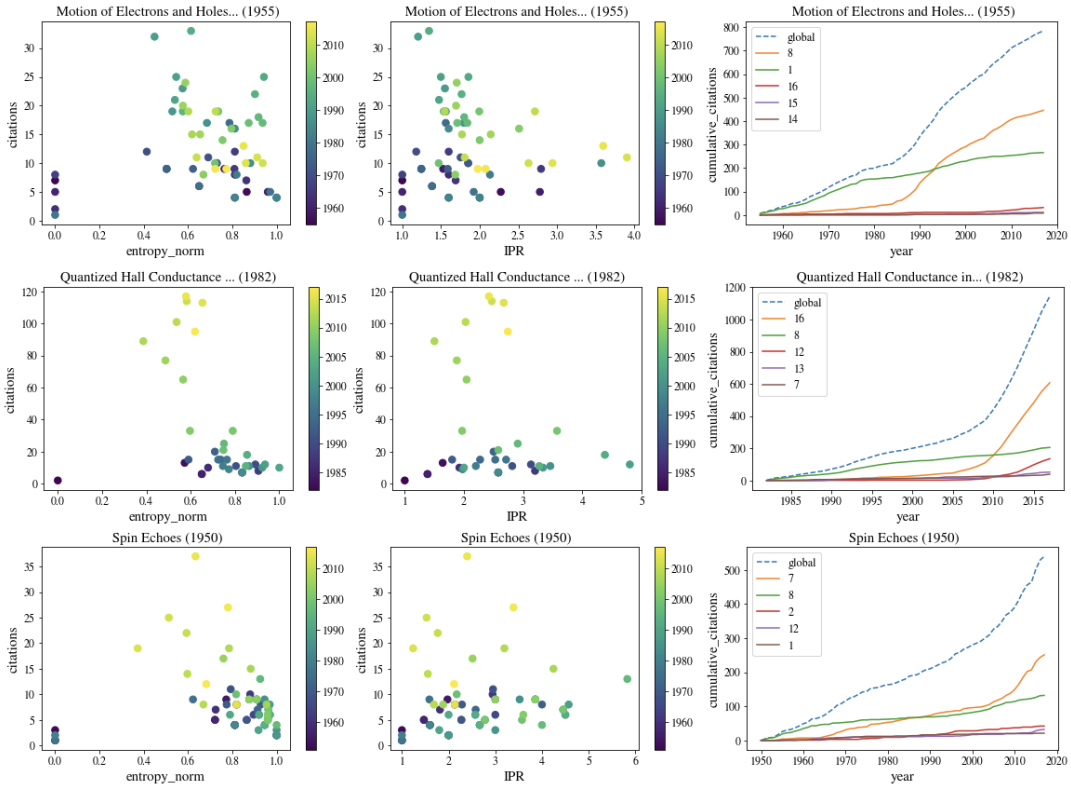


Figure 5: **More examples of exaptation patterns.** Three articles are shown: Refs. [36, 48, 22], one in each row. In the left and centre panels, we display their citation versus FNE and IPR, respectively. Patterns are similar to what was found in Fig. 3. In panels to the right, we show the time evolution of their cumulative citations in the respective clusters, whose id number is given in the legend. Their vertical order reflects their ranking in 2017, i.e., their relative importance in citing the selected paper.

We found 10 additional articles whose FNE and IPR patterns are similar to those of the graphite/graphene paper. Among those, we mention two notable examples, i.e., *Motion of Electrons and Holes in Perturbed Periodic Fields* [36] and *Quantized Hall Conductance in a Two-Dimensional Periodic Potential* [48]. A third article, that was not in the list of the most 200 cited papers but was highly cited from outside the APS, was chosen on the basis of our personal experience, *Spin Echoes* [22]. The first [36], is mainly cited by articles belonging to two clusters, the "Quantum dots / Quantum wells" cluster (QDQW) and the ESP cluster. These clusters are denoted by id=8 and id=1 respectively, and their mutual importance is sketched in the plots on the right panels of Fig. 5. Note how after 1980, the QDQW cluster starts to become more important than ESP, so that [36] acquires more importance in that field. A similar situation occurs with reference [48] (second row of the panel), where now the graphene cluster (id=16) competes with the QDQW cluster. Interestingly, from year 2010 on, also the "Bose-Einstein condensate" cluster (id=12) starts to get some importance.

The third paper on spin echoes was chosen since we expected its importance on Nuclear Magnetic Resonance (NMR) would become prominent in time (third row of the panel). We do not observe a clear NMR cluster in APS, rather we find an exaptation pattern, where the

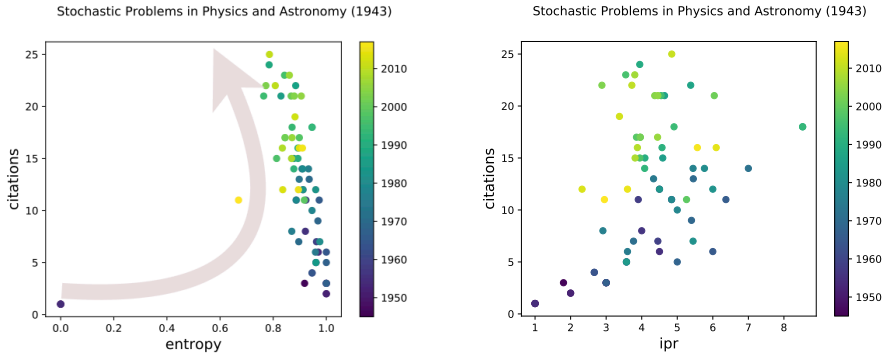


Figure 6: **FNE and IPR pattern for a typical review article in APS.** Review articles exhibit a distinctive dynamical behaviour slightly different than the exaptation patterns discussed in Figs. 3. and 5. Left: the forward normalised entropy increases quickly and remains at high values over time, while the number of citations per year grows. Right: the number of clusters citing a review article grows over time showing an increased interest in several sub-disciplines.

"Entanglement" cluster (id=7) emerges on the detriment of the QDQW cluster (id=8), as shown in the bottom right panel. We think that the NMR community cites this paper from outside the APS community, i.e., by papers not published by the APS journals.

Although not directly related to the problem of detecting exaptation in APS articles, it is worthwhile looking at the time evolution of the FNE and the IPR for review articles. We observe the same kind of pattern consistently in all highly cited APS review papers. As an example, we show the Chandrasekhar's review *Stochastic problems in physics and astronomy* [11]; the corresponding plots for FNE and IPR are reported in Fig. 6. Both FNE and IPR increase over time, witnessing the number of different scientific fields interested in the review. The manuscript itself [11] features four papers, the problem of random flights, the theory of the Brownian motion, probability after effects, and probability methods in stellar dynamics. The IPR value reaches the value four soon and oscillates around it in time. On the other hand, the FNE, after reaching its maximum value, starts to decrease, mainly because one citing field increases its importance over the others. The IPR and FNE behaviour of highly cited reviews is essentially different from the patterns found for co-opted articles presented in Figs. 5 and 3. This finding further reassures us of the robustness of the proposed indicators, as well as of the overall methodology.

4 Discussion

The concept of exaptation has not been appropriately quantified. Only very few systematic quantitative analyses are available in the literature. Here we focused on the exaptation phenomena in the framework of scientific progress as it can be observed in the APS article collection database. Our approach consists of looking for signatures of exaptation in the adoption (as reflected in citation patterns) of specific results by other scientific communities than the original scientific area of belonging. To make the analysis quantitative, we propose a method based on two main components: clustering of publications based on citation relations, and two observables, the Forward Normalised Entropy (FNE) and the Inverse Participation Ratio (IPR). These measures allow to single out patterns that are related to the exaptation phenomena in scientific evolution.

The evolution of feathers is a classical example of exaptation. Feathers were initially adapted to protect against thermal excursions and were later co-opted for flying. As soon as birds learn to fly, their fitness improves since they have higher chances of surviving predators. We extend this reasoning with a thought experiment (*Gedankenexperiment*) to the exaptation of scientific knowledge, where citations represent the fitness and the inverse participation ratio represents the number of functions a publication acquires over time. The example of graphene constitutes a popular instance of exaptation in physics, particularly in contributing to the development of two related sub-fields, electronic structure properties, and graphene. It shows that the 'survival' of a field depends on how pre-existing concepts are re-used or applied by communities in new domains resulting in new functionalities.

These results should be seen as preliminary steps towards a quantitative theory of exaptation; they show that is in principle possible to quantify exaptation, once it becomes possible to get a quantitative handle on the evolving context of a field. A better theory of exaptation is necessary to explain how something emerges and evolves by exploring its *creative potential* [2]. The creative potential has been described as the functional shift of a particular component that could open up an evolutionary path different from its original and perhaps intentional trajectory [2, 24]. We think that this paper could stimulate a new wave of studies in this promising area of scientific research.

Acknowledgements

This work was supported by the Austrian Research Promotion Agency FFG under grant 857136.

References

[1] P. Andriani, A. Ali, and M. Mastrogiorgio. Measuring exaptation and its impact on innovation, search, and problem solving. *Organization Science*, 28(2):320–338, 2017.

[2] P. Andriani and G. Cattani. Exaptation as source of creativity, innovation, and diversity: Introduction to the special section. *Industrial and Corporate Change*, 25(1):115–131, 2016.

[3] P. Andriani and J. Cohen. From exaptation to radical niche construction in biological and technological complex systems. *Complexity*, 18(5):7–14, 2013.

[4] W.B. Arthur. The structure of invention. *Research policy*, 36(2):274–287, 2007.

[5] A. Barve and A. Wagner. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, 500(7461):203, 2013.

[6] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[7] W.J. Bock. Preadaptation and multiple evolutionary pathways. *Evolution*, 13(2):194–211, 1959.

[8] G. Bonifati. 'more is different', exaptation and uncertainty: three foundational concepts for a complexity theory of innovation. *Economics of Innovation and New Technology*, 19(8):743–760, 2010.

[9] A.H. Castro Neto, F. Guinea, N.M.R. Peres, K.S. Novoselov, and A.K. Geim. The electronic properties of graphene. *Rev. Mod. Phys.*, 81:109–162, Jan 2009.

[10] G. Cattani. Preadaptation, firm heterogeneity, and technological performance: A study on the evolution of fiber optics, 1970–1995. *Organization Science*, 16(6):563–580, 2005.

[11] S. Chandrasekhar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15:1–89, Jan 1943.

[12] W.M. Cohen and D.A. Levinthal. Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly*, 35(1):128–152, 1990.

[13] C. Darwin. *On the origin of species, 1859*. Routledge, 2004.

[14] N. Dew and S.D. Sarasvathy. Exaptation and niche construction: behavioral insights for an evolutionary theory. *Industrial and Corporate Change*, 25(1):167–179, 2016.

[15] N. Dew, S.D. Sarasvathy, and S. Venkataraman. The economic implications of exaptation. *Journal of Evolutionary Economics*, 14(1):69–84, 2004.

[16] R.D. Dewar and J.E. Dutton. The adoption of radical and incremental innovations: An empirical analysis. *Management science*, 32(11):1422–1433, 1986.

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

[18] L. Fleming and O. Sorenson. Science as a map in technological search. *Strategic Management Journal*, 25(8-9):909–928, 2004.

[19] APS Data Sets for Research. `http://journals.aps.org/datasets`.

[20] L. Gabora, E.O. Scott, and S.A. Kauffman. A quantum model of exaptation: Incorporating potentiality into evolutionary theory. *Progress in Biophysics and Molecular Biology*, 113(1):108–116, 2013.

[21] S.J. Gould and E.S. Vrba. Exaptation—a missing term in the science of form. *Paleobiology*, 8(1):4–15, 1982.

[22] E.L. Hahn. Spin echoes. *Phys. Rev.*, 80:580–594, Nov 1950.

[23] C.A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[24] S.A. Kauffman. *The origins of order: Self-organization and selection in evolution*. OUP USA, 1993.

[25] S.A. Kauffman. *Investigations*. Oxford University Press, 2000.

[26] R. Klavans, K.W. Boyack, A.A. Sorensen, and Chaomei Chen. Towards the development of an indicator of conformity. In *14th international society of scientometrics and informetrics conference. ISSI*, 2013.

[27] P. Klimek, R. Hausmann, and S. Thurner. Empirical confirmation of creative destruction from world trade data. *PloS ONE*, 7(6):e38924, 2012.

[28] P. Klimek, S. Thurner, and R. Hanel. Evolutionary dynamics from a variational principle. *Phys. Rev. E*, 82:011901, Jul 2010.

[29] T.S. Kuhn. The structure of scientific revolutions. *Chicago and London*, 1962.

[30] V. Larivière and Y. Gingras. On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology*, 61(1):126–131, 2010.

[31] V. Larivière, S. Haustein, and K. Börner. Long-distance interdisciplinarity leads to higher scientific impact. *Plos one*, 10(3):e0122565, 2015.

[32] D.A. Levinthal. The slow pace of rapid technological change: gradualism and punctuation in technological change. *Industrial and corporate change*, 7(2):217–247, 1998.

[33] J.M. Levitt and M. Thelwall. Is multidisciplinary research more highly cited? a macrolevel study. *Journal of the American Society for Information Science and Technology*, 59(12):1973–1984, 2008.

[34] E.A. Lloyd and S.J. Gould. Exaptation revisited: changes imposed by evolutionary psychologists and behavioral biologists. *Biological Theory*, 12(1):50–65, 2017.

[35] V. Loreto, V.D.P. Servedio, S.H. Strogatz, and F. Tria. Dynamics on expanding spaces: modeling the emergence of novelties. In *Creativity and universality in language*, pages 59–83. Springer, 2016.

[36] J.M. Luttinger and W. Kohn. Motion of electrons and holes in perturbed periodic fields. *Phys. Rev.*, 97:869–883, Feb 1955.

[37] J. Mokyr. Evolutionary biology, technological change and economic history. *Bulletin of Economic Research*, 43(2):127–149, 1991.

[38] M.E.J. Newman. The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, 86(6):68001, 2009.

[39] E.J. Rinia, T.N. Van Leeuwen, and A.F.J. Van Raan. Impact measures of interdisciplinary research in physics. *Scientometrics*, 53(2):241–248, 2002.

[40] E.M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[41] M.F. Rosenman. Serendipity and scientific discovery. *Journal of Creative Behavior*, 22(2):132–38, 1988.

[42] J.A. Schumpeter. *Business cycles*, volume 1. McGraw-Hill New York, 1939.

[43] V.D.P. Servedio, P. Buttà, D. Mazzilli, A. Tacchella, and L. Pietronero. A new and stable estimation method of country economic fitness and product complexity. *Entropy*, 20(10):783, 2018.

[44] P. Stephan, R. Veugelers, and Jian Wang. Reviewers are blinkered by bibliometrics. *Nature News*, 544(7651):411, 2017.

[45] C.R. Sugimoto and V. Larivière. *Measuring research: what everyone needs to know*. Oxford University Press, 2018.

[46] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries' fitness and products' complexity. *Scientific reports*, 2:723, 2012.

[47] Siang Yong Tan and Yvonne Tatsumura. Alexander fleming (1881–1955): discoverer of penicillin. *Singapore medical journal*, 56(7):366, 2015.

[48] D.J. Thouless, M. Kohmoto, M.P. Nightingale, and M. den Nijs. Quantized hall conductance in a two-dimensional periodic potential. *Phys. Rev. Lett.*, 49:405–408, Aug 1982.

[49] S. Thurner, P. Klimek, and R. Hanel. Schumpeterian economic dynamics as a quantifiable model of evolution. *New Journal of Physics*, 12(7):075029, jul 2010.

[50] S. Thurner, W. Liu, P. Klimek, and S.A. Cheong. The role of mainstreamness and interdisciplinarity for the relevance of scientific papers. *ArXiv*, abs/1910.03628, 2019.

[51] F. Tria, V. Loreto, V.D.P. Servedio, and S.H. Strogatz. The dynamics of correlated novelties. *Scientific reports*, 4:5890, 2014.

[52] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

[53] A.F.J. Van Raan. Sleeping beauties in science. *Scientometrics*, 59(3):467–472, 2004.

[54] C.S. Wagner, J.D. Roessner, K. Bobb, J.T. Klein, K.W. Boyack, J. Keyton, I. Rafols, and K. Börner. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of informetrics*, 5(1):14–26, 2011.

[55] P.R. Wallace. The band theory of graphite. *Phys. Rev.*, 71:622–634, May 1947.

[56] P. Watson. *Convergence: The idea at the heart of science*. Simon and Schuster, 2017.

[57] WordClouds.com. Free online word cloud generator and tag cloud creator. `http://wordclouds.com`. accessed Jan. 2020.

Chapter 4

The geography of knowledge diffusion

Scale-free growth in regional scientific capacity building explains long-term scientific dominance

Vito D. P. Servedio, Márcia R. Ferreira, Niklas Reisz, Rodrigo Costas, and Stefan Thurner

Published in: Chaos, Solitons & Fractals, 167:113020, 2023

Chaos, Solitons and Fractals 167 (2023) 113020

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

journal homepage: www.elsevier.com/locate/chaos

Scale-free growth in regional scientific capacity building explains long-term scientific dominance

Vito D.P. Servedio [a,1], Márcia R. Ferreira [a,1], Niklas Reisz [a], Rodrigo Costas [b,c], Stefan Thurner [a,d,e,*]

[a] Complexity Science Hub Vienna, Josefstädter Str. 39, 1080 Vienna, Austria
[b] Centre for Science and Technology Studies (CWTS), Leiden University, Willem Einthoven Building, Kolffpad 1, 2333 BN Leiden, The Netherlands
[c] DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, RW Wilcocks Building, Stellenbosch, Western Cape, South Africa
[d] Section for Science of Complex Systems, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, A-1090 Vienna, Austria
[e] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

ARTICLE INFO

Keywords:
Science of science
Preferential attachment
Critical mass
Scaling

ABSTRACT

The regional capability of performing front-running research and technological development has been identified as a necessary condition for future wealth creation. The quality and amount of scientific capabilities in specific fields vary dramatically across different world regions. Capabilities, knowledge, and skills are embodied by scientists working in research institutions or companies. The conditions for the emergence of a leading regional scientific environment – and the resulting early technological leadership – are poorly understood. The existence of a critical mass – the threshold above which a region can build comparably strong scientific capabilities – of scientists is often assumed. Using a unique dataset of global scientific activity and researcher mobility over several decades, we present empirical evidence in three scientific areas (semiconductor research, embryonic stem cells, and Internet research) that the process of scientific knowledge accumulation is remarkably general and applies to practically all regions. Regional knowledge accumulation data obtained from an analysis of scientists' geolocated trajectories follow a preferential attachment mechanism characterized by a sub-linear kernel with a robust growth exponent. Scale-free growth patterns suggest that regions that move early into new technologies tend to dominate the corresponding scientific fields. We find no evidence that critical mass is required to achieve prolonged scientific dominance. We propose a simple preferential attachment model that explains the empirical data and allows us to understand deviations from the growth exponent as focused interventions to strategically attract scientists at regional level. We demonstrate this explicitly for China in the three scientific fields examined.

1. Introduction

It has long been known that a region's ability to attract scientists is key to its future scientific and economic development [1]. Far less is known about how new fields of science are created and how a leading edge is achieved, established, and maintained. Are there optimal strategies for regional institutions and stakeholders to achieve global leadership and thus economic benefits? In this context, the success story of Silicon Valley comes to mind, in particular, what led to the development of semiconductor science that later made the region the largely unchallenged leader in these technologies for many decades [2]. Why could not other regions do the same? To explain what it takes for regions to accumulate scientific capacity and sustain decades of dominance, it is often argued that a minimum number of researchers is required to grow and create a self-sustaining capacity-a *critical mass* [3,4].

Global scientific mobility and international collaboration [5] further reinforce the spatial concentration of knowledge, favoring regions that are becoming major scientific hubs [6]. Silicon Valley is often cited as a canonical example of the regional growth and dominance of high-tech industry and scientific knowledge. It developed in the 1950s when the U.S. military nurtured companies in the state of California with multibillion-dollar contracts. Electrical and electronics companies then settled in Santa Clara to take advantage of the city's location near defense-related aircraft, missile, and space markets [2]. Before this,

* Corresponding author at: Section for Science of Complex Systems, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, A-1090 Vienna, Austria.
E-mail address: stefan.thurner@meduniwien.ac.at (S. Thurner).
[1] These authors contributed equally to the paper.

https://doi.org/10.1016/j.chaos.2022.113020
Received 20 July 2022; Received in revised form 8 November 2022; Accepted 10 December 2022

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

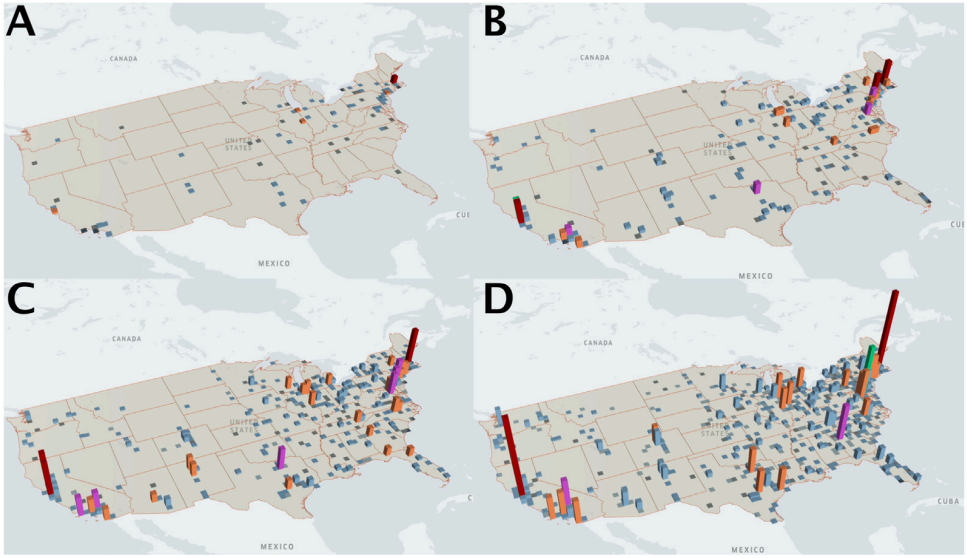


Fig. 1. Regional evolution of the number of in-flowing researchers into the field of semiconductors in the US (height of bars) Panels A–D correspond to time periods 1954–1970, 1970–1986, 1986–2002, and 2002–2020, respectively. In the fifties and sixties, only the Boston area plays a role, some researchers are present in Chicago and Silicon Valley. From the seventies on, the Boston area and Silicon valley are dominant; other regions are building up some expertise but never get into a position of becoming a challenger. The five colors (blue, orange, purple, green, and red) indicate steps of 20% of the aggregated sum of authors' flows value of the inflows at each time period. The respective maximum values for panels A–D are: max = 57, 207, 1086, and 8270. For the definition of a region, see methods Section 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the San Francisco Bay Area was already a place where the electronics industry existed, which shows that the Bay Area needed some time to build momentum [7]. While there is little doubt that Silicon Valley has evolved from a productive environment to one of the leading areas for high-tech industries and scientific knowledge, producing important new technologies and industrial clusters, the underlying mechanisms and critical parameters for this development are still unclear.

With ever more complete bibliographic databases available, it is possible to reconstruct the historical buildup of regional capacities in specific fields in all regions of the world. This is possible by reconstructing the mobility of individual scientists and their productivity in terms of individual papers in specific scientific fields [5,8]. Researchers can be reliably tracked in time and space by the publication date of articles and the history (temporal sequence) of the affiliations where they produced their scientific output. Recent advances in bibliographic databases have also greatly improved the consistency and quality of metadata extracted from publications, with particular attention to author-affiliation links [5,8].

Here, we use the *Dimensions* database, see Appendix A. It contains millions of publications and data spanning several decades. The database includes about 20 million disambiguated researchers associated with a researcher ID. The database contains links to the Global Research Identifier Database, which currently covers more than 98,000 research institutions worldwide [9,10]. *Dimensions* currently covers more local journals than any other large-scale bibliographic database such as *Web of Science* or *Scopus*. Its broader scope allows us to include more organizations with better local resolution [9].

With the *Dimensions* dataset, we can reliably reconstruct the changes in affiliations of all researchers in a specific field and infer how they move between geographic regions. We can see which region attracts researchers with certain scientific skills in a time-resolved manner. We see how quickly different regions accumulate specific types of skills over time, which allows us to define regional growth rates of scientific skills. A schematic picture of semiconductor researchers moving into different regions over seven decades is shown in Fig. 1. The reasons why researchers switch locations are plentiful. In addition to the attractiveness of a region and the everyday surroundings, the scientific environment and the innovation potential also play an important role [6,11]. This also includes the number of scientists already present in a region [12].

This paper studies the growth mechanism leading to the observed temporal and regional distributions of accumulated scientific skills in three specific scientific fields: Semiconductor Research, Embryonic Stem Cells, and Internet Research. We identify the underlying mechanism as the so-called "rich-get-richer" effect and find no evidence of a critical mass of scientists. In particular, in our analysis of empirical flows of scientists (and their skills), we find strong evidence for a sub-linear preferential attachment (PA) growth mechanism, meaning that a region becomes increasingly attractive to scientists as the number of existing scientists in a field increases.

Preferential attachment mechanisms (or rich-get-richer or Mathew effect) have been identified in a wide variety of social [13–16] and network phenomena [17–19]. For a given quantity of interest (integer), $x(t)$, at time t, growth following a PA mechanism means that the probability of the quantity gaining one unit within the next timestep is

$$\mathcal{P}[x(t+1) = x(t) + 1] \propto x(t)^{\alpha} . \tag{1}$$

Preferential attachment can appear in linear, sub- and super-linear versions, depending on whether the growth exponent $\alpha = 1$, $\alpha < 1$, or $\alpha > 1$, respectively. Sub-linear PA growth has been reported, for example, in the actor collaboration network and scientific co-authorship networks [18,19], as well as in the friendship network formation in a massive multiplayer online computer game [20]. The microscopic mechanisms underlying a sub-linear PA dynamics may be of various types, for example it could be an ageing effect, where scientists prefer newborn institutions, or it could be a fitness-related attachment, where researchers prefer institutions with a high impact, or it could be hiring selection rules, where institutions may lower the entry requirements for the candidates. A few publications consider these effects in their modeling efforts [21–24] but do not explicitly relate them to a sub-linear *effective* PA kernel. In our context we observe that a PA mechanism does not need the knowledge of the entire network for the researchers to move to a region. Picking random scientists (colleagues) and copying their behavior leads to a PA dynamics. In our study, we do not consider

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

Table 1
Descriptive features of the data set used in the three studied scientific fields.

	Semiconductors	ESC	Internet
Starting year	1941	1941	1956
End year	2019	2019	2019
Researchers	2,011,170	752,575	109,098
Articles	5,062,639	1,083,100	246,953
Regions	1633	1161	1032
Heaps' exp. γ	0.39	0.37	0.48
PA exponent α	0.79	0.84	0.79

the details of the causes leading to an effective power-law sub-linear PA. We focus on how to measure it and demonstrate its importance for explaining the regional growth of scientific fields.

To test if the conclusion of an underlying effective sub-linear PA mechanism is valid, we design a simple global model of the regional attractiveness of regions to scientists. The model not only replicates the empirical frequency distribution of researchers in different areas but also helps us understand the peculiar behavior of regions that are late adopters of new scientific fields. We illustrate the case of late adopters in the context of semiconductor research, where Silicon Valley intellectually dominated the scene for a long period, from the end of the 1940s on [25]. Only very late China entered the scene, starting in the early 1980s.

From 2006 on, China became the most dominant player in semiconductor research, also thanks to the Chinese "Medium- and Long-Term Plan for the Development of Science and Technology" [26]. We find that the cumulative number of scientists moving to China and publishing scientific papers exceeds those of California from 2007 on. Understanding how this take-over was possible needs special attention — a naive explanation with PA growth will not be sufficient. Change of leadership is only possible if special action is taken to locally change the PA mechanism with the help of strategic interventions that make regions radically more attractive by investments in technology [27] or research and development [28]. We show that our model can implement these interventions and we quantify how large these efforts have to be to allow regional latecomers to become dominant.

2. Results

We first characterize the mobility of scientists, in particular the growth (and change) of numbers of scientists of a given field working in different world regions. This is based on analyzing the temporal sequences of their publications and affiliations. We define a region as the first-level administrative division of a country. In the case of the United States, such a region corresponds to a state. For the data collection procedure, see Appendix A; for an overview of the data set, consult Table 1.

To arrive at a practical data structure for the analysis, we define an event, $E(T)$, as a publication of an article in a given field of research at time T (date of publication) by a scientist. It consists of four entries, $E(T) = (S, R, T, F)$, the scientist, S, the region, R, with the condition that S has published in the region R for the first time (i.e., the pair (S, R) has never occurred before, and the scientific field is F. We skip an event for which this condition is not fulfilled. We ensure that we only record the scientists that move into a new region or who start their careers. Our work focuses on academic mobility from the perspective of receiving countries and regions. Since we mainly focus on how regions receive new people in specific scientific domains, we do not account for returned mobility (i.e., people who return to a location where they previously have become "attached" to). Post-migration retention, attrition, and returning are highly significant dimensions of the knowledge transfer equation and are ideal for follow-up studies focusing on questions of *brain circulation* and strategies to mitigate the effects of *brain drain*. We do not deal with these themes here.

Table 2
Sequence of publication events, E, in the field of physics. Intrinsic time increases by one each time a scientist publishes an article in a given region for the first time. From this data, all necessary sequences can be extracted. In the specific example, Bob has already published in Arizona at intrinsic time $t = 2$, so all his future publications in Arizona no longer appear in the stream. Using initials for regions, this example generates the regional stream, $\mathcal{R} = (A, A, A, C, A, D, C)$.

Paper	Real time	intr. time	Scientist	Region	Field
Article 1	1960	1	Alice	Arizona	phys.
Article 1	1960	2	Bob	Arizona	phys.
Article 1	1960	3	Charlie	Arizona	phys.
Article 2	1961	4	David	California	phys.
Article 3	1961	–	Bob	Arizona	phys.
Article 3	1961	5	David	Arizona	phys.
Article 4	1962	6	Alice	Delaware	phys.
Article 5	1963	–	Bob	Arizona	phys.
Article 5	1963	–	David	Arizona	phys.
Article 5	1963	–	Alice	Arizona	phys.
Article 6	1964	7	Bob	California	phys.

One publication may generate multiple events according to the number of scientists in the author list. Here, we consider a full counting of all authors and their affiliations (and regions) recorded in the publications. Note that a "fractionalization" of researchers' and regions' contributions to papers, where researchers and regions are assigned a relative weight, makes little sense in the present context since we build sequences of events. Moreover, fractional counting is usually not considered when the analysis is conducted for a single discipline, nor when studying mobility flows, as in our case [29–31]. After sorting events, $E(T)$, ascending in time, we get a sequence, E_t, where t is an index that indicates the event's position in the sequence. We call t the *intrinsic time*. For every index, t, there is an associated time, T, given by a non-decreasing function $T(t)$. Intrinsic time is a convenient way for treating sequences of events whenever their rate of appearance in real-time is not constant. This is indeed the case since, in most fields, the number of published articles increases exponentially in real-time [21], and, correspondingly, the real-time difference between two consecutive articles decreases exponentially. We can consider the derivative of intrinsic time with respect to real-time, as a proxy for attractiveness of new locations. We increase intrinsic time whenever someone moves to a new location in their life. This means that the field has to be sufficiently attractive to justify a person with their family to move or to justify the addition of a new affiliation to one's list. From the sequence, E_t, we extract the corresponding sequences of regions, R_t, which we call *regional stream*, $\mathcal{R}$. For example, see Table 2.

Given a regional stream, one defines two useful quantities. First, the cumulative number of new scientists who moved into the region i (or started their publishing career there), before time t

$$k_i(t) = \sum_{\tau=1}^{t} \delta(R_\tau, i)\,, \tag{2}$$

where $\delta(x, y)$ is the Kronecker symbol, $\delta(x, y) = 1$ if $x = y$ and zero, otherwise. The second quantity is the number of different regions appearing in the regional stream before time t

$$D(t) = \sum_{\tau=1}^{t} \delta(k_{R_\tau}(\tau), 1)\,. \tag{3}$$

$D(t)$ is the number of regions appearing at least once and resembles the "regional diversity" of a field. We only focus on the number of papers published within a region and ignore their impact and quality.

The sub-linear PA kernel. From the time evolution of the quantities, $k_i(t)$, we can estimate the probability for a scientist to move to a new location of work (and publish there for the first time) that has already had k scientists up to time $t - 1$:

$$P_k(t) = \mathcal{P}\left[k_i(t) = k + 1 \mid k_i(t-1) = k\right]\,. \tag{4}$$

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

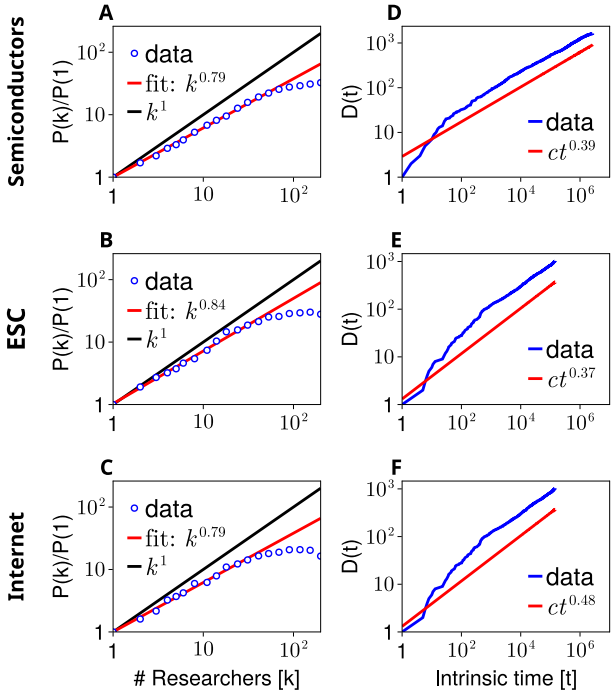


Fig. 2. Growth statistics of the three fields. Panels A-C show the probability for a scientist to move into a region where k researchers are present. The attachment probability, $P(k)/P(1)$ (PA kernel), (blue) shows a sub-linear power-law increase as a function of k. The least-square fits are shown in red, the black line indicates the linear exponent, $\alpha = 1$. (A) shows the case for semiconductor science, B for ESC, and (C) for Internet research. For the sake of readability, in panels A-C we show only data for small k where we find an approximately linear relation in the log–log plot. We do not show data for $k > 200$ where curves drop due to insufficient statistics. Panels D–F shows the increase of regional diversity, $D(t)$, occurring in the regional stream, $\mathcal{R}$, as a function of intrinsic time, t (blue). After an initial transient, $D(t)$ approaches an approximate power-law (Heaps' law). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

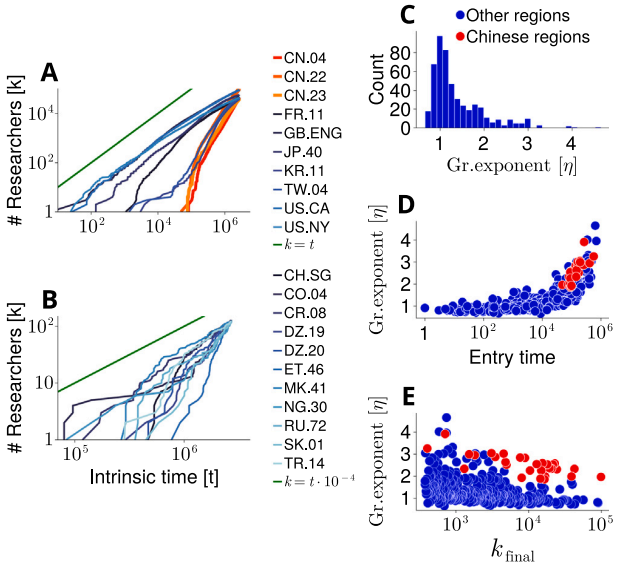


Fig. 3. Panels A and B show the cumulative number of scientists coming into a region (lines) in intrinsic time in semiconductor research; (A) shows the 10 most visited regions. Most regions grow sub-linearly. Chinese regions are shown with red thick lines and increase super-linearly. In (B), we see 11 poorly visited regions ranked from 700 to 710 regarding the number of scientists. Curves are (additively) shifted to the left so that their first point is placed at coordinates (1,1); In both (A) and (B), the green straight lines on the top represent the linear regime. (C) Histogram of growth exponents, η, for all regions; (D) Growth exponents, η, of all regions, i, as a function of their entry time, t_i (intrinsic time). Note that Chinese regions (red) come in late (after timestep 10^5) and have higher exponents than other regions (blue). (E) Growth exponents, η, as a function of the cumulative number of scientists in 2019, k_{final}. The high exponents of Chinese regions are visible (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For simplicity, we assume that $P_k(t)$ does not explicitly depend on time, t, but on the number of occurrences, k, of the regions in the regional stream, and we write $P_k(t) \approx P(k)$. We will see that this approximation already explains the experimental data well. We estimate $P(k)$ with the running histogram method [18], briefly described in the methods Section 4.

In Fig. 2 panels A, B, and C, we show the empirical settlement probabilities, $P(k)$, for a scientist (in the fields of semiconductor research, embryonic stem cells, and Internet research, respectively) to move to a new location of work and publish there for the first time — as a function of the number of scientists who already work in that region, k. The probability is normalized by $P(1)$; we call $P(k)/P(1)$ the effective *attachment kernel* [32].

Clearly, with increasing numbers of already present scientists, a power-law increase of the attachment kernel is visible. The attachment exponents, defined in Eq. (1), are $\alpha \sim 0.79$ for semiconductor and Internet research, and 0.84 for ECS (see Table 1). For reference, the black lines indicate the linear PA mechanism, i.e., $\alpha = 1$.

Panels D, E, and F of Fig. 2 show the number of different regions appearing in the stream $D(t)$ as a function of intrinsic time. After a brief linear transient that ends around $t \approx 100$ publications in all three fields, D increases as a power-law, $D(t) \approx t^{\gamma}$, with $\gamma < 1$. The initial linear growth, which is the fastest growth possible in intrinsic time, reflects the fast diffusion of the three scientific fields across the world at their onset. As time goes on, the initial constant rate, $\frac{d}{dt}D(t)$, turns into a time decreasing regime with an exponent $\gamma - 1$. We observe no saturation effects at high intrinsic times despite the number of different regions in the three cases has crossed (semiconductors, embryonic stem cells), or is close to (Internet), the 50% of the total number of regions in the world, i.e., 2092 in the year 2019.

A power law increase in novel entries in a stream has been associated with the existence of the so-called *Heaps' law* that often appears in evolutionary time series [33–37]. In the present case, Heaps' law shows how fast a new scientific field spreads in regions around the globe.

Growth of regional capacity. In Fig. 3, we show the cumulative number of semiconductor scientists working in several selected regions in intrinsic time. We do not distinguish those scientists leaving a region for a new one from those who start in a region for the first time in their career in the semiconductor field. We fit with a least square procedure the regional growth curves with $k(t) \sim t^{\eta}$, where η is the *growth exponent*, that can be interpreted as the *regional growth capability*. In all the three scientific fields we find $R^2 > 0.9$ for the η fits, with very few values between 0.85 and 0.90.

In panel A, every curve represents a region that belongs to one of the 10 most successful ones (cumulative number of scientists in the order of 10^5). Most curves grow sub-linearly with the exception of the three regions in China (denoted by CN.{04,22,23}); these grow faster than linear (left-bent). Both California (US.CA) and England (GB.ENG) have an exponent of approximately 0.8 (yellow and green curves practically overlap), while the Chinese region around Beijing (CN.22) shows an exponent of about 1.8. This means that Chinese regions follow a very different growth pattern. Note that Chinese regions entered the scene only after about 90,000 papers were published in the field, corresponding to the year 1977.

In panel B, we show the growth curves of 11 low-ranked regions (rank 700–710). These all have a final cumulative number of scientists of about 120. All of them grow super-linearly with an exponent of

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

approximately 1.6. The general trend is that well-established regions in a field generally grow sub-linearly (curves bent to the right), while late adopters grow super-linearly.

Fig. 3C shows the histogram of the growth exponents, η, for the semiconductor case. We find a distribution with mean $m = 1.55$, standard deviation $s = 0.69$ and skewness $b = 1.52$. For the other fields the situation is similar: Internet: $m = 1.36$, $s = 0.59$ and $b = 1.48$; embryonic stem cells: $m = 1.56$, $s = 0.86$ and $b = 1.73$.

In Fig. 3D we show the growth exponents, η, of all world regions, i, as a function of their entry time, t_i, defined as the time (measured in intrinsic time) at which a region appears in the affiliation list of an article for the first time. Exponents are larger the later a region enters. Late adopter regions seem to bring in scientists faster than regions where the field started. Yet, in most cases, higher exponents are not enough to catch up and challenge the leaders in the field. Chinese regions are marked in red. They come in late (10^5) and have exponents higher than other regions (blue).

Finally, in Fig. 3E we see the regional growth exponents, η, as a function of the cumulative number of scientists up to year 2019 in a region, k_{final}. The visual general decay means that the more scientists are present in a region, the lower is its growth capability. Note again the exceptionally high growth exponents for the Chinese regions.

Relations between the three exponents. The exponents α, γ, and η are related. The corresponding functional expressions can be estimated with simple reasoning, see Appendix B, or are obtained through an approximate analytic solution of the model, see Appendix C.

A simple generative model. To understand the observed statistical features presented in Figs. 2 and 3 we devise a simple model. We first specify a scientific field, e.g., semiconductor research. We use the empirical regional stream, $\mathcal{R}$, as the baseline to build a new synthetic stream, $\mathcal{S}$. We start at $t = 0$ with the first region, R_0, where a scientist's affiliation in an article appeared for the first time. We insert this region in $\mathcal{S}$ as its first element S_0. For each following intrinsic time, $t > 0$, we add an element S_t in $\mathcal{S}$ in the following two ways according to whether R_t has already appeared in $\mathcal{R}$ or not:

- if region R_t has never appeared in $\mathcal{R}$ before, i.e., $k_{R_t}(t) = 1$, we also insert it in $\mathcal{S}$, so that $S_t \equiv R_t$;
- if region R_t has already appeared in $\mathcal{R}$, i.e., $k_{R_t}(t) > 1$, we randomly select a region s, from those regions already in the stream $\mathcal{S}$, with a preferential attachment probability

$$\begin{aligned} P_k(t) &= \mathcal{P}\left[k_s(t) = k+1 \mid k_s(t-1) = k\right] \\ &= Z^{-1}k^{\alpha}, \end{aligned} \tag{5}$$

where $Z(t) = \sum_{s=1}^{D(t)} k_s^{\alpha}(t)$ is a normalization term. In other words, we choose regions with a probability proportional to a power, α, of their number of occurrences in $\mathcal{S}$ until time $t-1$.

Since the entry times of new regions in the two streams, $\mathcal{R}$ and $\mathcal{S}$, are the same by construction, the number of different regions in time, $D(t)$, coincides in the two streams, meaning that also Heaps' law is the same (Fig. 4B). The model has one free parameter, the PA attachment exponent, α, of the sub-linear PA mechanism. Note that this model is similar to the one presented in [38] where, however, the PA was linear. The model can be approximately solved analytically; see Appendix C.

For the bulk of the model simulations, we chose α as measured in the data, i.e., $\alpha = 0.79$, 0.79, and 0.84 for the Internet, semiconductors, and stem cell areas, respectively; see Table 1. For this choice, a run of the model for the semiconductor research followed by a simple linear regression yields $\gamma \approx 0.39$ and $\eta \approx 0.74$.

Understanding exceptional super-linear regions. To model the exceptional super-linear growth of Chinese regions in the field of semiconductors, visible as the red curves in Fig. 3A, we use as an input a slightly larger PA exponent $\alpha_{\text{CN}} = 0.915$ for these regions only, while for all the other regions we keep $\alpha = 0.79$, see the methods Section 4.

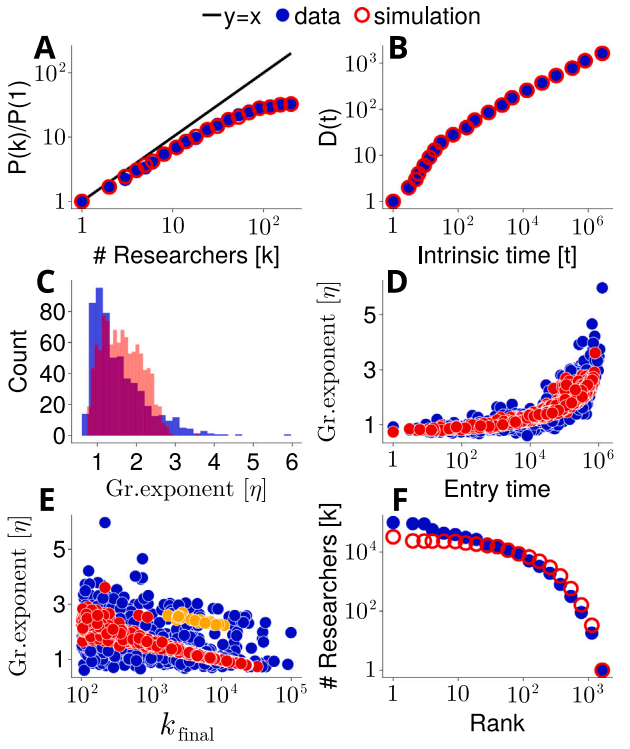


Fig. 4. Comparison of model results (red) with experimental data (blue) for the field of Semiconductors. (A) PA kernel; data and simulation practically coincide; the black line depicts a linear kernel. (B) Heaps's law; the two curves coincide by construction. (C) Histograms of the regional growth exponent, η. Simulations show a slightly higher mean. (D) Exponents, η, reproduce the empirical increase as a function of entry time. (E) The decrease of η as a function of the cumulative number k_{final} of scientists in the year 2019 is well captured by the model. Note the fact that Chinese regions (orange circles) are predicted with higher than usual values. (F) Also, the frequency-rank distributions of the regions appearing in the two streams practically coincide except for very small ranks. In panels C, D, and E, only regions with $k_{\text{final}} > 100$ are taken into account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 4, we compare the model results (red) with the empirical data (blue). In panel A, we show the PA kernel, as in Fig. 2, panels A, B, and C. In the simulation, we use the empirical value of the PA exponent of $\alpha = 0.79$ for all the 1633 regions (except the 31 Chinese regions for which we use $\alpha_{\text{CN}} = 0.915$). The two curves almost coincide even for $k > 60$, i.e., outside the interval used to fit α. This confirms the goodness of the fit.

In Fig. 4B, we plot the number of distinct regions, $D(t)$, as a function of intrinsic time, in the regional, $\mathcal{R}$, and the synthetic stream, $\mathcal{S}$. Since the entry time of new regions is the same in both streams, the two curves are identical.

Fig. 4C compares the distribution of the growth exponents, η. In the simulation, the distribution has a somewhat higher mean than the data and lower skewness; we get mean $m = 1.63$, standard deviation $s = 0.49$ and skewness $b = 0.28$. For the data we found $m = 1.55$, $s = 0.69$ and $b = 1.52$. The model under-populates the large exponents and underestimates the small ones.

In Fig. 4D, the growth exponent at the entry time of a given region is plotted against the entry time in intrinsic time. The simulation results explain the data, particularly that latecomers tend to have higher growth exponents.

Fig. 4E shows the comparison for the growth exponent, η, as a function of the cumulative number of researchers in a region in 2019, k_{final}. The more scientists are present, the smaller the growth exponent. Fewer scientists correspond to higher entry times and, thus, to higher

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

exponents. The red circle marks the Chinese regions in the model to which we assigned higher α values. Also, for these, the model practically reproduces the data.

Finally, Fig. 4F gives the frequency rank distribution of the regions appearing in the two streams. The model also reproduces the distribution obtained from data well. A consequence of a sub-linear PA is that the corresponding frequency-rank of the cumulative number of distinct scientists who worked in a region is not an exact power-law [17] – a fact that we hereby verify.

In Appendices E and F we show the results for the embryonic stem cells and Internet research in the same manner as in Figs. 3 and 4.

3. Discussion

Using metadata from scientific publication databases, we reconstructed the regional cumulative number of publishing scientists across the world regions. We find that researchers tend to move to regions according to a sub-linear PA mechanism, where the settling probability for a region is proportional to a power of the cumulative number of scientists already working in that region. In other words: We find evidence for the rich-get-richer phenomenon of scientific mobility of a sub-linear type. This means that regions that move early-moving are likely to dominate the field (or become one of the few dominant regions) and then retain that dominance for a long time.

We find power-law-like regional growth curves that indicate the absence of any specific scale that would reflect the presence of a critical mass. We find no signs of a critical mass (or density) of scientists in a particular topic necessary before the field takes off regionally; a simple PA model is sufficient to explain the main features of the data.

The number of researchers in early-moving regions tends to grow sub-linearly in intrinsic time. The more researchers there are, the lower the regional growth rates. Latecomer regions, which start attracting researchers at a later stage (after the first several hundred articles have been published), are characterized by a number of researchers that can be several orders of magnitude smaller than that of pioneer regions. We find evidence that latecomer regions tend to grow faster, typically even superlinearly, in their initial phase, but with numbers so small that they can practically never even approach dominance. The exponents of regional growth increase with the delay of entry into the field relative to the age of the discipline. Since regions that have been in research for a long time have higher cumulative numbers of scientists than regions that enter late, regional growth exponents tend to decrease as more scientists populate the region. We demonstrate the existence of the same generic mechanism in three different scientific domains over many decades.

For these three domains, we find that Chinese regions behave differently from all the others. They do not follow the sub-linear growth of the pioneering regions but grow super-linearly (in intrinsic time), starting after 500,000 articles have been produced in the field. This suggests that Chinese regions follow a different pattern. The simplest way to explain the differences is to assign them a higher PA exponent — still being sub-linear. This means that as compared to the other regions, Chinese ones attract scientists with higher probability. We speculate that Chinese regions become more attractive thanks to strategic efforts, including proactive hiring policies, massive funding, and by training a huge reservoir of scientists abroad and bringing them back. We emphasize that the PA exponent does not depend on the multiplicity of Chinese scientists (population effect) since it determines how researchers attach to a region. In the first approximation, the PA mechanism does not depend on the number of attaching nodes. Of course, there could be an indirect effect by which the high number of scientists requires the introduction of *ad hoc* policies. Fig. 3E shows how the growth exponents decrease with the final number of scientists, with Chinese regions being the outlier red circles. Introducing a higher PA exponent for Chinese regions in our model yields the correct massively larger growth exponents, η, that are observed. The deeper reasons why Chinese regions behave differently by attracting researchers more strongly than others – still under a PA scheme – have to be investigated in future research.

Note that the study considers all regions as having the same PA exponent, except for the Chinese regions that have a larger one. This is a gross simplification. However, it still explains the data reasonably well. Nevertheless, the shortcoming of identical exponents should be addressed in future work by a more precise description that assigns different exponents to different regions, ideally on an empirical basis. How this can be done is not so clear.

Finally, we mention that, remarkably, we find that the rate at which regions join a specific field follows an approximate power-law. This feature has been referred to as Heaps' law that occurs in numerous evolutionary processes [33–35,37]. Heaps' law is intimately connected to the idea of the adjacent possible [36], where the discovery of new things or ideas leads to other discoveries and results that were impossible to achieve before [39].

In summary, our simple PA model captures the essence of the underlying preferential attachment process. To a large extent, it explains the empirical data of the historical evolution of the three studied disciplines. It captures both the growth of scientific capacity in individual regions and the observed super-linear initial growth of latecomer regions. The model is simple enough to be analytically tractable, which allows for a detailed understanding of the relations between the three power-law exponents involved in the process. In particular, it explains why faster growth of latecomer regions is observed.

The popular explanation that attaining (or catching up to) leadership in a scientific discipline requires building a critical mass of scientists is not supported by data on scientific mobility. We conclude that there are two ways to secure leadership and scientific dominance in a field: One is to have a strong presence among the earliest and earliest players, which is entirely consistent with the theory of increasing returns [40]. The other is to ensure the continuation of exceptionally high, superlinear growth rates through strategic interventions that must be sustained over long periods of time (decades). The latter has been most evident in some areas of Chinese science: The beginning of the catch-up process in the late 1970s led to a dominant role today.

4. Methods

Estimating the PA kernel. We estimate the PA attachment probability, $P(k)$, by means of a method originally devised for co-authorship networks [18]. Following this method, we go through the regions in the time-sorted regional stream, $\mathcal{R}$, one by one and build a histogram from which we estimate the attachment kernel. If a PA process is realized, at each point in time t, the probability of a region already occurred k times to appear again is given by

$$\Pi(k,t) = P(k)\frac{n(k,t)}{D(t)}, \tag{6}$$

where $n(k,t)$ is the number of regions already appeared k times at time t, and $D(t)$ is the total number of regions already appeared irrespective of their number of occurrences. Then, the attachment probability, $P(k)$, can be estimated from a histogram where each contribution is weighted with the inverse frequency $\frac{D(t)}{n(k,t)}$. The resulting histogram starts to deviate from a straight line (in double logarithmic scale) at around $k \approx 60$ (Fig. 2 panels A–C). This is a well-known behavior of the method. Therefore, we resort to the first 60 points to estimate the exponents with a least squares method in all three scientific fields studied. In all the three fields we find $R^2 > 0.99$.

Determining the PA exponent for Chinese regions. We obtain the special value of $\alpha_{CN} = 0.915$ in the following way. We first determine from the data the average growth exponent, η, exclusively for the Chinese regions obtaining $\langle\eta\rangle_{CN} \approx 2.53 \pm 0.09$. Then, from a starting value of $\alpha_{CN} = 0.79$ – the value of all the remaining regions – we increment α_{CN} in the simulation in steps of 0.005 until we find an

agreement between the empirical $\langle \eta \rangle_{CN}$ and the one estimated from the stream of regions generated by the simulation.

Figs. 3C–E and 4 panels C–E are obtained by first removing all regions with a cumulative number of scientists at the year 2019 less than 100. For those regions with a lower number of scientists, the curves of their growth in intrinsic time become noisy, and the estimated exponent is unreliable.

CRediT authorship contribution statement

Vito D.P. Servedio: Analyzed the data, Devised the model, Studied it analytically, Writing – original draft. **Márcia R. Ferreira:** Conceptualized the idea, Wrangled and curated the data, Analyzed the data, Writing – original draft. **Niklas Reisz:** Analyzed the data. **Rodrigo Costas:** Wrangled and curated the data. **Stefan Thurner:** Conceptualized the idea, Devised the model, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We acknowledge the Centre for Science and Technology Studies (CWTS) of Leiden University for providing access to their in-house version of the Dimensions database and support by the Austrian Research Promotion Agency FFG under grant 882184. All authors contributed to the final version.

Appendix A. Establishing the data set

The Dimensions data. We access the *Dimensions* database through the Centre for Science and Technology Studies (CWTS), Leiden University. Dimensions covers millions of research publications connected by more than 1.6 billion citations, supporting grants, datasets, clinical trials, patents, and policy documents. In this work, we only use research publications. The publications database of Dimensions contains publication data spanning several decades. The database comprises about 20 million disambiguated researchers assigned to a researcher ID. It also comprises affiliation linkages to the Global Research Identifier Database that currently covers more than 98,000 research institutions worldwide [9,10]. *Dimensions* is produced by Digital Science and launched in January 2018. For further references, see the Dimensions.ai website.[2]

Preparing the data. We start by collecting documents (i.e., all document types) in three different topics, i.e., Semiconductors, Embryonic Stem Cell (ESC) research, and Internet research. We chose these areas since they are sufficiently large and important to provide a meaningful statistical analysis. Delineating research areas is crucial to study their growth. Different types of data and techniques have been proposed to delineate research areas such as document co-citation [41], author co-citation [42], co-word analysis [43], and journal-based mapping [44]. The semiconductors, ESC, and Internet research areas are defined by collecting three sets of documents and associated groups of disambiguated authors, their affiliated institutions, and the geographical regions where the institutions are located. We use a combination of term matching, citing and cited links, and a selection of specific Fields of Research (FOR) to identify documents that belong to those areas of research. The method we employ for defining the three research areas in this paper can be divided into three steps:

Step 1: Creating a lexicon of key technical terms. We started by extracting keywords from the literature and validating them with domain experts. Terms identified as relevant within each topic by experts were kept for further analysis. This step involved identifying the core technical jargon that authors use in each of the areas under study. As a result, we prioritized technical terminology over popular terms, as the former were more likely to be used and recognized by our target research community. The vocabulary was expanded further using the VOSviewer software [45] to identify very frequently co-occurring concepts in addition to the initial terms, and the newly extracted concepts were validated by domain experts during a second round of consultation. The list of the lexicon used can be found in Appendix D.

Step 2: Retrieving items using term matching and Dimensions' concepts. In the second step of our approach, we searched the Dimensions database for all document types that matched the terminology established in the previous step using a combination of exact and fuzzy term matching. We matched the phrases to pre-extracted *concepts* from Dimensions publications that had an abstract between 1941 and 2019. According to the Dimensions documentation, concepts are normalized noun phrases that describe the core concepts of a document and are derived automatically from the publication's abstracts [46]. Additionally, we used Dimensions' four-digit FOR categories to ensure that the concepts correspond to a rather consistent collection of documents. For example, several concept matches can yield publication items associated with FOR categories such as "Anthropology" or "Law". We excluded these domains from our document search due to our strong focus on technical knowledge in semiconductors, ESC, and Internet research. In Appendix D, we list the whole vocabulary that was used to retrieve the documents.

Step 3: Extracting cited and citing items. Cited works stand for or symbolize works that have been inspired by past publications, while citing works symbolize works that have inspired subsequent items. Therefore, we also collected all citing and cited items to the publication sets retrieved in step 2. Both cited and citing items are restricted to the overall time spans in 1 and FOR categories.

Algorithms for author name disambiguation and institutional registries have been implemented and linked to the majority of large bibliometric databases, including Scopus from Elsevier and Dimensions from Digital Science. Most of these algorithms leverage open systems for uniquely identifying scholars, such as ORCID, or for identifying institutional affiliations, such as GRID [8].

Using this method, we can extract a representation of research areas that includes core documents in the international scientific literature in a period between 1941 and 2019. We use the affiliation of researchers to assign an article to one or more world regions. Regions are considered at the granularity of the first level of administrative regions, equivalent to provinces (e.g., US.MA [Massachusetts, USA], GB.ENG [England, Great Britain]). We then select a chosen scientific field and sort all the corresponding publications in ascending temporal order, whereas articles sharing the same year of publication are listed randomly. We checked that the reshuffling of the publications inside the same year did not change any of the results. Since we are interested in the cumulative growth of the number of scientists in a region, we discard all those events where scientists publish a paper with one of their old affiliations under which they had already published in the past (see Table 2 for an example of the procedure). Finally, we build the sequence of regions in the order they appear in time. The position in the sequence defines an intrinsic time that runs faster than real-time. One tick of intrinsic time corresponds to a region in the sequence. The relationship between intrinsic and real-time is an inverse stretched exponential, i.e., for all the three scientific fields considered, the relation between the two is approximately of the type $t \approx \exp(\sqrt{\frac{T-T_0}{\tau}})$ with t intrinsic time, T real time, T_0 starting year reported in Table 1, and τ representing a sort of characteristic time that gets values around 130–160 days in all three cases.

[2] https://www.dimensions.ai/.

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

Appendix B. Scaling relations between exponents

The important quantities we consider in this study are, asymptotically, for large values of intrinsic time t

$$\begin{aligned} &D(t) \approx t^{\gamma} \qquad k_i(t) \approx t^{\eta} \\ &P_k \propto k^{\alpha} \qquad Z(t) = \sum_{i=1}^{D} k_i^{\alpha} \approx t^{\sigma}, \end{aligned} \tag{B.1}$$

where $D(t)$ is the number of distinct regions (Eq. (3) in the main text); $k_i(t)$ is the cumulative number of scientists in region i (Eq. (2) in the main text); P_k is the preferential attachment probability (Eq. (5) in the main text); $Z(t)$ is the normalization term in the PA probability definition.

One can deduce some identities with simple approximate reasoning. Since $\sum_{i=1}^{D} k_i = t$ and we have D terms in the sum, we can write $t^{\gamma+\eta} \approx t$, hence $\eta = 1-\gamma$. Similarly, $Z(t) \approx t^{\gamma} t^{\alpha\eta} = t^{\gamma+\alpha(1-\gamma)}$, hence $\sigma = \alpha + \gamma(1-\alpha)$ (we rearranged the terms to highlight the role of α). These crude approximations are confirmed by a less straightforward analytic solution, which can also provide the behavior of $k_i(t)$ when t is close to the first introduction of region i, i.e., $k_i(t \approx t_{0,i}) \approx (1+c(t-t_{0,i}))^{\frac{1}{1-\alpha}}$. Since $0 < \alpha < 1$, the exponent $1/(1-\alpha)$ is larger than 1, in accordance with the observed super-linear growth of latecomer regions. Practically, the number of occurrences of regions that enter late in the system starts to grow super-linearly and eventually reaches the asymptotic sub-linear regime.

Appendix C. Approximate analytical solution of the model

We have a stream of geographical regions from empirical data and from it, we build a new synthetic stream. We shall generically refer to geographical regions as *tokens* in the following since the reasoning below can be applied to any kind of stream with sub-linear PA. Each time a brand new token appears in the real stream, we put it as it is in our own created stream; each time an already occurred token appears in the real stream, we extract a token from our synthetic stream with a probability that is sublinear in the number of its occurrences k so far, i.e., according to the probability P_k defined in the main text in Eq. (5).

Definition of parameters

Our model contains one free parameter: the exponent α of the sublinear rich-get-richer mechanism. We take into account Heaps' exponent automatically by adopting the real stream of new tokens. We need to introduce two parameters more, which eventually will be connected to the PA and Heaps' exponents, i.e., the asymptotic exponent σ of the PA normalization term and the asymptotic exponent η of the growth of tokens in intrinsic time. The definition of the exponents is that of Eq. (B.1). We call the normalization $Z(t)$ as partition function in the following. A rough estimate of the parameters based on one run of the model in the semiconductor field and simple linear regressions gives:

$$\gamma \approx 0.39 \qquad \eta \approx 0.74 \qquad \alpha \approx 0.79 \qquad \sigma \approx 0.85. \tag{C.1}$$

The value of η was inferred from the most populated token.

Occurrence of token i in intrinsic time

In the approximation of continuous time we can write

$$\frac{dk_i}{dt} = \frac{k_i^{\alpha}}{Z(t)} \tag{C.2}$$

where we know that asymptotically $Z(t) \approx ct^{\sigma}$ (with c a multiplicative constant). We solve the previous equation

$$\int_1^k \frac{dk_i}{k_i^{\alpha}} = c \int_{t_{0,i}}^{t} \frac{d\tau}{\tau^{\sigma}} \tag{C.3}$$

with t_0 the entry time of the token, to get

$$k = \left(1 + c\frac{1-\alpha}{1-\sigma}(t^{1-\sigma} - t_{0,i}^{1-\sigma})\right)^{\frac{1}{1-\alpha}}. \tag{C.4}$$

For $t \gg t_0$ we get (we drop the index i for simplicity)

$$k \approx t^{\frac{1-\sigma}{1-\alpha}} \quad \text{i.e.} \quad \boldsymbol{\eta = \frac{1-\sigma}{1-\alpha}}. \tag{C.5}$$

If $t = t_0 + x$ we get by expanding around t_0

$$k \approx \left(1 + c\frac{1-\alpha}{t_0^{\sigma}} x\right)^{\frac{1}{1-\alpha}}, \tag{C.6}$$

with a super-linear exponent $\frac{1}{1-\alpha}$.

Probability density function of occurrences

Let us call $N_k(t)$ the number of tokens having already occurred k-times at time t. After dropping the index i we can write the following master equation for $k > 1$:

$$N_k(t+1) = N_k(t) + N_{k-1}\frac{(k-1)^{\alpha}}{Z(t)} - N_k\frac{k^{\alpha}}{Z(t)}, \tag{C.7}$$

which can be written in a continuous approximation as the PDE

$$\frac{\partial N_k}{\partial t} = -\frac{1}{Z(t)}\frac{\partial (k^{\alpha} N_k)}{\partial k}. \tag{C.8}$$

Some important relations to keep in mind:

$$\sum_{i=1}^{D} k_i = \sum_{k=1}^{k_M} k N_k = t \tag{C.9}$$

$$\sum_{i=1}^{D} 1 = \sum_{k=1}^{k_M} N_k = D \approx ct^{\gamma} \tag{C.10}$$

with k_M representing the maximum value of the k_i. In order to normalize N_k and get the probability ρ_k of finding tokens with k occurrences we divide by D

$$N_k \approx ct^{\gamma}\rho_k. \tag{C.11}$$

We substitute the previous relation and $Z(t) \approx st^{\sigma}$ into the PDE. After simplifying the derivatives explicitly, we get

$$c\gamma t^{\gamma-1}\rho_k \approx -\frac{ct^{\gamma-\sigma}}{s}\left(\alpha k^{\alpha-1}\rho_k + k^{\alpha}\frac{d}{dk}\rho_k\right) \tag{C.12}$$

and eventually, obtain the following ODE

$$\frac{d}{dk}\rho_k = -s\gamma k^{-\alpha} t^{\sigma-1}\rho_k - \frac{\alpha}{k}\rho_k. \tag{C.13}$$

Its solution is proportional to

$$\rho_k \propto k^{-\alpha} e^{-\frac{s\gamma t^{\sigma-1}}{1-\alpha}k^{1-\alpha}} \tag{C.14}$$

which asymptotically at large t, gives

$$\rho_k \approx k^{-\alpha}. \tag{C.15}$$

Since the decay in time is pretty slow ($\sigma \approx 0.85$ in our systems), the limit distribution is attained over long times. Moreover, we recall that $0 < \alpha < 1$ so that ρ_k would have no thermodynamic limit without the stretching exponential term. For determining the scaling relations between the exponents, we stick to the power-law limit distribution and call k_M the maximum value of k at a certain large value of time t. By normalizing the ρ_k we find

$$\rho_k = \frac{1-\alpha}{k_M^{1-\alpha} - 1} k^{-\alpha}. \tag{C.16}$$

Partition function in time

We recall the partition function

$$Z(t) = \sum_{i=1}^{D} k_i^{\alpha} = \sum_{k=1}^{k_M} N_k k^{\alpha} \approx t^{\sigma}. \tag{C.17}$$

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

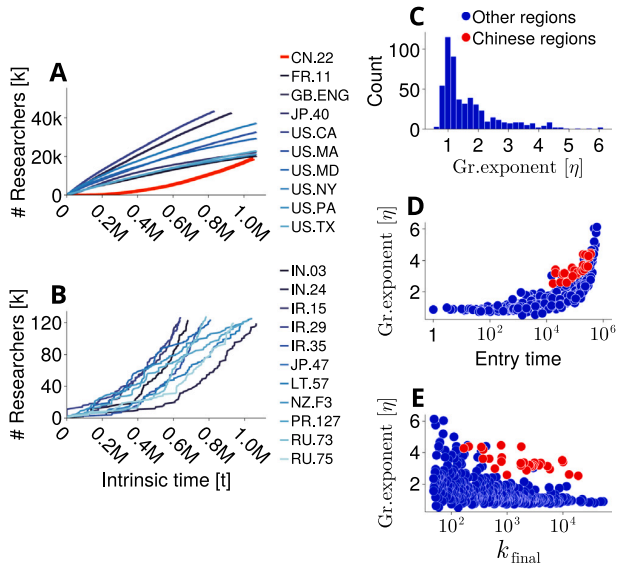


Fig. E.5. Panels A and B show the cumulative number of scientists coming into a region (lines) in intrinsic time in embryonic stem cell research; with respect to Fig. 2A and B, the axes are in linear scale to appreciate the different curvature of the growths at high intrinsic times; (A) shows the 10 most visited regions. Most regions grow sub-linearly. Chinese regions are shown with red thick lines and clearly increase super-linearly. In (B), we see 11 poorly visited regions ranked from 700 to 710 in terms of the number of scientists. Curves are (additively) shifted to the left so that their first point is placed at coordinates (1,1); (C) Histogram of growth exponents, η, for all regions; (D) Growth exponents, η, of all regions, i, as a function of their entry time, t_i (intrinsic time). Note that Chinese regions (red) come in late (after timestep 10^5) and have higher exponents than other regions (blue). (E) Growth exponents, η, as a function of the cumulative number of scientists in the year 2019, k_{final}. The high exponents of Chinese regions are visible (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

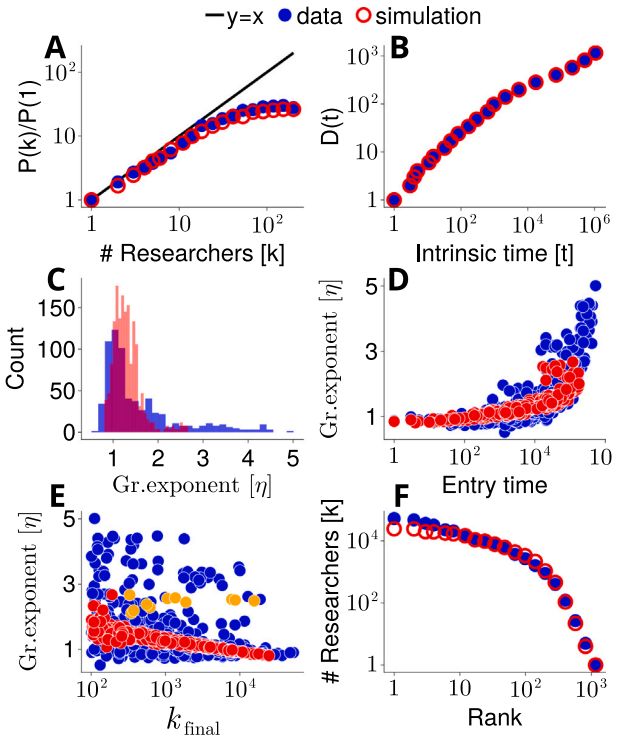


Fig. E.6. Comparison of model results (red) with experimental data (blue) for the field of embryonic stem cells. (A) PA kernel; data and simulation practically coincide; the black line depicts a linear kernel. (B) Heaps's law; the two curves coincide by construction. (C) Histograms of the regional growth exponent, η. The distributions of the empirical values have average $m = 1.56$, standard deviation $s = 0.86$ and skewness $b = 1.73$, while the distribution of the simulated values has $m = 1.31$, $s = 0.31$ and $b = 1.70$. (D) Exponents, η, reproduce the empirical increase as a function of entry time. (E) The decrease of η as a function of the cumulative number k_{final} of scientists in the year 2019 is well captured by the model. Note the fact that Chinese regions (orange circles) are predicted with a higher than usual values. (F) Also the frequency-rank distributions of the regions appearing in the two streams practically coincide except for very small ranks. In panels C, D, and E only regions with $k_{\text{final}} > 100$ are taken into account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

If $\alpha \to 0$ then $Z(t) = D \approx t^{\gamma}$; if $\alpha = 1$ then $Z(t) = t$. Therefore we expect $\gamma < \sigma < 1$. We can write

$$\begin{aligned} Z(t) &\approx \sum_{k=1}^{k_M} t^{\gamma} \rho_k k^{\alpha} \\ &\approx \sum_{k=1}^{k_M} t^{\gamma} \frac{1-\alpha}{k_M^{1-\alpha}-1} k^{-\alpha} k^{\alpha} \\ &\approx t^{\gamma} \frac{1-\alpha}{k_M^{1-\alpha}-1} \sum_{k=1}^{k_M} 1 \approx t^{\gamma} k_M^{\alpha}. \end{aligned} \tag{C.18}$$

If we now recall that the occurrences of a token grow as t^{η}, and therefore also k_M does, we find

$$\sigma = \gamma + \alpha\eta = \gamma + \alpha \frac{1-\sigma}{1-\alpha}, \tag{C.19}$$

which after solving for σ gives:

$$\boldsymbol{\sigma = \alpha + \gamma(1-\alpha)}. \tag{C.20}$$

By inserting the above expression of σ in the expression of η of Eq. (C.5) we also get

$$\boldsymbol{\eta = 1 - \gamma}. \tag{C.21}$$

Recap of results

After substituting the result for σ we get:

- $Z(t \gg 1) \approx t^{\alpha+\gamma(1-\alpha)}$
- $k_i(t \gg t_{0,i}) \approx t^{1-\gamma}$
- $k_i(t \approx t_{0,i}) \approx (1 + c(t - t_{0,i}))^{\frac{1}{1-\alpha}}$.

Note how the number of occurrences of tokens that enter late in the system starts to grow super-linear and eventually reaches the asymptotic sublinear regime.

Appendix D. List of terms

In the following, we list the lexicon of terms used to select the articles according to their scientific fields.

Semiconductor research transistor, analog circuits, bipolar junction transistor, bipolar transistor, carbide, carborundum, cat s-whisker detector, conductivity, crystal diode, darlington transistor, discrete device, doped monocrystalline silicon grid, electrical conduction, electrical conductivity, electron–hole pairs, electronic band structure, field effect junction transistor, field-effect transistor, four-terminal devices, gallium arsenide, germanium, gunn diode, hall effect sensor, heterojunctions, impatt diode, insulated-gate bipolar transistor, integrated circuit, iotatron, laser diode, light-emitting diode, light-emitting diode, metal rectifier, metal–oxide–semiconductor, metal–oxide–semiconductor field-effect transistor, microchip, microprocessor, mixed-signal circuits, mos transistor, mosfet, n-type semiconductor, optocoupler, photocoupler, photodiode, phototransistor, photovoltaic solar cells, pin diode, p–n junctions, p–n–p point-contact germanium,

V.D.P. Servedio et al. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena 167 (2023) 113020

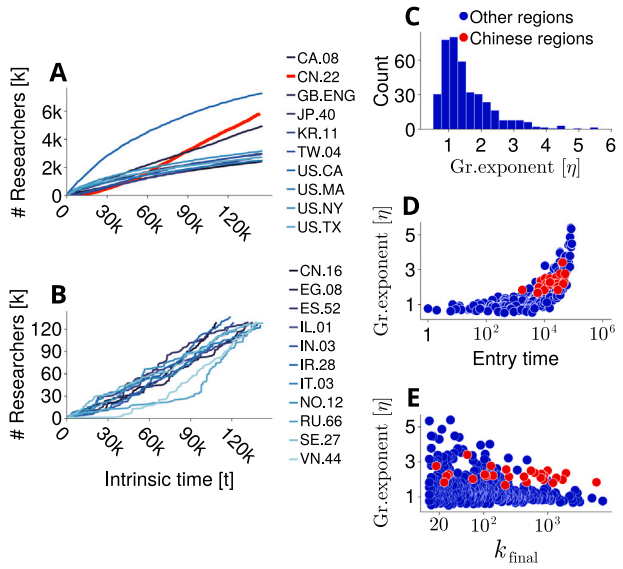


Fig. F.7. Panels A and B show the cumulative number of scientists coming into a region (lines) in intrinsic time in Internet research; with respect to Fig. 2A and B, the axes are in linear scale to appreciate the different curvature of the growths at higher intrinsic times; (A) shows the 10 most visited regions. Most regions grow sub-linearly. Chinese regions are shown with red thick lines and clearly increase super-linearly. In (B) we see 11 poorly visited regions ranked from 700 to 710 in terms of the number of scientists. Curves are (additively) shifted to the left so that their first point is placed at coordinates (1,1); (C) Histogram of growth exponents, η, for all regions; (D) Growth exponents, η, of all regions, i, as a function of their entry time, t_i (intrinsic time). Note that Chinese regions (red) come in late (after timestep 10^5) and have higher exponents than other regions (blue). (E) Growth exponents, η, as a function of the cumulative number of scientists in the year 2019, k_{final}. The high exponents of Chinese regions are visible (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

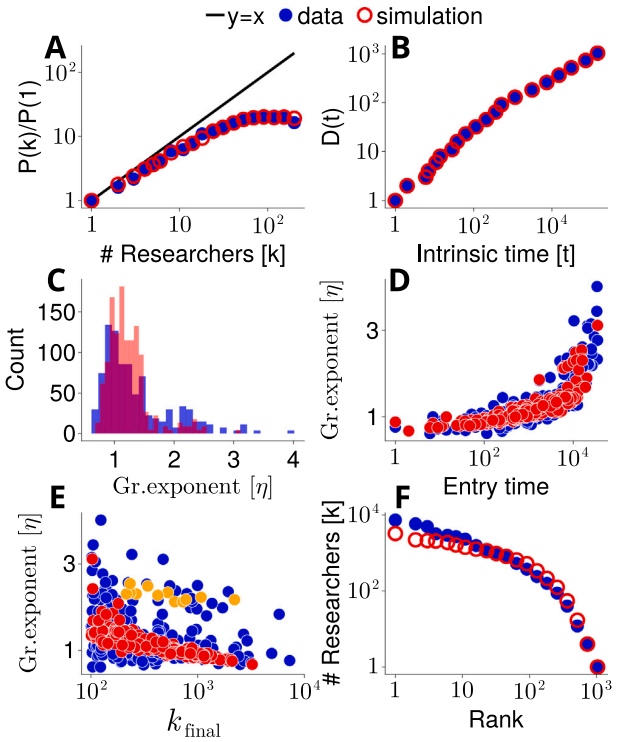


Fig. F.8. Comparison of model results (red) with experimental data (blue) for the field of Internet. (A) PA kernel; data and simulation practically coincide; the black line depicts a linear kernel. (B) Heaps's law; the two curves coincide by construction. (C) Histograms of the regional growth exponent, η. The distributions of the empirical values has average $m = 1.36$, standard deviation $s = 0.59$ and skewness $b = 1.48$, while the distribution of the simulated values has $m = 1.22$, $s = 0.35$ and $b = 1.93$. (D) Exponents, η, reproduce the empirical increase as a function of entry time. (E) The decrease of η as a function of the cumulative number k_{final} of scientists in the year 2019 is well captured by the model. Note the fact that Chinese regions (orange circles) are predicted with higher than usual values. (F) Also, the frequency-rank distributions of the regions appearing in the two streams practically coincide except for very small ranks. In panels C, D, and E, only regions with $k_{\text{final}} > 100$ are taken into account. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

polysilicon, p-type semiconductor, rectifier diode, reverse-biased p–n junction, schottky diode, selenium sulfide, semiconductor, semi-insulators, silicon-controlled rectifier, solid triode, three-terminal devices, thyristor, transconductance, transient-voltage-suppression diode, transistron, triode, tunnel diode, two-electrode vacuum tube, unijunction transistor, vaccum tube, varistor, vcsel, zen diode, zener diode.

Embryonic Stem Cells blastocyst, blastoid, embryonic AND stem AND cell, embryonic fibroblast, embryonic germ layers, hescs AND stem AND cell, mescs AND stem AND cell, pluripotency, pluripotent AND stem AND cell, undifferentiated pluripotent state.

Internet research arpanet, atm networks, atm switch, bitnet, broad-based electronic communication, classless interdomain routing, csnet, darpa internet, distributed network protocol, domain name system, end-to-end circuit, end-to-end transmission, esnet, exterior gateway protocol, federal internet exchanges, gateway protocol, gateways router, global information infrastructure, gopher protocol, ground-based packet, hop wireless network, host-to-host protocol, hypertext transfer protocol, ibm rscs protocol, input-queued packet, interior gateway protocol, internet infrastructure, internet protocol, internet router, internet routing, internetting, internetwork, internetworking architecture, ip service, ipv4, ipv6, lan technology, local area network technology, message switching AND internet, minitel, mmdf protocol, multihop networks, networked computers, network control protocol, network emulator, network traffic protocol, npl network, open systems interconnection, open-architecture network, optical packet, osi reference model, packet radio, packet satellite, packet switch, packet traffic, packetization, packet-switching, radio transmitting system, relay network, remote machines, simple mail transfer protocol, store-and-forward switching, switched network, tcp performance, tcp/ip, tcp-based applications, telecommunications infrastructure, telenet, time-sharing internet, time-sharing network, transmission control protocol, trans-oceanic circuits, transport layer protocol, user datagram protocol, virtual circuit model, virtual circuits, wireless atm AND internet, wireless sensor network AND internet, x.25 protocol.

Appendix E. Results for embryonic stem cells

Figs. E.5 and E.6 show the results for the field of embryonic stem cells and share the same design of Figs. 3 and 4.

Appendix F. Results for internet research

Figs. F.7 and F.8 show the results for the field of Internet and share the same design of Figs. 3 and 4.

References

[1] Gibbons M, Johnston R. The roles of science in technological innovation. Res Policy 1974;3(3):220–42.
[2] Saxenian A. The genesis of silicon valley. Built Environ 1983;(1978-):7–17.
[3] Harrison M. Does high quality research require critical mass. Question R&D Spec: Pers Policy Implic. 2009;57–9.
[4] Johnston R. Effects of resource concentration on research performance. Higher Educ. 1994;28(1):25–37.

[5] Sugimoto CR, Robinson-García N, Murray DS, Yegros-Yegros A, Costas R, Larivière V. Scientists have most impact when they're free to move. Nat News 2017;550(7674):29.
[6] Trippl M. Islands of innovation and internationally networked labor markets. magnetic centers for star scientists?. 2009.
[7] Sturgeon TJ. How silicon valley came to be. Underst Silicon Val: Anat Entrepreneurial Reg 2000;15–47.
[8] Macháček V, Srholec M, Ferreira MR, Robinson-Garcia N, Costas R. Researchers' institutional mobility: bibliometric evidence on academic inbreeding and internationalization. Sci Public Policy 2021.
[9] Hook DW, Porter SJ, Herzog C. Dimensions: Building context for search and evaluation. Front Res Metr Anal 2018;3:23.
[10] Herzog C, Hook D, Konkiel S. Dimensions: Bringing down barriers between scientometricians and data. Quant Sci Stud 2020;1(1):387–95.
[11] Menter M, Lehmann EE, Klarl T. In search of excellence: A case study of the first excellence initiative of Germany. J. Bus Econ 2018;88(9):1105–32.
[12] Zucker LG, Darby MR. Star scientists, innovation and regional and national immigration. Working paper 13547, National Bureau of Economic Research; 2007.
[13] de Solla Price D. A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 1976;27(5):292–306.
[14] Simkin MV, Roychowdhury VP. Re-inventing willis. Phys Rep 2011;502:1–35.
[15] Bol T, de Vaan M, van de Rijt A. The matthew effect in science funding. Proc Natl Acad Sci 2018;115(19):4887–90.
[16] Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. Phys Rev E 2006;74:036116.
[17] Krapivsky PL, Redner S, Leyvraz F. Connectivity of growing random networks. Phys Rev Lett 2000;85:4629–32.
[18] Newman MEJ. Clustering and preferential attachment in growing networks. Phys Rev E 2001;64:025102.
[19] Jeong H, Néda Z, Barabási AL. Measuring preferential attachment in evolving networks. Europhys Lett (EPL) 2003;61(4):567–72.
[20] Szell M, Lambiotte R, Thurner S. Multirelational organization of large-scale social network in an online world. Proc Natl Acad Sci 2010;107(31):13636–41.
[21] Reisz N, Servedio VDP, Loreto V, Schueller W, Ferreira MR, Thurner S. Loss of sustainability in scientific work. New J Phys 2022;24(5):053041.
[22] Wang D, Song C, Barabási A-L. Quantifying long-term scientific impact. Science 2013;342(6154):127–32.
[23] Jin C, Song C, Bjelland J, Canright G, Wang D. Emergence of scaling in complex substitutive systems. Nat Hum Behav 2019;3(8):837–46.
[24] Gleeson JP, Cellai D, Onnela J-P, Porter MA, Reed-Tsochas F. A simple generative model of collective online behavior. Proc Natl Acad Sci 2014;111(29):10411–5.
[25] Rao A, Scaruffi P. A history of silicon valley: the greatest creation of wealth in the history of the planet. 2nd ed.. Omniware Group; 2013.
[26] Lazonick W, Li Y. China's path to indigenous innovation. 2012.
[27] Li Y. The semiconductor industry: A strategic look at China's supply chain. In: The new Chinese dream: industrial transition in the post-pandemic era. Cham: Springer International Publishing; 2021, p. 121–36, Chapter 8.
[28] Tollefson J. China declared world's largest producer of scientific articles. Nature 2018;553:390.
[29] Robinson-Garcia N, Sugimoto CR, Murray D, Yegros-Yegros A, Larivière V, Costas R. The many faces of mobility: Using bibliometric data to measure the movement of scientists. J Informetr 2019;13(1):50–63.
[30] Waltman L, van Eck NJ. Field-normalized citation impact indicators and the choice of an appropriate counting method. J Informetr 2015;9(4):872–94.
[31] Moed HF. Citation analysis in research evaluation. Springer Dordrecht; 2005.
[32] Pham T, Sheridan P, Shimodaira H. PAFit: A statistical method for measuring preferential attachment in temporal complex networks. PLoS One 2015;10(9):e0137796.
[33] Serrano MÁ, Flammini A, Menczer F. Modeling statistical properties of written text. PLoS One 2009;4(4):1–8.
[34] Tria F, Loreto V, Servedio VDP, Strogatz SH. The dynamics of correlated novelties. Sci Rep 2014;4:5890.
[35] Mazzolini A, Grilli J, De Lazzari E, Osella M, Lagomarsino MC, Gherardi M. Zipf and heaps laws from dependency structures in component systems. Phys Rev E 2018;98:012315.
[36] Tria F, Loreto V, Servedio VDP. Zipf's, Heaps' and Taylor's laws are determined by the expansion into the adjacent possible. Entropy 2018;20(10).
[37] Simini F, James C. Testing Heaps' law for cities using administrative and gridded population data sets. EPJ Data Sci 2019;8(1):24.
[38] Zanette D, Montemurro M. Dynamics of text generation with realistic Zipf's distribution. J Quant Linguist 2005;12(1):29–40.
[39] Kauffman SA. Investigations. New York/Oxford: Oxford University Press; 2000.
[40] Arthur WB. Competing technologies, increasing returns, and lock-in by historical events. Econ J 1989;99(394):116–31.
[41] Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am Soc Inf Sci 1973;24(4):265–9.
[42] White HD, Griffith BC. Author cocitation: A literature measure of intellectual structure. J Am Soc Inf Sci 1981;32(3):163–71.
[43] Callon M, Courtial J-P, Turner WA, Bauin S. From translations to problematic networks: An introduction to co-word analysis. Soc Sci Inf 1983;22(2):191–235.
[44] Leydesdorff L. Clusters and maps of science journals based on bi-connected graphs in journal citation reports. J Doc 2004.
[45] Van Eck NJ, Waltman L. Software survey: Vosviewer, a computer program for bibliometric mapping. Scientometrics 2010;84(2):523–38.
[46] Working with concepts in the dimensions API. 2022, https://api-lab.dimensions.ai/cookbooks/1-getting-started/7-Working-with-concepts.html, accessed 2022-01-26.

Chapter 5

Knowledge diffusion in social networks

To what extent homophily and influencer networks explain song popularity

Niklas Reisz, Vito D. P. Servedio, and Stefan Thurner

Preprint version: arXiv:2211.15164, 2022

To what extent homophily and influencer networks explain song popularity

Niklas Reisz[a], **Vito D. P. Servedio**[a], **and Stefan Thurner**[1abc]

[a]Complexity Science Hub Vienna, Josefstädter Strasse 39, A-1080 Vienna, Austria; [b]Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, A-1090, Austria; [c]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 85701, USA

November 29, 2022

Forecasting the popularity of new songs has become a standard practice in the music industry and provides a comparative advantage for those that do it well. Considerable efforts were put into machine learning prediction models for that purpose. It is known that in these models, relevant predictive parameters include intrinsic lyrical and acoustic characteristics, extrinsic factors (e.g., publisher influence and support), and the previous popularity of the artists. Much less attention was given to the social components of the spreading of song popularity. Recently, evidence for musical homophily –the tendency that people who are socially linked also share musical tastes– was reported. Here we determine how musical homophily can be used to predict song popularity. The study is based on an extensive dataset from the *last.fm* online music platform from which we can extract social links between listeners and their listening patterns. To quantify the importance of networks in the spreading of songs that eventually determines their popularity, we use musical homophily to design a predictive *influence* parameter and show that its inclusion in state-of-the-art machine learning models enhances predictions of song popularity. The influence parameter improves the prediction precision (TP/(TP+FN)) by about 50% from 0.14 to 0.21, indicating that the social component in the spreading of music plays at least as significant a role as the artist's popularity or the impact of the genre.

social networks | scaling | preferential attachment | complex networks

While you read this paper's abstract, more than 50 new songs were released worldwide. Global music production has long reached such high levels that it is no longer possible to listen to every new song. On the world's largest music streaming platform *Spotify*, roughly 136 days worth of music are published daily (1, 2). This means that more music is produced in a year than one could listen to in an entire lifetime. In this ocean of new songs, there is a growing need for efficient selection and filtering, and the markets for attention have become increasingly competitive. This becomes apparent in increasingly imbalanced distributions of song popularity. While the majority of songs receive little to no attention, a small fraction become hits that are listened to billions of times (3). Predicting which songs have the potential to become one of these hits has become an increasingly important task for music publishers (4). The issue has sparked substantial commercial interest as publishers focus their marketing activities on high-potential songs and artists (5). For quantitative predictions, a multitude of different approaches has been explored. The vast volume of data made available by streaming platforms in the past decade triggered a boost of data-driven approaches.

In the early 2000s, a discussion started on the possibility of quantitatively predicting song popularity. Using traditional machine learning classifiers on lyrical and acoustic song attributes, it was attempted to identify hits prospectively (6). It was concluded that predictions were significantly better than random, with the lyrical features slightly outperforming the acoustic ones. In (7), it was claimed that the science of predicting hits, the so-called *hit song science*, was not yet a science as they demonstrated that the usual machine learning methods were not able to forecast the success of songs. This assertion was soon challenged in (8), where the authors argued that, given a relevant set of acoustic song features, forecasting song popularity was possible with more specific machine learning approaches. In these early works, the emphasis was primarily on *intrinsic* acoustic and lyrics features. Later, other studies improved outcomes with more sophisticated methods and expansive datasets. Specific song attributes such as "happiness", "partyness" and repetitiveness were found to increase the chances of songs becoming hits (9, 10). Deep convolutional structures and machine learning regressions were convincingly shown to have predictive power (11–13).

While these studies focused exclusively on *intrinsic* properties of songs, newer studies have included *extrinsic* features such as metadata or artist popularity. It was claimed that extrinsic factors often carry increased predicting power, the most predictive feature being the previous popularity of the artist (5). (14, 15) confirmed this result by finding that both, the previous popularity of the artist and the support of the publisher, have a significant impact on the success chances of songs. In addition to having increased chances of making it into the charts, (16) found that songs of well-known artists and

Significance Statement

Machine learning models for the prediction of song popularity have become popular in the music industry. While acoustic features, genre, or artist popularity are known to drive popularity, less attention has been given to social aspects of music perception. Here we use the concept of homophily –the idea that social ties form predominantly between people with similar features– to design a new predictive *influence*-parameter. Using this parameter in machine learning models allows us to improve the prediction precision of whether a new song is likely to become popular by up to 50%. We find that social influence is an essential factor in how listeners choose new songs and is at least as relevant as artist popularity or genre information.

NR, ST, VDPS conceptualized the work and designed the model. NR collected and prepared the data and performed the analysis. NR and ST wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: stefan.thurner@meduniwien.ac.at

big publishers also stay there longer. Reference (17) showed that superstars significantly dominate the market share and measured a positive correlation between song success and the number of songs an artist has released previously. For *extrinsic* song attributes, (18) found that Support Vector Regression (SVR) performs slightly better than neural networks.

Predictive indicators were also found in Twitter data (19) where music-related hashtags such as *#nowplaying* were counted. The number of daily tweets about specific songs and artists is highly correlated with the listening trend of that song. Counting the same tags, (20) found a moderate correlation between the number of tweets and the chart position of a song, as well as a weak correlation with the sentiment of the tweets. These Twitter-based correlations hint at a social component to the success of songs and that this information might be encoded in social networks. A similar thought was followed in (21), where the spread of song popularity was modeled with a SIR disease spreading model. There, it is argued that social processes that lead to the spread of music are similar to those of disease spreading. For some genres, for which social connectivity is higher, songs appear to spread faster.

The fact that social networks co-create homophily –the tendency that people that share common treats are more likely to form social ties– has been observed in a variety of contexts, ranging from obesity (22) to performance in schools (23). Also, in music listening behavior, the concept of homophily has been studied (24). Recently, homophily's importance in relation to music preference was confirmed in a study on 1144 early adolescents, where music preference plays a significant role in selecting friends (25). The online music-listening platform and social network *last.fm*[*] offers an excellent ground for studying homophily as it simultaneously provides access to both social links and music listening records of users (26). In (27), the authors study *last.fm* data to predict friendship links. They find that music preference alone rarely leads to friendships. In most cases, sharing friends is the best predictor of future friendships, i.e. triadic closure that has been predicted in (28) and explains basic structures of social multilayer networks (29). Reference (30) confirmed that people with similar music preferences tend to cluster, indicating that friends tend to listen to similar music. Homophily in music listening was found in explorative behavior (31). By estimating how mainstream, novel, or diverse listening records of users are, they find that highly explorative people tend to look for friends that are similar. Reference (32) confirmed that result and designed a model that explains the finding, stressing that highly explorative users tend to be friends with users with similar discovery rates. While (31) use homophily to predict friendship links on *last.fm*, in (33, 34) it is investigated how information about new artists spreads through social links and quantified to which extent user behavior is copied. It is demonstrated how homophily can be leveraged to improve song recommendations by recommending new songs to users shortly after close friends are listening to them.

While acoustic and metadata-based success factors for predicting song popularity have been studied extensively, much less attention has been given to the influence of social interactions. It has been shown that these influence listening behavior significantly. In this work, we quantify how social interactions, in particular homophily, impact song popularity. We demonstrate how homophily emerges as users influence their friends to listen to specific songs, which drives the discovery and popularity of new songs through cascades of song recommendations. We quantify this user influence by state-of-the-art machine learning predictions of song popularity by estimating to what extent its predictions can be improved.

Homophily. To estimate homophily, we derived a dataset from the online music-listening platform and social network *last.fm*. Our data includes 300 million individual listening events of 10 million songs. It is enriched by the (undirected and unweighted) friendship network, where nodes represent users and links represent bidirectional friendships. It consists of 2.7 million nodes and 15 million links with an average degree of 11.6. The network is connected with a diameter of 6. This dataset allows us to link friendships to music-listening histories. For more details, see SI text A.

We first determine the music preferences of every user. Music preferences can be assessed in *last.fm* through user-defined tags. Users define and add tags such as *Rock*, *80s*, or *Acoustic* to songs, artists, or albums. Here, we focus on artist tags because they are more abundant than song or album tags and offer more reliable statistics. Every artist, a, can have multiple tags, t, which are represented as weights, w_{at}, ranging between 1 and 100, based on their frequency of appearance (100 corresponds to the most frequent tag for that artist). We then compute a music preference matrix, m_{ut}, by summing over the weights, w_{at}, for each tag, t, of each artist, a, for each time a user, u, is listening to a song by that artist. The music preference matrix, M, with elements $m_{u,t}$ is hence defined as

$$m_{ut} = \sum_{a,s} l_{us} i_{sa} w_{at} \,, \qquad [1]$$

where l_{us} is the number of times user, u, listened to song, s. $i_{sa} = 1$ if artist, a, interpreted song, s, and is zero otherwise. In the music preference matrix M, each row corresponds to a vector of music preference of one user, expressed by the weighted artist tags. To compare the music preference vectors of users i and j, we use the cosine-similarity, defined as

$$\cos(\theta_{ij}) = \frac{\sum_t m_{u=i,t} m_{u=j,t}}{\sqrt{\sum_t m_{u=i,t}^2}\sqrt{\sum_t m_{u=j,t}^2}} \,. \qquad [2]$$

The idea will be to compare the alignment of music preference vectors for users that are friends to the alignment between randomly picked users (that are very likely not friends).

Influencer network. As a next step, we define the influencer network. Influence is the number of times a user's friends copied the listening behavior of that user within a specified time window. We define it algorithmically. Whenever a friend, u_j, of a user, u_i, listens to a song for the first time at time t_j shortly after user, u_i, listened to that song at time t_i, the influence, I_{ij}, of user, i, on user, j, increases by one if $t_j - t_i < \Delta$ is smaller than some threshold Δ. A sketch of the concept and the derived influencer network are shown in Fig. 1. The influencer network (orange) is a directed network where users are nodes, and a weighted link represents the strength of a user's influence over another. In our case, it consists of

[*] https://www.last.fm/

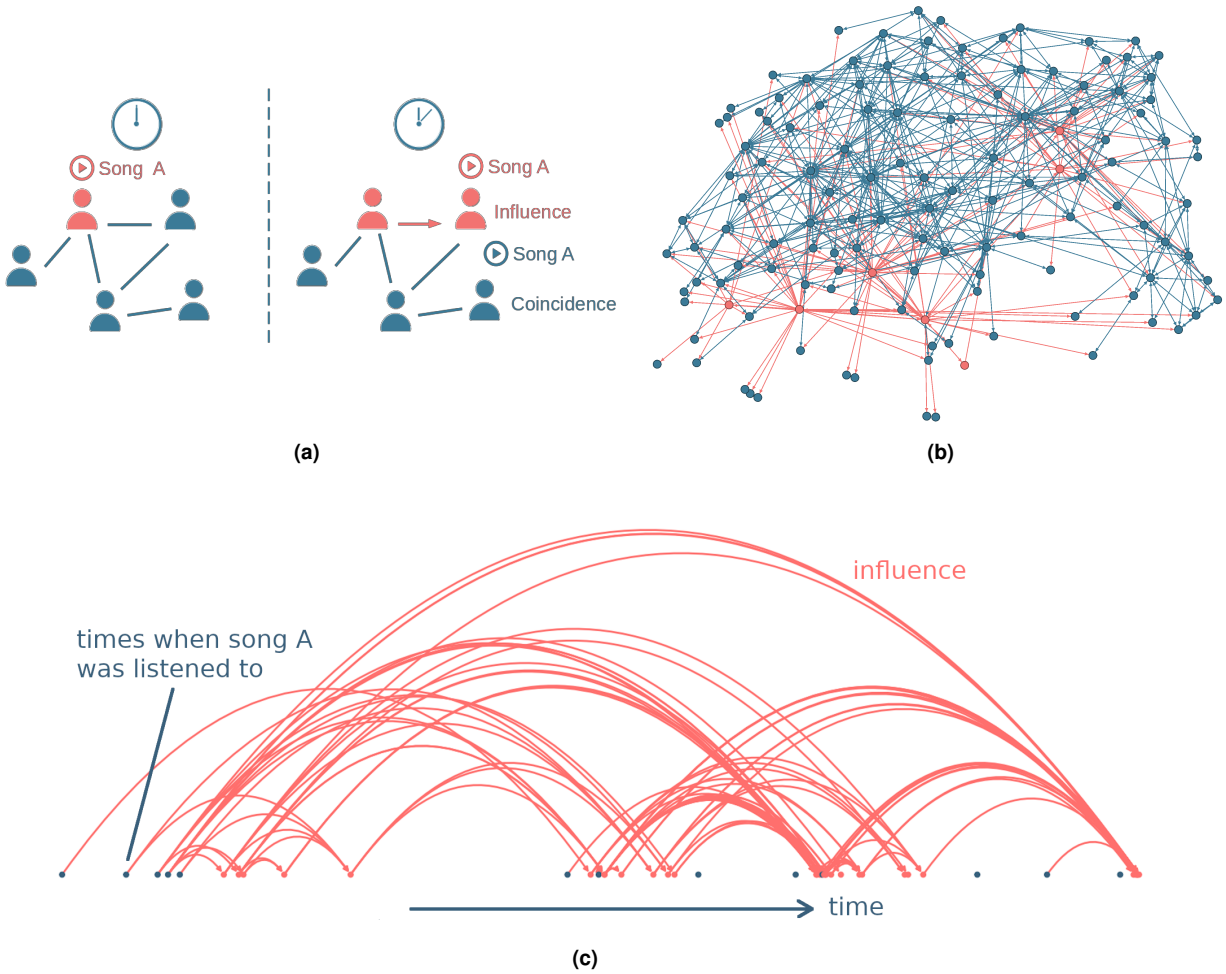


Fig. 1. (a) Schematic depiction of user influence. If a user listens to a song at time t_1 and a friend of that user listens to the same song for the first time at time t_2 with $t_2 - t_1 < \Delta$ smaller than some threshold Δ, the second user is said to have been influenced by the first with respect to that song. The influencer network is shown in orange, the friendship network in blue. The influencer network is a sub-network of the friendship network. (b) A fraction of the influencer network of *last.fm* from a 24h time window. Several strong influencers are visible (influencers and influencing links in orange) within the friendship network (blue). (c) Example of a timeline of a single song. Dots along the x-axis represent different users that listen to a song for the first time (location of dot). Dots are ordered in time, from left to right. Blue dots are users that found the song on their own, while orange dots represent users that were influenced by their friends. Arrows mark influencing events, where one user influences another into listening to a song.

9200 nodes and 190,000 links, with an average degree of 21. The influencer network is connected and has a diameter of 11. Since friendship is a requirement for influence, the influencer network is a sub-network of the friendship network (blue), all nodes and links present in the influencer network also exist in the friendship network. Influence on *last.fm* is possible because users can see which songs their friends are listening to and can give specific song recommendations to them.

Prediction strategy. To simplify the prediction task of the future success of a song, we only classify songs into average songs and future hit songs. We define hit songs as songs with at least 1000 listenings in any one month, which corresponds to the top 1% of songs. We do this classification based on an initial sample of features or parameters derived from the first 200 times a new song has been listened to. We define three classes of predictive parameters: preferential-attachment-based, time-based, and homophily-based. The preferential-attachment-based parameters are (i) the previous popularity of the artist and (ii) the genre's popularity. The time-based parameters are (iii) the time needed to reach 200 listenings, (iv) the tendency of users to re-listen to a song, and (v) the temporal trend quantified by the area under the curve of the cumulative listening count as a function of time. We also include three variations of these parameters, (vi) the number of users that listen to the song at least twice within a week, (vii) a normalized variation of (v), where we take the average y-value instead of the area, and (viii) a variation of (v) where we subtract the y-value from the diagonal and then compute the area. These constitute the basic eight parameters of the model. In addition to these parameters, we define homophily-based parameters and compute them for every song. For these, we identify the users that contributed to the first 200 times a song has been listened to. We use (ix) the influence scores, as defined above, as well as (xii) the cosine similarity between users and their friends as homophily-based parameters. Finally,

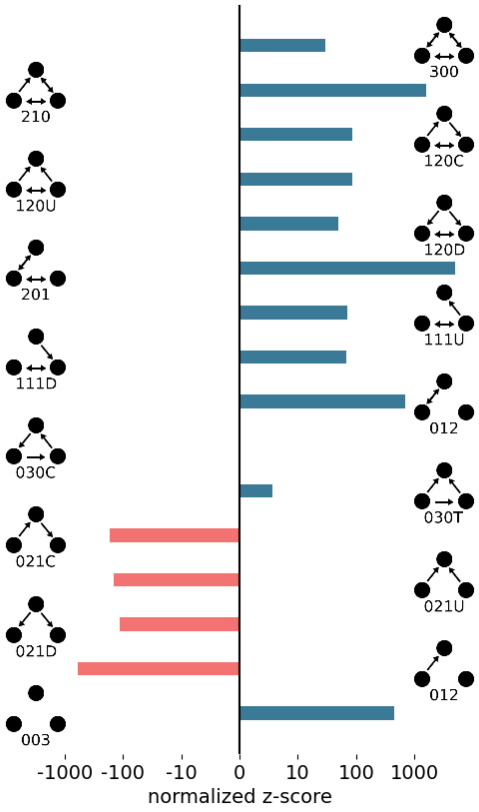


Fig. 2. Triadic census of the influencer network. Comparison to a shuffled network (configuration model), averaged over 100 random shufflings. Triad names follow the convention of (35).

we compute (xiii) the degree, (xiv) the PageRank, (xv) the nearest neighbor degree, and (xvi) the clustering coefficient both on the friendship network and the influencer network (xvii-xxi). All parameters are described in detail in SI text B. We use these parameters as input for a machine learning ensemble, see methods.

Results

Influencer network. Analyzing the triad statistics of the influencer networks, we find high levels of reciprocity in the influencer network, see Fig. 2. This is seen by the fact that triangles that include reciprocal links are strongly over-represented (blue bars) with respect to a random graph with the same number of nodes and (directed) links (shuffled network, configuration model). Triangles with no reciprocity are under-represented (red). This finding indicates that, generally, influence goes in both directions. Users who discover a new song from a friend (getting influenced) often influence other friends into listening to the song, leading to cascades, such as shown in Fig. 1 (c). In the SI text C, we identify the "influencers" and "followers"

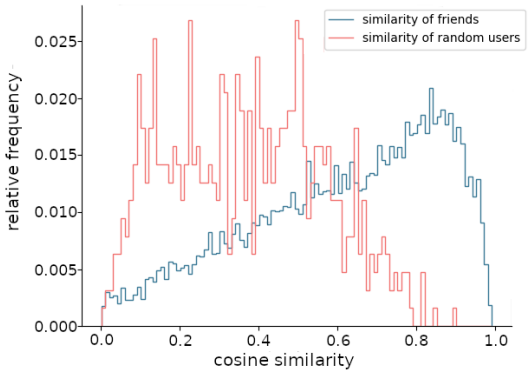


Fig. 3. Cosine similarity distribution for users and their friends (blue), and between random users (red) that are typically not befriended. From the distribution, it becomes apparent that very similar people are almost certainly friends on *last.fm*.

by comparing the number of times users got influenced with how often they influenced others.

Homophily results. We confirm the presence of strong homophily among users that are friends on *last.fm* by comparing the musical preferences of 1000 randomly picked users to those of their friends. We find an average cosine-similarity of musical-preference vectors $\vec{m}_u$ of $\cos(\theta) = 0.58$. In contrast, when comparing the musical preference vectors to an equal number of randomly picked users, the average cosine-similarity drops to $\cos(\theta) = 0.25$. A histogram of the respective cosine-similarities is shown in Fig. 3. The t-test for independent samples has a test statistic of 32 and a p-value of $p \leq 10^{-200}$. If the entries of the music preference vectors of the randomly picked users are shuffled in a random fashion, similarity decreases to levels below 10^{-4} and essentially disappears.

We find that both the user's influence and the tendency to get influenced correlate with the average similarity in musical preferences. When comparing users to their friends, this correlation is highly significant with a $p \leq 10^{-4}$. In contrast, when comparing random users, we do not find any correlation. For more details, see SI text D.

Improving predictions. Both the homophily score and the influence score correlate with song popularity. Parameters that correlate with popularity can be tested if they also predict it. The Pearson correlation coefficients, r, between all parameters and the song popularity are found in Tab. 1. The strongest (anti) correlation for song popularity we find for the area under the temporal listening trend with $r = -0.22$. Other time-based parameters have comparable correlations. The influence score correlates better ($r = 0.17$) with song popularity than the previous popularity of the artist ($r = 0.14$). Also, influencer network based parameters correlate similarly as artist popularity.

Table 2 shows the prediction results for three different models quantified by accuracy, precision, and recall. We predict if a song becomes a hit-song or an average song based on information from the 200 first listenings and the structure of

Table 1. Pearson correlation coefficients between prediction parameters and song popularity. Bold numbers show the strongest correlations for homophily-based and baseline parameters. The prediction parameters were computed on the first 200 listenings for each song. Song popularity was defined as the maximum number of times a song was listened to *last.fm* in its best month. All p-values are below 10^{-200}. Only the best-performing variations of parameters are listed here. All other variations can be found in SI text B.

parameter/feature	r
degree (friendship network)	0.090
degree (influencer network)	0.132
PageRank (influencer network)	0.110
influence	**0.174**
homophily	0.081
time to reach 200 listenings	-0.201
genre average	0.069
time between repeated listenings	-0.193
area under the trend curve	**-0.223**
previous popularity of artist	0.135

Table 2. Classification results for three different cases: (a) the combined model (preferential attachment, time, and homophily), (b) the model with only homophily-based social network parameters, and (c) the baseline model without explicit social network information. We see that a combined model (a) performs best, with the highest accuracy, hit precision, and non-hit recall. The homophily-based model (b) performs well on its own but is outclassed by the baseline model.

model	prediction category	
(a) combined model	accuracy	**0.98**
	hit precision	**0.21**
	non-hit precision	1.0
	hit recall	**0.60**
	non-hit recall	**0.99**
(b) social network only	accuracy	0.95
	hit precision	0.05
	non-hit precision	1.0
	hit recall	0.40
	non-hit recall	0.96
(c) without social network	accuracy	**0.97**
	hit precision	**0.14**
	non-hit precision	1.0
	hit recall	**0.70**
	non-hit recall	**0.98**

the influencer and friendship networks. A hit song is defined as a song that is listened to at least 1000 times in its best month, putting it in the top 1% of all songs. For the classification task, we use a machine learning ensemble including classifiers based on Support Vector Classification, Random Forest, Ada Boost, Gradient Boost, K-Neighbors, and a multilayer perceptron neural network, see Methods. Based on the input parameters, the machine learning models try to classify each song into one of two classes: hit songs or regular songs. In the first model, which we refer to as the combined model (a), we use the combined parameter sets with preferential-attachment-, time-, and homophily-based parameters. In the second, the social networks only scenario (b), we use the homophily-based parameters only. In the third, we combine the preferential-attachment and time-based parameters without using the homophily-based parameters and refer to it as the baseline model (c). The baseline model is based on commonly used parameters from the literature and forms the baseline against which we want to compare the results of models (a) and (b).

The strongest result is the improvement of hit-precision by 50% in the combined model (a) when compared to the baseline model (c). In addition, non-hit recall and accuracy improved by one percentage point at an already high level. On the downside, hit-recall decreased by around 14%. Adding the homophily-based influence score as a predictive parameter significantly improves accuracy and precision, as well as non-hit recall, at the cost of hit-recall. This suggests that the combined model (a) is able to highly reduce the number of false positives at the cost of missing a small number of true positives. The homophily-based model (b) on its own manages to already identify 40% of all hit songs correctly and reaches an accuracy of 95%. We see that while conventional models like our baseline model (a) are superior to the homophily-based model (b), the homophily-based model performs already on a high level, and a combination of both leads to the best results.

Discussion

We demonstrated how social interactions can be used to enhance song popularity predictions using a large dataset collected from the online platform *last.fm*. From an influencer network we derive an influence score for every user that captures their tendency to influence others to imitate their listening behavior. Based on a small sample of first listenings, we compute several metrics on the influencer network and use these as the predictive parameters in a machine learning classifier ensemble to categorize songs into potential hits and average songs. Our model exhibits up to 50% improved precision over a baseline model that uses common preferential attachment and time-based parameters.

The used concept of influencing is closely related to homophily. Using the music preference vectors we estimate the similarity of tastes and find that people that are related through friendship links tend to have aligned tatses – i.e. strong homophily is confirmed, consistent with earlier findings of (30, 33, 36, 37). Influence represents the degree to which users can make their friends listening to the same music. This increases the similarity in the music preference vectors and drives homophily. We show that while some people tend to pioneer new music tastes and others tend to follow these pioneers, in most cases, influencer interactions go in both directions. This means that users that get influenced by their friends can in turn influence multiple other friends, leading to network structures that enable cascading spreading of new songs, fueling the new song's popularity.

Influence between users contributes to preferential attachment. The more people listen to a song, the more likely it gets recommended to friends, thus the probability of it being listened to increases. Parameters that relate to either preferential attachment, such as the previous popularity of the artist, or to forgetting, such as the listening trend, have been widely used in previous works to predict the popularity of songs (5, 14–16). In this study, we focus on these extrinsic properties and contribute new parameters that have the potential to improve the performance of existing models.

There are several severe limitations. Obviously, *last.fm* only provides a partial view on what is going on in music listening

and recommendations. There are many other channels where people listen to music and exchange information about new songs which leads to a "cold-start" problem: a song that is new on the *last.fm* platform doesn't need to be new to its users. Hence, song popularity predictions might get offset by events that are beyond the scope of the available data. This is in part related also to the issue of comparability. A multitude of different datasets is used across the literature, each of them limited in some aspect, for instance, to a specific platform or a geographic region (5, 9, 11, 12). This is coupled with an apparent lack of a consistent definition of hit songs. In addition to that, in models with many parameters and hyperparameters, performance might be optimized with regard to different metrics. Altogether, specific results are difficult to compare across different studies. For this reason, we chose to use our own baseline model as a benchmark and aim for precise and intuitive definitions of popularity and hit songs, such that different machine learning models might be compared in a consistent way. Further, to some degree, popularity predictions might be self-fulfilling prophecies. It has been observed that people tend to reproduce perceived song popularity that is presented to them, even if these have been heavily modified (38–40). Finally, there is also an ongoing discussion on whether artists should focus on improving popularity over other goals. Most popular doesn't necessarily imply most enjoyable or most relevant (41). Rather, it indicates high commercial relevance, which might not necessarily be the top priority. However, even given these shortcomings, in this work, we were able to compare the predictions within a closed framework. Our main result is that we are able to find a substantial relative increase in predictability by including social information. The concept presented here can straightforwardly be transferred to other domains such as movies, books, posts, or even physical goods. Given the growing availability of social network data, comparable approaches might further uncover the social underpinnings of consumer behavior in our society.

Methods

For song popularity predictions, we use a machine learning ensemble. The ensemble includes classifiers based on Support Vector Classification, Random Forest, Ada Boost, Gradient Boost, K-Neighbors, and a neural network. The different classifiers hold a majority vote on the classification of each song. Each song is classified as either a future hit-song or an average song, where hit-songs are defined as songs in the top 1% of songs, which equates to being listened to at least 1000 times in the best month in our dataset. In the following, we give a brief overview of the machine learning models used, the specific implementation, and their hyperparameter settings.

Support Vector Classification is a supervised learning model that tries to map training data into a higher dimensional space with the aim of maximizing the gap between points of different classes in that space (42). In this study, we use the Python sklearn implementation (43) with balanced class weights and a value of $C = 1$ for the regularization hyperparameter.

Random Forest classification is a supervised learning model that builds multiple randomized decision trees based on the training set (44). The classification outcome is then the majority vote of the individual trees. In this study, we use the Python sklearn implementation (45) with 100 estimators and balanced class weights.

Ada Boost, short for adaptive boosting, is a classifier that iteratively learns from the mistakes of weak classifiers, turning them into strong classifiers (46). In this study, we use the Python sklearn implementation (47) with 100 estimators.

Gradient Boost classifier is a classifier that works as an ensemble of weak classifiers, typically decision trees. These weak classifiers are gradually added during the learning process while aiming for maximum correlation with the negative gradient of the loss function (48). In this study, we use the Python sklearn implementation (49) with 100 estimators, a learning rate of 1, and a maximum depth of 1.

K-nearest neighbor classification is based on a multidimensional feature space that is populated by the feature vectors of the training set and their labels (50). During classification, one looks at the nearest k neighbors of an element and attaches the label to it that is most common among its neighbors. In this study, we use the Python sklearn implementation (51) with the number of nearest neighbors k equal to 5.

A multilayer perceptron is a fully connected, feed-forward artificial neural network that consists of at least three layers: an input, a hidden layer, and an output layer (52). All nodes in each layer are connected to all other nodes of the following layer. It uses non-linear activation functions and learns through back-propagation. In this study, we use the Python sklearn implementation (53) with a maximum number of iterations of 1000.

These machine learning models are each trained on (the same) 60% of the data and then used to classify the other 40%. The data is classified according to the majority vote of the different models.

References

1. T Ingham, Over 60,000 tracks are now uploaded to spotify every day. that's nearly one per second (https://www.musicbusinessworldwide.com/over-60000-tracks-are-now-uploaded-to-spotify-daily-thats-nearly-one-per-second/) (2021) Accessed: 2022-08-25.
2. B Houghton, 60,000 tracks are uploaded to spotify every day (https://www.hypebot.com/hypebot/2021/02/60000-tracks-are-uploaded-to-spotify-every-day.html) (2021) Accessed: 2022-08-25.
3. HB Hu, DY Han, Empirical analysis of individual popularity and activity on an online music service system. *Phys. A: Stat. Mech. its Appl.* **387**, 5916–5921 (2008).
4. S Rosen, The economics of superstars. *The Am. economic review* **71**, 845–858 (1981).
5. J Pham, E Kyauk, E Park, Predicting song popularity. *Dept. Comput. Sci., Stanf. Univ., Stanford, CA, USA, Tech. Rep* **26** (2016).
6. R Dhanaraj, B Logan, Automatic prediction of hit songs in *proceedings of the international conference on music information retrieval (ISMIR)*. pp. 488–491 (2005).
7. F Pachet, P Roy, Hit song science is not yet a science. in *ISMIR*. pp. 355–360 (2008).
8. Y Ni, R Santos-Rodriguez, M Mcvicar, T De Bie, Hit song science once again a science in *4th International Workshop on Machine Learning and Music*. (Citeseer), (2011).
9. M Interiano, et al., Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Soc. Open Sci.* **5**, 171274 (2018).
10. A Lassche, F Karsdorp, E Stronks, Repetition and popularity in early modern songs in *DH 2019: Proceedings of the 2019 Digital Humanities Conference*. (2019 Digital Humanities Conference), (2019).
11. LC Yang, SY Chou, JY Liu, YH Yang, YA Chen, Revisiting the problem of audio-based hit song prediction using convolutional neural networks in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 621–625 (2017).
12. K Middlebrook, K Sheik, Song hit prediction: Predicting billboard hits using spotify data. *CoRR* **abs/1908.08609** (2019).
13. D Martin-Gutierrez, GH Penaloza, A Belmonte-Hernandez, FA Garcia, A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access* **8**, 39361–39374 (2020).
14. N Askin, M Mauskapf, What makes popular culture popular? product features and optimal differentiation in music. *Am. Sociol. Rev.* **82**, 910–944 (2017).
15. S SHIN, J PARK, On-chart success dynamics of popular songs. *Adv. Complex Syst.* **21**, 1850008 (2018).

16. H Im, H Song, J Jung, A survival analysis of songs on digital music platform. *Telematics Informatics* **35**, 1675–1686 (2018).
17. ST Kim, JH Oh, Music intelligence: Granular data and prediction of top ten hit songs. *Decis. Support. Syst.* **145**, 113535 (2021).
18. H Yu, Y Li, S Zhang, C Liang, Popularity prediction for artists based on user songs dataset in *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, ICCAI '19. (Association for Computing Machinery, New York, NY, USA), p. 17–24 (2019).
19. Y Kim, B Suh, K Lee, #nowplaying the future billboard: Mining music listening behaviors of twitter users for hit song prediction in *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14. (Association for Computing Machinery, New York, NY, USA), p. 51–56 (2014).
20. E Tsiara, C Tjortjis, Using twitter to predict chart position for songs in *Artificial Intelligence Applications and Innovations*, eds. I Maglogiannis, L Iliadis, E Pimenidis. (Springer International Publishing, Cham), pp. 62–72 (2020).
21. DP Rosati, MH Woolhouse, BM Bolker, DJD Earn, Modelling song popularity as a contagious process. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **477**, 20210457 (2021).
22. NA Christakis, JH Fowler, The spread of obesity in a large social network over 32 years. *New Engl. J. Medicine* **357**, 370–379 (2007).
23. I Smirnov, S Thurner, Formation of homophily in academic performance: Students change their friends rather than performance. *PLOS ONE* **12**, e0183473 (2017).
24. LSL Miller McPherson, JM Cook, Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
25. A Franken, L Keijsers, JK Dijkstra, T ter Bogt, Music preferences, friendship, and externalizing behavior in early adolescence: A SIENA examination of the music marker theory using the SNARE study. *J. Youth Adolesc.* **46**, 1839–1850 (2017).
26. P Mechant, T Evens, Interaction in web-based communities: a case study of last.fm. *Int. J. Web Based Communities* **7**, 234–249 (2011).
27. K Bischoff, We love rock 'n' roll: Analyzing and predicting friendship links in last.fm in *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12. (Association for Computing Machinery, New York, NY, USA), p. 47–56 (2012).
28. MS Granovetter, The strength of weak ties. *Am. journal sociology* **78**, 1360–1380 (1973).
29. M Sadilek, P Klimek, S Thurner, Asocial balance—how your friends determine your enemies: understanding the co-evolution of friendship and enmity interactions in a virtual world. *J. Comput. Soc. Sci.* **1**, 227–239 (2017).
30. R Guidotti, G Rossetti, "know thyself" how personal music tastes shape the last.fm online social network in *Formal Methods. FM 2019 International Workshops*, eds. E Sekerinski, et al. (Springer International Publishing, Cham), pp. 146–161 (2020).
31. T Duricic, D Kowald, M Schedl, E Lex, My friends also prefer diverse music: Homophily and link prediction with user preferences for mainstream, novelty, and diversity in music in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21. (Association for Computing Machinery, New York, NY, USA), p. 447–454 (2022).
32. G Di Bona, et al., Social interactions affect discovery processes (2022).
33. R Pálovics, AA Benczúr, Temporal influence over the last.fm social network. *Soc. Netw. Analysis Min.* **5** (2015).
34. R Pálovics, AA Benczúr, Temporal influence over the last.fm social network in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13. (Association for Computing Machinery, New York, NY, USA), p. 486–493 (2013).
35. S Wasserman, K Faust, *Social Network Analysis: Methods and Applications*, Structural Analysis in the Social Sciences. (Cambridge University Press), (1994).
36. Z Zhou, K Xu, J Zhao, Homophily of music listening in online social networks of china. *Soc. Networks* **55**, 160–169 (2018).
37. H Bisgin, N Agarwal, X Xu, A study of homophily on social media. *World Wide Web* **15**, 213–232 (2011).
38. MJ Salganik, PS Dodds, DJ Watts, Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
39. MJ Salganik, DJ Watts, Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Soc. Psychol. Q.* **71**, 338–355 (2008).
40. FB Lynn, MH Walker, C Peterson, Is popular more likeable? choice status by intrinsic appeal in an experimental music market. *Soc. Psychol. Q.* **79**, 168–180 (2016).
41. B Monechi, P Gravino, VDP Servedio, F Tria, V Loreto, Significance and popularity in music production. *Royal Soc. Open Sci.* **4**, 170433 (2017).
42. Cw Hsu, Cc Chang, CJ Lin, A practical guide to support vector classification. *J. Mach. Learn. Res.* (2003).
43. scikit-learn developers, Svc in python sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html) (2022) Accessed: 2022-11-22.
44. G Biau, E Scornet, A random forest guided tour. *TEST* **25**, 197–227 (2016).
45. scikit-learn developers, Random forest in python sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) (2022) Accessed: 2022-11-22.
46. RE Schapire, Explaining AdaBoost in *Empirical Inference*. (Springer Berlin Heidelberg), pp. 37–52 (2013).
47. scikit-learn developers, Ada boost in python sklearn (https://scikit-learn.org/stable/modules/ensemble.html#adaboost) (2022) Accessed: 2022-11-22.
48. A Natekin, A Knoll, Gradient boosting machines, a tutorial. *Front. neurorobotics* **7**, 21 (2013).
49. scikit-learn developers, Gradient tree boosting in python sklearn (https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting) (2022) Accessed: 2022-11-22.
50. T Cover, P Hart, Nearest neighbor pattern classification. *IEEE Transactions on Inf. Theory* **13**, 21–27 (1967).
51. scikit-learn developers, K neighbors classifier in python sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html) (2022) Accessed: 2022-11-22.
52. T Hastie, R Tibshirani, J Friedman, *The Elements of Statistical Learning*. (Springer New York), (2009).
53. scikit-learn developers, Multi-layer perceptron classifier (https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html) (2022) Accessed: 2022-11-22.

Supplementary information

A. SI 1: Data availability of last.fm. Table 3 shows the number of songs, listenings, albums, artists and users in the dataset that we collected from last.fm for this study. Users are further broken down into users for which the full friendship data i.e. the account names of all their friends are known and users for which we collected the full listen history. Figure 4 shows the timeline of last.fm and marks the time periods for which data was fetched. The bootstrap phase was used to bootstrap the popularity of artists.

Table 3. Available data that was collected on the last.fm for the purpose of this study. Numbers are rounded down.

songs	10,000,000
listenings	300,000,000
albums	200,000
artists	1,000,000
users	2,500,000
with friendship data	100,000
with listen history	17,500

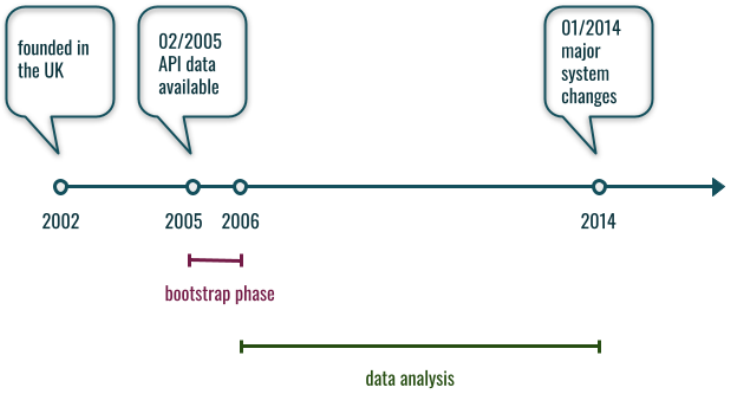


Fig. 4. last.fm timeline and data availability. The company was founded in 2002 in the UK. User data is available on the API starting from February 2005. In 2014 there was a major change to the system - users were not able anymore to listen to music directly on last.fm but rather could connect their last.fm account to other streaming services such as Spotify.

B. SI 2: prediction parameters. We test several social-network based parameters in our prediction model. For these we use two networks: the friendship network and the influencer network. The friendship network of *last.fm* consists of nodes that represent users and links that represent bidirectional friendships. The influencer network is a directed network where users are nodes, and a weighted link represents the strength of a user's influence over another. On these two networks, we define the following user metrics:

(a) the PageRank of each user (node), computed on the friendship network
(b) the nearest neighbor degree of each user, computed on the friendship network
(c) the degree of each user, computed on the friendship network
(d) the clustering coefficient of each user, computed on the friendship network
(e) the PageRank of each user, computed on the influencer network
(f) the nearest neighbor degree of each user, computed on the influencer network
(g) the degree of each user, computed on the influencer network
(h) the clustering coefficient of each user, computed on the influencer network
(i-k) user influence with time windows of 6h, 12h and 24h. The influence score of each user is divided by the number of songs where a user influenced another user thus giving us the number of influencing events per song. This is equivalent to the out-degree of the user in the influencer network divided by the number of songs.
(l) homophily as defined here by the average cosine similarity of the music preference vectors between a user and their friends

Each of these user metrics is computed for the users that contribute the first 200 listenings to a song and averaged per song. The average is then the value of the predictive parameter for that song.

The performance of these parameters is compared to the performance of commonly used parameters that form our baseline. Specifically, we test the following parameters:

(m) the time, Δt_j, it takes a song, j, to reach the first 200 listenings in last.fm, given by

$$\Delta t_j = t_{j,200} - t_{j,1} \quad [3]$$

where $t_{j,i}$ is the timestamp in seconds when song, j, was listened to for the ith time.

(n) the average number of listenings, that songs of the same genre receive in our dataset. We define the genre of a song, as the value of the the user-defined artist-tag with the largest weight (100).
(o) the average time that passes between two listenings of the same user, measured in seconds. Again, we only look at the first 200 listenings here. Whenever the same user listens more than once to the song withing those first 200 listenings, we compute the time that has passed in between. We then average these timespans per song.
(p) the number of users among the first 200 listenings that listen to the song again within a timespan of at most one week. This is similar to (o), but instead of looking at the time that passes in between, we simply count how many users listen to the song more than once.
(q) the area, A_j, under the curve if the cumulative listenings of a song, j, up to a number of 200 are plotted vs time. This is given by

$$A_j = \int_{t=t_{j,1}}^{t_{j,200}} l_j(t)dt \quad [4]$$

where $t_{j,i}$ is the timestamp in seconds when song, j, was listened to for the ith time and $l_j(t)$ is the total number of times song, j, has been listened to at time, t.

(r) the same area as in (q), but subtracted from the area under the diagonal. This is given by

$$A'_j = \frac{(t_{j,200} - t_{j,1})(l_j(t = t_{j,200}) - l_j(t = t_{j,1}))}{2} - \int_{t=t_{j,1}}^{t_{j,200}} l_j(t)dt \quad [5]$$

where $t_{j,i}$ is the timestamp in seconds when song, j, was listened to for the ith time and $l_j(t)$ is the total number of times song, j, has been listened to at time, t.

(s) the same area as in (q), but divided by the length of the x-axis (total time passed between the first and the 200th listening). This is given by

$$A''_j = \frac{\int_{t=t_{j,1}}^{t_{j,200}} l_j(t)dt}{t_{j,200} - t_{j,1}} \quad [6]$$

where $t_{j,i}$ is the timestamp in seconds when song, j, was listened to for the ith time and $l_j(t)$ is the total number of times song, j, has been listened to at time, t.

(t) previous popularity of the artist. For this, we take a snapshot in time when the new song is released and calculate the average number of listenings that songs of the same artist have received in the past.

Table 4 shows the correlation coefficients for the parameters listed above and the popularity of songs.

Table 4. Pearson correlation coefficients between prediction parameters and song popularity. The prediction parameters were computed on the first 200 listenings for each song. Song popularity was defined as the maximum number of times a song was listened to *last.fm* in its best month. All p-values are below 10^{-50}.

(a) pagerank (Friendship NW)	-0.054
(b) nearest neighbor degree (Friendship NW)	-0.081
(c) degree (Friendship NW)	0.090
(d) clustering coefficient (Friendship NW)	0.069
(e) pagerank (Influencer NW)	0.110
(f) nearest neighbor degree (Influencer NW)	0.090
(g) degree (Influencer NW)	0.132
(h) clustering (Influencer NW)	0.062
(i) influence 6h	0.159
(j) influence 12h	0.169
(k) influence 24h	0.174
(l) homophily	0.081
(m) time to reach 200 listenings	-0.201
(n) genre average	0.069
(o) time between repeated listenings	-0.193
(p) number of repeated listeners	0.039
(q) area under the trend curve	-0.223
(r) area under the trend curve (subtracted from diagonal)	-0.085
(s) area under the trend curve (normalized)	-0.044
(t) previous popularity of artist	0.135

C. SI 3: Influencers and followers. In Fig. 5, we report some heterogeneity of users in their influencing behavior. While some users tend to discover new songs early and (directly or indirectly) promote them in their friendship network –acting as "influencers"– others tend to follow these users and adopt their discoveries. We find that the higher the influence score of early song listeners, the quicker we can expect information about that song to spread in the friendship network. In the main text, we show how this information can improve machine learning models that predict the popularity of songs.

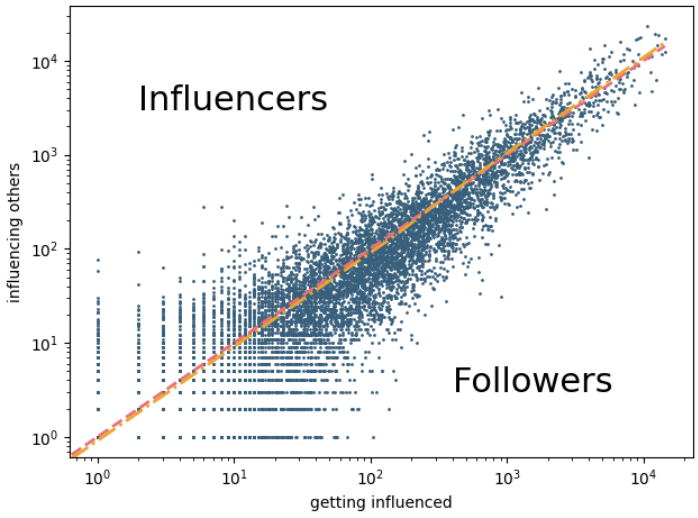


Fig. 5. Influencing others vs getting influenced by others. Each dot represents a user. The number of times a user was influenced into listening to a song is marked on the x-axis. The number of times a user influenced another user is marked on the y-axis. The red dashed line marks x=y. The yellow dashed and dotted line marks a power-law fit with $\alpha = 1.009$. Users far above the red line tend to act as "influencers", while users far below tend to act as "followers".

D. SI 4: Homophily vs influence. In figure 6 we show a scatterplot of how the cosine similarity of users relates to their influence score.

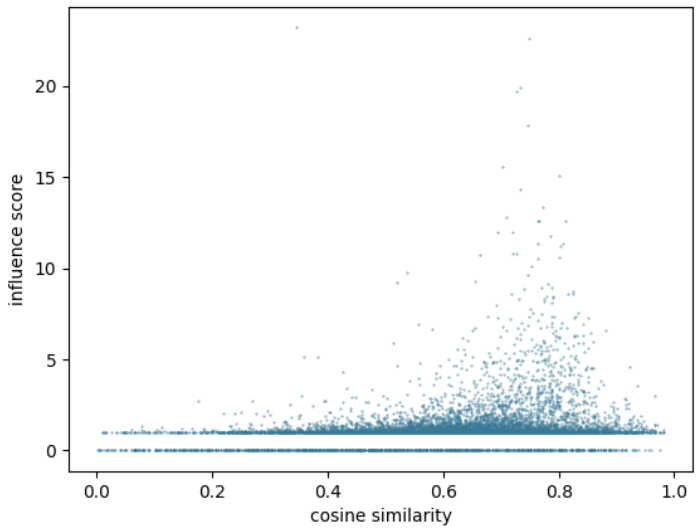


Fig. 6. User influence score vs average cosine similarity between a user and all of their friends. Each dot represents a user. The apparent gap between zero and one is due to to normalization of the influence score.

Chapter 6

Knowledge flows on short time scales

System, method, and computer-readable medium for tracking information exchange

Vittorio Loreto, Vito D. P. Servedio, Niklas Reisz, Stefan Thurner

Published in (pending): US Patent Application, number 2020/0320458 A1, 2020.

Information sharing tracking systems, methods and computer-readable storage media

Vittorio Loreto, Vito DP Servedio, Niklas Reisz, Stefan Thurner

Published in: Japan Published Patent, number JP7035105B2, 2022.

Note that these two patent applications have the same content and are listed separately here only for reference. The second application is a Japanese translation of the first, with a slightly different title.

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2020/0320458 A1**
Loreto et al. (43) **Pub. Date: Oct. 8, 2020**

(54) **SYSTEM, METHOD, AND COMPUTER-READABLE MEDIUM FOR TRACKING INFORMATION EXCHANGE**

(71) Applicants: **SONY CORPORATION**, Tokyo (JP); **Complexity Science Hub Vienna**, Vienna (AT)

(72) Inventors: **Vittorio Loreto**, Stuttgart (DE); **Vito D.P. Servedio**, Vienna (AT); **Niklas Reisz**, Vienna (AT); **Stefan Thurner**, Vienna (AT)

(73) Assignees: **SONY CORPORATION**, Tokyo (JP); **Complexity Science Hub Vienna**, Vienna (AT)

(21) Appl. No.: **16/375,017**

(22) Filed: **Apr. 4, 2019**

Publication Classification

(51) **Int. Cl.**

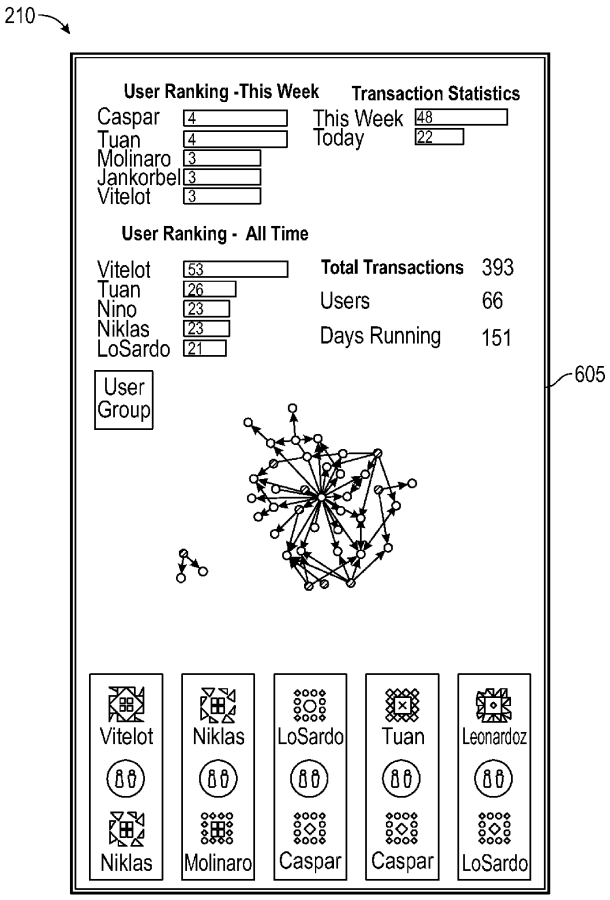
G06Q 10/06	(2006.01)
G06F 16/23	(2006.01)
H04L 29/06	(2006.01)
G06Q 20/36	(2006.01)
G06Q 50/00	(2006.01)

(52) **U.S. Cl.**

CPC ... ***G06Q 10/06398*** (2013.01); ***G06F 16/2365*** (2019.01); ***G06Q 50/01*** (2013.01); ***H04L 63/102*** (2013.01); ***G06Q 20/363*** (2013.01); ***G06F 16/2379*** (2019.01)

(57) **ABSTRACT**

An information exchange tracking system includes a server including processing circuitry configured to track an information exchange transaction, identify contributors in the information exchange transaction, determine a level of contribution of each contributor, the level of contribution being secured by a blockchain, and award each contributor based on their level of contribution. Accordingly, collaborative processes of co-creation are securely recorded using the blockchain and can be used to monitor the collective dynamics of the whole creative process while giving proper credits to all the relevant actors according to their specific contributions.



Patent Application Publication Oct. 8, 2020 Sheet 1 of 13 US 2020/0320458 A1

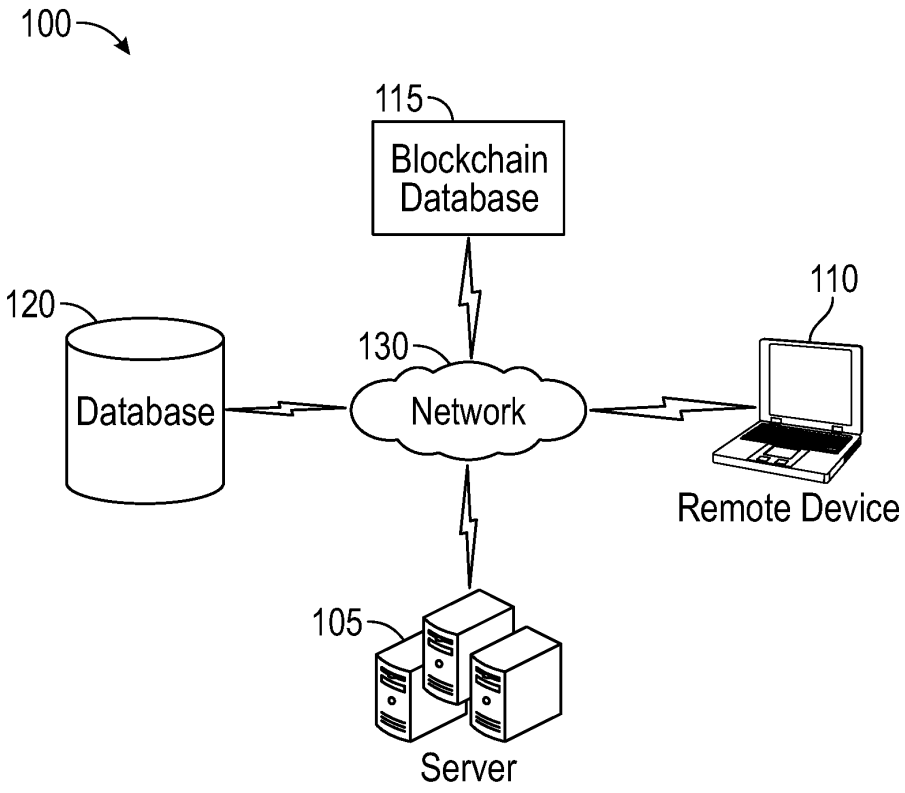


FIG. 1

Patent Application Publication Oct. 8, 2020 Sheet 2 of 13 US 2020/0320458 A1

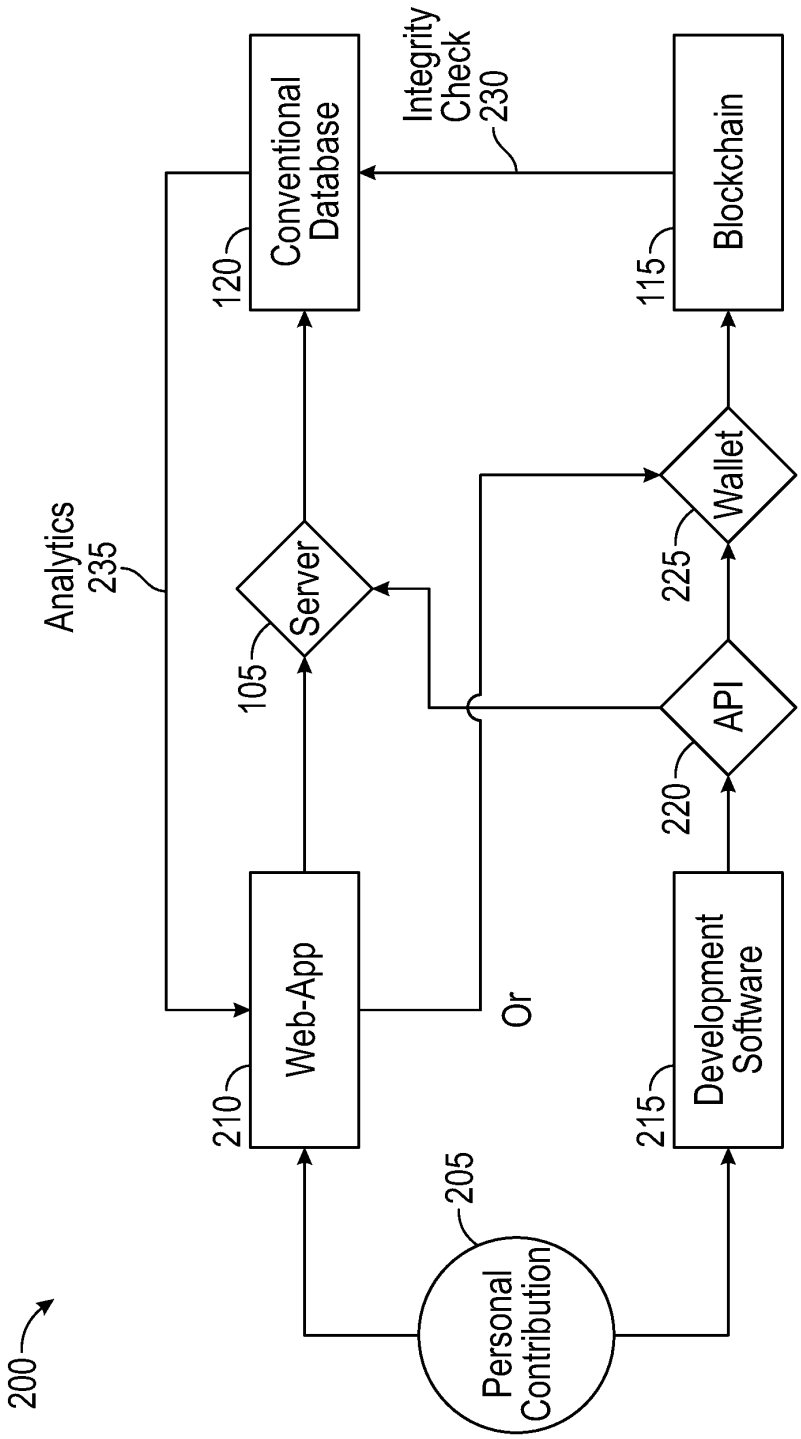


FIG. 2

Patent Application Publication Oct. 8, 2020 Sheet 3 of 13 US 2020/0320458 A1

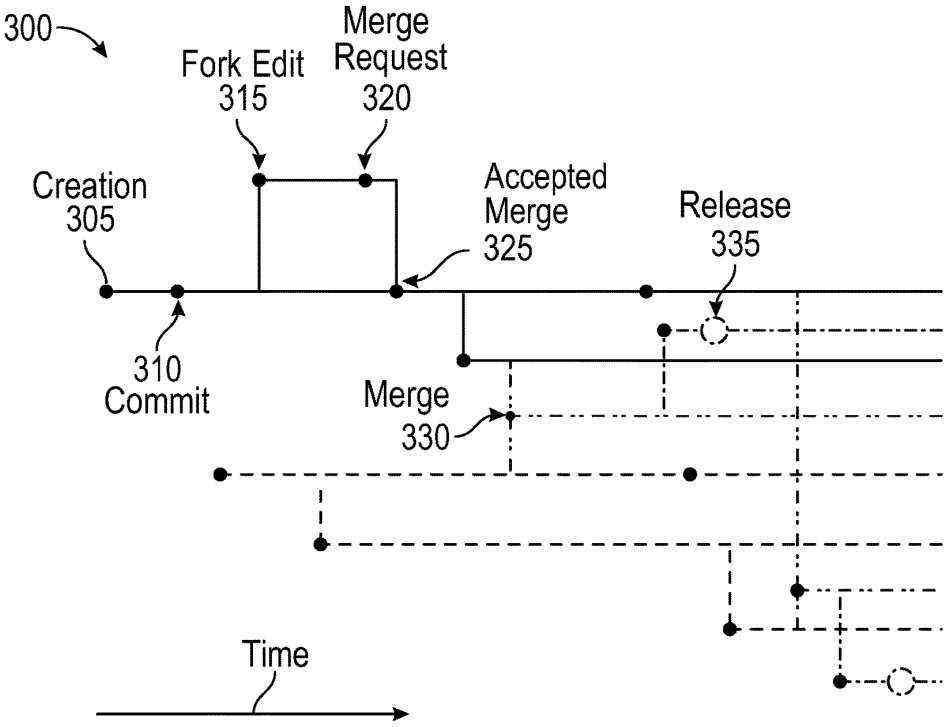


FIG. 3

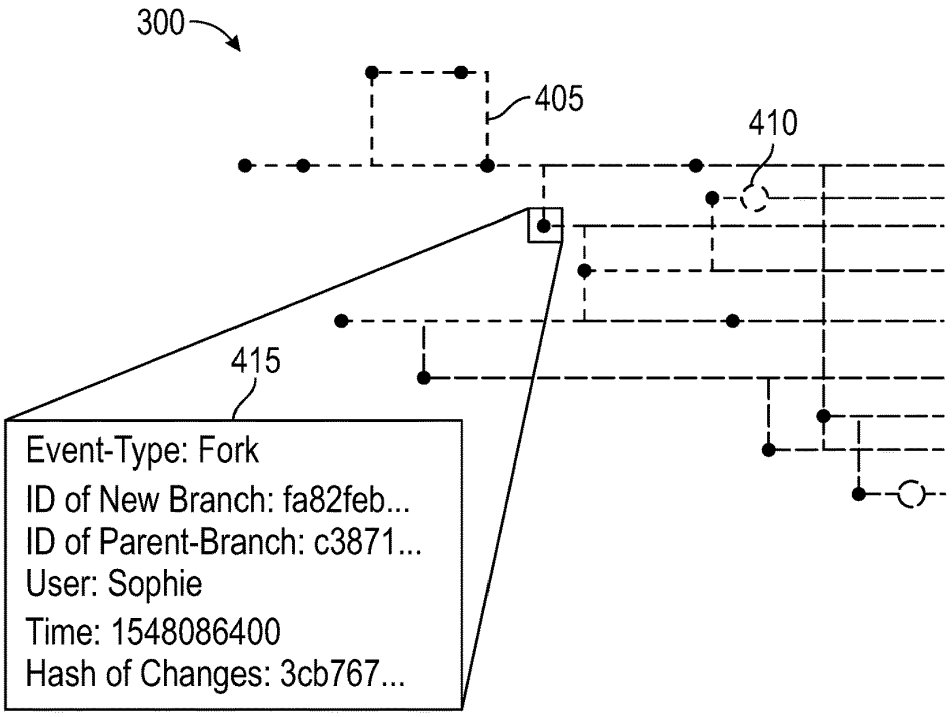


FIG. 4

Patent Application Publication Oct. 8, 2020 Sheet 4 of 13 US 2020/0320458 A1

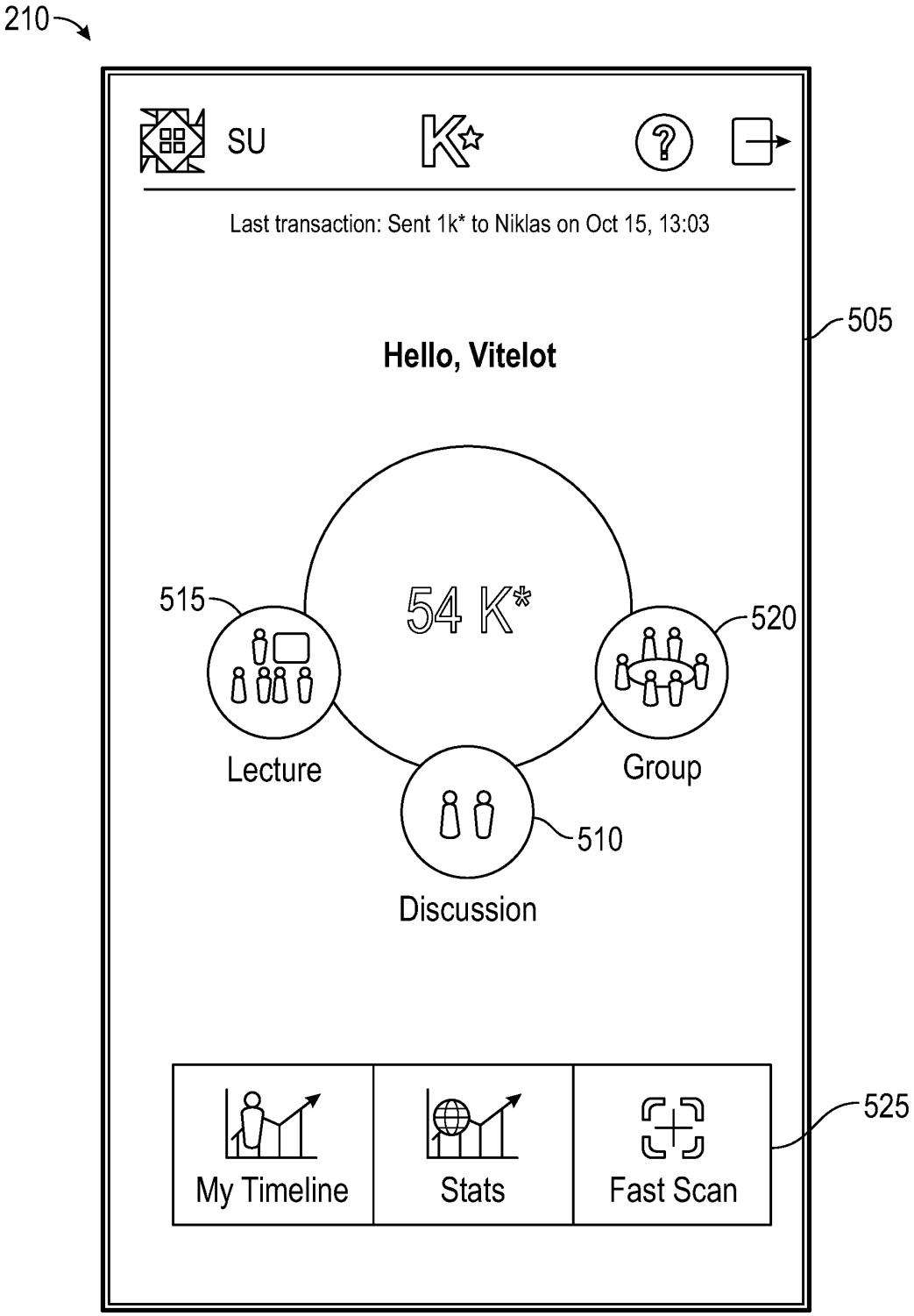


FIG. 5A

Patent Application Publication Oct. 8, 2020 Sheet 5 of 13 US 2020/0320458 A1

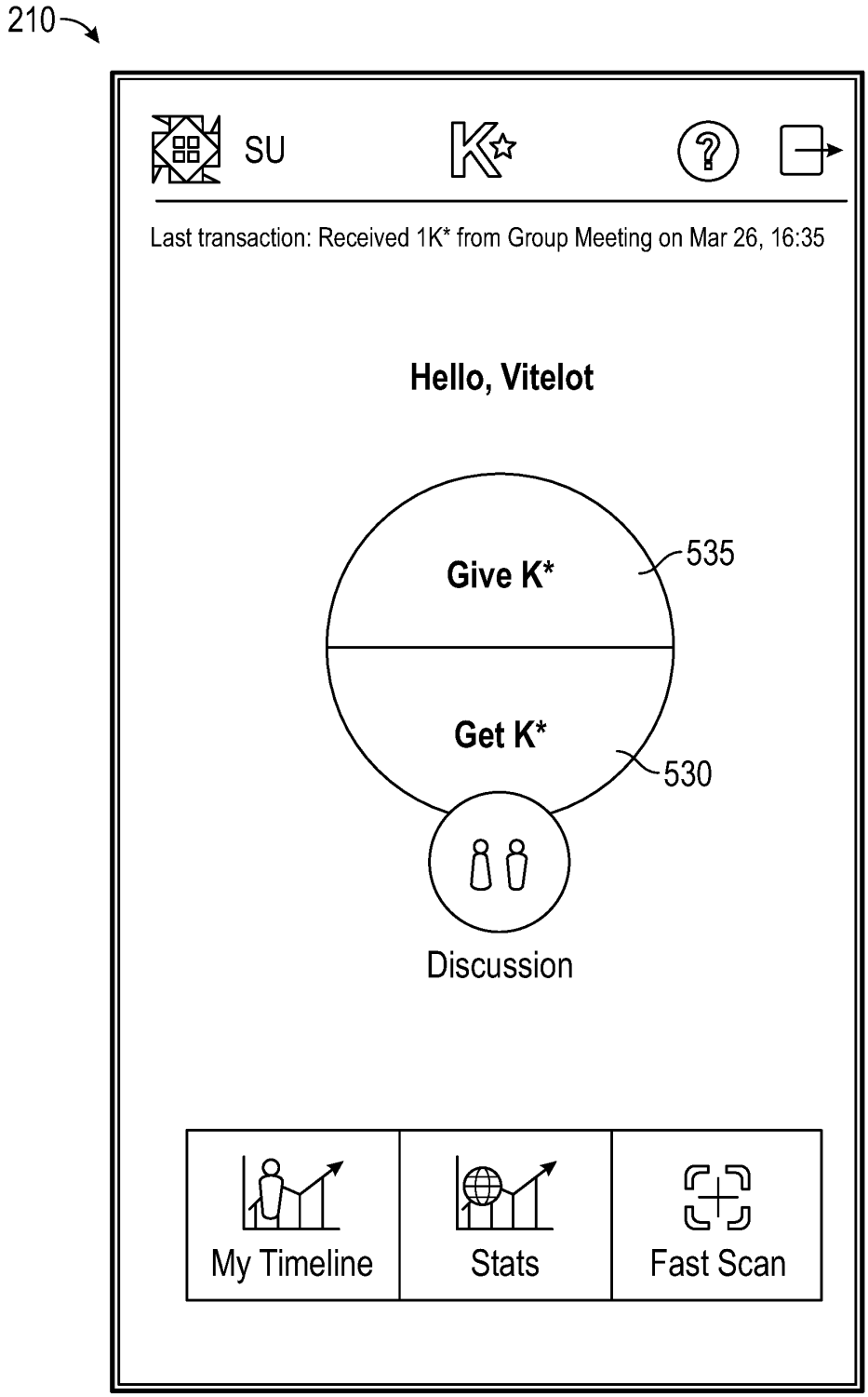


FIG. 5B

Patent Application Publication Oct. 8, 2020 Sheet 6 of 13 US 2020/0320458 A1

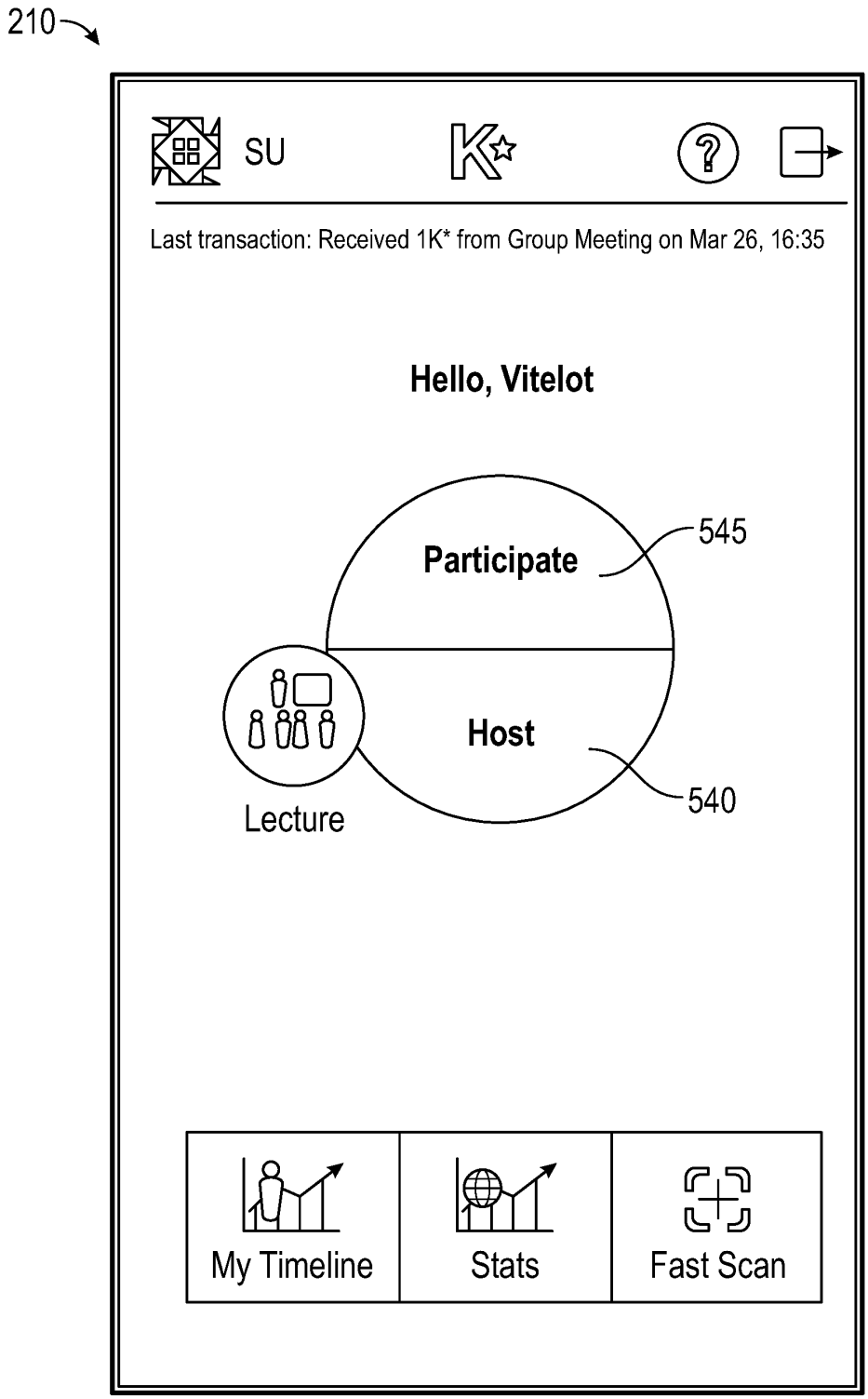


FIG. 5C

Patent Application Publication Oct. 8, 2020 Sheet 7 of 13 US 2020/0320458 A1

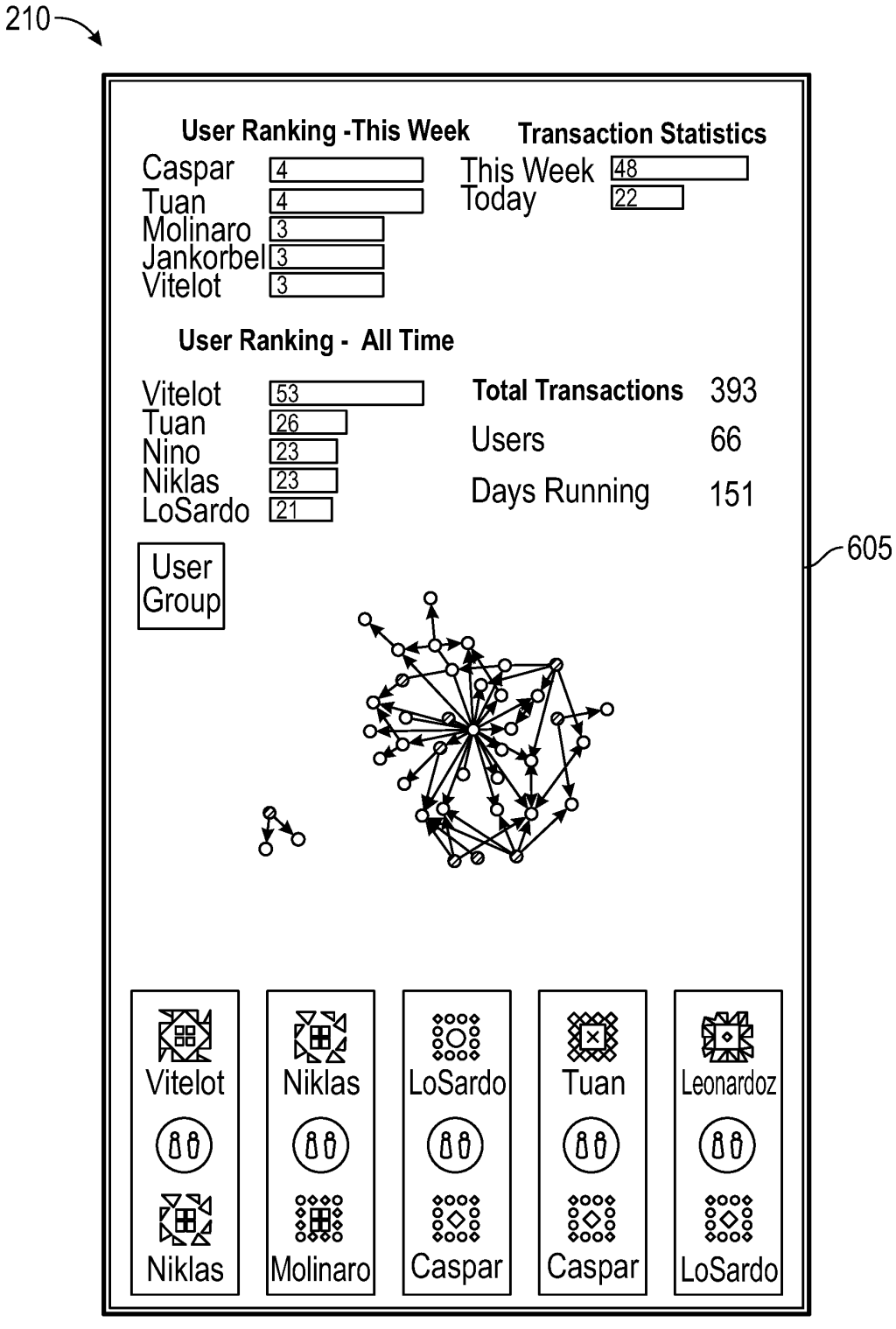


FIG. 6

Patent Application Publication Oct. 8, 2020 Sheet 8 of 13 US 2020/0320458 A1

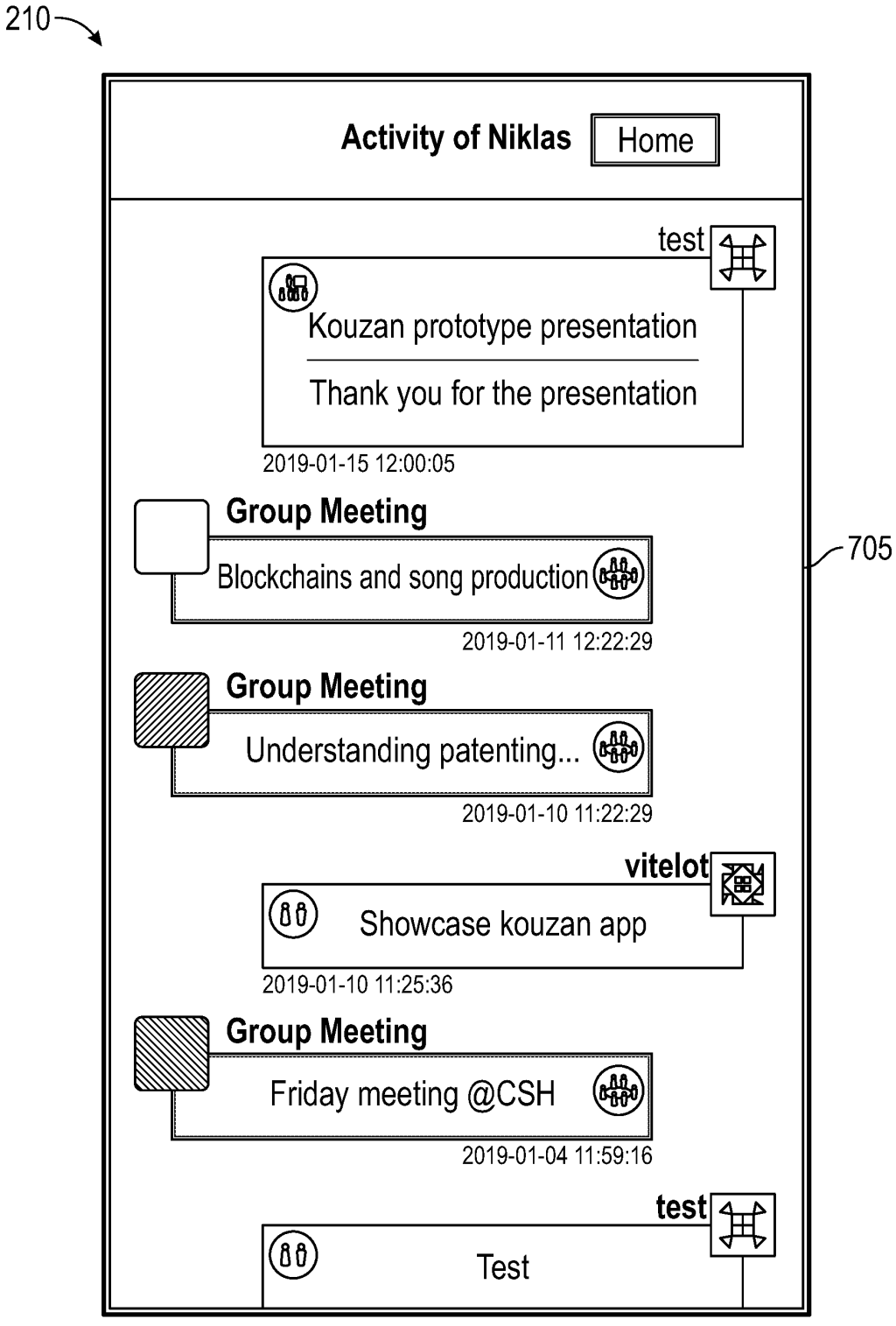


FIG. 7

Patent Application Publication Oct. 8, 2020 Sheet 9 of 13 US 2020/0320458 A1

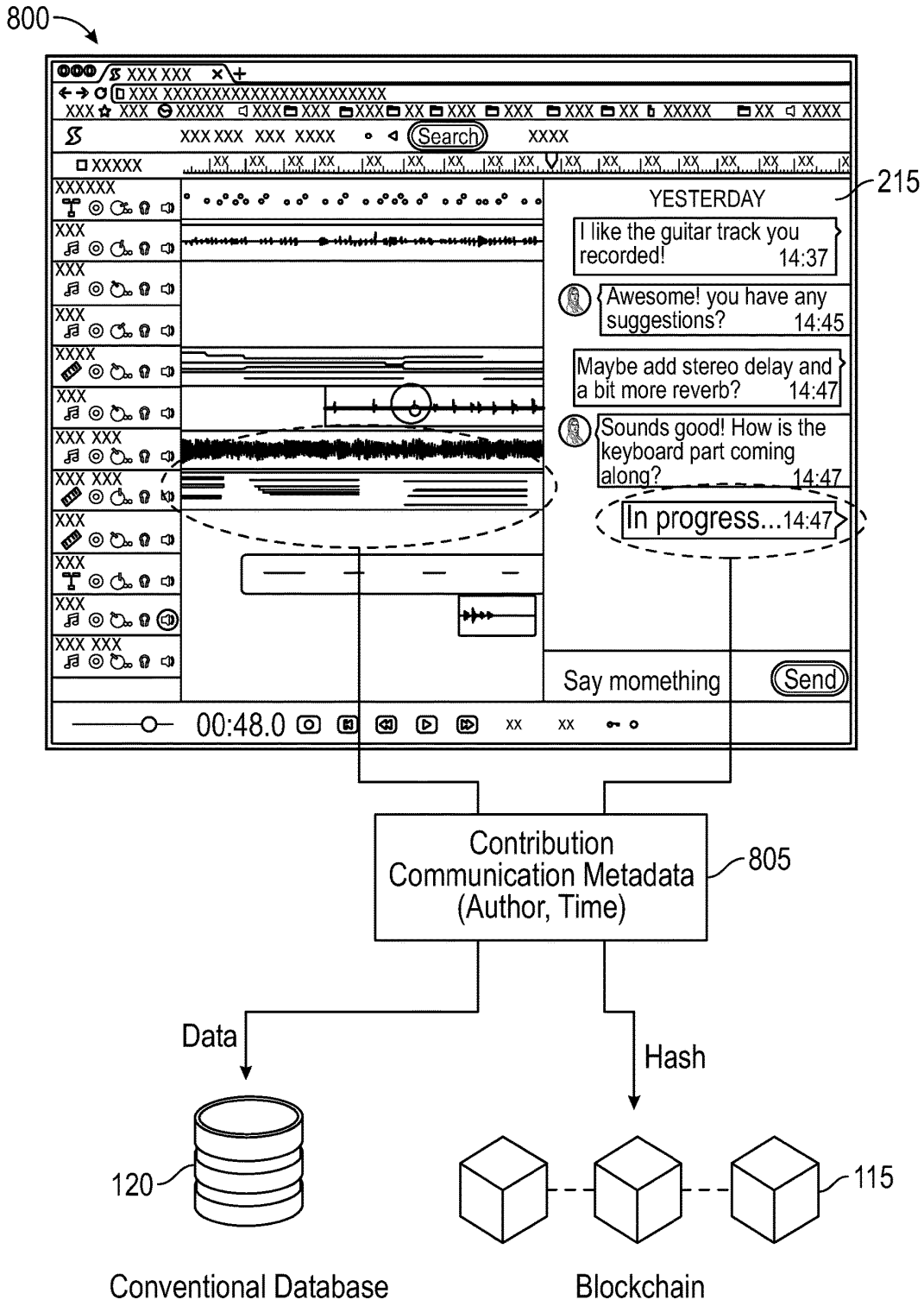


FIG. 8

Patent Application Publication Oct. 8, 2020 Sheet 10 of 13 US 2020/0320458 A1

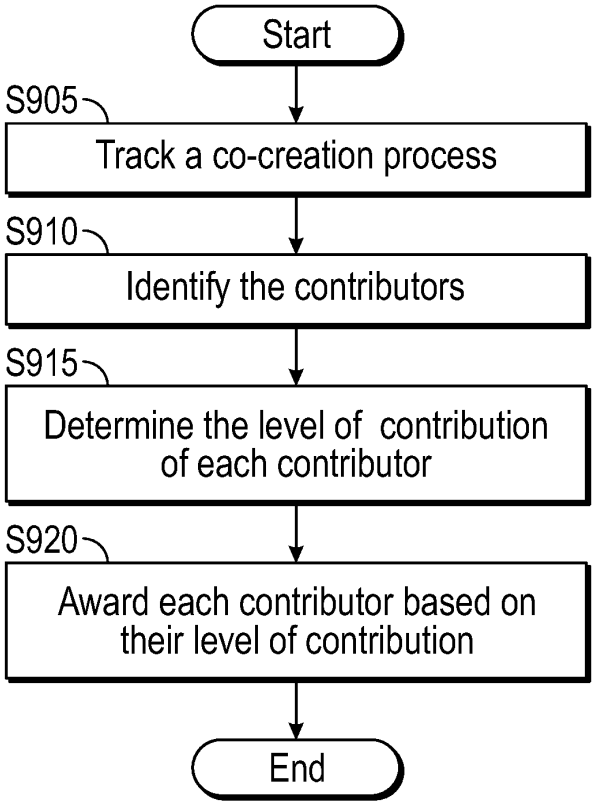


FIG. 9

Patent Application Publication Oct. 8, 2020 Sheet 11 of 13 US 2020/0320458 A1

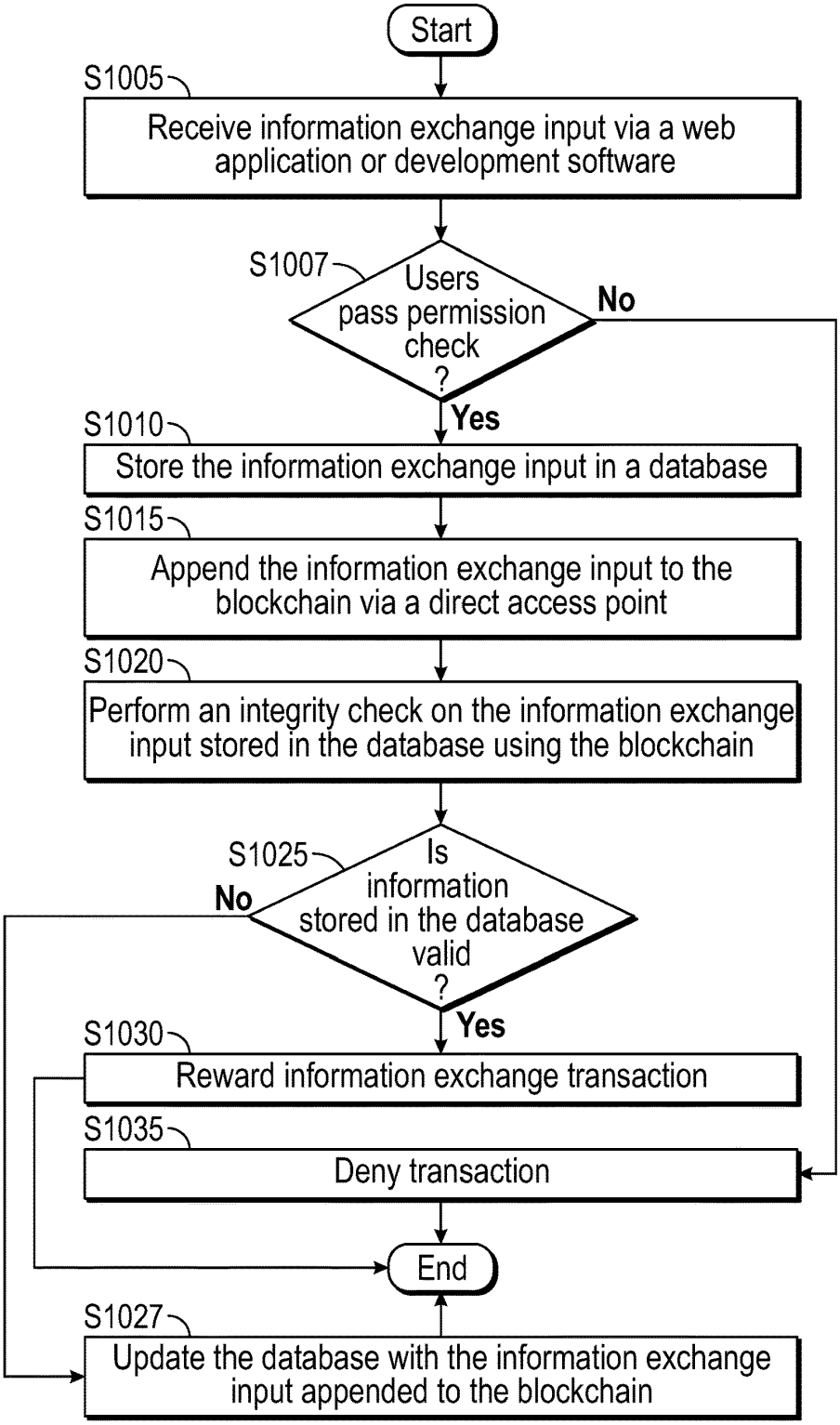


FIG. 10

Patent Application Publication Oct. 8, 2020 Sheet 12 of 13 US 2020/0320458 A1

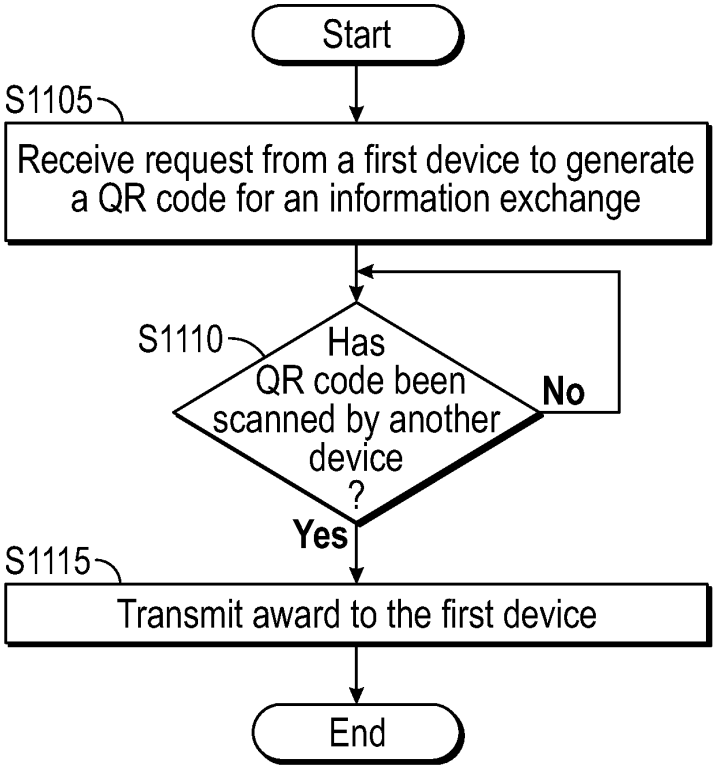


FIG. 11

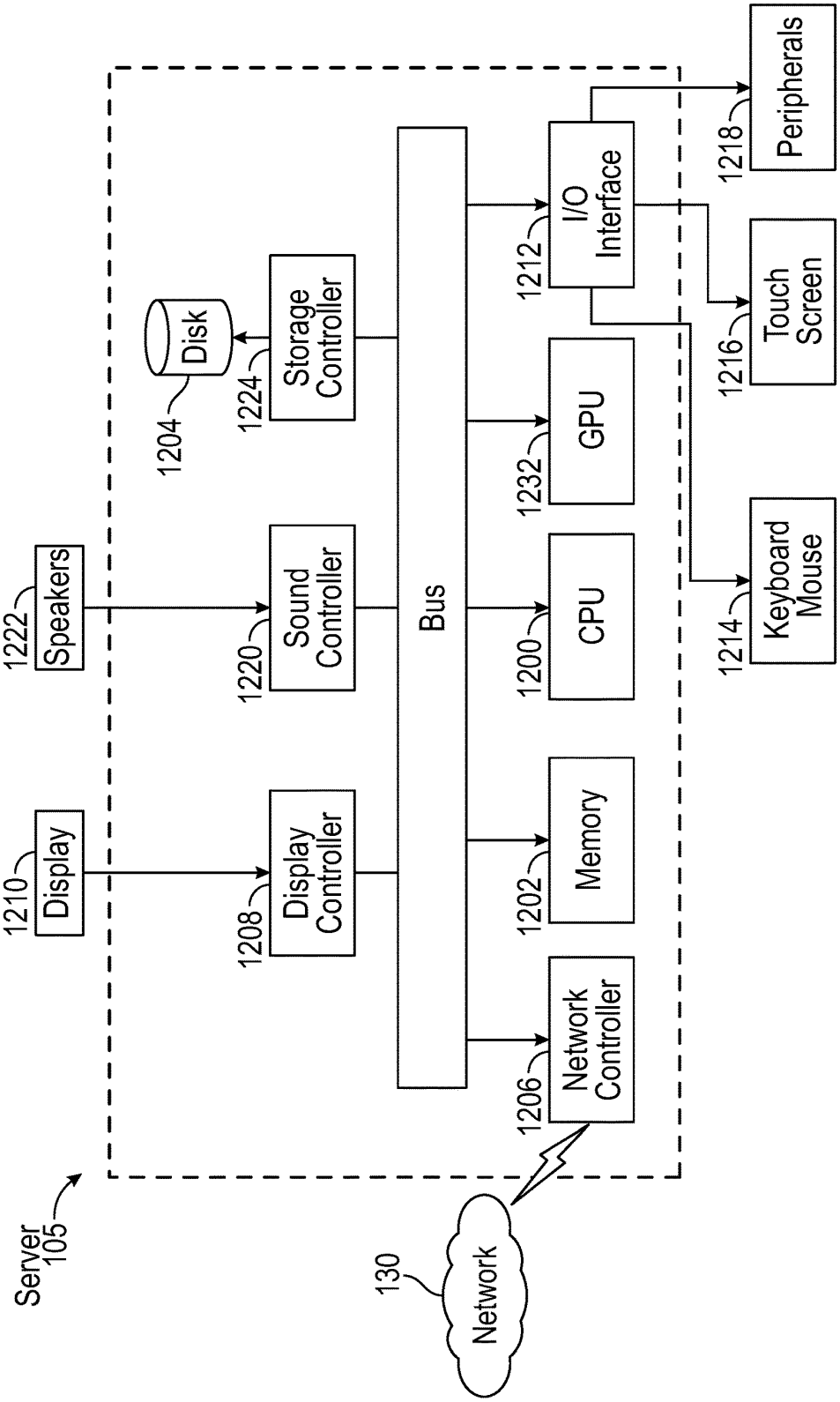

Server
105
Network
130
Bus
Display
1210
Display Controller
1208
Speakers
1222
Sound Controller
1220
Disk
1204
Storage Controller
1224
Network Controller
1206
Memory
1202
CPU
1200
GPU
1232
I/O Interface
1212
Keyboard Mouse
1214
Touch Screen
1216
Peripherals
1218

FIG. 12

US 2020/0320458 A1 Oct. 8, 2020

SYSTEM, METHOD, AND COMPUTER-READABLE MEDIUM FOR TRACKING INFORMATION EXCHANGE

BACKGROUND

[0001] Exchanges of information often go untraced. For example, consider all informal interactions, formal meetings, and brainstorming sessions in which no digital monitoring is in place and the actual information content transferred is very difficult to grasp. Accordingly, the outcome of an industrial, creative, or scientific project is often the result of the combined efforts of its contributors whose effective contributions cannot be easily assessed. To further complicate things, due to the ubiquitous reachability of the world wide web, many actors may contribute remotely to building complex projects including methods to manufacture physical goods that can be sold in the global market, creating intangible goods like songs or computer programs, authoring scientific publications, and the like. While these projects are in progress, contributions can get lost or attributed incorrectly, and even when the contributions get attributed to the correct contributors, distribution of rewards often does not correspond to the contributed value.

[0002] The "background" description provided herein is for the purpose of generally presenting the context of the disclosure. Work of the presently named inventors, to the extent it is described in this background section, as well as aspects of the description which may not otherwise qualify as prior art at the time of filing, are neither expressly or impliedly admitted as prior art against the present invention.

SUMMARY

[0003] According to aspects of the disclosed subject matter, an information exchange tracking system includes a server including processing circuitry configured to track an information exchange transaction, identify contributors in the information exchange transaction, determine a level of contribution of each contributor, the level of contribution being verified via a blockchain, and award each contributor based on their level of contribution. Accordingly, collaborative processes of co-creation are securely recorded using the blockchain and can be used to monitor the collective dynamics of the whole creative process while giving proper credits to all the relevant actors according to their specific contributions.

[0004] The foregoing paragraphs have been provided by way of general introduction and are not intended to limit the scope of the following claims. The described embodiments, together with further advantages, will be best understood by reference to the following detailed description taken in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] A more complete appreciation of the disclosure and many of the attendant advantages thereof will be readily obtained as the same becomes better understood by reference to the following detailed description when considered in connection with the accompanying drawings, wherein:

[0006] FIG. **1** illustrates an exemplary information exchange tracking system according to one or more aspects of the disclosed subject matter;

[0007] FIG. **2** illustrates an information exchange workflow according to one or more aspects of the disclosed subject matter;

[0008] FIG. **3** illustrates a general collaborative process according to one or more aspects of the disclosed subject matter;

[0009] FIG. **4** illustrates an evolutionary line that led to a final product release in the collaborative process according to one or more aspects of the disclosed subject matter;

[0010] FIG. **5**A illustrates an exemplary main page of a user interface of the web application according to one or more aspects of the disclosed subject matter;

[0011] FIG. **5**B illustrates an exemplary interface of the web application for a one to one interaction according to one or more aspects of the disclosed subject matter;

[0012] FIG. **5**C illustrates an exemplary interface of the web application for a one to many interaction according to one or more aspects of the disclosed subject matter;

[0013] FIG. **6** illustrates an exemplary statistics interface of the web application according to one or more aspects of the disclosed subject matter;

[0014] FIG. **7** illustrates an exemplary activity monitoring interface of the web application according to one or more aspects of the disclosed subject matter;

[0015] FIG. **8** illustrates an exemplary co-creation process according to one or more aspects of the disclosed subject matter;

[0016] FIG. **9** is an algorithmic flow chart of a method for accurately tracking a co-creation process according to one or more aspects of the disclosed subject matter;

[0017] FIG. **10** is an algorithmic flow chart of a method for accurately rewarding an information exchange transaction according to one or more aspects of the disclosed subject matter;

[0018] FIG. **11** is an algorithmic flow chart of a method for distributing rewards for an information exchange transaction according to one or more aspects of the disclosed subject matter; and

[0019] FIG. **12** illustrates a hardware block diagram of a server according to one or more exemplary aspects of the disclosed subject matter.

DETAILED DESCRIPTION

[0020] The description set forth below in connection with the appended drawings is intended as a description of various embodiments of the disclosed subject matter and is not necessarily intended to represent the only embodiment(s). In certain instances, the description includes specific details for the purpose of providing an understanding of the disclosed subject matter. However, it will be apparent to those skilled in the art that embodiments may be practiced without these specific details. In some instances, well-known structures and components may be shown in block diagram form in order to avoid obscuring the concepts of the disclosed subject matter.

[0021] Reference throughout the specification to "one embodiment" or "an embodiment" means that a particular feature, structure, characteristic, operation, or function described in connection with an embodiment is included in at least one embodiment of the disclosed subject matter. Thus, any appearance of the phrases "in one embodiment" or "in an embodiment" in the specification is not necessarily referring to the same embodiment. Further, the particular features, structures, characteristics, operations, or functions

may be combined in any suitable manner in one or more embodiments. Further, it is intended that embodiments of the disclosed subject matter can and do cover modifications and variations of the described embodiments.

[0022] It must be noted that, as used in the specification and the appended claims, the singular forms "a," "an," and "the" include plural referents unless the context clearly dictates otherwise. That is, unless clearly specified otherwise, as used herein the words "a" and "an" and the like carry the meaning of "one or more." Additionally, it is to be understood that terms such as "left," "right," "top," "bottom," "front," "rear," "side," "height," "length," "width," "upper," "lower," "interior," "exterior," "inner," "outer," and the like that may be used herein, merely describe points of reference and do not necessarily limit embodiments of the disclosed subject matter to any particular orientation or configuration. Furthermore, terms such as "first," "second," "third," etc., merely identify one of a number of portions, components, points of reference, operations and/or functions as described herein, and likewise do not necessarily limit embodiments of the disclosed subject matter to any particular configuration or orientation.

[0023] Referring now to the drawings, wherein like reference numerals designate identical or corresponding parts throughout the several views.

[0024] FIG. **1** illustrates an exemplary information exchange tracking system **100** (herein referred to as the system **100**) according to one or more aspects of the disclosed subject matter. The system **100** can include a server **105**, a remote device **110**, a blockchain database **115**, a database **120**, and a network **130**.

[0025] The server **105** can represent one or more servers communicably coupled to the remote device **110**, the blockchain database **115**, and the database **120** via the network **130**. The server **105** can be configured to perform various processing for the system **100** as further described herein. Additionally, the server **105** can represent a dedicated bank of servers, cloud-based processing, and/or a serverless computing system corresponding to a virtualized set of hardware resources.

[0026] The remote device **110** can represent one or more remote devices communicably coupled to the server **105**, the blockchain database **115**, and the database **120** via the network **130**. The remote device **110** can be a computer, laptop, smartphone, tablet, PDA, and the like. The remote device **110** can be operated by a user to interact with the system **100** as further described herein. For example, any remote device **110** can be used to connect to the system **100** through a web app and/or while using a software tool that communicates with the system **100** via an API.

[0027] The blockchain database **115** can be communicably coupled to the server **105**, the remote device **110**, and the database **120** via the network **130**. The blockchain database **115** can be configured to store information corresponding to user activities and interactions in an immutable ledger as further described herein. A blockchain is an immutable, cryptographically secured, distributed database. Accordingly, blockchain and blockchain database may be used interchangeably herein. In one example of a blockchain, a blockchain stores information in uniform sized blocks. Each block contains the hashed information from the previous block to provide cryptographic security. The hashing can be SHA256, for example, which is a one-way hash function. This hashed information is the data and digital signature from the previous block, and the hashes of previous blocks that goes all the way back to the very first block produced in the blockchain called a "genesis block". That information is run through a hash function that then points to the address of the next block. Once the block has been added to the blockchain, the information is immutable and transparent to all. Alternatively, or additionally, a blockchain can be operated in a private or permissioned environment, such as within a company. It should be appreciated that the blockchain is described as "immutable" because, despite a scenario where it's possible to alter content of the blockchain in theory, making retrospective changes to the blockchain requires very significant amounts of computing power and is thus in most cases impossible and certainly unfeasible. Accordingly, because it is practically unfeasible, the blockchain is referred to herein as immutable.

[0028] The database **120** can represent one or more database communicably coupled to the server **105**, the remote device **110**, and the blockchain database **115** via the network **130**. The database **120** can be configured to store information for the system **100** as further described herein. For example, the database **120** can store user interactions captured via the remote device **110** for fast data access and analytics. Additionally, the interactions stored via the database **120** can be appended to the blockchain (e.g., via the blockchain database **115**) in parallel, as further described herein.

[0029] The network **130** can be a public network, such as the Internet, or a private network, such as an LAN or WAN network, or any combination thereof and can also include PSTN or ISDN sub-networks. The network **130** can also be wired, such as an Ethernet network, or can be wireless such as a cellular network including EDGE, 3G, 4G, and 5G wireless cellular systems. The wireless network can also be Wi-Fi, Bluetooth, or any other wireless form of communication that is known.

[0030] Generally, the system **100** can assess and quantify all the individual contributions to a shared project in a trusted way.

[0031] More specifically, the system **100** can be a combined system including a web application with a user interface that can be accessed and interacted with via the remote device **110**, for example. The web application can be communicably coupled to a blockchain database (e.g., the blockchain database **115**) which will allow users to store their contributions in a distributed and immutable way. The user interface can allow tracking of both physical (i.e., in person) and virtual (i.e., via the web) interactions of one or more users. The web application can record user activities and interactions. And, the blockchain database **115** can receive the information corresponding to all the interactions and store them in an immutable ledger, guaranteeing immutability and trust.

[0032] The contributions can be both concrete or intangible (e.g., knowledge transfer processes including all the processes where people exchange information, knowledge, know-how or concrete intellectual outcomes). As a result, collaborative processes of co-creation are recorded and can be used to monitor the collective dynamics of the whole creative process while giving proper credits to all the relevant actors according to their specific contributions.

[0033] For example, the system **100** can record the flow of information in a collaborative scientific environment using a web application connected to a blockchain. With the web

application, users (e.g., collaborators in the scientific environment) can interact with each other by creating and scanning QR codes on the smart phones and/or sending URLs. The different types of interactions that can be recorded can include peer to peer (e.g., discussions, emails, etc.), one to many (e.g., lecture, book, paper, etc.), many to many (e.g., group meetings, etc.), projects (e.g., ongoing collaborations), and the like. After a user creates a QR code for an interaction (e.g., lecturer), the recipients of the information (e.g., audience attending the lecture) can scan the QR code to initiate a transaction. Each transaction can be annotated, and the recipient of the information can be awarded virtual coins, for example. The transaction can be stored redundantly in a conventional database (e.g., the database **120** (e.g., a SQL database)) and in a blockchain (e.g., the blockchain database **115**). The blockchain can be distributed and public or used as a permissioned blockchain inside a company or research group. The exchange of virtual currency and the evaluation of each user contribution by the use of a reputation system can be handled by smart contracts on the blockchain, for example.

[0034] In another example, the system **100** can be used for collaborative co-creation in the music industry (e.g., see FIG. **8**).

[0035] The system **100** includes several advantages including recording the activity of individuals in a distributed and immutable way while providing them with a personal record of activities along with a suitable credit scheme for their contributions. In scientific activities, for example, this can be adopted to build accurate resumes and CVs. In the creative industry, for example, the system **100** can be used to incontrovertibly allow the share of dividends of a final product according to the effective individual workload performed. The system **100** can also be deployed within companies where individual activity of employees is bound to responsibility and reward, for example.

[0036] More specifically, the advantages of the system **100** can include assessing the individual efforts of the participants to a project, which can be public, private, scientific, artistic, etc., in order to better quantify rewards; assessing individual responsibilities in collective chain workflow; building individual trusted records of expertise based on real achievements; displaying, visualizing, and statistical processing of personal and global knowledge transfer processes; understanding innovation, knowledge creation, collaboration on both a company and global level in order to boost creative production, and the like.

[0037] In other words, knowledge transactions recorded by the system **100** are decentralized and immutable, which means they are fully transparent and cannot be censored or manipulated as ensured by the blockchain. The system **100** benefits individual users by keeping a trusted and immutable record of their individual contributions to work projects. The record can be used as a portfolio to supplement job applications, for example. Further, it aids companies in understanding communication and collaboration dynamics on a deeper level and can be used to understand and improve the catalyzing conditions under which innovations are conceived.

[0038] FIG. **2** illustrates an information exchange workflow **200** according to one or more aspects of the disclosed subject matter. Personal contributions **205** can correspond to tracked information exchanges as user input. The personal contributions **205** can come from a web application **210** and/or an existing development software **215** with suitable APIs (e.g., API **220**). Once the input is received (e.g., the personal contributions **205** via the web-app **210** and/or development software **215**), the contributions can be sent to the server **105**. The contributions can then be stored in a conventional database (e.g., the database **120**) for fast data access and analytics **235**. In parallel, the contributions are appended to the blockchain **115** via a direct access point (e.g., a digital wallet **225**). The access point is user controlled and removes the threat of censorship. The blockchain **115** is used to guarantee the integrity and validity (e.g., integrity check **230**) of the data stored in the conventional database **120**. If a less centralized solution is required, a distributed file storage system (e.g., Inter Planetary File System (IPFS)) can be used, for example, instead of the conventional database **120**.

[0039] In other words, the workflow **200** illustrates that the system **100** can leverage high levels of security and trust of the blockchain **115**. In one embodiment, the blockchain **115** can be the Ethereum blockchain. Each user can access the blockchain **115** from a wallet, effectively eliminating the threat of censorship or fraud. If specific use cases require a more centralized approach, the Hyperledger Fabric blockchain, for example, can be an effective alternative.

[0040] Additionally, the web application **210** is intended to be exemplary and it should be appreciated that the web application **210** can also represent a mobile application or a desktop application, for example.

[0041] In one embodiment, the system **100** can include AI facilitators that can be in constant dialogue with users (e.g., composers, musicians, writers, scientists, businessmen, etc.) to help them to carve their way in complex landscapes and accompany them through unexplored paths, suggesting new uncommon avenues still close to the sensibility of the creator. Throughout this process, AI systems will constantly construct representations of the space explored adapting them to the emerging space of co-creations.

[0042] Additionally, in one embodiment, the system **100** can support suitably conceived credit schemes for users. These schemes will allow the system **100** to provide users with the fair acknowledgement of their contributions as witnessed by the distributed blockchain technology. This has several advantages and represents a paradigmatic shift from a scheme with a lot of unrecognized "ghosts" contributors to a scheme where each contributor will be rewarded according to his/her actual impact to the final product.

[0043] FIG. **3** illustrates a general collaborative process **300** as tracked by the system **100** according to one or more aspects of the disclosed subject matter. The collaborative process **300** includes several types of events. For example, creation **305** can correspond to a user creating a new element. Commit **310** can correspond to a user making changes to one of his or her contributions. A fork edit **315** can correspond to a user editing an element of another user which creates a new line of evolution while the old element (the element that the fork edit came from) keeps existing in its own evolutionary line. A merge request **320** can correspond to a user requesting his or her changes to be incorporated back into the original element. An accepted merge **325** can correspond to a merge request **320** being accepted. As a result, the elements are merged by incorporating the changes into the original version, which closes the forked evolutionary line. A merge **330** can correspond to a user merging two elements to create a new element and a new

evolutionary line (e.g., music and vocals). And the old elements can still keep evolving along their own evolutionary lines. A release **335** can correspond to an element that is released as a finished product. When a finished product is released, the system **100** can highlight all users who contributed by tracking the specific contributions in the evolutionary lines that led to a final product. A specific example of tracking the evolutionary lines that led to a final product is illustrated in FIG. **4**.

[0044] FIG. **4** illustrates an evolutionary line **405** that led to a final product release **410** in the collaborative process **300** according to one or more aspects of the disclosed subject matter. Additionally, each node of the complex structure is associated with a series of metadata **415** illustrating the type of event responsible for that node, different IDs, the contributing user, the time stamp, the hash of the actual content represented by the node, and the like.

[0045] FIGS. **5**A-**5**C illustrate exemplary views of a user interface of the web application **210** according to one or more aspects of the disclosed subject matter.

[0046] FIG. **5**A illustrates an exemplary main page **505** of a user interface of the web application **210** according to one or more aspects of the disclosed subject matter.

[0047] FIG. **5**B illustrates an exemplary interface of the web application **210** for a one to one interaction according to one or more aspects of the disclosed subject matter.

[0048] FIG. **5**C illustrates an exemplary interface of the web application **210** for a one to many interaction according to one or more aspects of the disclosed subject matter.

[0049] Referring again to FIG. **5**A, the web application **210** can be used to record the transfer of knowledge occurring between people. For example, whenever information flows from a user A to a user B, user B sends a virtual coin named K* (pronounced K-star) to user A as a sign of appreciation. Users can interact by generating and scanning a QR code, for example, on their mobile devices and adding a brief description of the information exchange. All activity can be recorded on the blockchain **115** where it is publicly visible. It should be appreciated that the K* virtual currency is exemplary and should not be limiting

[0050] The web application **210** can capture various types of interactions including one to one interactions, talks, seminars, lectures, group meetings, and the like. The following examples are simply meant to illustrate use of the web application **210** and are not intended to be limiting. For example, in a one to one interaction, a user A can help a user B in solving a problem or just delivers meaningful information to B. Then, user B can send a K* coin to A with the web application **210** and add a comment specifying the type and content of the interaction. To specify the type and content of the interaction, user A can select a Discussion icon **510** on the main page **505** and choose the "get K*" option **530** (see FIG. **5**B), for example. User A will generate a QR code containing the details to start the transaction. User B opens the Discussion section as well and chooses a "give K*" option **535** (see FIG. **5**B) to activate the optical QR-code scanner. After scanning the QR-code, user B may add a comment to annotate the interaction. The details of the transactions are stored in an internal database (e.g., the database **120**) and at the same time are submitted to the blockchain **115** for validation.

[0051] For interactions like talks, seminars, and lectures (e.g., one to many interactions), the lecturer has to generate the QR code for the event by pressing the Lecture icon **515** on the main page **505** and then a Host button **540** (see FIG. **5**C). The lecturer may also add a description of the event and set a QR expiration time. The description might also include useful details like a link to supplementary material, datasets, etc. Users attending the event would press the Lecture icon **515** from their web application and then a Participate button **545** (see FIG. **5**C). After scanning the QR-code, the attendees may add a comment that could be useful for the lecturer (e.g., feedback on the lecture). The lecturer will then receive a K* for each confirmed participant.

[0052] For interactions like a group meeting, the aim of the group meeting functionality is to record situations where information is shared equally by two or more users. For example, one of the participants can generate the QR code for the meeting by pressing the Group icon **520** and then a Create button, for example, on a subsequent screen to confirm the creation of the group meeting. A description of the meeting can be added and a QR expiration time can be set. Users participating in the meeting can also press the Group icon **520** on their web application and then a Participate button, for example, to confirm their participation in the group meeting. All users participating will receive a K*. The user who created the QR code will receive a K* as soon as at least another user joins the meeting by confirming the transaction, for example.

[0053] Users can scan QR codes directly by interacting with the Fast Scan icon **525**, for example. As soon as a QR code is available, users can scan it directly without clicking the icons for Discussion **510**, Lecture **515**, or Group Meetings **520**. In other words, the Fast Scan icon **525** can be a short cut to access the QR-code scanning camera. Alternatively, or additionally, external QR code scanners can be used to scan the QR codes generated by the web application **210** because the QR code embeds an encoded URL that includes all the details of the transaction.

[0054] Additionally, the QR-code used to activate a transaction contains the following information: user id, time stamp, and a hash of transaction generated after appending to these data a secret string known only to the server **105**. As a result, if any of the information was altered in an attempt to inject false information in the system, the hash will not match and the transaction will be denied from the server **105**. For example, in a peer to peer interaction, if one participant attempts to inject false information, the transaction will be denied. In this case, from the point of view of the system **100**, the transaction never took place and no one is rewarded credits. In a one-to-many or group interaction, the malicious transaction will be denied and thus the malicious contribution will be ignored. However, in the one-to-many or group scenarios, all other valid contributions and the corresponding credits are not affected by this event.

[0055] FIG. **6** illustrates an exemplary statistics interface **605** of the web application **210** according to one or more aspects of the disclosed subject matter. The statistics interface **605** can include transaction statistics and corresponding rankings, for example. Additionally, the statistics interface **605** can include a datapoint cloud to illustrate the exchange of information between users.

[0056] FIG. **7** illustrates an exemplary activity monitoring interface **705** of the web application **210** according to one or more aspects of the disclosed subject matter. The activity monitoring interface **705** can be a feed that lists various interactions where the web application **210** was used to track

information exchanges, for example. The interactions listed in the activity monitoring interface **705** can include a timestamp.

[0057] FIG. **8** illustrates an exemplary co-creation process **800** according to one or more aspects of the disclosed subject matter. The co-creation process **800** includes a music creation software tool (e.g., development software **215**). Additionally, the software includes a chat window. The system **100** can record all information corresponding to exchange **805** (e.g., contributions, communications, metadata (author, time), etc.) via the API **220**, for example, and store the information in a conventional database (e.g., the database **120**). In parallel, the hash of the information is added to the blockchain **115**. It should be appreciated that the software tool and co-creation process are intended to be exemplary, and co-creation processes and software tools can vary throughout the modern creative industries including multimedia productions generally.

[0058] In other words, various development software **215** can be used with the system **100** to record the collective creative process by securing the full sequence of actions in a blockchain-based immutable ledger. Additionally, users can be guided in their creative exploits by letting them navigate in real-time the complex structure of contributions, recommend interesting chunks to be merged, noting ideas to be further developed, and identify new paths to be taken. Additionally, the system **100** can accompany creators through suitable AI assistants that will produce and propose suitable segments of music or text adapted to the specific needs. The system **100** can also allow users to act as supervisors by continuously evaluating the current status of the project, validating and rating the relevance of the different contributions, and merging and polishing portions of music or text and higher-order compounds. The system **100** can also give proper credits to all the relevant contributors according to their specific contributions to the final product. As described above, the credits attribution is based both on human contributions and on algorithms evaluating the full network structure that mirrors the co-creation process.

[0059] The system **100** has several advantageous applications. For example, blockchain-based applications like the system **100** can revolutionize the way in which textual content is created as the outcome of collaborative efforts. Traditional publications are frozen in their final version. They are clearly associated to one or more authors who get the credits of their work through the acknowledgement given by an external actor, i.e., a publishing house, a journal, etc. This process is leading to an extreme proliferation of publications, with the obvious drawback that relevant information is often difficult to retrieve, and the credits system can become highly biased and unfair. In addition, the creative potential of the community of writers is not fully leveraged. The possibility to rely on a distributed and immutable ledger like the blockchain **115** in the system **100** allows for liquid publications. A liquid publication is a fluid object that evolves through a continuous series of individual contributions leading to a complex system of variants.

[0060] The possibility of tracking and assessing individual contributions decouples the need of releasing the product from the process of giving credits. Credits, in fact, can be given based on actual contributions (as has been described herein), also independently of the final release of the product. Applications can range from liquid textbooks, liquid news reports, liquid creative compositions (e.g., interactive screenplays), and the like. The system **100** can be used for this application by having an effective platform for co-creation, AI facilitators suggesting content and collaborators, blockchains for the acknowledgement of credits, and suitably conceived credit schemes for users.

[0061] Another advantage of the system **100** is trustworthy credit schemes. For example, a key ingredient to the system **100** is contributors expressing their creativity can be represented by reliable schemes to assign credits in a trustworthy way. Once the whole creative process is tracked (e.g., who did what at what time and under what circumstances) and represented through complex topological structures (e.g., FIG. **3**), a well-defined mathematical problem emerges of how to distribute the credits among all the contributors. For instance, in FIG. **4** the evolutionary line **405** that led to a final product release **410** represents the ensemble of contributions to the final product release and the credit assignation schemes will allow to redistribute the revenues in a fair way. The problem is made non-trivial by the need to assess the relative importance of very different contributions (e.g., an important change compared to an extensive polishing edit). The schemes will rely on a combination of human contributions (also collected and assessed through the blockchain) and state-of-the-art tools inspired to network theory and complexity science. The combination of trustworthy platforms and reliable credit schemes is crucial to revolutionize the creative industry.

[0062] Another advantage of the system **100** is optimizing organizational information flow. For example, many human activities rely on complex collective processes where the workload is subdivided in different tasks committed to different individuals. This process is often not easily parallelized, meaning that the different tasks cannot proceed independently. Very often instead, the efficiency of collective processes depends on a complex interplay among all the tasks. This generates a network of dependencies that implies that certain tasks cannot be completed unless other tasks are, and so on. For these reasons, work organization, i.e., the way that tasks are distributed amongst the individuals in an organization and the ways in which these are then coordinated to achieve the final product or service, it is not an easy task itself. Work organization may be determined by a number of factors including the composition of the working teams, the nature of the technology used to coordinate the tasks, the structure of the information flow, i.e., the way in which information is shared within and across working teams, managerial choices, bottom-up adjustments of tasks and their distribution within the teams, etc.

[0063] The system **100** can help in work organization through a thorough assessment of the information flows within complex collective processes. Knowing who is interacting with whom, for which task, at what time, and for how long, can identify bottlenecks and clogs in the process. In addition, a mapping of the complex structure of tasks and contributions given by individuals can determine the best way to form teams, making them more creative and engaged.

[0064] In other words, the system **100** is a framework that enables the tracking of knowledge transfers in all their forms. With the system **100** those things that are intangible, such as information, are made tangible and measurable. This information flow is recorded in an immutable and decentralized ledger, enabling users to easily access their records at any time to provide proof of and legitimize their contri-

butions. This feature opens the possibility to a panoply of potential applications. The system **100** can, for example, track co-creation processes by leveraging platforms for liquid publications, or assess individual contributions to complex projects and distributing credits in a fair way. Moreover, the system **100** enables the mapping of knowledge flows in working groups like large institutions or corporations with many different working sectors. This new possibility to assess individual contributions in a trustworthy way in the complex network of knowledge transfers is significantly advantageous to allow people to express their creativity and their potential in their individual and collective activities because, using the system **100**, contributors do not need to be concerned about a fair acknowledgement of their work. They can safely share their creations and access the content of others who have also shared theirs, enabling trouble-free collaborations. Reassured by a fair acknowledgement of their credits, possibly helped by personal AI assistants, creators can work on an immersive environment where their only concern will be that of expressing their talent.

[0065] FIG. **9** is an algorithmic flow chart of a method for accurately tracking a co-creation process according to one or more aspects of the disclosed subject matter.

[0066] In S**905**, the server **105** can track an information exchange transaction. The information exchange transaction can be a co-creation process (e.g., the co-creation process **800**), for example. Alternatively, or additionally, the information exchange transaction can be a one on one transaction, a lecture, a group project, and the like.

[0067] In S**910**, the server **105** can identify the contributors via each contributor's interaction with the system **100**. For example, each interaction can include metadata which can include user IDs and other identifiable information regarding the information exchange transaction.

[0068] In S**915**, the server **105** can determine a level of contribution of each contributor (the level of contribution can be confirmed using the blockchain, thus securing the level of contribution from being manipulated maliciously, as further described herein). The information exchange transaction can be processed at any time to determine the level of contribution. The level of contribution can be determined through information tracked throughout the information exchange. In other words, based on information about the transactions, the level of contribution can be calculated. For example, each node in an evolutionary line that led to a final product is associated with a series of metadata illustrating the type of event responsible for that node, different IDs, the contributing user, the time stamp, the hash of the actual content represented by the node, and the like (e.g., see FIG. **4**). Additionally, it should be appreciated that the timing for these events can depend on the use case. In one embodiment, the server **110** can calculate the level of contribution at the time of the transaction and then immediately issue a reward based on the result of the calculation. In one example, however, there can be some time between the transaction and the calculation of the level and contribution. For instance, in the case of co-creating music, contributions are calculated and rewards are distributed once a final product is released. In that case, the content of a transaction is added to the database and blockchain at the time of transaction and used at a later point for distributing rewards.

[0069] In S**920**, each contributor can be awarded based on their level of contribution determined in S**915**. In the system **100**, the award can be K* coins, for example. However, even though K* may have value, the accurate determination of contribution can also be awarded in practical ways like royalty payouts in music industry, for example. In other words, the award can correspond to a payment of money/coins, an acknowledgment of contributions, a share in the resulting object (e.g., royalties for a song), and the like. This is advantageous as it fairly distributes credit in a trust worthy way as further described herein. After each contributor is accurately awarded, the process can end.

[0070] FIG. **10** is an algorithmic flow chart of a method for accurately rewarding an information exchange transaction according to one or more aspects of the disclosed subject matter.

[0071] In S**1005**, the server **105** can receive information exchange input via the web application **210** or the development software **215**. The information exchange input can correspond to contributions of users, for example.

[0072] In S**1007**, the server **105** can check permissions of the contributing users. For example, when two or more users want to add information to the database, a permissions check is performed to make sure that the users have the rights to add the information. In one example, this is accomplished with user accounts. The user is able to log into their account only when they know the password. Additionally, the user is only able to create a QR code for a transaction or scan someone else's QR code from their account. For example, user A and user B can add a transaction between A and B if they are logged into their accounts. However, they cannot create a fake transaction between user C and user D without having access to their accounts. If the contributing users clear the permissions check, a new transaction is added to the database. The permissions check can be done with cookies, for example. Accordingly, the transaction can be either accepted or denied based on user rights, timeframe, and the like. If the contributing users do not pass the permissions check, the transaction can be denied in S**1035**.

[0073] In S**1010**, the server **105** can store the information exchange input in a conventional database (e.g., the database **120**).

[0074] In S**1015**, the information exchange input can be appended to the blockchain (e.g., the blockchain **115**) via a direct access point (e.g., a digital wallet). The information exchange input can be appended to the blockchain **115** in parallel with the information exchange input being stored in the database **120** (S**1010**). For example, even if the database **120** in S**1010** fails (e.g., due to server downtime), the information is still added to the blockchain **115**.

[0075] In S**1020**, the server **105** can perform an integrity check on the information exchange input stored in the database **120** using the immutable distributed ledger of the blockchain **115**. The integrity of the database **120** can be checked at any time by comparing it to the blockchain **115**. If there is a mismatch for the database **120** and the blockchain **115**, the transaction information in the database **120** can be replaced with the information from the blockchain **115** as it is assumed that the blockchain **115** has the correct information due to its immutability. In other words, the server **110** can access the data from the database **120** which can be compared to the data on the blockchain **115**. Ideally, there should be a **1**:**1** match of the data. However, in the case of malicious manipulation to the database **120**, there will be a mismatch. In such an event, the data stored on the blockchain **115** can be trusted as it is considered immutable.

In this case, the manipulated data in the database **120** can be replaced with the data stored on the blockchain **120**. To limit the amount of data stored in the less efficient blockchain **115**, the hashes of the data can be stored instead of the data itself, for example. By comparing the hashes, the same integrity check can be performed. Accordingly, the blockchain **115** is useful to guarantee that the data in the database **120** was not manipulated after it was initially added to the database **120**. Additionally, in one embodiment, the reward (K*) can be sent immediately after appending the transaction information to the database+blockchain. In this case, the user receiving the information can reward the sender of information with a coin (K*). This reward may not be based on the other transactions stored in the database/blockchain. The information in the database can then at a later point be used to calculate a different type of "award" like a reputation rating and transaction statistics, for example. In another embodiment (e.g. a collaborative music platform), the distribution of rewards may not be triggered by the transaction. In that case, transactions can be checked for permission and added to the database/blockchain or are denied. Then, at any point in time, there might be an event such as the release of a song that can trigger the calculation and distribution of the contributions based on the database which is secured by the blockchain.

[0076] In S**1025**, it can be determined if the information stored in the database **120** is valid based on the integrity check performed in S**1020**. If the integrity check is valid, the server **105** can reward the information exchange transaction in S**1030**, and then the process can end. However, if the information stored in the database is not valid (e.g., the hashes do not match because someone tried to maliciously manipulate the data), the database **120** can be updated with the information exchange information appended to the blockchain in S**1027**, and then the process can end.

[0077] In S**1030**, the server **105** can reward (e.g., provide an award to the user) the information exchange transaction when the information stored in the database **120** is valid (although the timing on distributing the award/reward can vary as further described herein). In other words, as long as a user has not tried to manipulate the system **100** to receive an unfair amount of credit for a contribution, the information exchange transaction can be rewarded fairly. Additionally, it should be appreciated that for distributing rewards, both the permissions check (S**1007**) and the integrity check (S**1020**) are important as the credit scheme must be based on valid data. Accordingly, the data must have been valid when it was added (i.e. the users who added it had the permission to do so) and it must not have been manipulated (i.e. the database **120** must match the blockchain **115**.) Further, it should be appreciated that rewarding the information exchanging transaction can include providing an award (e.g., payment of money/coins, an acknowledgement of contributions, a share in the resulting object, and the like as further described herein).

[0078] In S**1035**, the server **105** can deny a transaction when the information stored in the database **120** is not valid based on the integrity check. In other words, if any of the information was altered in an attempt to inject false information in the system **100**, the hash will not match and the transaction will be denied from the server **105**. After either S**1030** or S**1035**, the process can end.

[0079] FIG. **11** is an algorithmic flow chart of a method for distributing rewards for an information exchange transaction according to one or more aspects of the disclosed subject matter.

[0080] In S**1105**, the server **105** can receive a request from a first device (e.g., a first remote device **110**) to generate a QR code for an information exchange. The QR code for the information exchange can include various information regarding the information exchange including the type of exchange and the content of the exchange. The request to generate the QR code can also include a time frame. For example, if the QR code was generated for a lecture or a group project, the QR code may have an expiration time to prevent users from scanning the QR code without having actually attended the lecture or contributed to the group project.

[0081] In S**1110**, it can be determined if the QR code has been scanned by another device (e.g., one or more other remote devices **110**). The QR code can be scanned by one or more devices depending on the type of transaction. For example, in a one on one transaction, the server **105** can be prepared for one scan. Alternatively, in a lecture or group transaction, the QR code may be scanned more than one time. If it is determined that the QR code has not been scanned, the process can continue checking whether or not the QR code has been scanned. In one example, the process can continue checking until the QR times out (e.g., based on the predetermined expiration time). If it is determined that the QR code has been scanned by another device or a plurality of other devices, the server **105** can transmit an award to the first device in S**1115**.

[0082] In S**1115**, the server **105** can transmit an award to the first device that generated the QR code that verifies the first device user's contribution to the information exchange transaction. Additionally, it should be appreciated that the award does not have to only go in one direction. For example, in a group project or co-creation process, the award can be distributed fairly according to the level of contribution of each contributor. After the award is transmitted to the first device, the process can end. Additionally, the QR codes are exemplary and are not intended to be limiting. For example, in one embodiment, a coworking platform may automatically send all contributions associated with a user account to the server **105**/blockchain **115**, and thus may not require QR codes. Alternatively, or additionally, the system can use any unique code that can be identified by another user device, for example.

[0083] Additionally, it should be appreciated that, in one embodiment, information about the transaction can also be added by the person scanning the QR code. For example, user A has a discussion with user B and gives user B valuable information. Now, they want to record their interaction. User A will receive credits for the interaction, therefore he wants to create a QR code. He asks the server for a QR code, specifying the type of interaction (e.g., discussion) and the expiration time. User A gets a QR code including a timestamp, the user name of user A, the type of transaction, and a hash that is generated based on this information plus a secret string that is not known to any user. The server can save the timeframe for this QR code. User B wants to give credits to user A, so user B scans the QR code. User B sends the information from the QR code to the server. The server can check if the present time is within the specified timeframe, or otherwise it will reject the transaction. Further, the

server will hash the information provided by user B using the secret string that is only known to the server. If the result of the hash-function matches the hash provided by user B, the information is validated. If user B changed any of the information, there will be a mismatch of hashes and the transaction will be rejected. If the transaction is validated, user B will be offered to add further information to the transaction in the form of a comment. The comment can ideally include a description of the interaction. User B sends the comment to the server, and this concludes the process.

[0084] Additionally, in a one-to-many interaction (e.g., a lecture), the process may not end after the QR code was scanned by one device. Instead, it can end once the expiration time of the QR code is passed because the QR code might get scanned by other devices within the predetermined time frame. For a group interaction, every participate is rewarded once, and the process also ends after the expiration time is reached.

[0085] Further, it should be appreciated that the reward may not be triggered by the transaction. Alternatively, or additionally, the reward may be triggered by a later event and it can get triggered more than once (e.g. if a contribution to a song is used in a second song, money can be distributed upon release of each song).

[0086] In the above description of FIGS. **2**, **9**, **10**, and **11**, any processes, descriptions or blocks in flowcharts can be understood as representing modules, segments or portions of code which include one or more executable instructions for implementing specific logical functions or steps in the process, and alternate implementations are included within the scope of the exemplary embodiments of the present advancements in which functions can be executed out of order from that shown or discussed, including substantially concurrently or in reverse order, depending upon the functionality involved, as would be understood by those skilled in the art. The various elements, features, and processes described herein may be used independently of one another, or may be combined in various ways. All possible combinations and sub-combinations are intended to fall within the scope of this disclosure.

[0087] Next, a hardware description of a server (such as the server **105**) according to exemplary embodiments is described with reference to FIG. **12**. The hardware description described herein can also be a hardware description of the processing circuitry. In FIG. **12**, the server **105** includes a CPU **1200** which performs one or more of the processes described herein. Additionally, either individually or in combination with the CPU **1200**, a GPU **1232** can perform one or more of the processes described herein. The process data and instructions may be stored in memory **1202**. These processes and instructions may also be stored on a storage medium disk **1204** such as a hard drive (HDD) or portable storage medium or may be stored remotely. Further, the claimed advancements are not limited by the form of the computer-readable media on which the instructions of the inventive process are stored. For example, the instructions may be stored on CDs, DVDs, in FLASH memory, RAM, ROM, PROM, EPROM, EEPROM, hard disk or any other information processing device with which the server **105** communicates, such as a server or computer.

[0088] Further, the claimed advancements may be provided as a utility application, background daemon, or component of an operating system, or combination thereof, executing in conjunction with CPU **1200** and an operating system such as Microsoft Windows, UNIX, Solaris, LINUX, Apple MAC-OS and other systems known to those skilled in the art.

[0089] The hardware elements in order to achieve the server **105** may be realized by various circuitry elements. Further, each of the functions of the above described embodiments may be implemented by circuitry, which includes one or more processing circuits. A processing circuit includes a particularly programmed processor, for example, processor (CPU) **1200**, as shown in FIG. **12**. A processing circuit also includes devices such as an application specific integrated circuit (ASIC) and conventional circuit components arranged to perform the recited functions.

[0090] In FIG. **12**, the server **105** includes a CPU **1200** which performs the processes described above. Additionally, the GPU **1232** can perform or assist in performing the processes described herein. The GPU **1232** can be specialized electronic circuit designed to rapidly manipulate and alter memory. The highly parallel structure of a GPU makes them efficient for algorithms that process large blocks of data in parallel. The GPU can be present on a video card, embedded on a motherboard, or embedded in a CPU, for example. The server **105** may be a general-purpose computer or a particular, special-purpose machine. In one embodiment, the server **105** becomes a particular, special-purpose machine when the processor **1200** and/or the GPU **1232** is programmed to track and reward information exchange transactions (and in particular, any of the processes discussed with reference to FIGS. **2**, **9**, **10**, and **11**).

[0091] Alternatively, or additionally, the CPU **1200** may be implemented on an FPGA, ASIC, PLD or using discrete logic circuits, as one of ordinary skill in the art would recognize. Further, CPU **1200** may be implemented as multiple processors cooperatively working in parallel to perform the instructions of the inventive processes described above.

[0092] The server **105** in FIG. **12** also includes a network controller **1206**, such as an Intel Ethernet PRO network interface card from Intel Corporation of America, for interfacing with network **130**. As can be appreciated, the network **130** can be a public network, such as the Internet, or a private network such as an LAN or WAN network, or any combination thereof and can also include PSTN or ISDN sub-networks. The network **130** can also be wired, such as an Ethernet network, or can be wireless such as a cellular network including EDGE, 3G and 4G wireless cellular systems. The wireless network can also be WiFi, Bluetooth, or any other wireless form of communication that is known.

[0093] The server **105** further includes a display controller **1208**, such as a graphics card or graphics adaptor for interfacing with display **1210**, such as a monitor. A general purpose I/O interface **1212** interfaces with a keyboard and/or mouse **1214** as well as a touch screen panel **1216** on or separate from display **1210**. General purpose I/O interface also connects to a variety of peripherals **1218** including printers and scanners.

[0094] A sound controller **1220** is also provided in the server **105** to interface with speakers/microphone **1222** thereby providing sounds and/or music.

[0095] The general-purpose storage controller **1224** connects the storage medium disk **1204** with communication bus **1226**, which may be an ISA, EISA, VESA, PCI, or similar, for interconnecting all of the components of the

server **105**. A description of the general features and functionality of the display **1210**, keyboard and/or mouse **1214**, as well as the display controller **1208**, storage controller **1224**, network controller **1206**, sound controller **1220**, and general purpose I/O interface **1212** is omitted herein for brevity as these features are known.

[0096] The exemplary circuit elements described in the context of the present disclosure may be replaced with other elements and structured differently than the examples provided herein. Moreover, circuitry configured to perform features described herein may be implemented in multiple circuit units (e.g., chips), or the features may be combined in circuitry on a single chipset.

[0097] The functions and features described herein may also be executed by various distributed components of a system. For example, one or more processors may execute these system functions, wherein the processors are distributed across multiple components communicating in a network. The distributed components may include one or more client and server machines, which may share processing, in addition to various human interface and communication devices (e.g., display monitors, smart phones, tablets, personal digital assistants (PDAs)). The network may be a private network, such as a LAN or WAN, or may be a public network, such as the Internet. Input to the system may be received via direct user input and received remotely either in real-time or as a batch process. Additionally, some implementations may be performed on modules or hardware not identical to those described. Accordingly, other implementations are within the scope that may be claimed.

[0098] Further, in one embodiment, the server **105** can be implemented in a cloud-based computing system and/or represent the virtualized hardware in a serverless environment programmed to track and reward information exchange transactions (and in particular, any of the processes of the system **100** discussed with reference to FIGS. **2**, **9**, **10**, and **11**). More specifically, the processing may be performed using a serverless architecture for scalability. For example, cloud computing aggregates physical and virtual computation, storage, and network resources in the "cloud." Serverless computing offers a high level of computational abstraction, with increased scalability. Serverless computing works by uploading code to a serverless computing environment (or serverless computing platform or serverless environment), and the serverless computing environment executes the code. Additionally, the serverless computing environment can be event driven, meaning that the code can execute or perform computations triggered by events. In some cases, the code can be executed on demand, or according to a predetermined schedule. In a serverless computing environment, the user is abstracted from the setup and execution of the code in the networked hardware resources in the cloud. In some cases, the serverless computing environments are virtualized computing environments having an application programming interface (API) (e.g., the API **220**) which abstracts the user from the implementation and configuration of a service or application in the cloud.

[0099] Having now described embodiments of the disclosed subject matter, it should be apparent to those skilled in the art that the foregoing is merely illustrative and not limiting, having been presented by way of example only. Thus, although particular configurations have been discussed herein, other configurations can also be employed. Numerous modifications and other embodiments (e.g., combinations, rearrangements, etc.) are enabled by the present disclosure and are within the scope of one of ordinary skill in the art and are contemplated as falling within the scope of the disclosed subject matter and any equivalents thereto. Features of the disclosed embodiments can be combined, rearranged, omitted, etc., within the scope of the invention to produce additional embodiments. Furthermore, certain features may sometimes be used to advantage without a corresponding use of other features. Accordingly, Applicant (s) intend(s) to embrace all such alternatives, modifications, equivalents, and variations that are within the spirit and scope of the disclosed subject matter.

1. A server, comprising:
processing circuitry configured to
track an information exchange transaction,
identify contributors in the information exchange transaction,
determine a level of contribution of each contributor, the level of contribution being secured by a blockchain, and
award each contributor based on their level of contribution.

2. The server of claim **1**, wherein the processing circuitry for determining the level of contribution of each contributor is further configured to
receive information exchange input via an application or development software,
check user permissions,
store the information exchange input in a database in response to a determination that the information exchange input was received with permission,
append the information exchange input to the blockchain via a direct access point,
perform an integrity check on the information exchange input stored in the database using the blockchain,
determine if the information stored in the database is valid, and
reward the information exchange transaction in response to a determination that the information stored in the database is valid.

3. The server of claim **2**, where in the processing circuitry is further configured to
deny the information exchange transaction in response to a determination that the contributors did not pass the permissions check, and
update the database with the information exchange input stored in the blockchain in response to a determination that the information stored in the database is invalid.

4. The server of claim **1**, wherein the information exchange input is stored in the database and appended to the blockchain in parallel.

5. The server of claim **2**, wherein the direct access point is a digital wallet.

6. The server of claim **1**, wherein the processing circuitry for awarding each contributor is further configured to
receive a request from a first device to generate a unique code for the information exchange transaction,
determine whether the unique code has been identified by another device, and
transmit an award to the first device in response to a determination that the unique code was identified by another device or a plurality of other devices.

7. The server of claim **6**, wherein the unique code for the information exchange transaction includes metadata includ-

ing a type of information exchange transaction and at least a portion of content of the information exchange transaction.

8. The server of claim **6**, wherein the type of information exchange transaction includes a one to one transaction, a one to many transaction, a many to many transaction, and ongoing collaboration.

9. A method of accurately rewarding an information exchange transaction, comprising:

tracking an information exchange transaction;

identify contributors in the information exchange transaction;

determining a level of contribution of each contributor, the level of contribution being secured by a blockchain; and

awarding each contributor based on their level of contribution.

10. The method of claim **9**, wherein determining the level of contribution of each contributor further comprises:

receiving information exchange input via an application or development software;

checking user permissions;

storing the information exchange input in a database in response to a determination that the information exchange input was received with permission;

appending the information exchange input to the blockchain via a direct access point;

performing an integrity check on the information exchange input stored in the database using the blockchain;

determining if the information stored in the database is valid; and

rewarding the information exchange transaction in response to a determination that the information stored in the database is valid.

11. The method of claim **10**, further comprising:

denying the information exchange transaction in response to a determination that the contributors did not pass the permissions check; and

updating the database with the information exchange input stored in the blockchain in response to a determination that the information stored in the database is invalid.

12. The method of claim **9**, wherein the information exchange input is stored in the database and appended to the blockchain in parallel.

13. The method of claim **10**, wherein the direct access point is a digital wallet.

14. The method of claim **9**, wherein awarding each contributor further comprises:

receiving a request from a first device to generate a unique code for the information exchange transaction;

determining whether the unique code has been identified by another device; and

transmitting an award to the first device in response to a determination that the unique code was identified by another device or a plurality of other devices.

15. The method of claim **14**, wherein the unique code for the information exchange transaction includes metadata including a type of information exchange transaction and at least a portion of content of the information exchange transaction.

16. The method of claim **14**, wherein the type of information exchange transaction includes a one to one transaction, a one to many transaction, a many to many transaction, and ongoing collaboration.

17. A non-transitory computer-readable storage medium storing computer-readable instructions thereon which, when executed by a computer, cause the computer to perform a method, the method comprising:

tracking an information exchange transaction;

identify contributors in the information exchange transaction;

determining a level of contribution of each contributor, the level of contribution being secured by a blockchain; and

awarding each contributor based on their level of contribution.

18. The non-transitory computer-readable storage medium of claim **17**, wherein determining the level of contribution of each contributor further comprises:

receiving information exchange input via an application or development software;

checking user permissions;

storing the information exchange input in a database in response to a determination that the information exchange input was received with permission;

appending the information exchange input to the blockchain via a direct access point;

performing an integrity check on the information exchange input stored in the database using the blockchain;

determining if the information stored in the database is valid; and

rewarding the information exchange transaction in response to a determination that the information stored in the database is valid.

19. The non-transitory computer-readable storage medium of claim **18**, further comprising:

denying the information exchange transaction in response to a determination that contributors did not pass the permissions check; and

updating the database with the information exchange input stored in the blockchain in response to a determination that the information stored in the database is invalid.

20. The non-transitory computer-readable storage medium of claim **17**, wherein awarding each contributor further comprises:

receiving a request from a first device to generate a unique code for the information exchange transaction;

determining whether the unique code has been identified by another device; and

transmitting an award to the first device in response to a determination that the unique code was identified by another device or a plurality of other devices.

* * * * *

Part III

Conclusions

Chapter 7

Conclusions and Outlook

In this thesis, we have analyzed information diffusion processes in complex networks in three publications, one book chapter, and one patent. The thesis can be split into four topics, information diffusion in scientific citation networks, geographic scientific capacities, information diffusion in social networks, and information diffusion in organizations on short time scales.

First, we have studied information diffusion in scientific citation networks. In our first paper [1], we used the complete temporal citation network of all *American Physical Society* journals to study how knowledge in physics is created and forgotten over time. We built a microscopic generative model of citation dynamics, where the probability for a paper to be cited is based on its past popularity and age. In this model, most papers are forgotten quickly, while a small number of scientific milestones manage to stay popular for decades. We found that these exceptional papers can be identified early by a scaling exponent α_c. If that exponent is above a critical value, the paper will likely be of lasting relevance. Our model also allowed us to forecast the citation landscape of the future, estimating that around 95% of papers cited in 2050 have yet to be written. With this model, we can show how knowledge diffusion behaves over extended periods of multiple years.

In our book chapter [2], we extended the analysis of citation networks by highlighting a particular class of rare and exceptional papers. Out of the bulk of papers that is quickly forgotten, a small fraction is rediscovered in what is called *exaptation*, an event where previously found insights in one field are rediscovered and repurposed in a different field. We discussed exaptation in the context of science, presented an entropy-based model for identifying potential exaptation candidates at an early stage, and discussed some examples of exaptation in physics.

In our second paper [3], we studied the movement of researchers between institutions, which is an essential driver of global knowledge diffusion. There, we found that while the scientific capacities in specific fields vary significantly between regions, the process of scientific knowledge accumulation is remarkably similar between regions and is best described by preferential attachment. Due to preferential attachment, early birds tend to dominate and establish themselves as leaders in their fields. Despite that, we found that given enough resources, any region might become attractive enough to outgrow its competitors. Contrary to the often assumed *critical mass* of researchers needed to compete in a field, we found no evidence for such a threshold.

In our third paper [4], we translated our findings about scientific citation networks to the domain of music, where we had the opportunity to study a unique dataset that we collected, and that lets us analyze information diffusion in social networks. There we found that on short timescales, information about new songs is exchanged along the links of friendship networks in a diffusion process. These microscopic interactions lead to a macroscopic increase of similarity between connected users, called homophily. We showed how homophily shapes the network topology to enable the cascade-like spreading of knowledge between friends and thus fuels the preferential attachment process. By leveraging homophily-based indicators for greatly enhancing state-of-the-art machine learning models' predictions of song popularity, we showed how our findings can be translated into tools with potential commercial

applications.
In the last part, we presented our patent [5] that describes a new system for recording multilayer social interaction networks. The system records information on different channels such as personal discussions, talks, group meetings, emails, chat programs, and collaborative online writing tools. The system was presented at public outreach events at the *Palazzo delle Esposizioni* in Rome and the *Maker Faire 2022* in Rome. The dynamic multilayer network made available by this tool adds the missing piece of information required to study knowledge flows in day-to-day interactions in organizations. While out of the scope of this thesis, the obvious next step would be to apply the system to collect and study large-scale data on organizational knowledge diffusion.
In this work, we studied the many facets of information flow processes on different length and time scales. We showed how principles of diffusion in networks and simulations of microscopic interacting particles that have been studied in physics for the last decades can be applied to various domains of social dynamics, touching the fields of science of science, hit song science, and organizational information flow. We showed that the mechanisms found are similar across multiple domains, hinting at some degree of universality in how knowledge flows, which might be independent of the domain. Thus the purpose of this study was not to exhaustively study information flow on all possible domains but rather to identify robust principles and methods that are flexible and scaleable that apply to knowledge flows across multiple domains. This work can be seen as a physics-inspired, interdisciplinary exploration of information flow processes. Our findings increase our theoretical understanding of attachment processes across different domains through their contribution of microscopic generative models on the one hand and insights into flow processes on networks on the other hand. In addition to that, they contribute accurate predictions that allow for better policy-making and decision-making in science, the music industry, and organizational communication.
Due to the rising importance of information in our society during its continued transformation into a knowledge-based society, information flow processes will likely continue to play a crucial role in our lives. With this thesis, we want to contribute to understanding the fundamental mechanisms that determine the dynamics of information flow and hope to inspire and enable further research on the topic.

Bibliography

[1] N Reisz, et al., Loss of sustainability in scientific work. *New J. Phys.* **24**, 053041 (2022).

[2] MR Ferreira, et al., *Quantifying Exaptation in Scientific Evolution.* (Springer International Publishing, Cham), pp. 55–68 (2020).

[3] VD Servedio, MR Ferreira, N Reisz, R Costas, S Thurner, Scale-free growth in regional scientific capacity building explains long-term scientific dominance. *Chaos, Solitons & Fractals* **167**, 113020 (2023).

[4] N Reisz, VDP Servedio, S Thurner, To what extent homophily and influencer networks explain song popularity (2022).

[5] V Loreto, VD Servedio, N Reisz, S Thurner, System, method, and computer-readable medium for tracking information exchange (2020) US Patent App. 16/375,017.

[6] PA David, D Foray, Economic fundamentals of the knowledge society. *Policy futures education* **1**, 20–49 (2003).

[7] EH Lieb, Density functionals for coulomb systems in *Inequalities.* (Springer), pp. 269–303 (2002).

[8] S Thurner, R Hanel, P Klimek, *Introduction to the theory of complex systems.* (Oxford University Press, London, England), (2018).

[9] M Newman, *Networks.* (Oxford University Press, London, England), 2 edition, (2018).

[10] A Fick, Ueber diffusion. *Annalen der Physik* **170**, 59–86 (1855).

[11] JD Noh, H Rieger, Random walks on complex networks. *Phys. Rev. Lett.* **92** (2004).

[12] FRK Chung, *Spectral Graph Theory*, CBMS regional conference series in mathematics. (American Mathematical Society, Providence, Rhode Island), pp. 43–58 (1996).

[13] P Bonacich, Power and centrality: A family of measures. *Am. J. Sociol.* **92**, 1170–1182 (1987).

[14] L Katz, A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).

[15] L Page, S Brin, R Motwani, T Winograd, The pagerank citation ranking: Bringing order to the web, (Stanford InfoLab), Technical Report 1999-66 (1999).

[16] SP Borgatti, Centrality and network flow. *Soc. networks* **27**, 55–71 (2005).

[17] MO Jackson, *Social and economic networks.* (Princeton university press), (2010).

[18] R Pastor-Satorras, A Vespignani, Epidemic spreading in scale-free networks. *Phys. review letters* **86**, 3200 (2001).

[19] RM May, AL Lloyd, Infection dynamics on scale-free networks. *Phys. Rev. E* **64**, 066112 (2001).

[20] Y Moreno, R Pastor-Satorras, A Vespignani, Epidemic outbreaks in complex heterogeneous networks. *The Eur. Phys. J. B-Condensed Matter Complex Syst.* **26**, 521–529 (2002).

[21] ME Newman, Spread of epidemic disease on networks. *Phys. review E* **66**, 016128 (2002).

[22] Y Wang, D Chakrabarti, C Wang, C Faloutsos, Epidemic spreading in real networks: An eigenvalue viewpoint in *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.* (IEEE), pp. 25–34 (2003).

[23] SB Chang, KK Lai, SM Chang, Exploring technology diffusion and classification of business methods: Using the patent citation network. *Technol. Forecast. Soc. Chang.* **76**, 107–117 (2009).

[24] T Pham, P Sheridan, H Shimodaira, PAFit: A statistical method for measuring preferential attachment in temporal complex networks. *PLOS ONE* **10**, e0137796 (2015).

[25] ME Newman, Clustering and preferential attachment in growing networks. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.* **64**, 4 (2001).

[26] M Perc, The matthew effect in empirical data. *J. The Royal Soc. Interface* **11**, 20140378 (2014).

[27] AL Barabási, R Albert, Emergence of scaling in random networks. *science* **286**, 509–512 (1999).

[28] S Bornholdt, H Ebel, World wide web scaling exponent from simon's 1955 model. *Phys. Rev. E* **64**, 035104 (2001).

[29] HA Simon, On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).

[30] R Albert, H Jeong, AL Barabási, Diameter of the world-wide web. *nature* **401**, 130–131 (1999).

[31] S Thurner, P Klimek, R Hanel, Schumpeterian economic dynamics as a quantifiable model of evolution. *New J. Phys.* **12**, 075029 (2010).

[32] P Klimek, R Hausmann, S Thurner, Empirical confirmation of creative destruction from world trade data. *PLOS ONE* **7**, e38924 (2012).

[33] ES Andersen, *Evolutionary economics: post-Schumpeterian contributions.* (Routledge), (2013).

[34] PP Saviotti, et al., Technological evolution, variety and the economy. *Books* (1996).

[35] A Sood, GJ Tellis, Technological evolution and radical innovation. *J. marketing* **69**, 152–168 (2005).

[36] R Sinatra, D Wang, P Deville, C Song, AL Barabási, Quantifying the evolution of individual scientific impact. *Science* **354**, aaf5239 (2016).

[37] AL Barabâsi, et al., Evolution of the social network of scientific collaborations. *Phys. A: Stat. mechanics its applications* **311**, 590–614 (2002).

[38] K Börner, JT Maru, RL Goldstone, The simultaneous evolution of author and paper networks. *Proc. Natl. Acad. Sci.* **101**, 5266–5273 (2004).

[39] Y Sun, V Latora, The evolution of knowledge within and across fields in modern physics. *Sci. reports* **10**, 1–9 (2020).

[40] U Alvarez-Rodriguez, et al., Evolutionary dynamics of higher-order interactions in social networks. *Nat. Hum. Behav.* **5**, 586–595 (2021).

[41] KM Page, MA Nowak, Unifying evolutionary dynamics. *J. theoretical biology* **219**, 93–98 (2002).

[42] J Hofbauer, K Sigmund, Evolutionary game dynamics. *Bull. Am. mathematical society* **40**, 479–519 (2003).

[43] S Kauffman, S Levin, Towards a general theory of adaptive walks on rugged landscapes. *J. theoretical Biol.* **128**, 11–45 (1987).

[44] SA Kauffman, ED Weinberger, The nk model of rugged fitness landscapes and its application to maturation of the immune response. *J. theoretical biology* **141**, 211–245 (1989).

[45] SA Kauffman, S Johnsen, Coevolution to the edge of chaos: Coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.* **149**, 467–505 (1991).

[46] P Bak, K Sneppen, Punctuated equilibrium and criticality in a simple model of evolution. *Phys. review letters* **71**, 4083 (1993).

[47] S Jain, S Krishna, A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proc. Natl. Acad. Sci.* **98**, 543–547 (2001).

[48] P Klimek, S Thurner, R Hanel, Evolutionary dynamics from a variational principle. *Phys. Rev. E* **82**, 011901 (2010).

[49] SA Kauffman, *Investigations*. (Oxford University Press), (2000).

[50] F Tria, V Loreto, VDP Servedio, SH Strogatz, The dynamics of correlated novelties. *Sci. reports* **4**, 1–8 (2014).

[51] V Loreto, VD Servedio, SH Strogatz, F Tria, Dynamics on expanding spaces: modeling the emergence of novelties. *Creat. universality language* pp. 59–83 (2016).

[52] F Tria, V Loreto, VD Servedio, Zipf's, heaps' and taylor's laws are determined by the expansion into the adjacent possible. *Entropy* **20**, 752 (2018).

[53] SJ Gould, ES Vrba, Exaptation—a missing term in the science of form. *Paleobiology* **8**, 4–15 (1982).

[54] JC Mitchell, Social networks. *Annu. review anthropology* **3**, 279–299 (1974).

[55] GC Robertson, Hobbes, thomas. *Encycl. Britannica* **13**, 545–552 (1911).

[56] GG Iggers, Further remarks about early uses of the term "social science". *J. Hist. Ideas* **20**, 433–436 (1959).

[57] L Boltzmann, Weitere studien über das wärmegleichgewicht unter gasmolekülen in *Kinetische Theorie II.* (Springer), pp. 115–225 (1970).

[58] P Ball, The physical modelling of society: a historical perspective. *Phys. A: Stat. Mech. its Appl.* **314**, 1–14 (2002) Horizons in Complex Systems.

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

[60] M Jusup, et al., Social physics. *Phys. Reports* **948**, 1–148 (2022).

[61] S Boccaletti, V Latora, Y Moreno, M Chavez, DU Hwang, Complex networks: Structure and dynamics. *Phys. Reports* **424**, 175–308 (2006).

[62] V Latora, M Marchiori, Is the boston subway a small-world network? *Phys. A: Stat. Mech. its Appl.* **314**, 109–113 (2002) Horizons in Complex Systems.

[63] A Cardillo, S Scellato, V Latora, S Porta, Structural properties of planar graphs of urban street patterns. *Phys. Rev. E* **73**, 066107 (2006).

[64] S Porta, P Crucitti, V Latora, The network analysis of urban streets: A dual approach. *Phys. A: Stat. Mech. its Appl.* **369**, 853–866 (2006).

[65] I Iacopini, S Milojević, V Latora, Network dynamics of innovation processes. *Phys. Rev. Lett.* **120** (2018).

[66] J Tang, M Musolesi, C Mascolo, V Latora, V Nicosia, Analysing information flows and key mediators through temporal centrality metrics in *Proceedings of the 3rd Workshop on Social Network Systems*, SNS '10. (Association for Computing Machinery, New York, NY, USA), (2010).

[67] R Sinatra, P Deville, M Szell, D Wang, AL Barabási, A century of physics. *Nat. Phys.* **11**, 791–796 (2015).

[68] J Menche, et al., Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).

[69] I Iacopini, G Petri, A Barrat, V Latora, Simplicial models of social contagion. *Nat. Commun.* **10** (2019).

[70] E Guney, J Menche, M Vidal, AL Barábasi, Network-based in silico drug efficacy screening. *Nat. communications* **7**, 1–13 (2016).

[71] S Thurner, P Klimek, R Hanel, A network-based explanation of why most covid-19 infection curves are linear. *Proc. Natl. Acad. Sci.* **117**, 22684–22689 (2020).

[72] N Haug, et al., Ranking the effectiveness of worldwide covid-19 government interventions. *Nat. human behaviour* **4**, 1303–1312 (2020).

[73] S GALAM, Sociophysics: A review of galam models. *Int. J. Mod. Phys. C* **19**, 409–440 (2008).

[74] S Galam, YG (Feigenblat), Y Shapir, Sociophysics: A new approach of sociological collective behaviour. i. mean-behaviour description of a strike. *The J. Math. Sociol.* **9**, 1–13 (1982).

[75] S Galam, S Moscovici, Towards a theory of collective phenomena: Consensus and attitude changes in groups. *Eur. J. Soc. Psychol.* **21**, 49–74 (1991).

[76] K Klemm, VM Eguíluz, R Toral, M San Miguel, Nonequilibrium transitions in complex networks: A model of social interaction. *Phys. Rev. E* **67**, 026120 (2003).

[77] S Wasserman, K Faust, *Social Network Analysis: Methods and Applications*, Structural Analysis in the Social Sciences. (Cambridge University Press), (1994).

[78] M Szell, S Thurner, Measuring social dynamics in a massive multiplayer online game. *Soc. Networks* **32**, 313–329 (2010).

[79] MS Granovetter, The strength of weak ties. *Am. journal sociology* **78**, 1360–1380 (1973).

[80] A Asikainen, G Iñiguez, J Ureña-Carrión, K Kaski, M Kivelä, Cumulative effects of triadic closure and homophily in social networks. *Sci. Adv.* **6**, eaax7310 (2020).

[81] NA Christakis, JH Fowler, The spread of obesity in a large social network over 32 years. *New*

Engl. J. Medicine **357**, 370–379 (2007).

[82] I Smirnov, S Thurner, Formation of homophily in academic performance: Students change their friends rather than performance. *PLOS ONE* **12**, e0183473 (2017).

[83] LSL Miller McPherson, JM Cook, Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).

[84] PF Drucker, The rise of the knowledge society. *The Wilson Q.* **17**, 52–72 (1993).

[85] UNESCO Institute for Statistics, How much does your country invest in R&D (`https://uis.unesco.org/sites/all/modules/custom/uis_applications/apps/visualisations/research-and-development-spending/`) (2022) Accessed: 2022-12-20.

[86] N Stehr, Knowledge societies. *The Wiley-Blackwell encyclopedia globalization* (2012).

[87] UNESCO, Unesco science report (`https://en.unesco.org/unesco_science_report/figures`) (2021) Accessed: 2022-12-20.

[88] L Bornmann, R Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J. Assoc. for Inf. Sci. Technol.* **66**, 2215–2222 (2015).

[89] A Edmunds, A Morris, The problem of information overload in business organisations: a review of the literature. *Int. journal information management* **20**, 17–28 (2000).

[90] A Hall, G Walton, Information overload within the health care system: a literature review. *Heal. Inf. & Libr. J.* **21**, 102–108 (2004).

[91] E Landhuis, Scientific literature: Information overload. *Nature* **535**, 457–458 (2016).

[92] DA Kronick, *History of scientific and technical periodicals.* (Rowman & Littlefield, Lanham, MD), 2 edition, (1976).

[93] M Fire, C Guestrin, Over-optimization of academic publishing metrics: observing goodhart's law in action. *GigaScience* **8** (2019).

[94] H Small, Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. for Inf. Sci.* **24**, 265–269 (1973).

[95] J Martyn, Bibliographiy Coupling. *J. Documentation* **20**, 236–236 (1964).

[96] ME Newman, Coauthorship networks and patterns of scientific collaboration. *Proc. national academy sciences* **101**, 5200–5205 (2004).

[97] S Thurner, W Liu, P Klimek, SA Cheong, The role of mainstreamness and interdisciplinarity for the relevance of scientific papers. *PLOS ONE* **15**, e0230325 (2020).

[98] F Battiston, et al., Taking census of physics. *Nat. Rev. Phys.* **1**, 89–97 (2019).

[99] M Szell, R Sinatra, Research funding goes to rich clubs. *Proc. Natl. Acad. Sci.* **112**, 14749–14750 (2015).

[100] P Chen, S Redner, Community structure of the physical review citation network. *J. Informetrics* **4**, 278–290 (2010).

[101] DJ de Solla Price, D Beaver, Collaboration in an invisible college. *Am. psychologist* **21**, 1011 (1966).

[102] M Golosovsky, S Solomon, Growing complex network of citations of scientific papers: Modeling and measurements. *Phys. Rev. E* **95**, 1–19 (2017).

[103] MEJ Newman, Prediction of highly cited papers. *EPL (Europhysics Lett.* **105**, 28002 (2014).

[104] KW Higham, M Governale, AB Jaffe, U Zülicke, Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields. *J. Informetrics* **11**, 1190–1200 (2017).

[105] AF Van Raan, Sleeping beauties in science. *Scientometrics* **59**, 467–472 (2004).

[106] MEJ Newman, The first-mover advantage in scientific publication. *Europhys. Lett.* **86**, 68001 (2009).

[107] VA Traag, L Waltman, NJ Van Eck, From louvain to leiden: guaranteeing well-connected communities. *Sci. reports* **9**, 1–12 (2019).

[108] KA Khoon, Serendipity in physics research. *Coll. Stud. J.* **42**, 1181–1183 (2008).

[109] CI Beckwith, *Greek Buddha*. (Princeton University Press, Princeton, NJ), (2015).

[110] M Czaika, S Orazbayev, The globalisation of scientific mobility, 1970–2014. *Appl. Geogr.* **96**, 1–10 (2018).

[111] M Trippl, Scientific mobility and knowledge transfer at the interregional and intraregional level. *Reg. Stud.* **47**, 1653–1667 (2013).

[112] L Ackers, Moving people and knowledge: Scientific mobility in the european union1. *Int. Migr.* **43**, 99–131 (2005).

[113] M Gibbons, R Johnston, The roles of science in technological innovation. *Res. policy* **3**, 220–242 (1974).

[114] A Saxenian, The genesis of silicon valley. *Built Environ. (1978-)* pp. 7–17 (1983).

[115] M Harrison, Does high quality research require critical mass. *The Quest. R&D specialisation: Perspectives policy implications* pp. 57–59 (2009).

[116] R Johnston, Effects of resource concentration on research performance. *High. Educ.* **28**, 25–37 (1994).

[117] F Berkes, J Colding, C Folke, Rediscovery of traditional ecological knowledge as adaptive management. *Ecol. Appl.* **10**, 1251–1262 (2000).

[118] DJ Levitin, Knowledge songs as an evolutionary adaptation to facilitate information transmission through music. *Behav. Brain Sci.* **44** (2021).

[119] T Ingham, Over 60,000 tracks are now uploaded to spotify every day. that's nearly one per second (`https://www.musicbusinessworldwide.com/over-60000-tracks-are-now-uploaded-to-spotify-daily-thats-nearly-one-per-second/`) (2021) Accessed: 2022-08-25.

[120] B Houghton, 60,000 tracks are uploaded to spotify every day (`https://www.hypebot.com/hypebot/2021/02/60000-tracks-are-uploaded-to-spotify-every-day.html`) (2021) Accessed: 2022-08-25.

[121] HB Hu, DY Han, Empirical analysis of individual popularity and activity on an online music service system. *Phys. A: Stat. Mech. its Appl.* **387**, 5916–5921 (2008).

[122] J Pham, E Kyauk, E Park, Predicting song popularity. *Dept. Comput. Sci., Stanf. Univ., Stanford, CA, USA, Tech. Rep* **26** (2016).

[123] A Franken, L Keijsers, JK Dijkstra, T ter Bogt, Music preferences, friendship, and externalizing behavior in early adolescence: A SIENA examination of the music marker theory using the SNARE study. *J. Youth Adolesc.* **46**, 1839–1850 (2017).

[124] R Guidotti, G Rossetti, "know thyself" how personal music tastes shape the last.fm online social network in *Formal Methods. FM 2019 International Workshops*, eds. E Sekerinski, et al. (Springer International Publishing, Cham), pp. 146–161 (2020).

[125] T Duricic, D Kowald, M Schedl, E Lex, My friends also prefer diverse music: Homophily and link prediction with user preferences for mainstream, novelty, and diversity in music in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21. (Association for Computing Machinery, New York, NY, USA), p. 447–454 (2022).

[126] G Di Bona, et al., Social interactions affect discovery processes (2022).

[127] EF Cabrera, A Cabrera, Fostering knowledge sharing through people management practices. *The international journal human resource management* **16**, 720–735 (2005).

[128] Palazzo delle esposizioni, Uncertainty. interpret the present, predict the future (`https://www.palazzoesposizioni.it/mostra/incertezza-interpretare-il-presente-prevedere-il-futuro`) (2021) Accessed: 2022-12-22.

[129] M Furka, Master's thesis (Vienna University of Technology) (2022).

[130] SC Han, T Lim, S Long, B Burgstaller, J Poon, Glocal-k: Global and local kernels for recommender systems in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. pp. 3063–3067 (2021).

Appendix A

Additional work

In this appendix, we briefly present additional work that is not covered in the main body of the thesis but nevertheless was done as part of the author's Ph.D. program at the University of Vienna and the Complexity Science Hub Vienna.

A.1 Test case at the Complexity Science Hub Vienna and Sony Computer Science Labs Paris

In chapters 2.4 and 6, we presented the *KOUZAN* application, used to record knowledge flows in organizations. Using this application, we conducted a test case in the style of a social experiment at the Complexity Science Hub Vienna and Sony Computer Science Labs Paris that lasted for almost three years. In this section, we give a short overview of the test case and the *KOUZAN* application in this context.

With regard to knowledge flows on short timescales, our main aim is to capture, model, and understand knowledge flows in collaborations. Understanding the knowledge flows can help us to address the many issues related to knowledge, that organizations are facing today, such as the loss of knowledge when experienced staff is retired that is caused by the increasing generational change, or the silo-like structures of large organizations, where knowledge is highly segmented and separated sectors of the organization don't have access to all the knowledge they might need.

These issues are difficult to address, mainly because of the intangible nature of knowledge. Very often, people acquire implicit knowledge or knowledge that is related to their job but is not accounted for in their CVs, publications, or in organizational diagrams. Addressing them would require making the intangible tangible, i.e., making knowledge flows visible and verifiable. For this purpose, we built the *KOUZAN* application. The application records interactions everywhere knowledge is shared, in discussions, meetings, lectures, talks, *Skype* calls, *Slack* chats, and more. *KOUZAN* builds a precise, time-resolved picture of who knows what, who teaches whom, and who learns from whom. This dynamic overview allows for the observation of changes over time and may thus be used to assess the efficiency of newly implemented initiatives such as workshops or talks. With methods discussed in chapter 1 such as triadic closure and centralities, we can study the entire network of connections in order to highlight virtuous or detrimental patterns. This can be used to identify persons who connect groups or those who are isolated in silos, for example.

During the test case, the application was introduced to the members of the Complexity Science Hub Vienna and Sony Computer Science Labs Paris. They were encouraged, to use the application during their work to record exchanges of knowledge. The test case was active between June 2018 and early 2021, and a total of 126 users participated, recording 3509 interactions. User feedback was collected to iteratively improve the application.

There were two major versions of the application during the course of the test case. The first version

was a prototype that Niklas Reisz developed in PHP and JavaScript and that the patent is based on. In this version, the application was a responsive, mobile-friendly website. The second version was flexible REST-API, that is, an interface with standardized input and output that different services, such as other websites or mobile applications, can connect to. Niklas Reisz developed the REST-API using the python-based *Django REST framework*, built the database in *Postgresql*, built connections to *Slack* and Sony's collaborative text editor *WeWrite*, and supervised the development of the front-end application. The front-end application was designed and developed by the independent contractors, Mag. Raphael Kieslinger and Dr. Martin Knecht.
Due to the different versions, the many improvements we made, and the features we added, the data we collected during the test case, is inhomogeneous or even inconsistent. While it strongly served its purpose of improving the application, we decided against using it for a scientific study. Nevertheless, the final version of the application that emerged at the end of the test case, is very well suitable for being used in a well-controlled scientific experiment. A possible experiment design could include two surveys at the beginning and at the end of a three-month time period, during which the *KOUZAN* application would be used to record all knowledge flows in an organization. If the survey answers are consistent with the results of the application, one could reasonably assume that knowledge flows were accurately recorded and could use the recorded data to study the structure and dynamics of the organizational knowledge network.

A.2 Exhibition Incertezza at the Palazzo delle Esposizioni in Rome

As part of this thesis, we created the *KOUZAN* application, used to record frequent knowledge flows in organizations. We present this application in chapters 2.4 and 6. In addition to the *KOUZAN* application, we created a second application called *KOUZAN EXPO*, which was adapted to the context of exhibitions. This application was used during the exhibition *Incertezza* at the famous *Palazzo delle Esposizioni* in Rome between October 2021 and February 2022 in collaboration with Sony Computer Science Labs Paris.
Incertezza. Interpretare il presente prevedere il futuro, or *Uncertainty. Interpreting the present, predicting the future* in its English translation, was an exhibition about uncertainty. Its description was as follows:

> *The exhibition Uncertainty. Interpreting the present, predicting the future illustrates some of the many facets of the idea of uncertainty, and the ways in which science deals with it. And from the awareness of the ineluctability of uncertainty arises the need to rethink and enhance the ideas and tools that allow us to understand and govern it, questioning science, art and culture at all levels.* [128]

The *KOUZAN EXPO* application was used during this exhibition to provide supplementary information for each exhibit and allow the users to rate exhibits and give written feedback. In addition to that, the application tried to guess the user behavior at each step, by predicting the most likely path and ratings for the user. The predictions were then shown to the user at the end of the visit in a prediction report. As such, the application was an exhibit on its own that was used to demonstrate the concept of recommender systems and path predictions to its users. It was accompanied by a visualization of visitor paths and ratings, that was projected onto the wall of the exhibition hall. The application was available in English and Italian. In figure A.1, the landing page of the application is shown.
The data collected by the application included a survey of demographics that each user completed upon signup. Together with the visit timeline and ratings, this data formed a solid basis for a study with the aim of designing recommender systems that optimize knowledge flows. This could be done, for instance by suggesting optimized paths to users with limited time and making personalized suggestions about the most interesting points to visit. This task was approached in a data science master's

thesis at the University of Technology Vienna, which is introduced in the next section. In addition to that, the data can be used to understand and optimize visitor dynamics, for instance by identifying the most common visitor paths, identifying the most common sequences in which exhibits are visited, or optimizing paths and occupancies with regard to Corona restrictions.

Niklas Reisz's contributions to this experiment included:

- requirements engineering
- stakeholder management and communication
- conceptualization of the application
- planning the software architecture
- backend development of REST API and database
- supervision of frontend development, which was done by Dr. Martin Knecht, and the UI/UX design which was done by Mag. Raphael Kieslinger
- deployment and supervision during the first days of the exhibition
- data management, preparation, and analysis

The experiment demonstrated that the *KOUZAN EXPO* application is able to record knowledge flows in exhibition settings in a meaningful way, which opens up opportunities to expand and scale up the concept in follow-up use cases. It allows for the collection of unique data that is needed to understand knowledge flows in organizations on a deeper level.

A.3 Co-supervision of a master's thesis on rating and path predictions in exhibitions

The data collected during the *Incertezza* exhibition at the *Palazzo delle Esposizioni* in Rome, was studied in a data science master's thesis at the University of Technology Vienna titled *Exhibit rating prediction and visitor path prediction in a museum setting* and written by Marek Furka in 2022 [129]. The aim of the study was to perform a comprehensive comparison of state-of-the-art methods for predicting user paths and user ratings in an exhibition setting. The studied methods included matrix factorization, collaborative filtering, standard and advanced machine learning models, as well as deep learning models. The best-performing algorithm for the prediction of user ratings was the GLocal-K algorithm [130], which is the state-of-the-art approach for the popular *MovieLens* dataset. For user path predictions, a deep learning model performed best.
The results are relevant for the optimization of knowledge flows and visitor happiness in exhibition settings. Niklas Reisz's contribution to this work includes data collection during the experiment and co-supervision of the thesis.

A.4 Exhibition at the Maker Faire Rome 2022

The *KOUZAN EXPO* application, which is a separate version of *KOUZAN*, introduced in chapters 2.4 and 6, that was built explicitly for the setting of exhibitions, was used during the *Maker Faire Rome* that took place between the 7th and 9th of October 2022. The exhibition consisted of several hundred exhibits that were displayed in tents or halls around a larger area and visited by thousands of guests. There, the application acted as a companion that provided supplementary information for each exhibit

and allowed the visitors to rate exhibits and give written feedback. As part of the experience, the application predicted user ratings and compared them to the actual ratings in a gamified fashion.
The data collected during the exhibition can be used to provide feedback to exhibitors and organizers, to visualize the routes of visitors and occupancies of exhibitions, or to train advanced recommender systems. Niklas Reisz's contributions to this experiment were in the supervision of software adaptations and the preparation and analysis of the data.
The experiment can be seen as a step towards more ambitious goals, recording and optimizing the knowledge exchange at events like conferences where there is a mix of exhibits, talks, and discussions. In such settings, where time is limited and attendees have to choose which talks to listen to, which exhibits to visit, or whom to talk to, an advanced version of the *KOUZAN EXPO* application could provide interest-based recommendations, suggest optimized paths while taking into account occupancies or Corona restrictions, and make suggestions for like-minded people to meet.

A.5 Corona task force at the Complexity Science Hub Vienna

During the onset of the Corona pandemic in Austria in early 2020, the Complexity Science Hub Vienna took on the challenge to support the Austrian ministries in their decision-making with regard to the Corona crisis. All regular work was suspended and researchers were forming a *Corona task force* focused on topics related to the Corona pandemic. Their aim was to collect data and provide statistical analysis and simulations to enable evidence-based political decision-making.
During the time of the Corona task force, the thesis' author was working on preliminary work toward simulating the Austrian food supply chains, in order to determine possible scenarios in which supply shortages might occur and identify actors and practices that might be at high risk of causing such shortages. The general idea of the project was, to model the country-wide imports, production, logistics, retailers, and consumption of food. For this, we intended to collect data from the biggest supermarket chains in Austria and their suppliers and map the flows of goods to match them with local consumption. Once modeled, the model would allow us to assess scenarios, such as the shutdown of a logistics company due to the staff getting infected with Corona. Based on these simulations, we would be able to identify critical actors that carry high systemic risk and develop strategies to lower their risk that let us protect vulnerable regions.
While the project continued until today and led to a whole new research unit at the Complexity Science Hub Vienna, Niklas Reisz's involvement was limited to the time between March and November 2020, during which the *Corona task force* existed. During that time, he contributed the following:

- He collected data from the Austrian government that allowed us to break down Austria into very small geographical units called *Zählsprengel*. For these units, he collected data on the number of inhabitants, such that one could estimate the consumption of goods.
- He collected a dataset that would let us convert products into nutrients, which would allow us to estimate if available products were sufficiently covering nutrient demand.
- He contributed to the design of a model for estimating the number of customers per supermarket and estimating how these customers spread to other supermarkets in case of a supply shortage or a full shutdown of the supermarket.

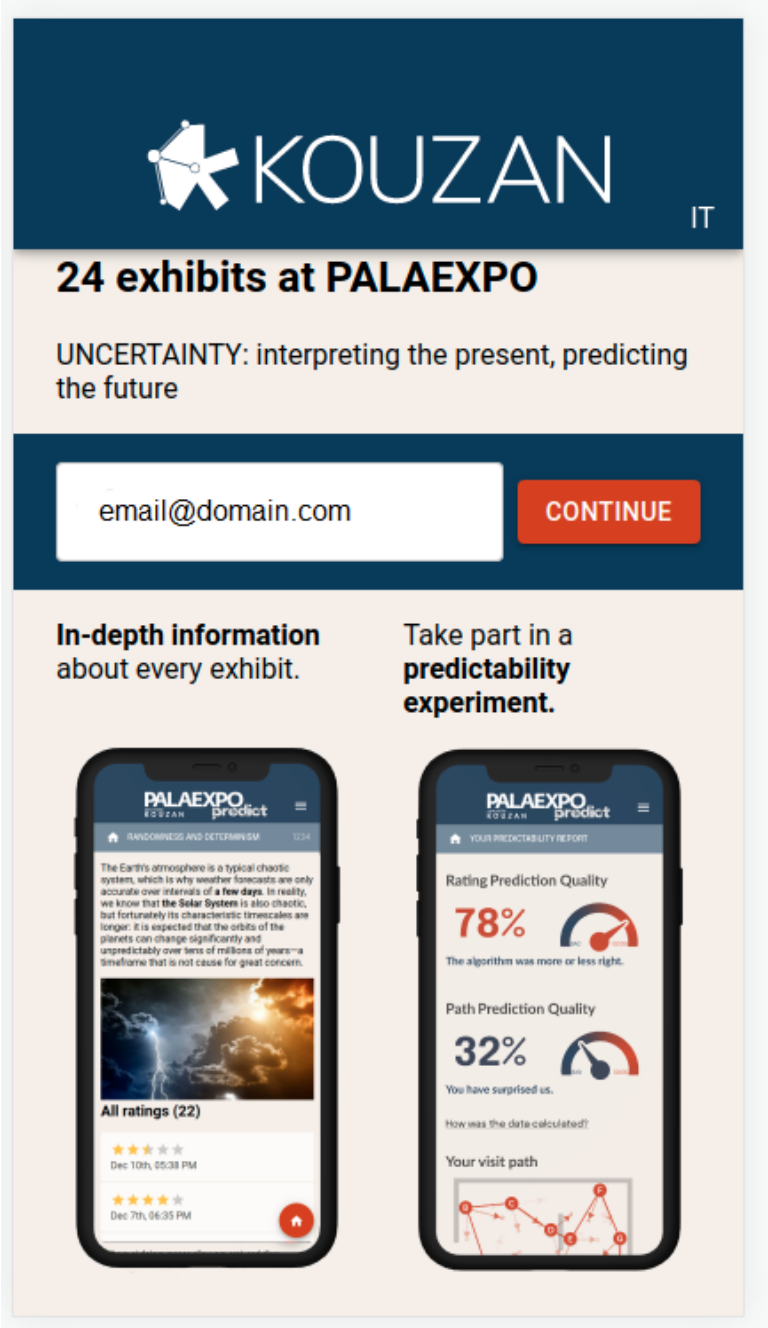


Figure A.1: An example landing page of the *KOUZAN EXPO* application used during the experiment at the *Palazzo delle Esposizioni* in Rome.

Appendix B

Supplementary information

B.1 The attachment kernel of songs

In this section, we present our findings on the attachment kernel of songs, that did not make it into the final paper.
From a birds-eye view, there are two main mechanisms that determine the success of a song, preferential attachment, and forgetting. Preferential attachment in our case indicates that the probability of listening to a particular song increases with the number of times that song has been listened to. Forgetting on the other hand indicates that the probability decreases over time. To demonstrate that these two mechanisms are dominant in *last.fm*, we apply a method introduced in [1]. We measure the attachment kernel in two dimensions and find a function $f(k,t) \sim \frac{k}{t}$ where f is the attachment probability of a song, k is the number of times a song has been listened to and t is the age of the song. The kernel is shown in figure B.1.
The second important factor we identified in the attachment kernel is time. If we look at the age of songs that people on *last.fm* are listening to, we find that people strongly prefer listening to new songs over older songs. We show this in figure B.2. This is directly related to the attachment kernel that we observe. The probability for a song to be listened to, declines as a power law of time.
While on longer timescales, song popularity typically declines as a power law, on shorter timescales there are exceptions. Different songs can exhibit different short-term trends which can tell a lot about the potential success of a song. Examples are shown in figure B.3. We used this information in the main paper [4], to give predictions about the future success of a song.

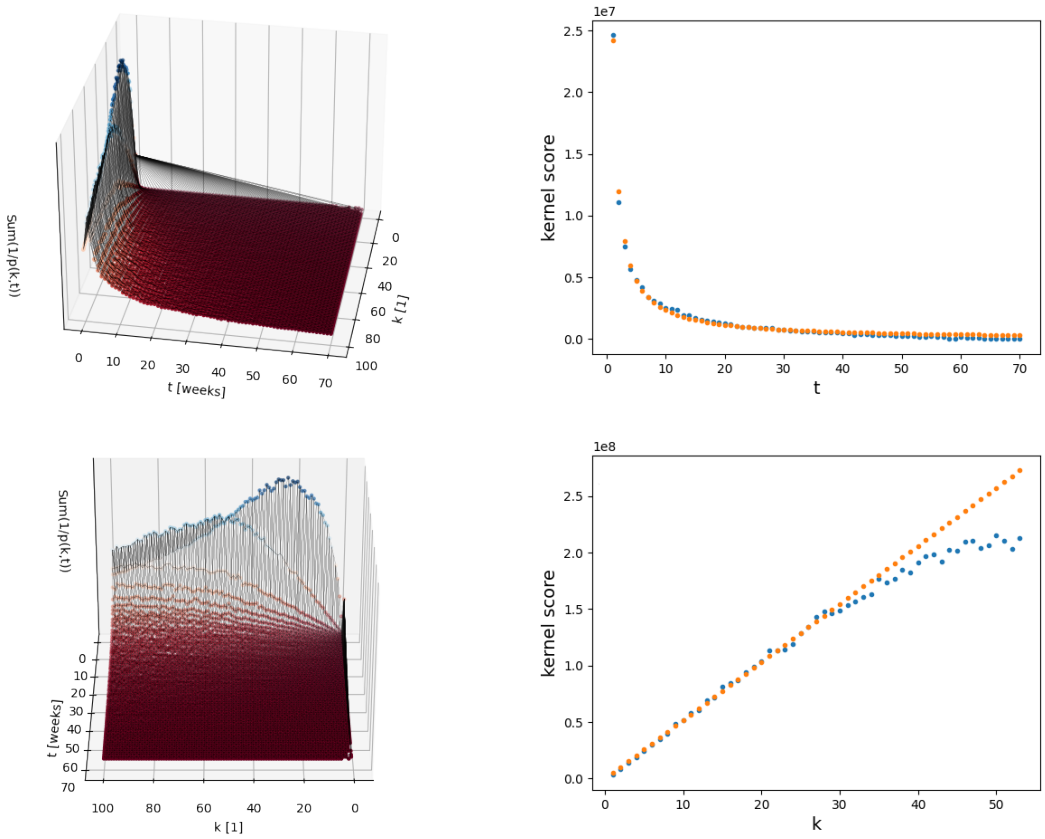


Figure B.1: The attachment kernel of songs in *last.fm*. The attachment kernel represents the likelihood of a song being listened to as a function of the parameters of that song. The two dominant parameters in *last.fm* are time (the time that has passed since the song was released) and the number of listenings the song has accumulated so far (k). On the left, two different view angles of the attachment kernel as a function of time and the number of listenings are shown. On the right, we show two slices through these planes. The top-right panel shows an example slice in time t, where the number of listenings k is held constant. The bottom-right panel shows an example slice in listenings k, where the time t is held constant. The slices reveal that the kernel can be approximated well by a shape $\sim \frac{k}{t}$.

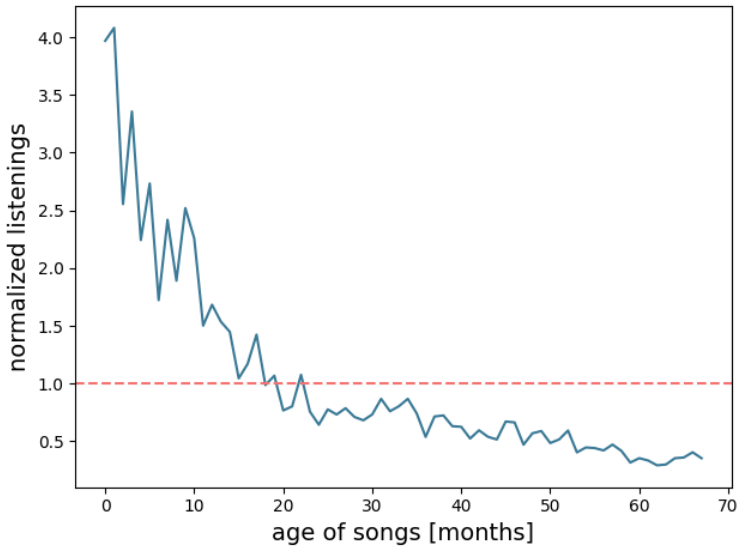


Figure B.2: The age of the music that last.fm users are listening to. The y-axis reports normalized counts, where a value of 1 represents the expected number of listenings in the case where people listen to songs independently of the song's age. Values greater than one mark over-representation and values lower than one under-representation respectively.

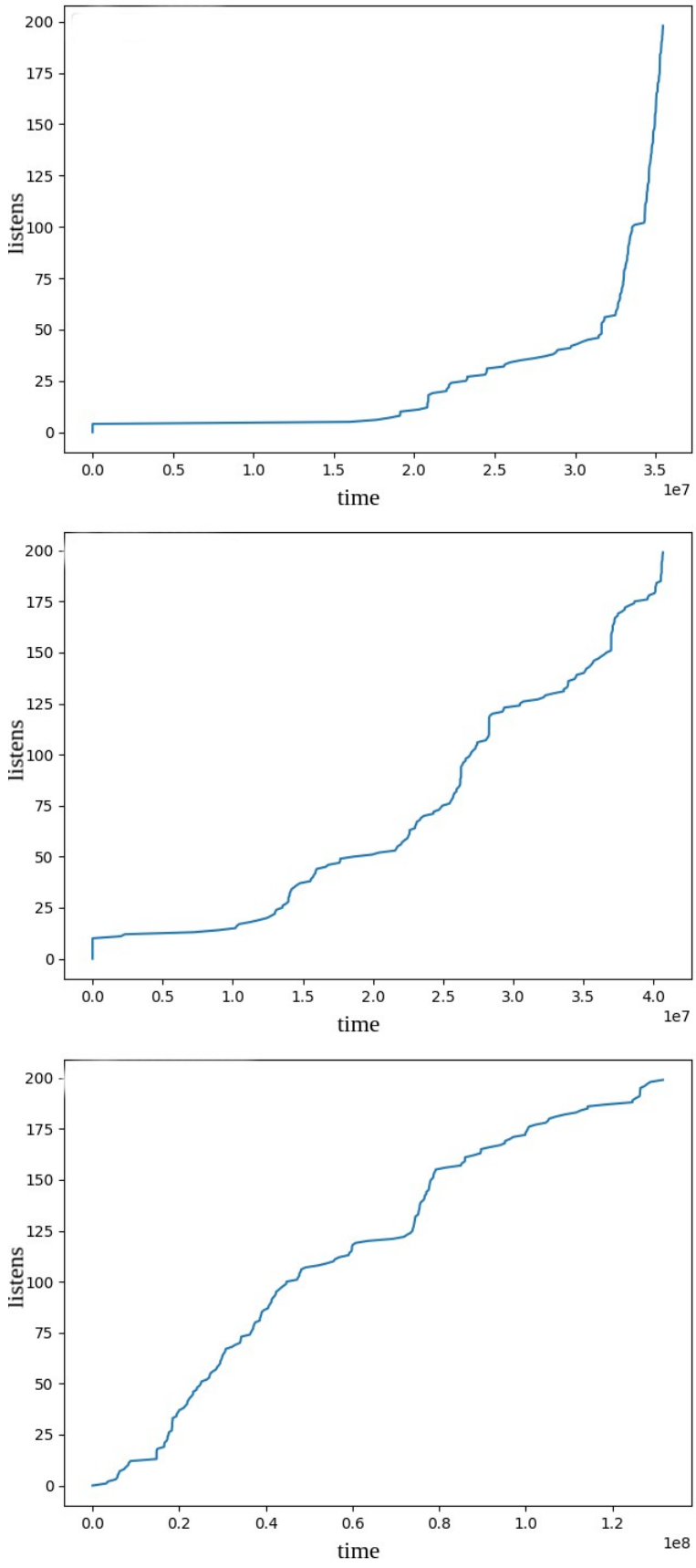


Figure B.3: Example listening curves for three different songs. The cumulative listenings of up to 200 listenings in total are plotted against time measured in seconds.