

Quantifying the Evolution of Success with Network and Data Science Tools



Milán Janosov

Department of Network and Data Science
Central European University

Principal supervisor: Prof. Federico Battiston
External supervisor: Prof. Roberta Sinatra
Associate supervisor: Prof. Gerardo Iñiguez

*A Dissertation Submitted in fulfillment of the requirements
for the Degree of Doctor of Philosophy in Network Science*

2020

Milán Janosov: *Quantifying the Evolution of Success with Network and Data Science Tools*, ©
2020
All rights reserved.

I Milán Janosov certify that I am the author of the work *Quantifying the Evolution of Success with Network and Data Science Tools*. I certify that this is solely my original work, other than where I have clearly indicated, in this declaration and in the thesis, the contributions of others. The thesis contains no materials accepted for any other degree in any other institution. The copyright of this work rests with its author. Quotation from it is permitted, provided that full acknowledgment is made. This work may not be reproduced without my prior written consent.

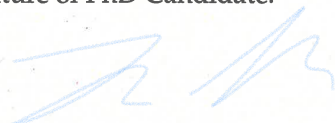
Statement of inclusion of joint work

I confirm that Chapter 3 is based on a paper, titled "Success and luck in creative careers", accepted for publication by the time of my thesis defense in EPJ Data Science, which was written in collaboration with Federico Battiston and Roberta Sinatra. On the one hand, the idea of using her previously published impact decomposition method to quantify the effect was conceived by Roberta Sinatra, where I relied on the methods developed by her. On the other hand, I proposed the idea of relating the temporal network properties to the evolution. I carried out the data collection, the numerical calculations and data analysis, and the data visualizations. All authors contributed to the writing of the paper on which the chapter is based and gave final approval for publication in this dissertation.

I confirm that Chapter 4 is based on a publication that was written with Federico Musciotto, Federico Battiston, and Gerardo Iñiguez, and has been accepted for publication under the title "Elites, communities and the limited benefits of mentorship in electronic music" in Scientific Reports, a journal of the Nature Research family. In this work, I proposed the idea of studying success and collaborations in electronic music, while Dr. Iñiguez proposed the used methodologies to understand the dynamics of the analyzed ranking. While I suggested the core methods used for the network analysis, both Dr. Musciotto and Dr. Battiston contributed to it. I suggested the idea of mentorship. I carried out the data collection, the numerical calculations and data analysis, and the data visualizations. All authors contributed to the writing of the paper on which the chapter is based and gave final approval for publication.

I confirm that Chapter 5 is based on a working paper which was written by me, Luca Maria Aiello, and Daniele Quercia based on my internship project at the Nokia Bell Labs in Cambridge in 2018. Dr. Aiello and Dr. Quercia conceived the idea of combining urban informatics and the science of success, while we collaboratively designed the study and developed the methodologies used. Dr. Aiello carried out the data collection, while I did the numerical calculations and data analysis. All authors contributed to the writing of the working paper and gave final approval for publication in this thesis.

Signature of PhD Candidate:



February, 2020

Signature of Dr. Federico Battiston, endorsing statement of joint work:



February, 2020

Signature of Dr. Roberta Sinatra, endorsing statement of joint work:



February, 2020

Signature of Dr. Federico Musciotto, endorsing statement of joint work:



Signature of Dr. Gerardo Iñiguez, endorsing statement of joint work:

A handwritten signature in black ink, appearing to be 'G. Iñiguez', with a stylized flourish at the end.

January, 2020

Signature of Dr. Luca Maria Aiello, endorsing statement of joint work:

A handwritten signature in black ink, appearing to be 'Luca Maria Aiello', with a stylized flourish at the end.

January, 2020

Signature of Dr. Daniele Quercia, endorsing statement of joint work:

A handwritten signature in blue ink, appearing to be 'D. Quercia', with a long horizontal flourish extending to the right.

January, 2020

ABSTRACT

In this thesis, I quantify several aspects of success, defined as the collective response to individual performance in five different domains: science, film, music, literature, and urban spaces. The general goal of my work is to provide new insights into the evolution of success in these five domains. To achieve that, I particularly focus on the role of networking behavior and the effect of the social fabric of the studied fields on the emergence of success.

Each chapter tackles different questions related to the temporal unfolding and the determinants of success in different social domains. First, I build on a previous modeling approach to describe the evolution of scientific careers and capture the role of randomness in science, music, literature, and film. In addition, I extend the discussion on the role of randomness to networking by analyzing the co-evolution of success and collaboration networks. Second, I zoom in to the field of electronic music. After uncovering the rise and fall of different DJ communities, I test a definition of mentorship and show the controversial effects mentoring has on the mentees' prospected success. Third, I analyze the time evolution of the popularity of urban venues and reveal that their success trajectories follow several substantially different shapes over time. Moreover, I present machine-learning-based modeling efforts on understanding what are the most influential urban features predicting what direction a venue's success will take. Finally, I outline the main contributions of my thesis work and discuss several possible real-world applications.

ACKNOWLEDGEMENTS

I would like to acknowledge several individuals who have, whether they realize it or not, significantly influenced this work.

First, I must thank my supervisors, Prof. Roberta Sinatra, who introduced me to the Science of Success and has been guiding me through the past three years, and Prof. Gerardo Iñiguez, who started advising me for the last year of my Ph.D., for their continued support and patience. I am exceptionally grateful to Prof. Federico Battiston for all his support and encouragement, even at the most critical moments. Their guidance has been invaluable.

I am grateful to the members of the Department of Network and Data Science at CEU: the head of the department, Prof. János Kertész, our beloved coordinator Olga Peredi, the fellow Ph.D. students; and the postdocs and visitors, who provided a vivid environment to learn and grow. I am particularly thankful for the helpful suggestions and comments to Dr. Srebrenka Letina, Manran Zhu, Dr. Johannes Wachs, Dr. Dávid Deritei, Prof. Rossano Schifanella, Luis Natera, Dr. Federico Musciotto, Matteo Neri, Prof. Silvia Fiarescu, Dr. Tamer Khraisha, Dr. Carl Nordlund, Júlia Perczel, Orsolya Vásárhelyi, Cory Cox, and Prof. Balázs Vedres.

I thank Dr. Thomas Rooney from the Center for Academic Writing at CEU for the many consultations and teaching me the skills to write this dissertation.

I would like to express my gratitude toward all my colleagues and fellow interns at the Bell Labs in Cambridge, UK, with whom I got to spend a fantastic summer internship and learn more than I could list here. In particular, I thank Dr. Luca Maria Aiello and Prof. Daniela Quercia for the opportunity, the time and patience, and the prosperous project they gave to me.

I am also grateful to Prof. Albert-László Barabási for hosting me in his Lab during the spring of 2019, where I got to learn a lot about doing world-class research from fantastic colleagues, such as Prof. Albert-László Barabási, Dr. Alexander J. Gates and Dr. Onur Valur. Besides, I thank Péter Ruppert for giving me the opportunities to test my skills outside of the academic world on exciting consultancy projects.

I am thankful to my external collaborators for broadening my perspective on data-driven research. In particular, to Prof. Robyn Radway, Dr. Miklós Orbán, Olga Zagovora, Prof. Claudia Wagner, Dr. Rui Sun, Dr. Miriam Redi, the Summer School Series on Methods for CSS organizer team, the COSTNET community, Maven7, and the Eötvös József Collegium.

I would like to acknowledge my former mentors without whom I would never have stepped into a Ph.D. program, let alone finish one. Thank you Prof. Tamás Vicsek and Dr. Gábor Vásárhelyi from Eötvös Loránd University for setting me on a path that led me to a Ph.D.. Dr. Péter Kozma, Dr. Csaba Sándor Daróczi and Dr. Péter Petrik from the Research Institute for Technical

Physics and Material Sciences who mentored and supported me from high-school through my undergrad years and introduced me to the realm of science.

I thank all my friends for your kind patience, encouragement, and support which was essential to keep me on track during even the most difficult times of my Ph.D. years, especially Dr. Ali Sayour, Dr. Bálint Koczor, Dr. János Zsiros, Manran Zhu, Dr. Péter Kozma, and Tamás Lukács.

I am lucky to have a loving family who provided me endless love and encouragement, and I hope they know how much this means to me. My mother, Angéla, my father Iván, and my grandmother, Erzsébet, have provided me incredible support throughout the years. I dedicate this work to them.

CONTENTS

Contents	i
1 Introduction	1
2 Related Work	5
2.1 Research on success originated in social sciences	6
2.1.1 Psychological research on excellence	6
2.1.2 Social network research and success	7
2.2 Data-driven research on success	9
2.2.1 Measuring success	9
2.2.2 Quantitative results on scientific impact	12
2.2.3 Measuring scientific fields	14
2.2.4 Capturing individual career success	15
2.2.5 Networks from the individuals' perspective	17
2.2.6 Networks of creative fields	20
3 Success and luck in creative careers	23
3.1 Introduction	23
3.2 Data description	25
3.2.1 Data collection	25
3.2.2 Measuring success	26
3.3 The Q-model: decomposing impact through luck and individual ability	29
3.3.1 The random-impact-rule	29
3.3.2 Introducing the Q-model	31
3.3.3 Predictions of the Q-model	32
3.4 Capturing the role of randomness in impact	36
3.4.1 Luck in the Q-model	36
3.4.2 Combining the Q-model with the classical test theory . . .	37
3.4.3 Randomness in creative careers	39
3.5 Role of randomness in collaborations	42

3.5.1	Constructing collaboration networks	42
3.5.2	Correlation between impact and network position	43
3.5.3	Randomness in networking	46
3.6	Discussion	49
4	Communities and mentorship in electronic music	51
4.1	Introduction	51
4.2	Data	52
4.3	Dynamics of the top 100 ranking list	53
4.3.1	Regime change in the ranking	54
4.3.2	Ranking and popularity	56
4.4	Co-release network in the DJ world	57
4.4.1	Building and describing the co-release network	57
4.4.2	DJ Communities	58
4.5	Mentorship in electronic music	64
4.5.1	The existence of mentorship	65
4.5.2	The benefits and limitations of mentorship	65
4.6	Discussion	68
5	Success evolution of urban venues	71
5.1	Introduction	71
5.2	Data	72
5.2.1	Data description	72
5.2.2	Data cleaning	74
5.2.3	Measuring urban success	75
5.3	Success trajectories	76
5.3.1	Measuring and transforming success trajectories	76
5.3.2	Clustering success trajectories	77
5.4	Describing urban spaces	80
5.4.1	Features describing the venues' characteristics (<i>what</i>)	80
5.4.2	Features describing the venues' neighborhoods (<i>where</i>)	81
5.4.3	Features describing the venues' visitors (<i>who</i>)	83
5.4.4	Features summary	85
5.5	Predicting venue success	87
5.5.1	Binary success classification	87
5.5.2	Six-shape prediction	90
5.6	Discussion	92
6	Conclusion	95

LIST OF FIGURES

2.1	Modeled productivity curve.	6
2.2	Prevalence of scientific collaborations.	8
3.1	Impact distributions modeled by log-normal functions for four example fields.	27
3.2	Rescaled impact distributions.	27
3.3	Career examples for four example professions.	28
3.4	Productivity distributions.	31
3.5	Measurements on the Q -model.	33
3.6	Comparison of Q parameters between early and final stages of careers in film, music, and literature.	33
3.7	Comparing the R -model and the Q -model to the data.	36
3.8	Rescaled cumulative impact distribution for four example fields.	37
3.9	Illustration of the classical test theory.	38
3.10	Results of the Mann–Whitney U test comparing the Q and p distributions.	39
3.11	Comparing the role of randomness across 28 professions	40
3.12	Temporal evolution of the collaboration networks.	43
3.13	Correlations of network measures.	44
3.14	Network position and timing of the biggest hit for movie directors.	45
3.15	Distribution of the shifting parameter τ	46
3.16	Success distributions for different networking behaviors	48
4.1	Visualizing the top 100 ranking over time.	54
4.2	Capturing regime change in the top 100 ranking.	55
4.3	Comparing the top and the bottom of the ranking.	56
4.4	Top 100 DJ network.	59
4.5	The evolution of the top 100 DJ network.	60
4.6	Temporal dynamics of communities in electronic music.	61

4.7	Average time scales of DJ communities.	61
4.8	Genre similarities and peak time differences of DJ communities.	63
4.9	Mentorship in electronic music.	65
4.10	Mentors in electronic music.	66
4.11	Mentee-mentor relationships.	67
5.1	Visitors' venue distribution and the DBSCAN method.	74
5.2	Measures of urban success.	75
5.3	Success trajectories of urban venues.	77
5.4	Clustered success trajectories.	79
5.5	Urban features describing the venues' characteristics (<i>what</i>).	81
5.6	Urban features describing the venues' neighborhoods (<i>where</i>).	82
5.7	Urban features describing the venues' visitors (<i>who</i>).	83
5.8	The social graph of Foursquare users resident in London.	86
5.9	Correlations between urban features.	88
5.10	Binarized success prediction results.	89
5.11	Feature importances in binary prediction.	90
5.12	Six-shape success prediction results.	91

CHAPTER 1

INTRODUCTION

Throughout history, exceptionally productive and successful people, from scientists to artists, have made a long-lasting impact on culture and society. However, our understanding of the emergence of their success and the evolution of their careers is still vague and incomplete. For example, Albert Einstein, who authored more than a hundred scientific papers from 1901-1955, published his four most influential articles in 1905 at a time he was 25 years old [1, 2]. On the other contrarily, George Orwell, a classic example of late-bloomers [3], published his two most famous novels, *Animal Farm* and *1984*, during the last years of his career [4]. Picasso paired exceptional productivity with success as he executed more than twenty-thousand paintings, while Mendel, the founder of modern genetics, wrote only seven publications [5–7].

These anecdotal examples illustrate that outstanding recognition and success in creative fields can come in diverse forms, apparently with less regard for experience, biological or career age, and overall productivity [8]. Therefore I ask: what are the major components that drive success? How much does success depend on the individuals themselves? What could be the role of external random factors? And how does the interplay – the network – of the individuals fit into the picture? How can we use the recently available large-scale databases combined with tools of network analysis and data science to answer these questions?

Social psychology has been studying career patterns of eminent artists and scientists for decades, attempting to capture common patterns throughout disciplines and ages [8–13]. While these early-stage studies produced several important and pioneering ideas and theories, they lack quantitative evidence, since most of them are based on case-studies or the analysis of small datasets covering only small, biased samples of high-achievers. Therefore, the generality of

these findings is questionable at the scale of entire populations of artists, scientists, etc. Fortunately, during recent years, the availability of online platforms and large datasets covering science (e.g., Web of Science [2]), film (e.g., Internet Movie Database, IMDb [14]), music (e.g., Discogs [15]), or even urban venues (e.g., Foursquare [16]) opened the door to test old theories and build new ones, which even led to the birth of a new data-driven field, computational social science [2, 14–21].

A driving area of the research aiming to quantify success under the umbrella of computational social science has been focusing on describing and modeling the success of scientists [22, 23]. These research projects include thorough analyses of how to measure the impact of scientific papers and how to extend those measurements to the success of individuals, most typically via various functions of the number of citations papers receive [24, 25]. In addition, interesting work has been published on understanding statistical properties and temporal evolution of scientific impact, just as much as its implications on research evaluation and policymaking [26–29]. Other researchers attempted to incorporate interdisciplinary differences to capture the boundaries of different fields, such as computer science and physics and discussed the effects of interdisciplinarity on success [30–33].

Researchers' interests have looked beyond science and started analyzing other creative domains as well, both by applying well-established methodologies and ideas developed for studying success in science, and inventing new ones, resulting in the emergence of the science of success [34]. For instance, Yin and co-authors proposed a mechanistic model of success and failure in the entrepreneurial ecosystem [35]. Lacasa et al. focused on individual success in the show business by using data from the IMDb [14, 36], while Yucesoy et al. took tennis as an example to capture and measure the differences between performance and success [37].

While fundamental work has been done on the study of entire disciplines by analyzing populations of individuals both within and outside of science, important contributions have been published on the level of individual success as well. For instance, previous work discussed the mobility of researchers with regards to impact and institutional prestige [38, 39], research interest evolution and topical changes of scientists [40], and how scientific credit is allocated for follow-up works [41]. Further research attempted to model the evolution and age dynamics of individual careers, both focusing on high achievers, entire populations, and the comparisons of these two directions [42–45]. The work of Sinatra et al. [43] introduced a modeling approach named the *Q*-model, which I am building on in this dissertation. This model allows us to decompose scientific impact into two components: one encoding the individuals' ability to generate

impact, and one representing all external fluctuations (luck). Both components align with existing works discussing the role and nature of individual merit and luck in success [46–48].

Investigating entire disciplines from science to music is just as important for understanding the universalities of success as modeling the dynamics of individuals; one represents the macroscopic while the other the microscopic perspective. Not surprisingly, researchers also started to combine the micro- and macroscopic aspects to track down the role of the individuals in the social fabric of their fields – in other words, taming the network effects in success. It has been shown that individuals' network position, for instance, captured by degree centrality or constraint, has a certain predictive power on success in creative fields like jazz or science [49–55]. Brian Uzzi and his colleagues reported that the collaboration network structure of artists working on Broadway musicals shows clear relations to both the financial and the artistic success of the field during 1945-1990 [56, 57]. These findings of Uzzi et al. were already pointing towards future research on team success [58, 59], while Bonaventura et al. discussed the role of the employee-networks and the flow of people between companies in the success of start-ups [60].

This thesis aims to contribute to the science of success by building on hypotheses from the social sciences, relying on large-scale datasets, and elaborating on previously introduced quantitative tools and methods of network and data science to fulfill multiple goals. First, I apply previously introduced methods to model creative career success on new data sources and creative fields. Then I use these results to decompose randomness' and the individuals' contribution to success to compare 28 different creative domains such as mathematics, pop music, and literature. After that, I compare the temporal dynamics of the individuals' career success to the underlying evolution of the network structure and relate that to the individuals' peak success. Next, I zoom on to the highly collaborative and widely popular field of electronic music and propose a rank-dynamics-based method to capture the existence of an all-time elite. I also present a network-based analysis of the occurrence, rise, and fall of the different electronic music artist communities. In addition, I propose a network-based mechanism, aligning well to the literature on mentorship, that may be responsible for the observed long-standing elite and the highly ephemeral rest of the DJ world. Finally, I adopt analytical tools on the temporal evolution of success to study the popularity of urban venues over time. By clustering the success trajectories of these establishments, I manage to identify six distinct shapes their success follows over time. I also rely on features supported by urban theories to build machine learning models and evaluate the best predictors of each success-shape, which at the end turn out to be metrics related to the social environment

of the urban spaces.

The main goal of this thesis is to give a deeper understanding of how success and popularity evolve in such diverse domains as science, music, and urban venues. For that, I build on the tools of network science and combine network analysis with several statistical and machine learning methods to provide new insights into the temporal evolution of success in various fields. My work presents novel results on quantifying success in several ways. I show generic results on the modeling and characteristics of the evolution of careers from the timing of careers, the causality and mentorship in electronic music, and the popularity dynamics of urban venues. In addition, I highlight the unique and universal role of social networks in achieving success. Social networks seem to affect success from predicting the timing of big hits and opening the gates of early-career individuals to the top to the success and prosperity of urban spaces, connecting seemingly strikingly different domains of life, further emphasizing the role of networks in social phenomena.

The thesis is structured as follows:

- Chapter 2: I review previous work and introduce the field of science of success, including foundational works on social psychology and data-driven research both on individual and group success.
- Chapter 3: I extend earlier work on mechanistic and statistical modeling of individual careers in creative fields to propose a framework for capturing the role of luck and demonstrate it during the analysis of 28 creative professions. In addition, I discuss the role of randomness in the networking behavior of individuals by analyzing the co-evolution of network positions and impact.
- Chapter 4: I focus on the field of electronic music, and first show the existence of a long-standing elite, then explain the existence and formation of this elite by studying the underlying co-release network of electronic music artists. As a dominant mechanism, I propose and study mentorship, a special form of networking behavior.
- Chapter 5: I study the temporal evolution of the popularity of urban venues and identify six distinct clusters of them. I also build urban theory-inspired machine learning models to extract the main predictors of venue success, with a particular focus on social features, such as the venues' visitors network.

CHAPTER 2

RELATED WORK

In this chapter, I review previous work on researching various aspects of success. On the one hand, research on the emergence of excellence was initially done in social psychology disciplines. On the other hand, in recent years, many data-driven quantitative works from network, data, and computational social science [20] started to tackle similar problems, revisited old hypotheses, and proposed new directions, even leading to the birth of the science of success [23, 34].

First, I introduce the results of several qualitative and small-scale studies on creative career success, in particular on its temporal evolution, that later served as inspiration for large-scale, more data-driven studies. After that, I connect these results to recent researches closer to network and data science by first introducing ways of defining, measuring, and characterizing success in the era of big data. I highlight the differences between direct success measures, e.g., the number of citations a scientific publication receives, and indirect measures, such as the network centrality of a paper in the citation network of articles. Next, I show how these measures can be used to reconstruct success and popularity trajectories as time series, and how these time series can be used to understand the evolution of impact over time by statistical analysis, mechanistic modeling, and machine learning techniques. Finally, in the last section of this chapter, I focus on the relationship between the networking behavior of creative individuals, and the success and prosperity of their entire fields, from science to music.

2.1 Research on success originated in social sciences

2.1.1 Psychological research on excellence

Career development psychology has been an active discipline for decades, with many researchers being interested in the understanding of the ideational process behind creative production. Since Lehman [9], some of the major directions of these studies have been the temporal evolution of creative careers (age or productivity curves) and the link between quality and quantity [61, 62].

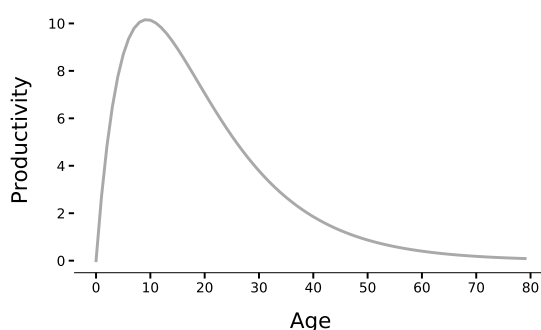


Figure 2.1. Modeled productivity curve.

Illustration of a theoretical age curve based on Eq. (2.1) showing the productivity of the individual as a function of its age based on Simonton's work [11, 12].

One of the first mathematical descriptions of creative career evolution was given by Simonton [11, 12], who explained how the age-curves shape over time (Figure 2.1). He built on a widely accepted theory by Campbell [10, 63]. Campbell's theory claims that the ideational process of creative work is Darwinian, meaning that the manner of the variation and selection of these combinations is unpredictable and disregards whether a certain combination is advantageous (and leads to a successful product) or not. With the help of Campbell's theory, Simonton proposed a differential-equation-based model on how the human mind constructs numerous combinations of ideas, concepts, and prior knowledge during creative work, which eventually results in creative products, like scientific papers, songs, and poems, presented to the public [11, 12]. The model's main working hypothesis is that the creative individuals have an initial level of "creative potential" C , that is consumed with a rate α during creative ideation and results in a collection of ideations, from which during the second step these "works-in-progress" became elaborated into finished products with a rate of β . In the next step, Simonton defined productivity (p), as the number

of contributions at time t , in the following form:

$$p(t) = c(e^{-\alpha t} - e^{-\beta t}). \quad (2.1)$$

Simonton also supported the validity of his model by empirical findings and compared the predictions of his model to smaller sets of data. For instance, by the analysis of individual careers, he found that the predicted number of patents over the career of Thomas Edison showed a 0.87 correlation to the actual number of patents Edison filed. Besides that, Simonton also found agreements between his model's predictions on peak productivity age, and the observations of Adams (1946) and Dennis (1966) based on several hundreds of creative individuals from professions such as lyric poetry, pure mathematics, and theoretical physics [8, 64]. Building further research on this conceptual framework, Simonton pointed out interdisciplinary differences and found zero correlation between eminence and age at the best contribution (while observed negative correlation at the age of the first, and positive correlation at the age of the last contribution) on a set of 2,026 scientists and inventors [65].

Simonton's model also served as empirical evidence in the debate between Dennis and Lehman [66, 67] on age effects on achievements, a finding that later became known in the psychology literature as the "equal-odds-rule" [6]. This rule states that during a creative career each product has the same chance of being the most successful. Simonton also argued that the linkage between quality and quantity is a probabilistic consequence of the "equal-odds-rule" [12]. His explanation, following Lehman's work [9], relied on the observation that the more average products someone creates, the higher her chance to create something exceptional. Feist arrived at similar results by studying an elite sample of 99 physicists and chemists, concluding that the quantity and the impact of research are positively related [68].

2.1.2 Social network research and success

Extending the theories on individual careers and teams, researchers like Bayer et al. started early work on understanding scientific collaborations [70]. Along these lines, the article of Beaver and Rosen [69] discussed how the prevalence of scientific collaborations (Figure 2.2) in the fields of biology, chemistry, and physics accelerated the professionalization of science and increased the quality of research mostly due to institutional changes and improvements on funding schemes [69, 71]. These findings were later supported by the analysis of the profiles of 443 scientists by Lee and Bozeman [72]. Further research also found that the number of authors significantly correlates to the papers' acceptance rate to prestigious journals showing direct evidence on the advantages of networking

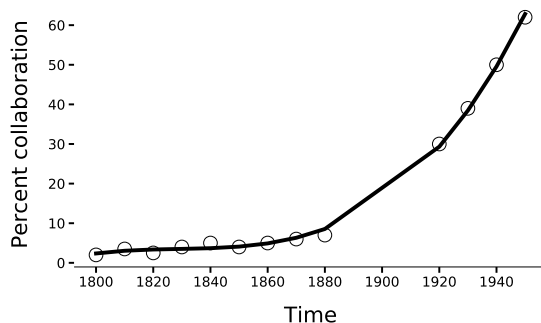


Figure 2.2. Prevalence of scientific collaborations.

Illustration on the increasing fraction of scientific collaborations over time based on the work of Beaver and Rosen [69].

in science [70, 73]. Following these results in Section 3.5, I am going to dive deeper into the effects of peers and networking on success amongst mathematicians, film directors, and pop musicians.

Kram et al. discussed the different types of network ties and the possible enhancing outcomes of both peer and mentor relationship pairs based on interviews the authors conducted [74, 75]. Later, Higgins et al. formalized these relationships in terms of the topology of the so-called developmental network structure: a network of junior individuals and their supporting senior mentors. Higgins et al. used this network construction to differentiate mentees based on two dimensions: the developmental relationship intensity (typical strength of ties to the mentors) and the developmental relationship diversity (a quantity proportional to the range within which the tie weights vary) [76]. Based on these two properties of the developmental relationship, strength and diversity, the authors proposed four protégé profiles: *i*) receptive (weak ties, low diversity); *ii*) traditional (strong ties, low diversity); *iii*) opportunistic (weak ties, high diversity); and *iv*) entrepreneurial (strong ties, high diversity). I am going to build on this knowledge in Chapter 4 and present a data-driven analysis of the benefits and limitations of mentorship via the example of electronic music.

As these few cited works illustrate, in the field of development psychology, pioneering theories and important findings had been published even decades ago [6, 77]. However, most of them were only studied on small and possibly heavily biased datasets of high achievers. Consequently, they lack general conclusions on the large populations of creative individuals. Yet, they serve as an excellent theoretical foundation and starting point for today's data-driven research, which I am going to review in the following section.

2.2 Data-driven research on success

Here I review publications that build on theoretical ideas and concepts on success and take advantage of the recent availability of large-scale data sets covering creative domains. First, I introduce ways of capturing the success of creative products, such as scientific publications and books. Then, I summarize research that goes beyond creative products and studies various aspects of individual careers represented as time series of creative products. Finally, I zoom out from individual-based research and discuss the works that cover the characteristics of entire fields or were carried out on the network aspects of success in creative domains.

2.2.1 Measuring success

Following the *The Formula: The universal laws of success* (Barabási, A.-L.) [34], I define success as a measurable response of a collective audience to the performance of the individual [37]. The notion of success varies across fields. For example, it can be expressed by the box office revenue of movies, the number of signatures of online petitions, the popularity of urban venues, the prizes of scientists, and the rankings of athletes. Despite the high diversity of fields where success can be defined, to carry out a quantitative analysis of success, and to capture universal laws and hidden patterns, we require success to be measurable. The most straightforward way of such measurements relies on pre-defined quantities, such as box office revenue and popularity. However, success can also be captured with network measures, for example, by the influence of a groundbreaking research paper or an old-time classic movie. In the next subsections, I am going to refer to these as direct and indirect success measures.

Measuring success directly

The emergence of success has been widely studied in science [23], due to the recent availability of databases such as the Web of Science [2] and Google Scholar [18]. To quantify the impact of scientific publications, the most widely used measure is the number of citations a publication receives, which naturally emerges and captures the number of future work relying on a particular paper. Citations are cumulative measures of impact, which also seem to follow the Matthew effect: the more citations a paper has, the more future citations it may receive. [78–81]. However, not all these citations represent positive endorsements but sometimes express criticism. Yet, the works of Radicchi et al. and Catalini et al. found that “there is no bad publicity” in science either, as criticized papers typically became highly cited [82, 83]. In addition, citation

counts vary broadly across different scientific fields due to the different citation traditions, which jeopardizes fair comparisons across disciplines. Therefore, the use of citation counts as a measure of success led to concerns, debates, and the emergence of alternative ways of capturing impact [84–87].

While the number of citations is a natural way of measuring academic impact, measures of different origins, such as social media and other on-line outlets, have occurred as well. They together are referred to as altmetrics [88, 88, 89]. These measures aim to connect academic research to the outer world and capture its societal impact. In some cases, these two may deviate from each other at a surprising level. For instance, Zuccala et al. [28] and Kousha et al. [90] both compared the citation-based (scholarly) and the altmetrics-based (non-scholarly) success of academic books across several fields. Afterward, they both reported low and moderate correlations, indicating that scholarly and non-scholarly impacts are considerably different.

Inspired by the examples of the previous paragraphs, I propose the following categorization of direct success measures, as I am going to build on these both in the upcoming part of the literature review and in the later chapters:

1. The origin of the measure can be based: *i*) on experts' (probably) less biased opinion, such as the Nobel-prize [91] or movie critics; or *ii*) on the opinion of a wider audience (which is more likely to be biased by external factors [92, 93]), such as the number of individuals listening to a certain song on a music providing service (as in Chapter 3-4) or the number of visitors an urban venue has (as in Chapter 5).
2. The statistical behavior of the measures can either be *i*) aggregated over time, such as average ratings of books [94]; *ii*) cumulated over time, like the number of citations from peer-reviewed articles scientific publications receive [24] (as in Chapter 3); *iii*) or have a linear distribution as derived from the previous ones in the shape of rankings [95–100], like that of the Top 100 DJs (as in Chapter 4).

As a final note on direct success measures, mostly represented by citation counts with a rather simple definition, I would like to reference a list of suggestions and warnings on avoiding the false judgment of evaluating success, which is of high importance due to its role in scientific research evaluation. These thoughts summarized in Nature News in an article titled "Bibliometrics: the Leiden Manifesto for research metrics" by Hicks et al. [101] in the form of the following ten principles listed in Table 2.1.

On the one hand, in this summary of direct success measures, I illustrated the diversity of success measures, and I pointed out a few issues, such as the rich-get-richer phenomenon. On the other, I highlighted the importance of these

- 1 Quantitative evaluation should support qualitative, expert assessment.
- 2 Measure performance against the research missions of the institution, group or researcher.
- 3 Protect excellence in locally relevant research.
- 4 Keep data collection and analytical processes open, transparent and simple.
- 5 Allow those evaluated to verify data and analysis.
- 6 Account for variation by field in publication and citation practices.
- 7 Base assessment of individual researchers on a qualitative judgment of their portfolio.
- 8 Avoid misplaced concreteness and false precision.
- 9 Recognize the systemic effects of assessment and indicators.
- 10 Scrutinize indicators regularly and update them.

Table 2.1. Warnings on research evaluation.

Ten suggestions on fair research evaluation in the work "Bibliometrics: the Leiden Manifesto for research metrics" by Hicks et al. [101]

measures for instance in individuals' performance evaluation, policy-making, and funding distribution, which not surprisingly led to the rapid growth of the field Scientometrics [102–107]. For the sake of simplicity, during the rest of this thesis, if not mentioned otherwise, I am going to refer to cumulative measures as impact.

Measuring success indirectly

While the previously introduced, more standardized methods directly define specific dimensions of success, indirect methods have been proposed as well. A widely studied group of these indirect measures are building on the tools of network science: define relations between creative products, such as citations of papers, or references of feature movies, and associate success with network centrality measures [108–116].

These relation-networks of creative products can be defined in different ways. Typically, the nodes of the networks are the individual scientific papers or other creative products, and the links between them represent similarities (e.g., topical overlaps or the number of shared collaborators) or causality (e.g., references or citations). In the latter case, the citations are not simply used to quantify the magnitude of success, but also the influence of the individual papers within their field as their centrality may emerge via the other connecting articles. This centrality can be measured in various ways, for instance, by modified versions of the PageRank [117] algorithm, such as DivRank [118], CiteRank [119], and PrestigeRank [120]. The advantages of these algorithms have been demonstrated: they allowed researchers to identify so-called "scientific gems" in the literature as well as groundbreaking patents that are highly influential yet poorly cited [26, 121, 122]. In addition, network approaches have been used in the world of cinema, by relying on cinematic citations between

more than 40,000 international feature films to identify the leading creations in the movie industry [123].

2.2.2 Quantitative results on scientific impact

The most common way of measuring the impact of scientific papers as of today is by the number of citations publications receive. However, there is not even explicit agreement on one of the most simple properties of impact, the shape of the distribution function of citations counts. Previous research, dating back to the middle of the '90s, reported contradicting findings, namely that the citation distribution follows log-normal, power-law, and even exponential distributions depending on the type of data and selection criteria used [124–128].

A frequently mentioned issue with raw citation count as a success metric is its high diversity across different disciplines due to different citation traditions of various fields. To account for these differences, Radicchi and his colleagues proposed an impact-transformation method in 2008, assuming log-normal citation distributions, and tested its statistical validity on the Web of Science and the American Physical Society data bases [2, 127, 129]. Their method rescales the citation distribution of different fields by dividing citations by the average number of citations on a particular field, such as mathematics, biology, and various branches of physics.

Another major criticism against the use of citation counts as the ultimate measure of success and research evaluation is its shortcomings on temporal differences. The reason is that the number of citations increases by a steady $\sim 4\%$ every year, which means that the total number of citations doubles every 12 years [29], possibly due to the rich-get richer-phenomenon. Building on the Web of Science database, Petersen et al. have published a correction method to account for this inflation and offer a way to factor out the first-mover advantage [2, 29, 130].

Connecting to the temporal behavior of citation counts, a major finding has been published by Wang et al. [27] modeling the long-term dynamics of scientific impact. They have proposed a mechanistic model, including *i*) preferential attachment [131] that models highly cited papers being more central; *ii*) aging effects capturing that scientific papers are cited subsequently; and *iii*) a fitness parameter that represents the intrinsic differences between papers in terms of their novelty and scientific importance. Based on these components, they construct, formulate, and solve a Master-equation. Then they find that the number of citations c a paper i receives by time t after its publication can be expressed

as:

$$c_i^t = m \left[e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1 \right], \quad \text{where} \quad (2.2)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy. \quad (2.3)$$

In Eq. (2.2) m shows the average number of references a new paper has, λ_i captures the relative importance of a paper, μ_i measures the immediacy (typical time needed for a paper to reach its citation peak), and σ_i captures the decay rate of the number of new citations over time. They tested the validity of this model on both papers published during 1950-1980 in the Physical Review corpus and those published in 1990 by 12 specific prominent journals, such as Science and Cell [27]. In addition, Yucesoy et al. have explicitly shown that Wang et al. 's model also holds for describing the success of books, captured by the number of copies sold over time [132].

The majority of the studied publications follow the dynamics of Eq. (2.2) and reach their peak in a few years after their publication, followed by a stable decay in their yearly citation count. However, a smaller fraction of research papers shows an entirely different behavior with a heavily delayed recognition. A classic example of these so-called sleeping beauties was the article that later became the starting point of modern genetics by Gregor Mendel published in 1866 [133]. In this case, it took more than three decades for the scientific community to recognize its importance [134].

Later on, many of the sleeping beauties turned out to be of major importance; therefore, researchers became interested in identifying them [135], for instance, by estimating the likelihood of a paper becoming a sleeping beauty. Such methods typically build on the temporal analysis of thousands of papers' citation evolution by quantifying attributes such as the awakening time of papers and the jump in impact right after their awakening happens [136, 137].

The temporal evolution of impact, even expressed by cumulative measures such as citation count, has been studied outside of science as well. For instance, Rossetti et al. have analyzed the temporal dynamics of the product sales at the retail company Coop, and the play count of songs on the music providing service LastFM [138, 139]. By clustering the popularity time series of these items, they reported the existence of two main clusters: early adopters, and late adopters with a shifted peak. These two reported shapes seem to align well with the scientific impact evolution proposed by Wang et al. [27] and to the observed dynamics of the sleeping beauties. I am going to build on these models, approaches, and ideas when studying the statistics of success measures of creative fields in Chapter 3, and when evaluating the temporal trends of the popularity of urban venues in Chapter 5.

2.2.3 Measuring scientific fields

During the history and evolution of science, different scientific fields emerged. While some fields are easy to distinguish, such as mathematics and political science, others are much closer to each other, like biology and medicine. These commonalities could imply potential mutual influence, spreading of results and knowledge, and collaborations, leading to the prosperity of innovation. Therefore not surprisingly, researchers asked: is it possible to capture the differences between scientific fields in a quantitative way [140, 141]? Is it possible to determine the clear boundaries of different disciplines?

One possible start of answering these questions is the analysis of the network structure of scientific papers, e.g., based on their citations or references. These networks have been studied since the 1960s [142]. However, one of the first large-scale analysis building on 1M articles aimed at extracting the backbone structure of science was only published in 2005 [143]. Later, Porter and Rafols defined a measure called integration score [144] and used this measure on the Web of Science database to estimate the level of interdisciplinarity. They reported the values of the integration score and its changes during 1975-2005 for six scientific fields. Further investigations of these networks, extended by analysis on paper titles, abstracts, and the Physics and Astronomy Classification Scheme (PACS) of the American Physical Society [2, 145], resulted in a deeper understanding of the structure and evolution of science and scientific disciplines. This includes the birth and decay dynamics of scientific fields, the identification of the boundaries of physical sciences, and trending research directions in physics [31, 33, 146, 147].

Interdisciplinary research indeed allows scientists to explore and learn about previously untouched areas in the hope of new findings and applications. However, are there any measurable advantages of combining different branches of science? Uzzi et al. have shown that references can proxy exceptional success: they have reported that atypical combinations of previous knowledge and novel ideas can boost the success of scientific publications, traced down by the disciplinarity distribution of the publications' references [148]. Later, Mukherjee et al. studied millions of papers and patents, and measured the age difference between the publication time of the papers and patents and the works they are referring [149]. They found that papers that cite previous works with a low average age and high variance in age double their chance of arriving into the 5% most cited articles. This finding seems to be universal across all studied branches of science and technology. Moreover, they reported that multi-authored papers are much more likely to reach this hotspot than single-authored articles.

Not only has the relationship between success and the combination of re-

search interests been studied so-far, but also the topical interests and fields of activities of the individuals. In an article published in 2017, Jia et al. have studied the evolution of the research interest of physicists by characterizing their research profiles using the well-structured PACS codes [40, 145]. For each scientist, they generated a topic (PACS) vector g with elements counting the number of times a certain topic occurred over the career of the individual, and slit it into the first m papers and the rest, characterized by g_i and g_f . From this, they defined the research interest change as:

$$J = 1 - \frac{g_i \cdot g_f}{||g_i|| ||g_f||}, \quad (2.4)$$

and measured its value over the population of $\sim 15,000$ scientists with at least 16 publications (reportedly, the exact choice of m does not influence the trends found). Moreover, they proposed a random-walk based mechanistic model that reproduced these measured values well.

These works show interesting examples and practical tools for capturing topical changes, which I am going to take advantage of when studying the genre-differences of communities of electronic music artists in Section 4.4.

2.2.4 Capturing individual career success

In the previous subsection, I reviewed the different approaches of describing and mapping the various branches of science by relying on large-scale data and the aggregated properties of tens of thousands of individuals. Here I aim to zoom in from this meta-level perspective into the micro-level and review the literature on the evolution of individual careers with a particular focus on science. My motivation to do this is that both Chapter 3 and Chapter 4 of this thesis is focusing on creative careers, while Chapter 5 tackles certain aspects of the temporal success evolution of urban venues. In the following, I am considering careers to be time series, for instance, as the sequence of publications each attached with various features such as impact and affiliation.

Derville et al. [38] have investigated the migration-patterns of physicists by analyzing 420,000 scientific papers, 237,000 individuals, and roughly 4,000 institutions. They reconstructed the researchers' careers defined as their affiliation sequence and used their relocation traces to build their migration network. In this network, the centralities of institutions naturally define a ranking between them, which they used to uncover that moving from an elite to a lower-rank place on average results in a modest decrease in performance while moving from a lower-ranked place to an elite institutions does not result in significant performance improvements. This finding suggests that already well-performing scientists do not gain much from institutional prestige; however,

research accommodated to the circumstances of an elite place suffers from a downgrading move.

An early way of measuring the success of individuals is based on their productivity (e.g., the number of papers scientists they published, N) [150]. However, as the work of Simonton and others uncovered, quality and quantity are not the same, just correlated. Hence productivity is not always a good proxy of quality. The most established methods on measuring individual success are rather based on the number of citations of the individuals' papers received, which, in practical terms, is similar to the number of ratings books received or the number of times people vote for a movie [151]. Multiple measures have been proposed derived from the researchers' citation history, just as the total number of citations or the total number of citations within a specific time window. Yet, these measures often suffer from similar shortcomings such that citation counts themselves, for instance, the rich-get-richer phenomenon. Another individual measure, the h -index [25], defined as the highest number h of a scientists' papers having at least h citations, was suggested to correct for some of these issues and is today widely used. Besides, a review article by Wildgaard et al. collected 108 different measures in 2014, including the h -index, to illustrate the diversity and complexity of measuring the scientific impact of individuals [152]).

One of the latest measures meant to capture individuals' excellence is the so-called Q parameter introduced by Sinatra et al. [43] based on a quantitative modeling approach. First, they proved for scientific careers at a large scale that the "equal-odds-rule" or "random-impact-rule" first proposed by Simonton [6] holds. The "equal-odds-rule" states that the probability of each paper a scientist publishes has the same probability of being the most successful one. To quantify this, they measured scientific impact as the number of citations a publication receives during the first ten years after its publication. Then, building on these findings, they proposed an impact decomposition method called the Q -model, that decouples the success $S_{i,\alpha}$ of paper α by author i into two components:

$$S_{i,\alpha} = p_\alpha Q_i, \quad (2.5)$$

where p_α corresponds to all external random factors (luck), while Q_i encodes the typical success of the individuals. According to the authors' findings [43], Q_i can be approximated in the following way:

$$Q_i = e^{\langle \log S_{i,\alpha} \rangle - \mu_p}, \quad (2.6)$$

where μ_p is the average of all the measured p_α values of a certain field. This method highlights that, on the one hand, it is possible to untangle the role of luck in science. On the other, it proposes a way to measure individual success

(Q) that is not sensitive to the inflation and accumulation of citations but captures the individual's intrinsic ability to generate work with impact of a certain level. It also turns out that the Q parameter has a surprisingly high predictive power on Nobel-prizes compared to previous measures and predictive models [153–155]. In Chapter 3, particularly in Section 3.3-3.4.2 of this dissertation, I build on this model and expand the review on the work of Sinatra et al. in more details.

While the random-impact-rule states that the biggest hit occurs at random, Liu et al., following the work of Sinatra et al., reported that the largest few hits, also known as hot streaks, occur together, showing a high degree of temporal correlation over a relatively short fraction of the entire career, yet accounting for the main body of total impact the individual has gathered [44]. To explain their findings, they also proposed a probabilistic model for career evolution, the "hot hand model", which shows good agreements with the observed hot streak patterns on large-scale career data of scientists, movie directors, and artists.

The Q -model captures both the driving and the luck component of the dynamics of career success. Connected to these thoughts, later work on career evolution also attempted to capture the shapes of age-curves by clustering them (e.g., late and early bloomers) using time series analysis tools [156, 157]; while the role and mechanism of luck on success by Gaussian agent-based models had been studied by Biondo et al. [158]. Petersen et al. proposed a stochastic model on career longevity that also reproduces the observed rich-get-richer effect [79]. Jones et al. also studied the shape of age-curves in terms of the timing of the great achievement over the career of Nobel laureates [159]. They have reported that the chronological time has more effect than the scientific discipline, underlined both an interdisciplinary difference and a temporal trend on how science as a whole evolves.

Taken together, several interesting features, such as geographical movements and the timing of the best hit have been studied for scientific careers. Moreover, mechanistic models have also been proposed to describe certain characteristics of these careers. For the work described in Chapter 3 I rely on these findings and further extend them.

2.2.5 Networks from the individuals' perspective

In the earlier subsections, I introduced the individual-based aspects of creative career success; however, these creative products are usually the results of collaborative efforts. This trend of collaborations is reflected in the dominance of multi-authored scientific papers above single-authored ones [69, 70, 160], and in the positive effects of larger team size on the expected impact [70, 73, 161], implying that network effects should be considered as well.

On the one hand, like earlier works, such as Higgins et al. [76] showed, the network roles, and consequently, contributions to a project can highly differ from junior to senior authors of scientific publications. On the other hand, the citation numbers, the basic units of academic success, are equally distributed across the co-authors of the papers. This equal distribution may have been aligned to the early trend of alphabetic author-orders [162]; however, the author-order of publications typically implies the level of involvement of the contributors during recent times. The lack of clarity on the level of contributions by the co-authors motivated several research projects to estimate the real credit of each author and try to approximate the actual share of the individuals [97, 163, 164]. Shen et al. [41] proposed an algorithm to compute the contributed share of each author to a given paper. Their method is based on a network-spreading assumption where they study which authors keep publishing on similar topics the most. They validated their result by using Nobel-prize winning papers, where they managed to identify the actual laureates with an accuracy nearing 90%.

This example on co-authorships already hints the importance of how networks can reveal far richer information about creative success than direct measures. In particular, Petersen studied 166,000 collaboration records and defined so-called super ties corresponding to the large (at least 50%) overlap in the publication history of the collaborators [51] – a type of collaboration similar to what Kram. et al. referred to as peers [74, 75]. Petersen's regression analysis revealed that these super ties come with above-average productivity and an increase of 17% in total citations showing clear quantitative proof of the advantages of life-long partnerships. These findings demonstrate the clear advantage of having strong ties, while classical social network theory argues for the "strength of weak ties [165]. The work of Pan and Saramäki further elaborates on the "strength of strong ties" in scientific collaborations [166]. They reported that scientific collaborations behave differently from regular social networks [165] by the strong ties being the information bottlenecks. Their reasoning is rooted in the mechanisms shaping the scientific social graph's structure, such as the underlying space of knowledge and ideas and the typical ways of sharing and processing information (such as group work). In addition, the famous friendship paradox [167] had been confirmed, moreover, generalized between co-authors [167, 168], stating that "your friends have on average higher characteristics than you have", where characteristics include the number of co-authors, the number of citations, and the number of publications.

Such mechanisms, for example, team formation, have also been studied in more detail to better understand how the need and desire to extend knowledge and increase team prosperity may work. For instance, Milojević showed

that the typical team size over time could be well-modeled by a Poissonian process [169]. Guimera and his colleagues [56] proposed a different probabilistic model on team formation with three simple parameters: the size of the teams, the fraction of new team members in new products, and the likelihood of incumbents to repeat earlier collaborations. By testing different values of these parameters, they observed two distinct phases of the global team structure: it either consists of multiple small components, or a giant component emerges, which also aligns well with the proposed phenomenon of invisible colleges [170]. The authors tested their model against real data covering the Broadway musical industry and several scientific disciplines, like social psychology and astronomy, and confirmed their findings.

Based on the analysis of 150 000 self-organized, online team projects Klug et al. reported that the most successful teams' members had the most diverse backgrounds and experience [171]. Similarly, Bonaventura et al. investigated the role of diversity, however, by studying the advantages of interdisciplinary collaborations in science, where they defined diversity as the entropy of the published articles' PACS codes [32]. On the one hand, they observed that the most successful scientists are either highly specialized (low diversity) or highly interdisciplinary (high diversity). On the other hand, they found that having a more heterogeneous collaboration network tends to be significantly more successful on average.

As work on scientific credit shares has shown, not all team players have the same role and importance. In fact, having good mentors can offer various advantages to the proteges' future success, such as having Nobel laureate mentors increases the chance of becoming a laureate in the future [172–174]. One of the first data-driven studies on mentorship focused on mathematics [175]. In that article, the authors studied the number of mentees each researcher had, discussed the effect of having prolific mentors on the future mentoring patterns of the used-to-be mentees, and found that the number of mentees correlates with the mentors' academic success [175]. Advantageous mentorship, also known as the Chaperone-effect, exists in other fields of science as well, according to Sekara. al. [176]. Their findings reveal that publications by chaperoned authors have a significantly higher impact than those who did not have such mentors. Yet, the effect varies across scientific disciplines being the highest in interdisciplinary research, and the lowest in mathematics.

Good mentors and prominent central authors not only affect the future success of mentees but show limited advantages for citation rates as well, as it has been shown on the careers of 450 high-profile researchers authoring 83,693 articles combined [177]. More specifically, it turned out that the reputation of the central author, measured as the total number of citations he/she ever received,

seems to boost citation rate until a citation threshold $c_x \approx 40$; however, this effect disappears for highly cited papers.

In this subsection, I reviewed several major findings on the individuals and their social surroundings to highlight what has already been discovered on the relationship between networks and success. I plan to further extend this knowledge by studying the dynamical aspects and the temporal co-evolution of success and network position for film directors, mathematicians, and pop musicians as a main component of Section 3.5 in Chapter 3, and focus on the particular networking behavior of mentoring in electronic music in Section 4.5 of Chapter 4. In the next subsection, I zoom out from the micro-level network scope and discuss global network properties and their relation to success.

2.2.6 Networks of creative fields

In this subsection, I review the network aspects from another angle: how does the embeddedness of the individual into the social fabric of its community relate to its success, and what does global network information tell on success?

One of the first large-scale analyses of collaboration networks covering physics, biomedical research, and computer science, was done by M. E. J. Newman in 2001. In his work, Newman defined scientific collaboration networks as social networks in which each node is a scientist, and the weight of the undirected link between each pair of them is the number of publications they co-authored [178–180]. His descriptive results measure basic local statistics, such as the nodes' degree distribution and the number of authors per paper, and observe typically power-laws describing these distributions. On the global scale, he observes a small-world phenomenon, meaning that the distance (number of intermittent edges) between each pair of authors is low and scales logarithmically by the number of nodes in the networks, which also served as a basis of an efficient way of computing betweenness centrality [181, 182]. His pioneering work served as a starting point of collaboration analysis on various fields, such as econophysics, finance, political science, and scientometrics [183–186].

Significant work has been done about quantifying success in the realm of art and culture. Gleiser and Danon studied the collaborations and community structure of jazz musicians and reported correlations between the community structure, locations, and even racial segregation [187, 188]. Park et al. studied the properties of both the collaboration and the similarity network of 32,377 contemporary pop musicians collected from a platform called Allmusic [189, 190]. They found only $\sim 15 - 50\%$ of overlap in edges between the two networks. Then, they compared the basic measures, such as degree- and betweenness distribution, between the social- and the similarity networks and conclude that the collaboration network gives more objective insights into the studied social sys-

tem. The authors later investigated the topology and evolution of western classical musicians as well, and reported a superlinear preferential attachment-like growing process [131, 191]. By processing and analyzing data of 5M releases by 3.5M artists present on Discogs, it also has been shown that the collaboration structure and the emerging communities in music can be related to musical genres and used as a basis of classification [192].

Uzzi and Spiro studied the small-worldness of Broadway musicals during 1945-1989, where they found a curvilinear relationship between the global clustering coefficient and both the artistic and financial success of the musicals. They reported that for low and high clustering values success of the industry stays low; however, moderate clustering is associated with prosperity [57, 193]. While clustering is based on triangular topology, Krumov reported that in the field of computer science, box motifs are a better proxy of success since they are associated with the highest mean citation per collaboration [194, 195]. Based on his findings in jazz music, Vedres argues that forbidden triads are the key to the emergence of such prominent artists like Miles Davis [53, 196]. At the same time, Budner and Grahl point out the importance of building a large number of bridges in the collaboration network of the artists who ever entered the Rolling Stone Magazine's list of "500 Greatest Albums of All Time" and the "1001 Albums You Must Hear Before You Die" lists [197].

Besides music, collaborations in the film industry have been targets of researchers as well. For instance, Auber et al. in 2003 showed that the small-world property holds for a smaller subset of actors being linked based on movies they costarred [198]. Kitsak et al. proposed a method based on the k-core [199] decomposition to capture the influence of the nodes during the spreading process, such as the ones described by the SIS and SIR epidemics models, in actors costarring network [200–202]. Later research also went on to study the core-periphery structure defined by the k-cores to infer success and performance in film. Cattani and Ferriani found that those with an intermediate position between the center and the periphery of the collaboration network are the most beneficial in terms of creative production in the Hollywood film industry [203]. To support their findings, they presented a regression-based model and used data covering 2137 high-profile movies from 1992-2003. Similar advantages of brokering between the core and the periphery have been shown in the Hungarian Film Industry by Juhász et al. [204].

Finally, I summarize several pieces done on predictive modeling of success that include network features. Previous research not only focused on the structure and description of collaboration networks and their relation to success but also presented time series and correlation analysis, regression models, and predictive modeling on the evolution of success based on the authors' network po-

sitions, captured by various centrality measures [205–209]. For example, Sarigöl et al. studied scientific collaborations via more than 100,000 publications from the field of computer science and discussed the link between network centrality and success measured by citation counts[50]. They have built machine learning models to predict whether a publication is going to be highly cited (fall into the top 10% most cited papers) five years after the publication. For this prediction, they relied on the degree centrality k-core centrality, eigenvector centrality, and the betweenness centrality of the authors in the co-publishing network. Jadidi et al. confirmed the predictive power of network features, such as closure and brokerage, to success as well. They measured success both by the *h*-index and the raw citations counts [210]. Jadidi et al. also tested for gender differences in success and did not report any significant differences between the relation of collaboration patterns to impact between different genders.

Wachs and co-authors discussed the role of the network positions of graphic designers in novelty and success in an online social network [54]. They found that designers producing novel work while having a highly constrained network are much more likely to be successful than those with an open, less cohesive network. Fraiberger et al. [211] quantified reputation and success in art by analyzing the exhibition history of half a million artists worldwide and reconstructed the directed co-exhibition network of institutions where these exhibitions took place. They show that network correlates to prestige, a measure determined by experts; and that early-career exhibitions at high-prestige institutions predict stable, longstanding success while starting at low-prestige institutions leads to high dropout rates and limited chances of success.

In the following chapters of this dissertation, I am going to build on the findings mentioned above first in Chapter 3, where I extract the temporal dynamics of the individuals' network positions based on field level collaborations. Next, in Chapter 4 I am going to analyze the evolution of electronic music via the co-release patterns of top DJs, while in Chapter 5 I relate the popularity of urban venues to the underlying social network of their visitors.

CHAPTER 3

SUCCESS AND LUCK IN CREATIVE CAREERS

3.1 Introduction

Luck is considered to be one of the crucial ingredients to achieve success in life, yet only a little research has tried to understand its quantitative nature so far [47, 48, 213]. For instance, in science, the movie industry, music, and art, the occurrence of the highest impact work and a hot streak within a creative career are challenging to predict. However, fundamental questions are still unanswered, such as: are there domains that are more exposed to luck than others? Here I aim to tackle this question and provide new insights on the role of randomness in impact in creative careers in three ways. First, I systematically untangle luck and individual ability to generate impact in the movie, music, and book industries, and in science. Second, I compare the luck factor between these fields and describe the different domains' characteristics. Finally, I show the role of randomness in the relationship between collaboration network positions and career hits and highlight the surprising presence of randomness in networking. Taken together, the analysis presented in this chapter suggests that luck consistently affects career impact across all considered sectors and improves our understanding in pinpointing the key elements in driving success.

This chapter is based on the article "Success and luck in creative careers" [212], where I analyzed the data and obtained all numerical analyses, Roberta Sinatra conceived the study, and the authors, Milan Janosov, Federico Battiston, and Roberta Sinatra collaboratively drafted, revised, and edited the manuscript. At the time of the submission of this dissertation, this publication was under the second round of revision at EPJ Data Science, while by the time of the thesis defense it was accepted for publication in EPJ Data Science.

As reviewed earlier in Section 2.1, research in developmental psychology has studied careers of prominent artists and scientists for decades, advocating the importance of chance for the successful unfolding of careers in various creative professions [9–12]. In recent years, the availability of big databases on scientific publications [2] and artistic records, from books to movies [36, 123, 132], has made it possible to test several previously proposed hypotheses on a large scale, as the short review in Section 2.2 also illustrates. For instance, previous work [43, 44] on the analysis of thousands of creative careers have shown that the biggest hit of an individual occurs randomly within one’s career, a finding earlier published as the equal-odds-rule [11]. This rule explains the variability in the occurrence of creative individuals’ best hits. Yet, career hits are not only the results of luck but also of other individual and team properties [23, 27, 56, 148, 210, 214, 215]. While previous literature suggests that luck and individual ability are both necessary to excel in art and science [79, 213, 216–218], the quantification of the role of luck across different creative domains is still lacking. In which creative fields are individuals more likely to go from rags to riches and vice-versa? Does the position of an individual in a network predict the occurrence and the timing of a hit, or does the hit forecast future centrality?

In this chapter, I propose a framework combining earlier theories to quantify and compare luck fluctuations in impact across creative careers from movies, music, literature, and science [127, 132]. Do these random fluctuations have the same magnitude across careers? To answer this, I build on the mathematical framework known as the *Q*-model proposed in Ref [43] and outlined in Subsection 2.2.4 to decompose the impact into two components, one representing external random fluctuations that can be interpreted as luck, and another depending only on the individuals’ success history. I show that this model is consistent with the classical test theory [219], also known as the true score theory [220], stating that the measured value of a particular observable attribute consists of the sum of its true – error-free – score, and a stochastic error term. I designed a specific index to capture *randomness* and found that the value of such randomness slightly varies depending on the creative fields. This implies that randomness plays a universal role in achieving success across the studied creative fields.

Such low heterogeneity in the role of randomness also aligns with my findings on the temporal relationship between networking and impact. Namely, I found that two networking behaviors co-exist: for some individuals success peaks first, and the central network role follows, while for others, it happens the other way around. However, illustrating the role of randomness, it occurs by chance which individual falls into which category. To carry out these analyses, I rely on a large-scale data set covering more than four million individuals

from the beginning of the 20th century up until 2017.

This chapter is organized as follows. First, I introduce the data sets I am using, including the possible ways of measuring success. Second, I review the modeling approach Q-model and test the validity of its requirements. Third, I connect the Q-model impact decomposition method with the classical test theory to quantify the role of luck within each field and discuss the observed differences across fields. Finally, I elaborate on the role of randomness, also affecting the collaboration patterns of the individuals via comparing the temporal evolution of their network position to their impact evolution.

3.2 Data description

3.2.1 Data collection

I collected information and built a database by using four data sources on the film, music, and book industries, and across scientific fields, covering 28 different types of creative professions total during the period of June-August 2017.

1. First, I scraped and processed individual profiles from the Internet Movie Database (IMDb [14]) and compiled a data set of 803,013 people working in the movie industry as movie directors, producers, art directors, soundtrack composers, and scriptwriters, altogether contributing to 1,297,275 movies. The IMDb offers different metrics, such as average rating, rating count, metascore [221], gross revenue, the number of user and critic reviews to evaluate the success of movies, which I collected, and attached to the individuals based on their profiles. During this process, I used the website's Advanced Title Search functionality¹ following the IMDb's policies to list all the identifiers of the relevant movies with at least five user ratings.
2. Using Discogs [15, 19] I collected the discographies of musicians from nine genres: electronic, rock, pop, funk, folk, jazz, hip-hop, and classical music. While Discogs provides good coverage on music discographies, it lacks success metrics. Therefore I extended my dataset with song play counts by matching song titles extracted from Discogs to the ones available on LastFM [138] via its public API. This way, I constructed a database of 379,366 musicians who released 31,841,981 songs.
3. I crawled the biography profiles of authors from Goodreads [222] to cover the book industry, where the success of books is measured by the average

¹<http://www.imdb.com/search/title>

user rating of a book, the total number of ratings, and the number of editions a book has, from which I used the rating count for further analysis. The book-dataset I compiled contains information about 2,069,891 authors and 6,604,144 books total.

4. I used Thomson Reuters' Web of Science database [2] to reconstruct the career trajectories of 1,204,688 scientists from the fields of chemistry, mathematics, physics, applied physics, space science and astronomy, zoology, geology, agronomy, engineering, theoretical computer science, biology, environmental science, political science, and health science, altogether authoring approximately 87.4 million papers. In this database, the success of creative products, scientific papers, is measured by the number of citations they received during the first ten years after their publication (to align to previous work on quantifying success [43]).

3.2.2 Measuring success

While the data sources provided several possibilities to measure success on different fields, I decided to measure the impact of creative products (movies, songs, books, and articles) by using the so-called cumulative measures introduced in Subsection 2.2.1 (e.g., citation count) and disregarded other measures (e.g., average rating of movies). The exact measures I picked are the following: the rating counts for movies and books; the play counts for songs; and the number of citations received within the first ten years after publication for scientific papers [24]. The reason behind this choice was that all these measures show similar statistical properties with the more widely studied citation counts (Subsection 2.2.2), which made further modeling and the adaptation of existing tools easier. Such a common property shared across the studied impact distributions is that these measures can be well-fitted by log-normal functions, which I show in Figure 3.1 for four example professions (film directors, pop musicians, book authors, and mathematicians). These fits, evaluated by R^2 values, showed that the log-normal function indeed is a good model to describe the impact distributions in my datasets (for the R^2 values see Table 3.1).

To carry out comparative measurements across creative professionals, I had to ensure that the impact values vary within the same range for the studied fields. For this reason I applied a linear min-max rescaling method similar to ones previously used [129] to transform the success measure distributions in the different fields (S^f):

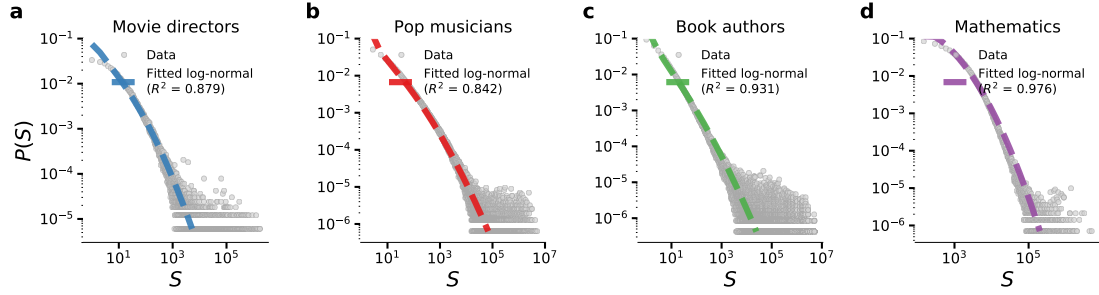


Figure 3.1. Impact distributions modeled by log-normal functions for four example fields.

Impact (S) distributions measured as the rating counts for movies and books, the play counts for songs, and the number of citations of mathematicians. The grey dots represent the data points, while the dashed colored lines show the binned trend (10 bins, percentile binning).

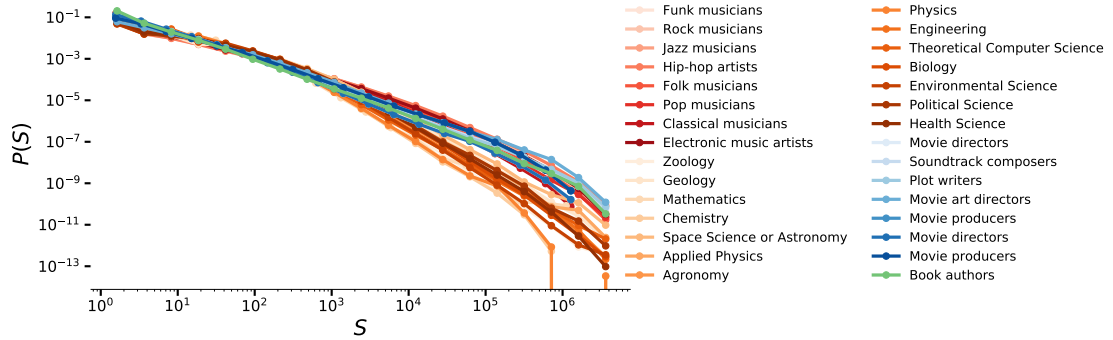


Figure 3.2. Rescaled impact distributions.

Rescaled impact distributions based on Eq. (3.1) for all the 28 studied fields.

$$S^f \rightarrow \frac{S^f - r_{\min}^f}{r_{\max}^f - r_{\min}^f} \cdot (t_{\max} - t_{\min}) + t_{\min}, \quad \text{where} \quad (3.1)$$

$$r_{\min}^f = \min S^f, \quad (3.2)$$

$$r_{\max}^f = \max S^f, \quad (3.3)$$

$$t_{\min} = \min \left(\bigcup_g S^g \right), \quad (3.4)$$

$$t_{\max} = \max \left(\bigcup_g S^g \right). \quad (3.5)$$

The results of the rescaling are shown on Figure 3.2.

Next, I associated the transformed measures of each creative product to their creators to reconstruct individual careers as time series consistently across professions. As practical examples, in Figure 3.3 I illustrate careers in the four different datasets: movie director Stanley Kubrick, pop singer Michael Jackson,

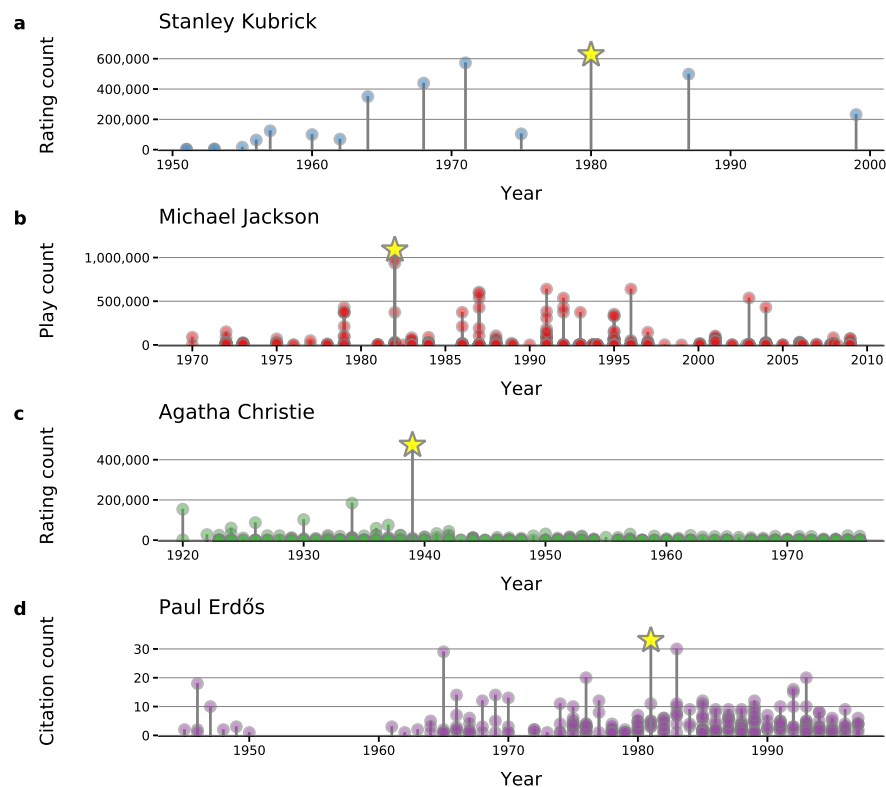


Figure 3.3. Career examples for four example professions.

a, Stanley Kubrick as a film director, where the horizontal axis shows the release year of his movies, and the vertical axis shows the impact of each movie, captured by the number of ratings they received. **b**, The career of Michael Jackson represented by the series of his releases and their songs' impacts captured by the total play counts on the music providing service LastFM. **c**, The career of Agatha Christie shows her publication dates and the their impact as the number of ratings they received on Goodreads. **d**, Publication history of the mathematician Paul Erdős based on his record in the Web of Science database. The papers' impact is measured by the total number of citations 10 years after publication.

writer Agatha Christie, and mathematician Paul Erdős. As the statistical descriptions and modeling during the upcoming sections are going to illustrate, this time series representation of careers offers a practical framework to conduct quantitative analysis on the careers, such as capturing the role of luck or comparing the co-evolution of impact and network position.

3.3 The Q-model: decomposing impact through luck and individual ability

In this section, I introduce an impact-decomposition method called the Q-model outlined in Subsection 2.2.4, developed to model scientific careers. I describe the basic assumptions the model builds on and test its requirements on all the data sets I am studying. Then I illustrate its predictive power in terms of the biggest hit of the individuals. Next, I integrate the Q-model with the true score theory. Finally, I use this combined framework to capture the role of randomness in creative success.

3.3.1 The random-impact-rule

As the examples in Figure 3.3 show, Kubrick's highest impact movie was one of his last ones, while Michael Jackson had his biggest hit at a relatively early stage. This suggests that a career's biggest hit may occur at any time, agreeing with former findings in the psychology literature [10–12]. Indeed, a rigorous analysis of my data revealed that any work in a career has an equal chance to be the highest impact work, following the *random-impact-rule*, consistently with what others previously found for large data sets of artists and scientists [43, 44]. I tested the validity of this rule in the following way: first, I denoted the productivity of the individual by N , while the best product has a chronological rank of N^* . In case the random-impact-rule holds for a certain set of individuals ($N - N^*$ pairs), $P(N^*/N)$ is well-approximated by a uniform $U(0, 1)$ distribution. Then I compared the measured $P(N^*/N)$ to the randomized case where I reshuffled careers randomly, conducted the same measurement on $P(N^*/N)$, repeated this randomized measurement a hundred times, and took the average distributions of these randomizations. In addition, to reduce noise, I tested the cumulative distribution function (CDF) of these distributions instead of the originals. Finally, I evaluated the random-impact-rule by measuring R^2 values between the data and the randomized case (Table 3.1).

For this measurement, I had to limit the analysis to careers with sustained productivity, which means the introduction of a filtering threshold based on the

Field	R^2_{random}	R^2_S	R^2_N	R^2_p	R^2_Q
Agronomy	0.99955	0.99955	0.988	0.9808	0.9419
Applied Physics	0.99979	0.99979	0.982	0.9572	0.9654
Biology	0.99967	0.99967	0.984	0.9884	0.9822
Book authors	0.97773	0.97773	0.894	0.9796	0.9796
Chemistry	0.99963	0.99963	0.989	0.9945	0.996
Classical musicians	0.96694	0.96694	0.963	0.9444	0.9933
Electronic music artists	0.99553	0.99553	0.947	0.9857	0.9813
Engineering	0.99973	0.99973	0.986	0.9776	0.9793
Environmental science	0.99969	0.99969	0.968	0.9932	0.9933
Folk musicians	0.96508	0.96508	0.923	0.981	0.973
Funk musicians	0.95236	0.95236	0.934	0.9916	0.988
Geology	0.99967	0.99967	0.982	0.9885	0.9824
Health Science	0.99962	0.99962	0.991	0.9689	0.9811
Hip-hop artists	0.94326	0.95512	0.882	0.9915	0.9875
Jazz musicians	0.96488	0.96488	0.869	0.989	0.9844
Mathematics	0.99969	0.99969	0.983	0.9856	0.9778
Movie art directors	0.97207	0.97207	0.922	0.9989	0.9994
Movie directors	0.9982	0.9982	0.96	0.9914	0.9875
Movie producer	0.99941	0.99941	0.916	0.9817	0.974
Physics	0.99972	0.99972	0.989	0.9961	0.9972
Political Science	0.99973	0.99973	0.984	0.9913	0.9844
Pop musicians	0.99407	0.99407	0.903	0.9742	0.9844
Rock musicians	0.95565	0.95565	0.948	0.9934	0.9952
Script writers	0.99957	0.99957	0.877	0.9841	0.9751
Soundtrack composers	0.99923	0.99923	0.974	0.9801	0.9822
Space Science or Astronomy	0.99959	0.99959	0.965	0.9928	0.9922
Theoretical Computer Science	0.99954	0.99954	0.982	0.9918	0.9886
Zoology	0.9996	0.9996	0.976	0.9807	0.9206

Table 3.1. Results of model fittings supporting the Q -model.

The table shows the goodness of the fit for the random-impact-rule against the randomized null model, and the goodness of the fit of the log-normal model to the impact (S), productivity (N), and the p and Q distributions.

number of creative products each individual has. For individuals in the film industry and science, I set limits to 10 movies and papers (except art-directors, for whom it was 20), while in literature, it was 50 books per author, and in music, 80 songs per artist. The latter, relatively high threshold for musicians, is due to the fact releases usually containing multiple songs.

3.3.2 Introducing the Q-model

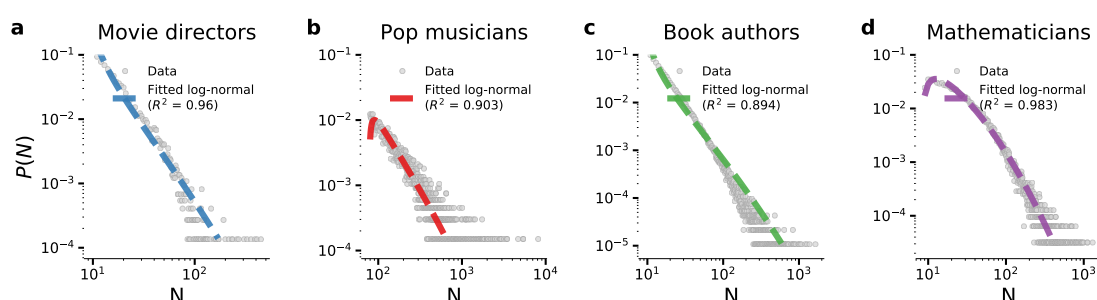


Figure 3.4. Productivity distributions.

The distribution of the individuals' productivity (N) for film directors, pop musicians, book authors, and mathematicians. On the figure the grey dots represent the data points while the dashed colored lines show the binned trend (10 bins, percentile binning). I evaluated the goodness of fit by the R^2 value.

As the heavy-tailed log-normally distributed impact measures in Subsection 3.2.2 illustrate, individual products' impact can differ broadly from each other. These broad differences are reproduced and explained for scientific careers by the modeling approach called Q-model, a mechanistic stochastic model on creative career evolution. According to this model, the impact of a certain creative product α created by an individual i ($S_{i,\alpha}$) can be decomposed as the product of two independent factors $S_{i,\alpha} = Q_i p_{i,\alpha}$, where Q_i is a variable specific to the individual and only depends on its career history, and $p_{i,\alpha}$ is a probabilistic variable, independently drawn for every creative product from a field-specific distribution.

In addition, the Q-model is building on the random-impact-rule and the log-normality of the impact- and productivity distributions (four example fields in Figure 3.1-3.4, and the goodness of the log-normal fit for all the studied fields in Table 3.1). Moreover, as seen in the following, the Q-model requires both N , Q , and p to be independent, and Q and p following log-normal distributions as well.

The Q-model assumes that when the distribution of the impact can be described by log-normal functions (as shown in Subsection 3.2.2), then the impact

can be expressed as the trivariate log-normal distribution of three variables. These three variables are (i) the productivity of the individuals (the number of papers they publish, N); (ii) an individual based "quality" parameter only depending on the individual's prior works' success (Q); and (iii) a random parameter representing stochastic outer factors (p). By transforming these variables to the logarithmic space ($\hat{N} = \log N$, $\hat{Q} = \log Q$, $\hat{p} = \log p$), the impact distribution $P(\hat{S})$ can be written as:

$$P(\hat{S}) = P(\hat{p}, \hat{Q}, \hat{N}) = \frac{1}{\sqrt{(2\pi)^3}} \exp \left(-\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right), \quad (3.6)$$

where $\mathbf{X} = (\hat{p}, \hat{Q}, \hat{N})$, $\boldsymbol{\mu} = (\mu_N, \mu_p, \mu_Q)$ is the average vector of the trivariate normal distribution, and $\boldsymbol{\Sigma}$ is its covariance matrix: $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_p^2 & \sigma_{p,Q} & \sigma_{p,N} \\ \sigma_{p,Q} & \sigma_Q^2 & \sigma_{Q,N} \\ \sigma_{p,N} & \sigma_{Q,N} & \sigma_N^2 \end{pmatrix}$.

To obtain the covariance matrix $\boldsymbol{\Sigma}$ of the trivariate log-normal distribution of Eq. 3.6, I used a CMA-ES algorithm [223, 224] (Covariance Matrix Adaptation Evolution Strategy), from which I obtained the parameters and summarized them in Table 3.2. The showed results are consistent with the reported findings of scientific careers in Ref [43]. By using these constants obtained by the optimization, I computed the values for Q and p for all the careers and creative products, from which I fitted the $P(Q)$ and $P(p)$ distributions by log-normal functions to validate the consistency of the analysis (Figure 3.5 and Table 3.1).

Finally, I showed that the Q parameter is robust over the course of individual careers in artistic domains as well, complementing the earlier analysis of Sinatra et al. on scientific careers [43]. These tests further support the use the Q -parameter of the individuals as reliable measures of success. I computed the correlation between Q parameters measured early and late in a career, and found a high correlation between the two, while I found low correlations between Q and N (Figure 3.6 and Table 3.3). These findings are in good agreement with the results reported about scientific careers [43].

3.3.3 Predictions of the Q -model

In the previous subsection, I introduced the Q -model and showed that its' requirements consistently hold for all the studied fields: the random-impact-rule, the low correlations between Q, p , and N , and the log-normality of these quantities alongside that of S . This way, I obtained a model that decomposes the observed impact S into two components, an individual-based quality parameter Q that captures the individuals' ability to generate high-impact work, and

3.3. The Q -model: decomposing impact through luck and individual ability 33

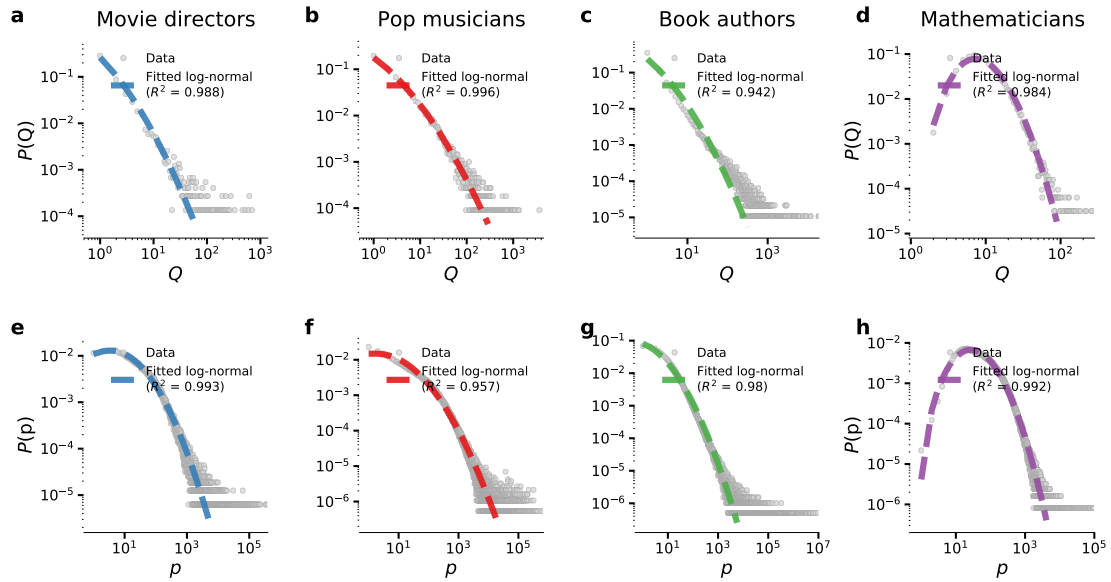


Figure 3.5. Measurements on the Q -model.

Figure a-d) and e-h) illustrate how the values of Q and p are distributed, showing the data points by grey dots and the fitted log-normals by dashed coloured lines for movie directors, pop musician, book authors, and mathematicians.

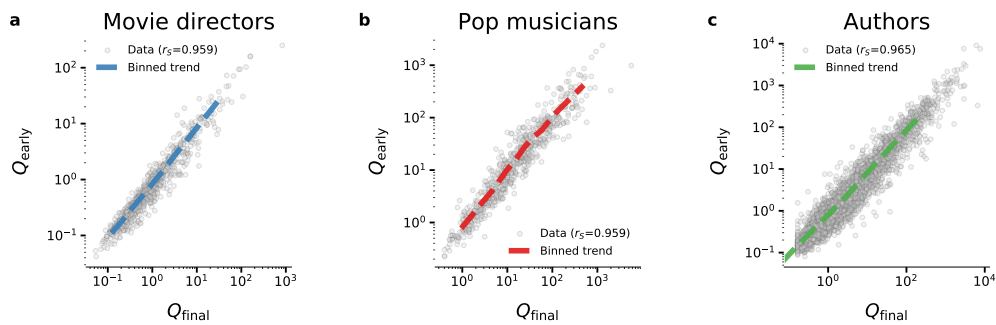


Figure 3.6. Comparison of Q parameters between early and final stages of careers in film, music, and literature.

Field	μ_N	μ_p	μ_Q	σ_N	σ_Q	σ_p	σ_{pQ}	σ_{pN}	σ_{QN}
Agronomy	4.467	2.148	4.147	1.711	0.452	0.514	-0.069	0.025	-0.005
Applied Physics	4.521	1.224	2.938	1.823	0.399	0.452	-0.046	-0.015	0.012
Movie art directors	4.214	4.557	3.877	1.606	0.437	0.465	-0.047	0.030	0.037
Biology	4.305	2.334	4.078	1.867	0.434	0.519	-0.072	-0.074	-0.027
Books authors	3.514	2.868	5.571	1.682	0.396	0.476	-0.054	-0.014	-0.062
Chemistry	4.366	2.462	4.586	1.629	0.440	0.485	-0.059	-0.132	0.083
Classical musicians	5.313	4.762	5.896	1.873	0.488	0.502	-0.072	0.024	0.003
Soundtrack composer	3.687	3.935	5.312	1.634	0.358	0.416	-0.030	0.005	-0.059
Movie directors	4.174	4.673	4.992	1.783	0.426	0.469	-0.053	0.097	-0.017
Electronic music artists	5.144	4.531	3.761	1.833	0.345	0.415	-0.035	0.026	0.038
Engineering	4.976	2.170	3.768	1.601	0.457	0.500	-0.063	0.017	0.046
Environmental Science	3.489	2.112	4.194	1.517	0.445	0.512	-0.063	0.034	-0.025
Folk musicians	5.246	4.071	6.260	1.744	0.434	0.475	-0.054	-0.026	-0.060
Funk musicians	5.608	3.612	6.081	1.656	0.447	0.477	-0.058	-0.076	0.011
Geology	4.592	4.749	4.441	1.802	0.374	0.435	-0.037	0.071	-0.056
Health Science	4.114	3.986	3.513	1.702	0.448	0.494	-0.061	0.048	-0.023
Hip-hop artists	5.354	4.703	5.433	1.472	0.481	0.495	-0.068	-0.072	-0.078
Jazz musicians	4.380	3.340	5.337	1.695	0.379	0.413	-0.030	-0.022	-0.099
Mathematics	4.557	4.662	4.212	1.647	0.378	0.434	-0.040	-0.030	-0.102
Physics	4.552	1.760	3.396	1.571	0.383	0.454	-0.046	0.055	0.037
Political Science	4.723	2.889	4.041	1.807	0.384	0.461	-0.055	-0.002	0.010
Pop musicians	6.071	2.553	4.421	1.911	0.417	0.454	-0.046	-0.024	0.018
Movie producers	3.823	4.424	4.010	1.499	0.397	0.476	-0.052	-0.081	-0.019
Psychology	4.565	4.696	4.514	1.000	0.737	2.013	0.220	0.009	-0.012
Rock musicians	5.518	3.858	3.643	1.572	0.451	0.509	-0.069	-0.017	0.046
Space Science Astronomy	5.231	2.164	2.523	1.672	0.371	0.462	-0.046	-0.056	0.094
Theoretical Computer Science	4.058	4.781	4.216	1.799	0.426	0.456	-0.045	-0.006	0.042
Script writers	3.024	2.944	4.128	1.620	0.461	0.516	-0.073	-0.005	-0.073
Zoology	4.058	3.655	3.523	1.723	0.377	0.429	-0.042	-0.028	-0.048

Table 3.2. Optimization results for the trivariate log-normal impact distribution.

The table reports the parameters of the $P(N)$, $P(Q)$, and $P(p)$ distributions obtained by a CMA-ES evolutionary optimization algorithm for all the studied fields.

3.3. The Q-model: decomposing impact through luck and individual ability 35

Field	Correlations _{NQ}	Correlations _{Q_{early}-Q_{late}}
Book authors	0.047	0.947
Classical musicians	0.184	0.956
Electronic music artists	0.063	0.947
Folk musicians	0.067	0.96
Funk musicians	0.022	0.953
Hip-hop artists	0.037	0.948
Jazz musicians	0.101	0.951
Movie art directors	-0.144	0.96
Movie directors	0.096	0.972
Movie producers	0.063	0.981
Plot writers	-0.005	0.974
Pop musicians	0.141	0.959
Rock musicians	-0.029	0.968
Soundtrack composers	0.106	0.955
hline		

Table 3.3. Robustness of the Q-model.

Correlations values between productivity (N) and the Q parameter; and the correlations between Q measured at early and late career stages for the different artistic fields.

a probabilistic term p that encodes random external fluctuations. After this, I tested the predictive power of the Q -model on how accurately it can reproduce the impact of the biggest hit of the individuals based on their quality parameter and the effects of luck. For that, I compared the scaling of the highest impact work with productivity as predicted by the Q -model, and showed that the Q -model gives significantly better results than a more simple model based on the random-impact-rule (R -model).

First, in the random model, I generated sets of careers on each field based on the random-impact-rule. To ensure that the set of synthetic random careers was directly comparable to the data, I constructed them by randomly reshuffling the time events of the individual careers from the data, repeated this randomization 100 times, and averaged the results. Second, I built a set of synthetic careers by using the Q -model. For that, I combined the given career length N_i and measured Q_i parameter of the individual i for every individual, and randomly redistributed the possible p_j parameters (picking exactly N_i p_j values for individual i) among the individuals' creative products. Then I computed the expected impacts of the synthetic careers by using the equation of the Q -model ($S_{i,\alpha} = Q_i p_{i,\alpha}$) for every creative product separately. Then I repeated this process 100 times. Finally, I measured the highest impact work as a function of career length on the data and the two types of synthetic careers. On the one hand, I found a good agreement between the prediction of the Q -model to the data. On the other hand, I reported that it performs much better than the simple

Field	R^2	Field	R^2
Funk musicians	0.636	Biology	0.991
Electronic music artists	0.700	Space Science or Astronomy	0.991
Movie art directors	0.704	Zoology	0.992
Script writers	0.808	Geology	0.993
Soundtrack composer	0.812	Applied Physics	0.994
Hip-hop artists	0.812	Engineering	0.994
Book authors	0.823	Theoretical Computer Science	0.994
Classical musicians	0.868	Chemistry	0.995
Rock musicians	0.871	Mathematics	0.996
Movie directors	0.918	Physics	0.998
Movie producers	0.925	Political Science	0.999
Pop musicians	0.968	Health Science	1.000
Agronomy	0.985	Environmental Science	1.000
Jazz musicians	0.987	Folk musicians	0.938

Table 3.4. Goodness of the predictions of the Q -model.

The table reports the goodness of the fit of the Q -model to the highest impact works, expressed by the R^2 values.

R -model on predicting the success of the highest impact work, suggesting that the Q -model is a sufficient approach to capture this phenomenon. These results are shown for four example fields in Figure 3.7, and the goodness of the fit for all the fields is summarized in Table 3.4.

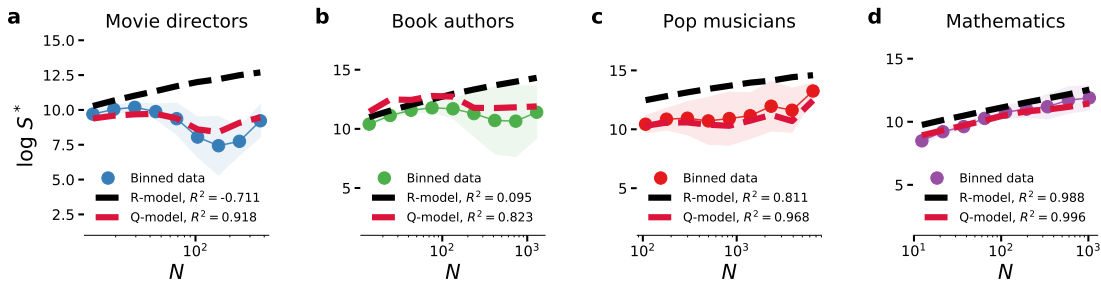


Figure 3.7. Comparing the R -model and the Q -model to the data.

The figure shows a comparison between the predictions of the Q -model on the highest impact works of the individuals as the function of their productivity to the random model and to the original data. The models are evaluated by the R^2 values relative to the data (Table 3.4).

3.4 Capturing the role of randomness in impact

3.4.1 Luck in the Q -model

Following the earlier results, the values of the covariance matrix of the trivariate impact-distribution (Eq. (3.3.2, Table 3.2) pinpoint a general conclusions about the behavior of the probabilistic term p . As the cross-terms of the covariance

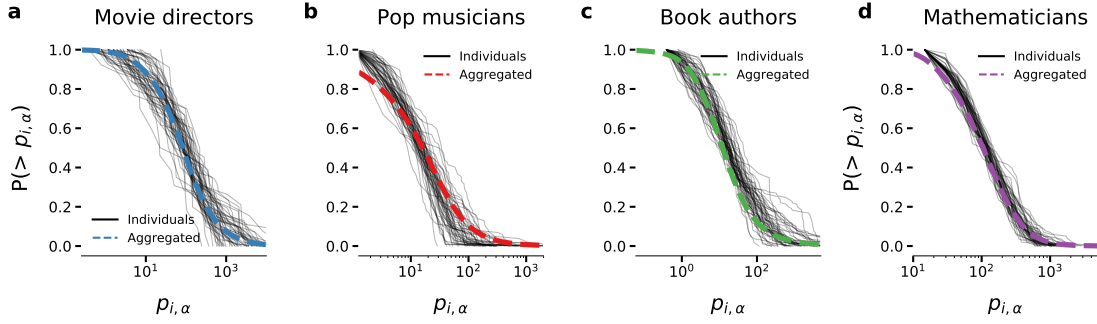


Figure 3.8. Rescaled cumulative impact distribution for four example fields.

The figures show the individuals' Q_i -rescaled impact distribution $P(p_{i,\alpha} = S_{i,\alpha}/Q_i)$ and how they collapse onto roughly the same aggregated curve, marked by continuous colored lines.

The distribution of 50 randomly chosen individuals is visualized by light grey lines.

matrix $\sigma_{p,Q}$ and $\sigma_{p,N}$ are close to zero (and significantly smaller than the other matrix elements), the distribution $P(p)$ does not depend on variables related to the individuals' careers (such as N and Q). This implies the impact rescaled by the individual parameter Q , i.e. $p_{i,\alpha} = S_{i,\alpha}/Q_i$, should collapse on the same distribution for all individuals on a given field. As turns out, this is the case for the studied data sets, illustrated in Figure 3.8. Since the rescaled $p_{i,\alpha} = S_{i,\alpha}/Q_i$ distribution is independent of any individual variables but universal for a given creative field, former research concluded that p can be interpreted as a “luck factor” contributing to impact [43].

3.4.2 Combining the Q -model with the classical test theory

After introducing the Q -model and validating it on the data sets I use, I combined it with the classical test (or true score) theory [220, 225–227] to show a use case of this impact decomposition method: the comparison of the fluctuations in luck and variations in the typical impact across different creative fields in a quantitative fashion.

First, recall the impact decomposition $S_{i,\alpha} = Q_i p_{i,\alpha}$ presented in Section 3.3 and the log-normality of its components, and transform the Q -model into the log-space in the following way:

$$\hat{S}_{i,\alpha} = \hat{Q}_i + \hat{p}_{i,\alpha}, \quad (3.7)$$

where $\hat{S}_{i,\alpha} = \log S_{i,\alpha}$, $\hat{Q}_i = \log Q_i$ and $\hat{p}_{i,\alpha} = \log p_{i,\alpha}$. I also note, based on earlier findings that Q_i and $p_{i,\alpha}$ are independent, since the covariance $\sigma_{pQ}^2 \approx 0$, then $\sigma_{\hat{p}\hat{Q}}^2 \approx 0$. After these transformations, Eq. (3.7) takes exactly the form proposed

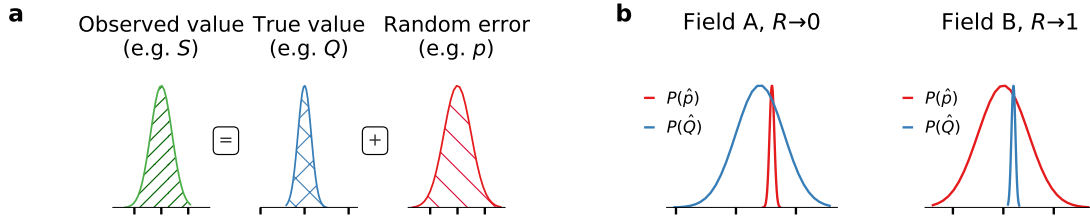


Figure 3.9. Illustration of the classical test theory.

a, According to the classical test theory, the normal distribution of an observed variable (green in the example, e.g., success S) can be decomposed as the sum of the distributions of the true score (blue, e.g., Q) and the error term (red, e.g., p). **b**, Illustrates the randomness index R on two extreme examples: on Field A the distribution of \hat{p} is narrow (has a low variance compared to \hat{Q}), therefore randomness has a negligible role ($R \rightarrow 0$). On the contrary, Field B has a narrow \hat{Q} and broad \hat{p} distribution meaning that the individual's luck dominates impact ($R \rightarrow 1$).

by the classical test theory [47, 48] for decomposing the observed value of a certain measurable. According to this statistical theory, the normally distributed observed value of an attribute, in this case \hat{S} , can be decomposed as the sum of two variables both following normal distributions under the condition of them being uncorrelated. One of these two variables encodes the true, error-free score of the observed quantity, in this case corresponding to \hat{Q} , and the other variable encodes a random error term, in parallel to \hat{p} . The two normal distributions of the variables \hat{Q}_i and $\hat{p}_{i,\alpha}$ are in line with previous studies, suggesting that any individual variables like merit and skill, and globally experienced quantities such as luck, are typically normally distributed [46–48, 213, 219, 228]. The principles of the true score theory are also illustrated in Figure 3.9a.

Building on Eq. (3.7), on the properties of normal distributions and on the measured attributes of Q and p distributions, I express the variance of the distribution of $\hat{S}_{i,\alpha}$, $\sigma_{\hat{S}}^2$, as:

$$\sigma_{\hat{S}}^2 = \sigma_{\hat{Q}}^2 + \sigma_{\hat{p}}^2, \quad (3.8)$$

where $\sigma_{\hat{p}}^2$ and $\sigma_{\hat{Q}}^2$ are the variance of the distributions of \hat{p} and \hat{Q} , respectively. This decomposition allowed me to measure the relative importance of the luck component compared to the individual component in determining impact fluctuations. Being inspired by previous work discussing the role of luck in sports and business [47, 48], my colleagues and I introduced the index R capturing the proportion of luck in the overall impact variance as:

$$R = \frac{\sigma_{\hat{p}}^2}{\sigma_{\hat{Q}}^2 + \sigma_{\hat{p}}^2} = \frac{\sigma_{\hat{p}}^2}{\sigma_{\hat{S}}^2}. \quad (3.9)$$

The R index reads as follows. When individuals in a domain have a similar ability to generate impact, captured by a narrow \hat{Q} distribution, differences in impact are mainly driven by the term corresponding to p , interpreted as luck, which leads to $R \rightarrow 1$. In contrast, when p has a low variance compared to S , then $R \rightarrow 0$, and luck plays only a small role (Figure 3.9b). Therefore this index allows the comparison of the role of randomness across different creative professions covered by the analyzed data sets. In addition, to ensure that the distinction between different fields based on R is sound, we conducted pairwise Mann-Whitney U tests to compare the distributions of Q and p between each possible pairing of professions. We found that there is not a single pair of fields in our dataset, which have both indistinguishable Q and p distributions at the same time (Figure 3.10).

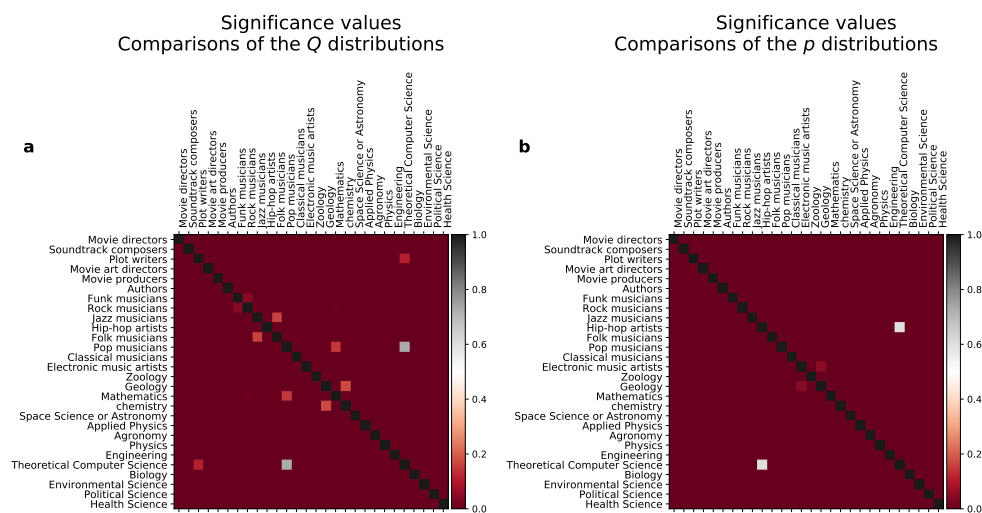


Figure 3.10. Results of the Mann-Whitney U test comparing the Q and p distributions.

The figure shows the pairwise p significance values of test comparing the **a**, Q and **b**, p distributions of the different creative professions.

3.4.3 Randomness in creative careers

Now that I managed to combine two existing approaches, the Q -model, and the true score theory in a mathematical framework, I use it to investigate the role of luck on different creative fields by answering the question: in which creative domains are impact-inequalities driven more by luck than by individual ability? Using the Q -model, first I measured σ_Q^2 and σ_p^2 for all the 28 creative professions from the film, music, and book industries, and in science. Figure

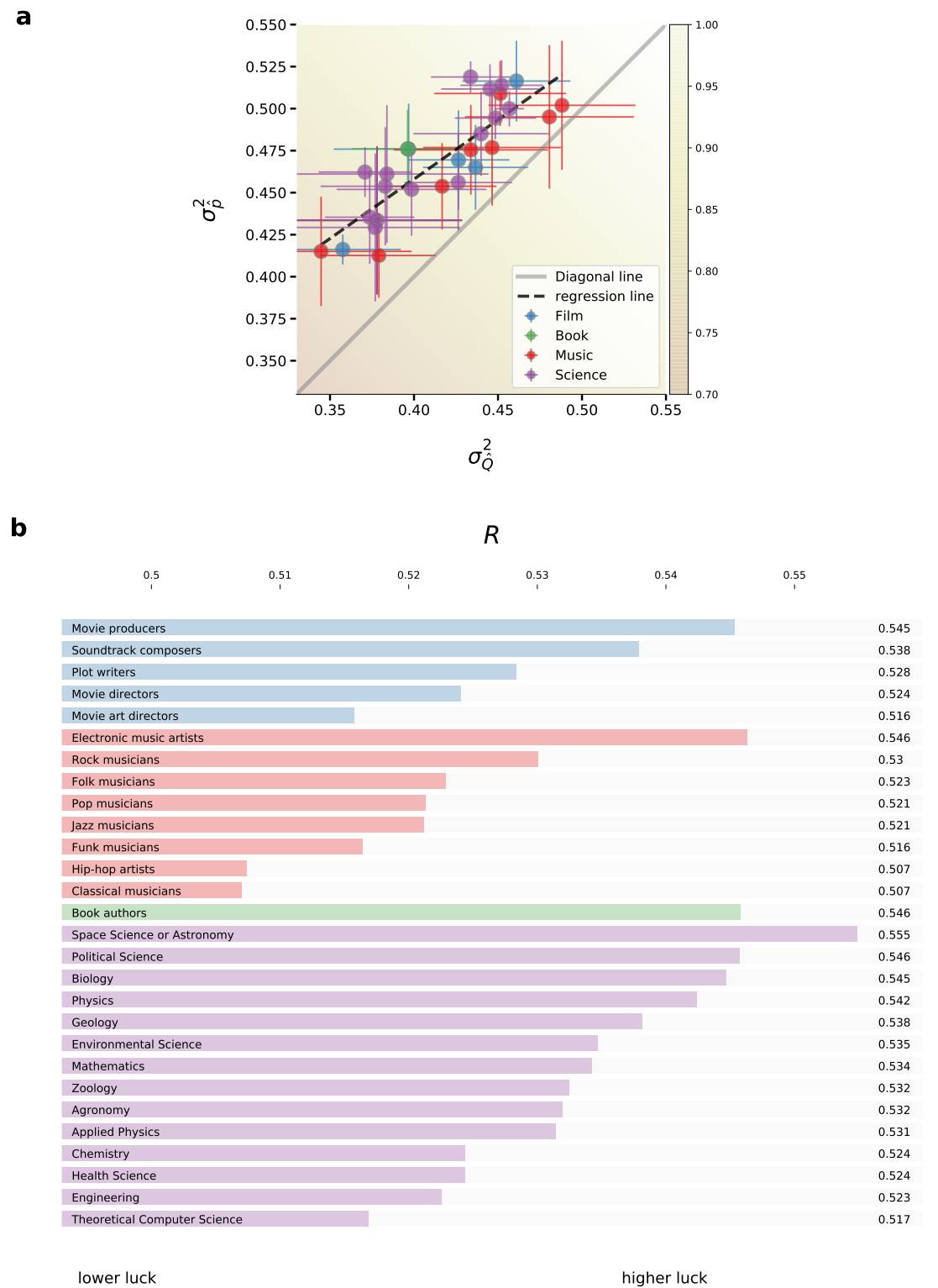


Figure 3.11. Comparing the role of randomness across 28 professions .

a, I show the studied 28 creative fields on the (σ_Q^2, σ_P^2) plane, marking fields from different data sets with different colors. I denoted a fitted line by continuous black line and added the diagonal as continuous grey line as a reference. The gradient-coloring of the background changes in a diagonal direction, illustrating that the points being on the same off-diagonal line have the same σ_S^2 . **b**, The table shows the values of the R randomness index based on Eq. (3.9) for the different fields.

3.11a shows all these professions placed alongside these two dimensions. Since the data points seem to be organized along a single line, I conducted a linear regression between σ_P^2 and σ_Q^2 (black dashed line on Figure 3.11a), where I determined a linear relationship with a slope of ≈ 0.71 , intersection of 0.18, and an r value of 0.85. Because $\sigma_S^2 = \sigma_P^2 + \sigma_Q^2$ and the regression slope is equal to the ratio σ_P^2 / σ_Q^2 , the measured slope being smaller than 1 indicates that as σ_S^2 increases (illustrated by the shading on Figure 3.11a), the value of σ_Q^2 increases faster than σ_P^2 . Hence large fluctuations in impact are dominated by large fluctuations in individual ability, captured by Q , rather than fluctuations in luck. Figure 3.11a offers several other interesting findings as well. On the one hand, I observed that all the studied fields are placed above the diagonal line ($\sigma_P^2 > \sigma_Q^2$), indicating that within each domain fluctuations in luck are broader than those in the typical career impact. On the other hand, I did not see any domain-specific clustering on the (σ_Q^2, σ_P^2) plane, which implies that the studied domains do not differ from each other in the role of luck.

From these computed variance values, it is straightforward to measure the randomness index R of Eq. (3.9) and to use it for a comparison of the characteristics of career success across domains (Figure 3.11b). The comparison reveals that first, within the movie industry, producers' careers are the most driven by luck, followed by composers; and second, being an art director is associated with the lowest R index, suggesting that achieving high impact as an art director is significantly less likely to happen by chance than in other careers within the movie industry. Since both professions rely on creative writing, it is interesting to compare the randomness index of scriptwriters ($R = 0.528$) and book authors ($R = 0.546$). The values of the indices show that writing for the movie industry is less determined by luck than in the book industry, probably reflecting the positive effects of working in a more collaborative environment.

In music, classical and hip-hop are the most robust against luck fluctuations as they are associated with the lowest randomness indices of our data set, $R = 0.507$. This could be explained by classical music being more dependent on skills, experience, deliberate practice, and musical training, as well as the more traditional schooling system of the style. Regarding hip-hop music, one could speculate that as being mostly an underground genre, it is less exposed to the rich-get-richer effect and is much less organized, leaving more space for rising juniors and a larger possibility to alternative ways of success. In contrast, the most popular genres, namely electronic music ($R = 0.546$) and rock music ($R = 0.530$) are on the other side of the range with top R values. These two genres contain the largest number of one-hit-wonder careers, and impact consistently has more pronounced fluctuations.

Regarding science, I found a wider range of randomness, with space science and astronomy ($R = 0.555$) and political science ($R = 0.546$) being at the luck-end of the range for the highest R -index fields. On the other hand, theoretical computer science ($R = 0.517$) and engineering ($R = 0.523$) are the least influenced fields by luck fluctuations.

3.5 Role of randomness in collaborations

In the previous sections, I reviewed the Q -model and analyzed the role of randomness in impact focusing on individual careers of 28 creative professions. However, typically, movies, songs, and scientific publications are the results of teamwork, as the overview in Subsection 2.2.5 points out as well. Therefore next, I ask: can collaborations between individuals improve, for instance, in the lights of previous findings (Subsection 2.2.6), our ability to predict the timing of career hits? How do success and network positions relate to each other? Earlier research already suggests a connection between network positions and the magnitude of the biggest hit [50, 82, 215, 229–231], which here I aim to extend to the temporal dimension and compare the co-evolution of these two alongside the career trajectories. For the sake of simplicity, here I focus on one example profession from each collaborative dataset: movie directors, pop musicians, and mathematicians.

3.5.1 Constructing collaboration networks

First, I reconstructed the temporal aggregated network of movie directors, pop musicians, and mathematicians at a yearly resolution from the beginning of the 20th century by using the cast lists of movies, lists of featured artists on pop releases, and co-authors of mathematics articles. In these weighted undirected networks, each individual is represented by a node, and the strength of the connection between two nodes at year T is proportional to the intensity of their collaboration. To compute the tie weights, first, I took the set of products, e.g., publications or movies, each individual i contributed to up until year T ($P_i(T)$). Then, I measured the weight of the connection between nodes i and j at year T as $w_{ij}(T)$, the Jaccard-index of the sets of works of the two individuals i and j :

$$w_{ij}(T) = \frac{|P_i(T) \cap P_j(T)|}{|P_i(T) \cup P_j(T)|}, \quad (3.10)$$

that is the number of works both individuals collaborated on, divided by the total number of works they contributed to until year T .

Based on this definition, the giant component of the final aggregated collaboration network of movie directors consist of 8,091,208 links between 184,220 people, the network of pop musicians measures as 52,366 artists connected by 8,232,349 edges, and for mathematicians, I obtained 94,755 connections between 27,401 scientists. The temporal growth of these networks measured as the number of edges over time is visualized in Figure 3.12a, showing a clear exponential trend with the network of mathematicians being the two orders of magnitudes smaller than the other two, most likely due to the way smaller typical time size in comparison to e.g., film. To further illustrate the network evolution of these fields, I visualized changes of the clustering coefficient over time in Figure 3.12b, which clearly shows that movie directors are increasingly more embedded in the social fabric of their field. In contrast, clustering in mathematics (most likely due to the slow increase in team sizes) increases, while pop music shows the opposite behavior.

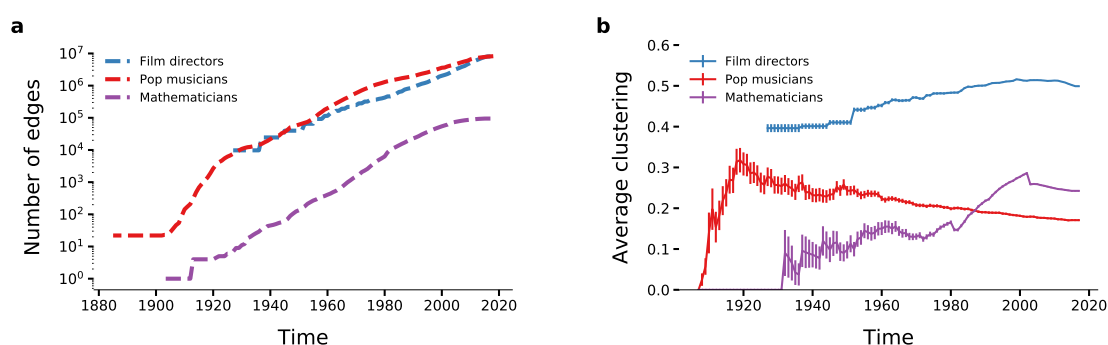


Figure 3.12. Temporal evolution of the collaboration networks.

The panel **a**, shows the size of the collaboration network between movie directors, mathematicians, and pop musicians over time, while panel **b**, illustrates how the average clustering of these networks evolves during the observation period.

3.5.2 Correlation between impact and network position

In this subsection, I analyze the temporal evolution of the network position of the film directors, pop musicians, and mathematicians, measured by the degree centrality, PageRank centrality, clustering coefficient, node strength, betweenness centrality, closeness centrality, network constraint [232], and coreness centrality [199]. In addition, I computed the correlations of the different network measures, which are shown in Figure 3.13. These correlation measurements show that while several measures, such as node degree and node strength, are highly correlated, others are not, covering significantly different angles of the

network perspectives, which also justifies the usage of these measures for further analyses.

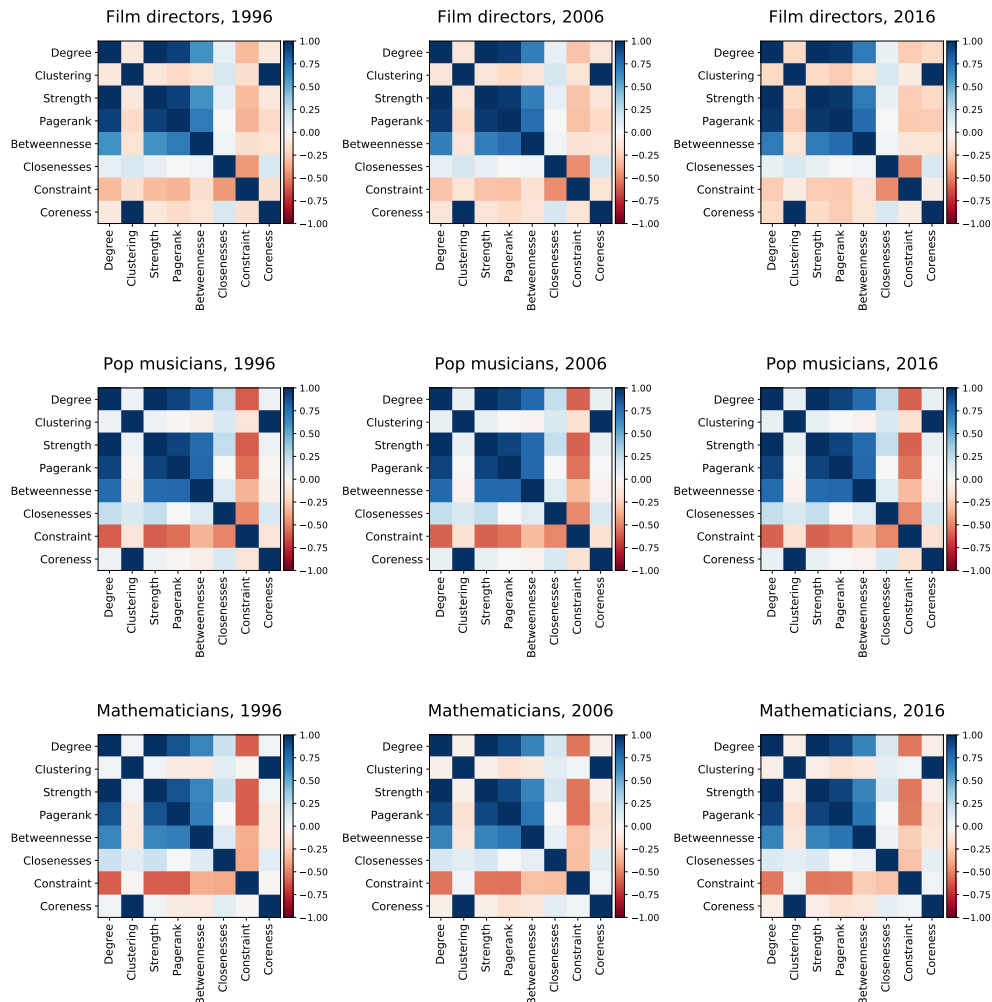


Figure 3.13. Correlations of network measures.

The figures show the correlations between the different network measures for the three studied creative professions at three example years (1996, 2006, 2016).

After computing the centralities in the networks over time, I created the network measure time series for each individual, where time events correspond to the works of the individual's career in chronological order. Then I compared the average-normalized network time series ($n(t)$) to the average-normalized impact time series ($i(t)$) on the level of the individuals and found that they are in general correlated. However, they are typically shifted. I estimated this

	Movie directors	Pop musicians	Mathematicians
Success peaks first	40.24 ± 0.07	44.19 ± 0.17	45.31 ± 0.18
Network peaks first	43.88 ± 0.08	45.39 ± 0.16	46.09 ± 0.19
Number of individuals	6,830	7,192	18,703

Table 3.5. Fraction of the individuals showing different networking patterns.

The table shows the fraction of individuals for which the network or the impact peaks first, averaged across the eight different network measures. The table also shows the number of individuals used for this analysis.

shifting parameter τ by aligning the network and impact time series so that the correlation between the two becomes maximal:

$$\tau = \arg \max_{\Delta\tau} \left([\text{Corr}(i(t + \Delta\tau), n(t))] \right). \quad (3.11)$$

For the sake of simplicity, I normalized these time series $n(t)$ and $i(t)$ by their average values. This is illustrated in Figure 3.14 by two examples. On the one hand, for the career of George Lucas as a movie director, I measured $\tau = -1$, meaning his success peaks first, and his network position followed by exactly one movie. On the other hand, the case of Francis Ford Coppola shows the complementary behavior: network peaks first, and success only comes later, delayed by a few movies as the observed $\tau = 5$ implies.

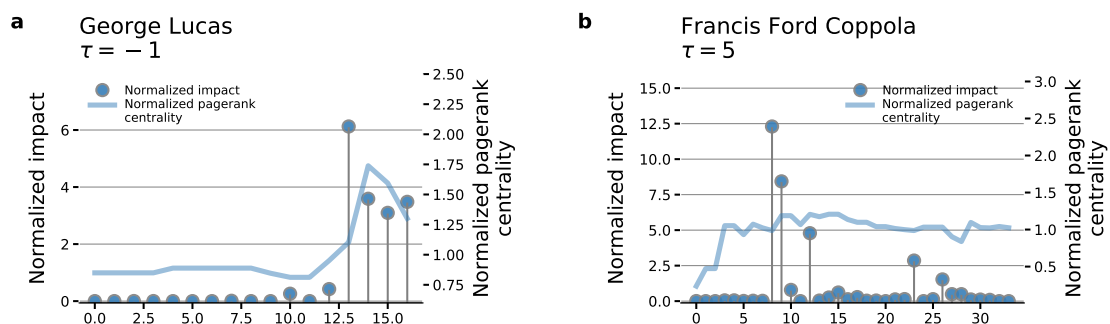


Figure 3.14. Network position and timing of the biggest hit for movie directors.

The figure shows the normalized impact of the movie directors compared to their normalized page rank centrality **a**, for George Lucas illustrating the case where the peak in impact is followed by the peak in the network position ($\tau = -1$) and **b**, for Francis Ford Coppola showing the opposite behavior when network centrality peaks first ($\tau = 5$), followed by the impact.

Following the idea of these two examples, I analyzed all individuals in my dataset and found that indeed, there are two major groups of individuals: those for whom the network measures peak before the highest impact work occurs,

	Pop music		Mathematicians		Film directors	
	d	p	d	p	d	p
PageRank	0.129	0.557	0.273	<0.001	0.142	0.600
Degree	0.129	0.515	0.269	<0.001	0.121	0.795
Clustering	0.114	0.698	0.292	<0.001	0.158	0.479
Strength	0.177	0.582	0.191	0.558	0.078	0.999
Betweenness	0.154	0.752	0.172	0.692	0.095	0.981
Closeness	0.143	0.839	0.156	0.772	0.068	1.000
Constraint	0.121	0.940	0.164	0.726	0.109	0.945
Coreness	0.146	0.805	0.179	0.613	0.059	1.000

Table 3.6. Results of the Kolmogorov-Smirnov test between the randomized and the measured network parameters for the τ distributions.

The network positions are captured by PageRank centrality, degree centrality, clustering coefficient, node strength, betweenness centrality, closeness centrality, constraint, and coreness centrality. The distributions of the corresponding values for the professions of pop music, mathematics, and movie directors can be found in Figure 3.15. The table reports both the measured KS distance values (d) and their statistical significance (p).

and those for whom the network peak occurs after, and only a significantly smaller fraction for whom network and success peaked at the same time. The exact number of individuals for whom this network analysis was possible to carry out and the fraction of the different networking behaviors are summarized in Table 3.5.

3.5.3 Randomness in networking

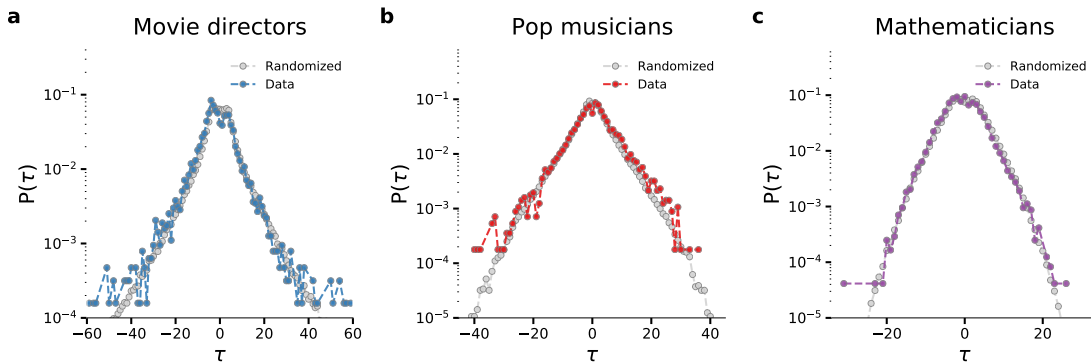


Figure 3.15. Distribution of the shifting parameter τ .

The panels show the distributions of the measured shift parameter τ (coloured lines) comparing to the randomized case (grey lines) between the film directors', pop musicians' and mathematicians' PageRank and impact time series. The corresponding Kolmogor-Smirnov tests' results are in Table 3.6.

As there seems to be two complementary networking behavior describing the studied creative professions, the question naturally arises: how may these

	Pop music		Mathematicians		Film directors	
	d	p	d	p	d	p
PageRank	0.074	<0.001	0.039	<0.001	0.041	<0.001
Degree	0.029	<0.1	0.085	<0.001	0.068	<0.001
Clustering	0.073	<0.001	0.029	<0.01	0.057	<0.001
Strength	0.131	<0.001	0.203	<0.001	0.078	<0.001
Betweenness	0.24	<0.001	0.075	<0.001	0.113	<0.001
Closeness	0.099	<0.001	0.061	<0.001	0.094	<0.001
Constraint	0.088	<0.001	0.071	<0.001	0.032	<0.001
Coreness	0.065	<0.001	0.083	<0.001	0.039	<0.001

Table 3.7. Results of the Kolmogorov-Smirnov test between the S distributions.

Individuals are split based on the sign of their τ parameter. We used the following network measures: PageRank centrality, degree centrality, clustering coefficient, node strength, betweenness centrality, closeness centrality, constraint, and coreness centrality. The table reports both the measured KS distance values (d) and their statistical significance (p).

	Pop music		Mathematicians		Film directors	
	d	p	d	p	d	p
PageRank	0.048	<0.1	0.022	<0.001	0.02202	<0.1
Degree	0.032	<0.1	0.076	<0.001	0.05878	<0.001
Clustering	0.073	<0.01	0.054	<0.001	0.00662	<0.001
Strength	0.131	<0.001	0.203	<0.001	0.078	<0.001
Betweenness	0.240	<0.001	0.075	<0.001	0.113	<0.001
Closeness	0.099	<0.001	0.061	<0.001	0.094	<0.001
Constraint	0.088	<0.001	0.171	<0.001	0.032	<0.001
Coreness	0.065	<0.001	0.083	<0.001	0.039	<0.001

Table 3.8. Results of the Kolmogorov-Smirnov test between the Q distributions.

Individuals are split based on kind of career, while their network positions are captured by their PageRank centrality, degree centrality, clustering coefficient, node strength, betweenness centrality, closeness centrality, constraint, and coreness centrality. The table reports both the measured KS distance values (d) and their statistical significance (p).

groups differ e.g., in their expected success, and can we tell the differences between them? To test these, I constructed synthetic careers where I randomly reshuffled both the network and the impact time series of each empirical career trajectory of the data sets and conducted the same measurement of the τ shifting parameter for each of them. Afterward, I compared the distribution of the shifting parameter measured both on the original and the randomized data. My surprising findings revealed that the two distributions are closely overlapping, as seen in Figure 3.15 in the case when network centrality is measured by PageRank. I confirmed this observation for all the eight studied network measures and three creative professions as well by conducting the double-sided Kolmogorov-Smirnov test (see the results in Tabel 3.6).

As my previous analysis showed, whether an individuals' network or success peaks first, it happens totally by chance, from which I expected their suc-

cess to be independent of this behavior as well. I tested this by comparing the distribution of success S and the individuals' Q parameters for the three studied fields between these two groups and by comparing the distributions using the Kolmogorov-Smirnov test, and I found no significant differences. In other words, regardless of whether success or network peaks first, the expected success of the individuals stays the same. The results for the case of PageRank centrality are visualized in Figure 3.16, while the results of a thorough comparison splitting the individuals based on all the centrality measures are in Table 3.7-3.8.

In conclusion, on the role of randomness, I have uncovered two complementary networking behaviors: individuals for whom network peaks first, and those for impact peaks first. However, this distinction seems to happen at random, showcasing another example of the role of randomness. This also implies that network properties do not improve our understanding of predicting the timing of the biggest hits, which is also in line with the random-impact-rule described earlier.

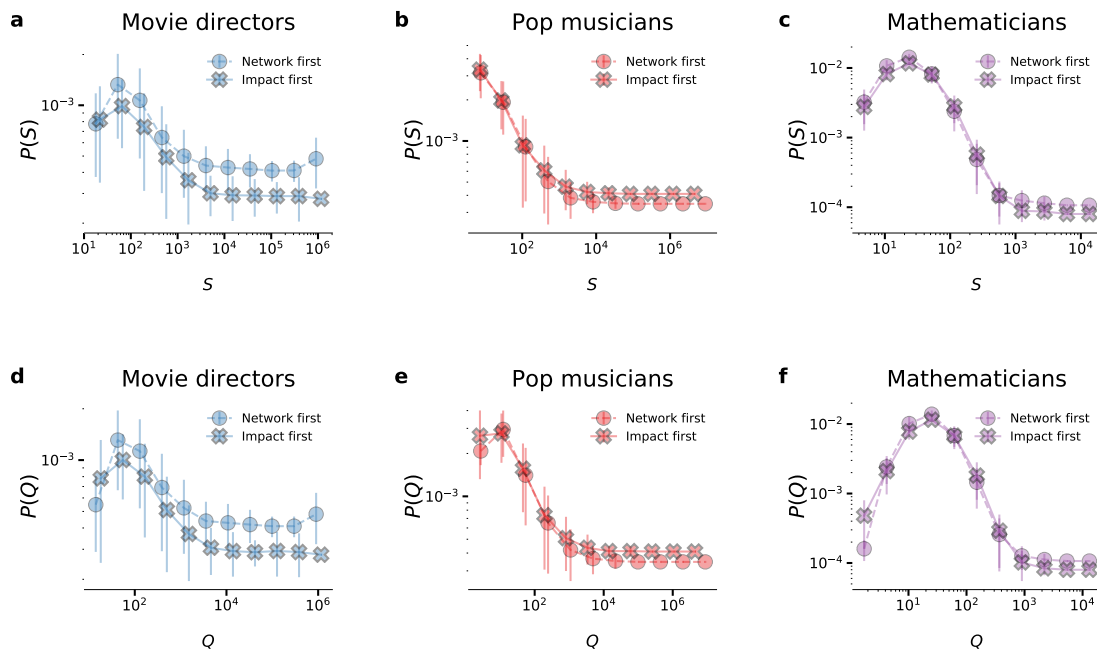


Figure 3.16. Success distributions for different networking behaviors

Success measured by impact (S) and Q distributions, where individuals are split based on their professions (director, pop musician, mathematician), while their network centralities are captured by their degree, PageRank, and clustering.

3.6 Discussion

In this chapter, I reviewed the modeling approach called Q -model and showed how its criteria are fulfilled on the four large-scale datasets I collected covering the film, book, and music industries and science. Then I applied this model to hundreds of thousands of creative careers reconstructed from my database to decompose the success of individuals into two independent components defined by the model. One of these is expressing the ability of an individual to have a consistently high or low typical impact, captured by the Q -parameter, and the other one is associated with random fluctuations, as shown during my analysis, corresponding to *luck*.

Connecting the Q -model to the classical test theory, I managed to compare the importance of the two components Q and p across the 28 studied creative professions and found that on average fluctuations in impact of single creative works are more influenced by luck than by individual ability and that the fluctuations in the individual parameter are more pronounced for fields with large fluctuations in impact. In addition, I found that professions within different domains are not clustered based on the relative magnitude of these fluctuations, which suggests that the magnitude of luck is not a distinctive feature of creative domains but instead is a universal attribute.

Further building on the combination of the Q -model and the classical test theory and earlier research on the role of luck, I defined the randomness index as the relative ratio of the variance of the random component to that of success. The analysis of this index shows that its values vary in a relatively narrow range, which further confirms the lack of distinct typical scales of random fluctuations associated with the four different domains investigated in the paper. In this narrow range of randomness, I have found that on the one hand, professions with the highest exposure to luck are those of film producers, electronic music artists, book authors, and scientists working in the fields of space science and political science. On the other hand, randomness has the lowest influence on fields such as hip-hop and classical music and theoretical computer science.

Finally, I studied the relationship between individuals' network position and success and found another surprising dimension of success where luck plays a major role. In fact, the temporal correlation between success and centrality in the collaboration network for movie directors, pop musicians, and mathematicians as a case study revealed that two distinct classes of creative careers exist regardless of their creative domain. Individuals belonging to the first group produce their most successful work first and become well-connected in the network only after the occurrence of the hit, while people falling into the second category show the inverse of this behavior and first build favorable connections,

and only produce their big hit afterward. However, as my comparative analysis showed, whether an individual falls into the first or the second category has no more regularities than the random case. This lays in line with earlier work, e.g., the random-impact-rule, illustrating the difficulties of predicting the timing of an individual's biggest hit. Furthermore, I found no difference between individuals' success and Q parameter and group the individual belongs to.

CHAPTER 4

COMMUNITIES AND MENTORSHIP IN ELECTRONIC MUSIC

4.1 Introduction

The emergence of success in creative professions, such as music, has been studied extensively, as the number of cited works in Section 2.2.5-2.2.6 also illustrates. However, the mechanisms connecting individual success to collaboration is not yet fully understood. This chapter contributes to this line of work by analyzing longitudinal data on the co-releasing and mentoring patterns of popular electronic music artists appearing in the annual *Top 100 ranking* of the DJ Magazine [99, 233].

Throughout history, music has been one of the most powerful forms of expressing culture and identity. Music is typically the result of teamwork, a collaborative effort of individuals with various backgrounds and behaviors, as also seen in Section 3.5. Consequently, the world of musicians is a complex social ecosystem, encompassing a great variety of genres, trends, tools, and audiences. Even earlier works, mostly originated in developmental psychology [234–236], started to analyze career patterns in music (for details see Section 2.1). Later, many researchers [21, 23, 34, 45, 237], including myself as detailed in Chapter 3, started to incorporate large-scale data sets and further investigate the roots of

This chapter is based on the article “Elites, communities and the limited benefits of mentorship in electronic music ” [231], that has been accepted for publication in Scientific Reports, a journal of the Nature Research family, before the submission of this dissertation. In this paper, I proposed the idea of the project and performed the data collection and analysis. All authors, Milan Janosov, Federico Musciotto, Federico Battiston, and Gerardo Iñiguez designed the measurements and contributed to the manuscript.

individual success in music. For example, previous research studied the role of forbidden triads and the relational field in jazz [53, 238]. Other researchers attempted to capture large-scale features of the musical world, such as extracting collaboration patterns and community structures or identifying genres of various scenes like classical music, jazz, and the Rolling Stone Magazine's list of '500 Greatest Albums of All Time' [188, 191, 192, 197]. More recent works have also analyzed the changes of trends and fashion cycles in music over time [239–241].

Despite these results, a clear connection between the success of individuals and their role in the social fabric of music scenes is lacking. Here I aim to fill this gap by investigating the well-defined ecosystem of artists working on electronic music. For this, I focus on the annual top 100 ranking of the DJ Magazine [233, 242] (DJs as disc jockeys) from 1997 to 2018, a period during which electronic music transitioned from the outskirts of music to become one of its most popular fields. Despite this recent popularity, electronic music has only produced a handful of stars, while the majority of DJs and producers remain unknown.

The goal of this chapter is to provide a deeper understanding of how superstar DJs and producers (since a significant fraction of DJs also act as producers) emerge by analyzing the interplay between individual success. For this, I quantify success as the DJs' positions in the top 100 DJ ranking lists, and the underlying interaction network is modeled by the co-releases of the artists based on Discogs [15, 19]. I also discuss the community structure and dynamics of the DJ world and show examples of the main success trends of the DJ communities. Moreover, similarly to studies on science and academic success [176], I uncover the existence of mentoring in electronic music, which both aligns with earlier research on peers and mentor-protégé relationships (Section 2.1) and shows controversial effects on the expected later-success of the mentees.

This chapter is structured as follows. First, I present an analysis of the dynamics of the DJ ranking list [98, 243–246] to measure the most stable subset of star DJs. Then I connect this dynamics with the underlying co-release network of electronic music artists and analyze their collaboration patterns. Finally, I provide and test a definition of mentorship and study its relations to success.

4.2 Data

To carry out the proposed analysis, I combined the following four data sources:

1. The annual top 100 rankings of DJs from the official website of the DJ Magazine [233] and related sources [247, 248]. This ranking list represents

a direct way of measuring success by the terms of Section 2.2.1, and is derived from the cumulative number of votes coming from the DJ Magazine's poll filled out by several million people [99] every year. While the raw vote counts are unpublished, the rankings are publicly available from the years 1997 to 2018 and contain ~ 540 electronic music artists.

2. Like in Chapter 3, I used discography profiles from Discogs [15, 19], an online crowd-sourced music discography platform that lists the production of 46,063 artists active on electronic music, comprising 1,103,769 releases up to December 2018. The discography data includes collaborations, featuring appearances, and remixes, yet it lacks information on the popularity of the produced songs.
3. To complement the lack of popularity measure, similarly to Chapter 3, I collected song play count information by matching the song titles extracted from Discogs to the public API of LastFM [138].
4. I collected genre information from Wikipedia about half of the DJs that have available profiles on the platform and extracted the consistent genre-keywords each profile contained. This way, I attached every artist with a set of specific genre tags, such as trance and deep house.

These sources combined allowed me to reconstruct the ranking trajectory and collaboration evolution of the DJs and associated both popularity numbers and genre tags to them as additional information.

4.3 Dynamics of the top 100 ranking list

During the 22 year-long history of the top 100 ranking of the DJ Magazine, more than five hundred DJs have made it to the top 100. Yet, the electronic music scene has only seen a handful of stars in the ranking list for extended periods. Eleven artists have been crowned as No. 1 DJ in the world during 1997-2018, a sign of the prominence of figures like Carl Cox, Tiësto, and Armin van Buuren. Conversely, success has been ephemeral for most of the artists: almost 170 DJs have been in the top 100 only once (with average rank $\langle r \rangle \sim 75.3$), and 99 just made it twice (with average rank $\langle r \rangle \sim 72.8$). The heterogeneous dynamics of the ranking are also shown in Figure 4.1, where DJs' careers are represented as ranking time series.

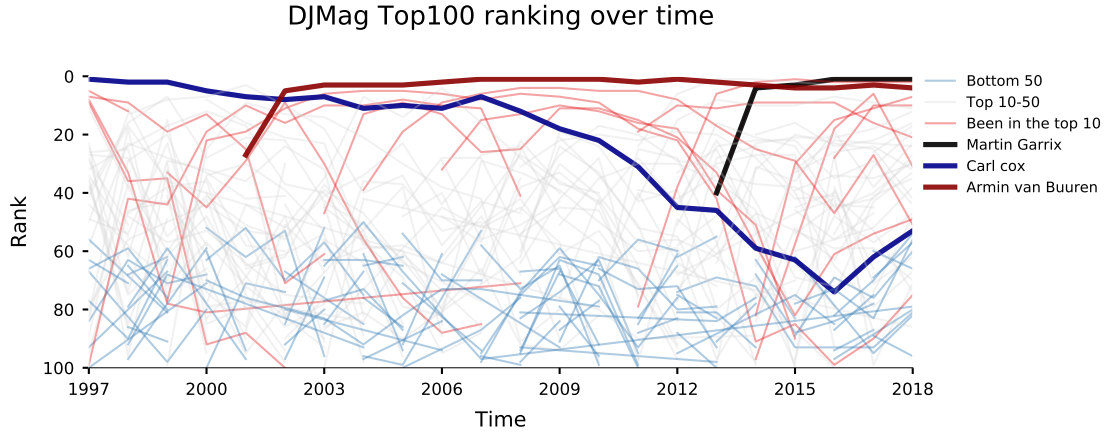


Figure 4.1. Visualizing the top 100 ranking over time.

Temporal evolution of the rank of the top 100 DJs [233]. Each artist is denoted by a single line, while colors indicate the part of the ranking list DJs have visited (red – top 10, blue – bottom 50, grey – in-between). The graph also shows a few examples by thickened lines, such as Armin van Buuren.

4.3.1 Regime change in the ranking

This strong heterogeneity raises questions, such as what positions and regions in the top 100 can be associated with well-established success, and where is success just ephemeral? Where is the boundary of being a star DJ, if any? For instance, DJ Magazine releases the names of all top 100 DJs, and the top 10 is often treated specially as well. While both 10 and 100 seem to be arbitrary thresholds of success, I managed to identify a real threshold that emerges naturally from the dynamics of the ranking. For that I computed a measure characterizing the ranking list called *rank diversity* $d(r)$ [243], which counts the number of different names that appear at a given rank r during the observation period, normalized by the length of this period (T). For instance, 11 different DJs have ever reached the No. 1 position during $T = 22$ years, therefore $d(1) = 11/22$. The distribution of the rank diversity in Figure 4.2a shows that a regime-change happens between the upper and lower parts of the ranking. The values of the rank diversity also show a significant level of noise due to limitations on the size of the dataset. Therefore, to capture the real threshold separating the top from the rest, I used the following method. First, I split the ranking into upper and lower tiers based on an arbitrary threshold of r . After that, I took the rank diversities separately over these two segments and measured their variances σ_i . Finally, I compared the lower and the upper tiers of the ranking based on the variance difference between them:

$$\Delta\sigma_d(r) = |\sigma_{d,\rho}(\rho \leq r) - \sigma_{d,\rho}(\rho > r)|. \quad (4.1)$$

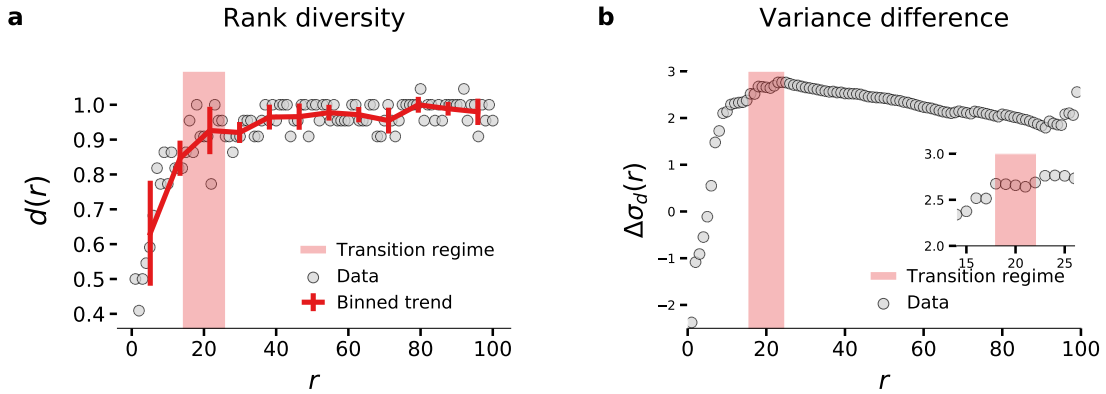


Figure 4.2. Capturing regime change in the top 100 ranking.

a, Rank diversity $d(r)$, defined as the number of individuals that have ever occupied rank r , normalized by the length of the observation window. **b**, Variance difference $\Delta\sigma_d(r)$ captured by rank diversity between the top and the bottom of the ranking (see Eq. 4.1). Inset: Zoomed excerpt of the transition regime between the boundary of the real top and the rest (axes are the same as in the main plot).

By computing Equation 4.1 for varying arbitrary thresholds (i.e., r values), I observed monotonously non-decreasing behavior for $\Delta\sigma_d(r)$ in ranks 1–18 and slowly decreasing behavior after rank 22 (Figure 4.2b). In addition, I neglected the end of the ranking due to boundary effects. Based on the transition between these two regimes (highlighted on the inset of Figure 4.2b as a maximum in variance difference), I estimated the best splitting threshold separating the top of the ranking from its bottom as $r^* \approx 20 \pm 2$. During the rest of this chapter, I will refer to the top 20 as the real top-tier of the ranking list and call top DJs those who have made it to the top 20 at least once.

Rank diversity $d(r)$ shows a clear difference between high-tier and low-tier DJs, implying that it is more difficult to break into the top 20 than to catch lower positions of the ranking. As a consequence, once DJs make it to the top 20, they usually are able to maintain their positions with more ease than those at lower ranks (i.e., large r). In particular, I found that the yearly rank difference Δr of DJs (Figure 4.3a) has different trends for top DJs than for those who never make it there: the chances of not changing rank (step size of zero) is twice as high for top DJs than for the rest. Surprisingly, all DJs have similarly low chances of extremely large rank jumps, even from the top (low r) where there is more rank-space to fall. Yet, unexpectedly, large jumps do exist. For instance, a great success of recent years, the American DJ duo The Chainsmokers, started at $r = 97$ in 2014 and jumped forward by 79 places to $r = 18$, while the Russian trio Swanky Tunes entered the top 100 at $r = 97$ in 2015, made a huge jump to $r = 27$ the year after, but then fell back to $r = 99$ in 2017. In addition, those

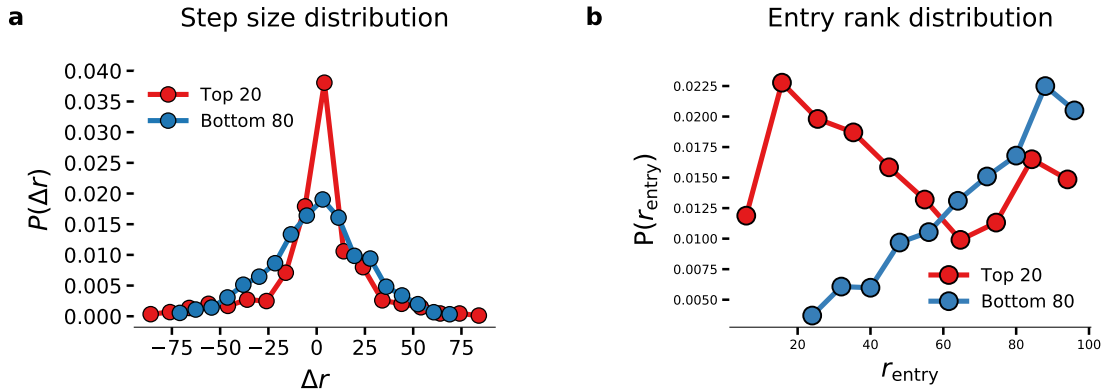


Figure 4.3. Comparing the top and the bottom of the ranking.

a, Yearly step size distribution (Δr), defined as the individuals' rank differences between two consecutive years, comparing top DJs who have ever been in the top 20 (red line) with the rest (blue line), where a positive sign means a drop to larger r values, i.e., decline in success. **b**, The entry rank distribution of the artists placing in the top and the bottom of the ranking (by splitting it at $r = 20$).

DJs who make it to the real top seem to have a different start, measured by their entry positions. The distribution of the entry positions is shown in Figure 4.3b: for later top DJs, it is skewed towards better ranks, while for bottom DJs, it is skewed towards lower ranks.

4.3.2 Ranking and popularity

The top 100 rankings are based on the results of a popularity poll resulting in a public opinion-based cumulative success measure. Therefore I expected other known measures, such as the play count of the artists' songs, to show similar trends and high correlations to the rankings. To test this hypothesis, I collected the song play count information of the top DJs based on LastFM [138] keeping track of both the total play count of their discography and the play counts of their songs throughout 2019 with five measurements (December 2018, February, March, April, and May 2019). Surprisingly, in each case (annual/total play counts) and measurement times, I found significantly low (< 0.4) correlations between play counts and rankings (results are summarized in Table 4.1). The unexpectedly low correlations between DJs' ranks and their total annual play counts on LastFM are suggesting a dichotomy between quality and success similar to findings in other creative domains [93, 211, 249]. From this observation, a question follows: if raw popularity is not enough for success, what else do DJs need to reach the top of their profession? To answer this question, I propose a network-based explanation based on the collaboration and co-release patterns

of the DJs.

	December 20	February 17	March 22	Apr 19	May 26
Total play count of songs in 2018	14.7M	17.9M	27.2M	17.8M	18.0M
Rank in 2018 year vs. play count of songs released in 2018	0.394 (0.0091)	0.393 (0.0008)	0.371 (0.0007)	0.346 (0.0017)	0.35 (0.0015)
Best rank ever vs. total play count over the career	0.262 (3.3·10 ⁻⁶)	0.321 (3.2·10 ⁻⁶)	0.332 (8.7·10 ⁻⁸)	0.318 (4.8·10 ⁻⁷)	0.319 (4.2·10 ⁻⁷)

Table 4.1. Play count and ranking correlations.

Correlation (and related p -values) between DJs' best and average ranks by 2018 and the play counts of their songs released in 2018. The measurements were conducted on different occasions with a few month differences during 2-3 day-long crawling periods (marked by their starting dates). The changing trend of total play counts shows how songs released in 2018 are losing popularity in 2019.

4.4 Co-release network in the DJ world

As the previous section shows, the top 100 DJs, considered to be a unique and elite club of artists already, is split into two: a real circle of long-standing stars (as my analysis shows, the top 20), and the rest with ephemeral success. Since my first hypothesis stating that this phenomenon can be explained by raw popularity failed, I asked: are there any network effects that keep this handful of stars at the top, and result in a faster dynamics of rank change at the bottom? What is the relationship between the social fabric of DJs and the observed dynamics of the top 100 rankings? To tackle these questions, I constructed and analyzed the co-release network of the top 100 DJs based on their profiles as electronic music artists on Discogs [15], which contain different types of the releases, such as singles and remixes, each artist contributed to.

4.4.1 Building and describing the co-release network

First, I extracted all the releases and their contributors from the electronic music category in Discogs. Then defined the network as follows: each DJ who made it to the top 100 is represented by a node, and the strength of the connection between each pair of artists is measured by the number of releases they co-occurred on. Due to a large number of remixes and best-of collections, this definition resulted in a dense network with a giant component of 15,403 edges distributed among 486 nodes. Therefore I applied a network filtering algorithm,

the recently introduced noise-corrected filter [250], with which I filtered out $\sim 88\%$ of edges while keeping $\sim 86\%$ of nodes. The backbone-filtered network is visualized on Figure 4.4.

Then, I computed the network centralities of the DJs to compare those metrics with success, following previous results (Subsection 2.2.5). I measured the network positions of the artists by their degree centrality, betweenness centrality, PageRank centrality, and their clustering coefficients, and I used both the best and the average rank of the DJs within the ranking as success measures. Unexpectedly, I found low (~ 0.2) correlations between network positions and success, as summarized in Table 4.2, which means that the most successful DJs are not the most central ones on average, contradicting with previous findings (Section 2.2.5). This, and the visual observation of Figure 4.4 indicated that success should be related to the community structure rather than the global network properties.

	Degree	Betweenness	PageRank	Clustering
Best rank	0.263 ($<10^{-6}$)	0.234 ($<10^{-6}$)	0.261 ($<10^{-6}$)	0.109 (0.012)
Average rank	0.207 ($<10^{-6}$)	0.172 ($<10^{-6}$)	0.202 ($<10^{-6}$)	0.09 (0.011)

Table 4.2. Correlation between centrality and success.

The correlations (and related p values) between the different centrality measures (degree centrality, betweenness centrality, PageRank centrality, and clustering coefficients) and the best and the average ranks of the top 100 artists.

4.4.2 DJ Communities

As Figure 4.4 shows, this network does not have a single core or a core-periphery structure but is rather polycentric with multiple smaller communities. To further investigate the modularity of the network, I used an established heuristic algorithm to extract communities [251] similar to earlier works about scientific and musical communities (Subsection 2.2.3-2.2.6). I conducted the community-detection only at the final stage of the network. This assumption has two major arguments. First, as turns out later, these communities are centered around top-tier DJs, whose communities therefor are fixed. Second, the rest of the DJs only stay in the ranking for a few years, and during such a short period it is very unlikely for an artist to change its musical profile so drastically that it changes community. With this method, I managed to extract seven communities covering 92% of the nodes in the giant component of the top DJ network. Surprisingly, each community includes one or two DJs who

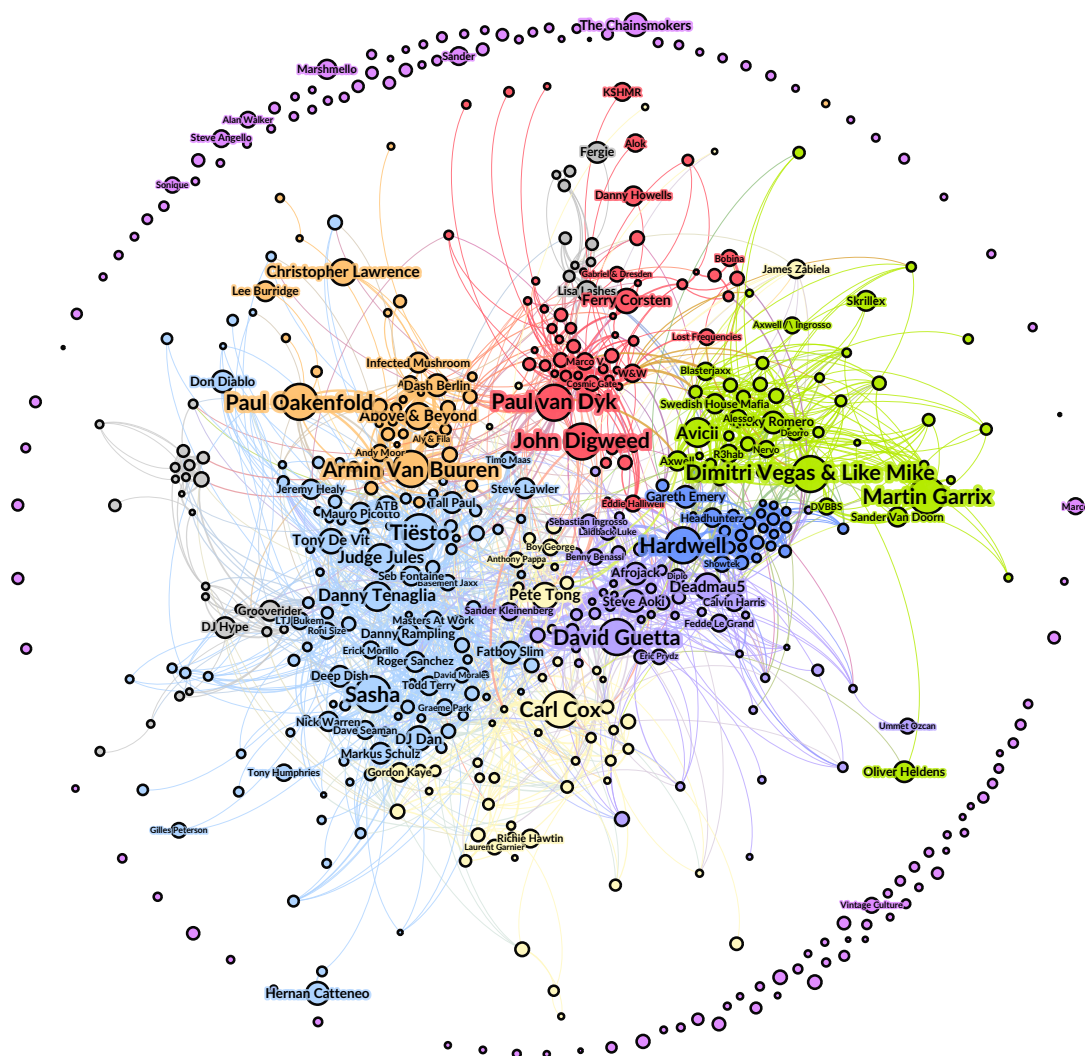


Figure 4.4. Top 100 DJ network.

Temporally aggregated and back-bone filtered [250] co-release network of top 100 DJs. DJs are represented by nodes and co-releases by links between them, with link width proportional to the number of releases the DJs collaborated on. Node size is proportional to the DJ’s best rank (larger size means lower r and higher success). Node colors show the detected music communities [251]. Top 20 DJs are labeled by their names.

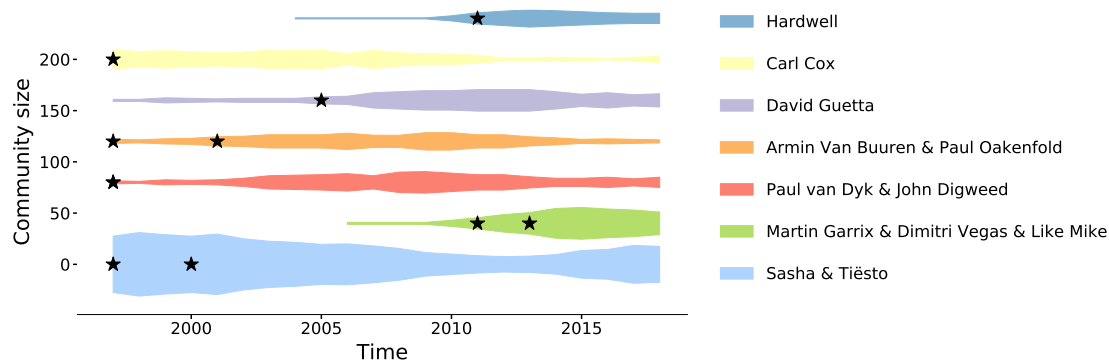


Figure 4.5. The evolution of the top 100 DJ network.

Temporal growth of DJ communities, with size measured as the number of DJs in a given year's top 100 ranking. Black stars denote the entry years of the named (later No. 1) DJs.

once earned the No. 1. DJ title – in other words, the communities seem to be clustered around star DJs.

Next, I studied the temporal evolution of DJ communities by measuring their size, defined as the number of top 100 DJs in each group, over time. I found that the communities, named after their leading artists, rise and fall over time distinctively, highlighting how new artist communities form and old ones fade away. Furthermore, I found that these former and current No. 1. DJs, as shown by the star symbols in Figure 4.5, are amongst the earliest members of their communities. These temporal trends are in agreement with recent findings on changes in fashion cycles and the roles of the elite in them [239].

In addition, I observed a connection between the growth of the communities and the success of their leading artists. For instance, how the mainstream of electronic music emerged in the community of Sasha and Tiësto, and how the latest electronic dance music trends started to grow around Dimitri Vegas & Like Mike and Martin Garrix. The size and success of the communities are shown in parallel in Figure 4.6, which pinpoints a significant correlation of ≈ 0.73 on average between the evolution of the size of communities and the average rank of the three most successful artists of each community. This observation further supports the major role leading artists play, besides being early members, in the growth of their community.

How do these DJ communities differ from each other? One possibility according to the literature is that musical genres (such as techno, house, and trance) reflect these differences [192, 197, 198]. To test this hypothesis, I collected genre information on the top DJs from Wikipedia. Out of the 420 artists present in the giant component, 251 have genre information, from a pool of 64 subgenres of electronic music, with 3.2 tags per DJ on average. Then I used a

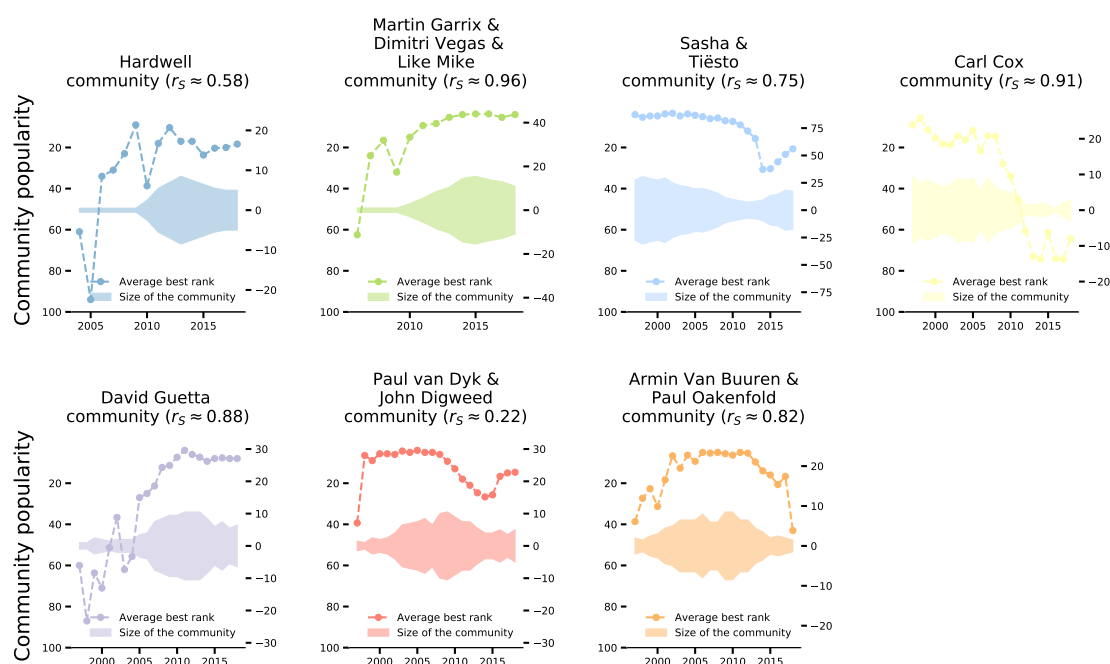


Figure 4.6. Temporal dynamics of communities in electronic music. Average popularity of the three highest-ranked DJs of each community (colored dashed lines), and the size of the community (shaded area) over time. Titles include the Spearman rank correlation r_s between the two quantities over time.

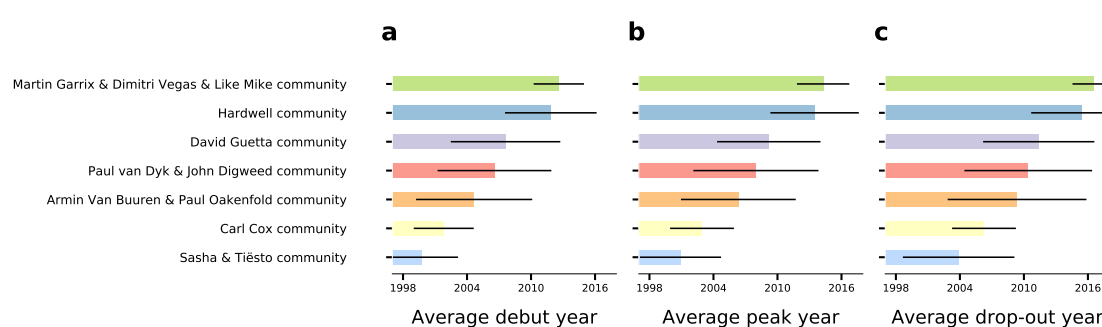


Figure 4.7. Average time scales of DJ communities. The average debut year, peak year (reaching the highest average rank), and drop-out year for the detected communities, different colors denoting the different communities.

mathematical formalization similar to Eq. (2.4) and reconstructed the genre-tag distribution of each community, characterized by the genre vector \mathbf{g}_i for community i such that $\mathbf{g}_{i,j}$ equals to the number of DJs in community i that are associated with genre tag j . For instance, if there are only two genres (e.g., techno and house music), and one community has 10 techno DJs, then its genre vector is $\mathbf{g} = (10, 0)$. However, if a community has five techno and five house DJs, then its genre vector is $\mathbf{g} = (5, 5)$. In this way, I computed the genre-similarity Γ of two communities, l and m , as the cosine-similarity of their genre vectors:

$$\Gamma_{l,m} = \frac{\mathbf{g}_l \cdot \mathbf{g}_m}{|\mathbf{g}_l| |\mathbf{g}_m|}. \quad (4.2)$$

These computed similarity scores are shown in Table 4.3. I found that the major genres in the newly emerging communities are usually moderately different, with an average cosine similarity of $\bar{\Gamma} \approx 0.445$, in agreement with recent results on changes in fashion trends [239].

Community ₁	Community ₂	Γ_{12}
Paul van Dyk	Armin Van Buuren & Paul Oakenfold	0.842
Sasha & Tiësto	Carl Cox	0.821
David Guetta	Martin Garrix & Dimitri Vegas & Like Mike	0.764
Sasha & Tiësto	David Guetta	0.691
Hardwell	Martin Garrix & Dimitri Vegas & Like Mike	0.61
Sasha & Tiësto	Armin Van Buuren & Paul Oakenfold	0.566
Carl Cox	David Guetta	0.558
Carl Cox	Armin Van Buuren & Paul Oakenfold	0.489
David Guetta	Armin Van Buuren & Paul Oakenfold	0.444
David Guetta	Paul van Dyk	0.443
Hardwell	David Guetta	0.414
Paul van Dyk	Martin Garrix & Dimitri Vegas & Like Mike	0.413
Paul van Dyk	Sasha & Tiësto	0.391
Carl Cox	Paul van Dyk	0.354
Hardwell	Paul van Dyk	0.342
Martin Garrix & Dimitri Vegas & Like Mike	Armin Van Buuren & Paul Oakenfold	0.324
Hardwell	Armin Van Buuren & Paul Oakenfold	0.276
Sasha & Tiësto	Martin Garrix & Dimitri Vegas & Like Mike	0.224
Sasha & Tiësto	Hardwell	0.166
Hardwell	Carl Cox	0.166
Martin Garrix & Dimitri Vegas & Like Mike	Carl Cox	0.138

Table 4.3. Cosine similarities of the genre distributions of top DJ communities.

The similarities show that the two most alike communities are both focused on trance and progressive, have a similarity score of $\Gamma \approx 0.84$, and are led by

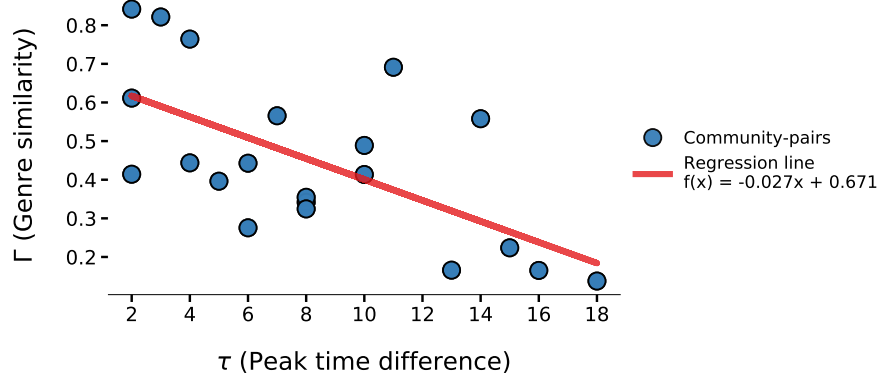


Figure 4.8. Genre similarities and peak time differences of DJ communities.

The graph shows the relationship between the debut year difference in years (τ) and genre-profile similarity (Γ) for each pair of communities with an average Spearman correlation $r_{\Gamma\tau} \approx 0.62$. The red line shows the results of a linear regression.

Paul van Dyk and Armin van Buuren. I also measured that these two communities are the closest in time, with average debut years of 2005 and 2006. In contrast, the two most different communities are led by Martin Garrix (joined in 2013) and Carl Cox (joined in 1997), with a similarity score of $\Gamma \approx 0.14$ and with more than a decade difference in typical debut years. While DJs in the former group are mostly playing house music, the latter is more focused on techno.

As these time differences already suggest, the further two communities peak (its DJs reach their highest average rank) from each other (at time $t_{p,i}$ for community i), the more different their genre profiles (\mathbf{g}) are. I quantified this effect by computing the time difference between peak years of the pair of communities l and m :

$$\tau_{l,m} = |t_{p,l} - t_{p,m}|, \quad (4.3)$$

and correlating these values with the genre-similarity score $\Gamma_{l,m}$. Visualized in Figure 4.8, I got an average Spearman correlation of $r_{\Gamma\tau} \approx 0.62$, supporting the claim that the closer two communities peak in time, the more similar their genre distributions are. After conducting a linear regression shown in Figure 4.8, it turned out that the typical relative change in genre-profiles is roughly about $\sim 3\%$ per year. The main trends of these genre differences, illustrated by genre tags, are summarized in Table 4.4: while in the late 1990s and early 2000s house and techno were the most popular genres, by the middle of the 2000s trance and progressive house started gaining popularity.

Overall, I can report that the top 100 DJs form different, temporarily separated communities, and these communities represent slight changes in musical trends. Each community typically has one or two leading figures, who are

Lead DJs (debut year)	Genre 1	Genre 2	Genre 3	Average debut year
Sasha (1997), Tiësto (2000)	house	electronica	techno	2000
Carl Cox (1997)	house	techno	electronica	2002
Armin van Buuren (2001), Paul Oakenfold (1997)	trance	progressive house	electronica	2005
Paul van Dyk (1997) , John Digweed (1997)	trance	progressive house	progressive trance	2006
David Guetta (2005)	house	electro house	progressive house	2008
Hardwell (2011)	hardstyle	progressive house	dutch house	2012
Dimitri Vegas & Like Mike (2011), Martin Garrix (2013)	electro house	progressive house	big room house	2013

Table 4.4. Genre distributions in DJ communities.

Name and debut year of the No. 1 DJs of each community, the three most frequent genres of the DJ communities, and the average debut year of the artists in each group.

one of the first and usually the most successful members of their communities. These observations suggest that top, central DJs act as gatekeepers by continually renewing the field of electronic music, and they shape both music trends and communities by bringing in new artists. In the rest of this chapter, I further explore the existence of such a mentorship effect.

4.5 Mentorship in electronic music

The previous results showed that most communities in the electronic music scene contain one or two No. 1 DJs who joined in the early stages of each community's life-cycle. How do these groups form? What are the major social forces shaping the DJ world? Do newcomers join existing groups independently, or are they more likely to be brought in by their former collaborators? In other words, does collaborating with the 100 DJs help new artists make it to the top 100 as well?

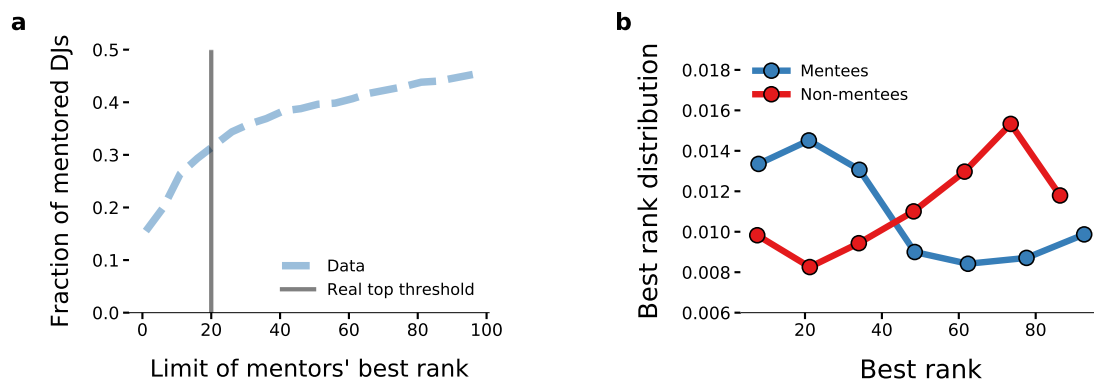


Figure 4.9. Mentorship in electronic music.

a, Fraction of DJs who have been mentored by artists with a best rank not lower than the limit rank measured on the horizontal axis. The vertical line represents the threshold of the top 20, who mentored more than 30% of all the mentored DJs. **b**, Comparison between the (percentile-binned) distribution of the best rank of the DJs who were mentored (blue line) and those who were not (red line).

4.5.1 The existence of mentorship

Known success stories and anecdotes, like the Rolling Stone magazine's take on Afrojack and David Guetta [252], suggest that mentoring plays an important role. In addition, most of the record labels of the star DJs have their demo-drop platforms to encourage young DJs' engagement [253, 254]. To investigate this hypothesis, also supported by findings introduced in Section 2.1, such as the work of Kram et al. [74, 75], I defined *mentorship* [76, 176] among top DJs in the following way. DJ₁ is the mentor of DJ₂ if they both made it to the top 100 ranking respectively at times t_1 and t_2 (with $t_1 < t_2$) and if they first appeared on the same release earlier than t_2 . My measurements revealed that about half of the DJs that ever made it to the top 100 had been mentored before, and about 30% of them were mentored by DJs with the best rank of 20 or better (Figure 4.9a). This implies that the role of the most successful individuals is central in community building by means of mentoring of new artists.

4.5.2 The benefits and limitations of mentorship

These results suggest that the most successful DJs build communities around themselves. Is this beneficial only for them, or does it also boost the expected success of their mentees? To answer this question, I compare the distribution of the best rank of top 100 DJs, differentiating between DJs that have been mentored before and those that have not. As Figure 4.9b shows, protégé DJs have a significantly higher chance of achieving top ranks, and a large fraction of them

even approaches the edge of the top 20. On the other hand, DJs who have not been mentored typically just show up at the tail of the top 100 and have much smaller chances of making it to the top 20.

One side of the formula is clear: mentorship boosts the expected success of newcomers, which aligns with previous findings on mentoring in science [175, 176]. However, the results also show a clear boundary between all-time stars and the rest, which makes one wonder whether star DJs are star mentors as well? I answered this question by comparing the average best rank of mentees to the best rank of their mentors. This comparison uncovered that protégés only profit slightly from having high-profile mentors since the mentees' expected best rank barely improves for highly successful mentors. This feature of mentoring was captured by the low correlation between the best rank of the mentors and the average best rank of their most successful mentees (Figure 4.10a). In other words, no matter how successful a mentor is, the expected success of their mentees is capped and is slightly below the real top, even for the best mentees.

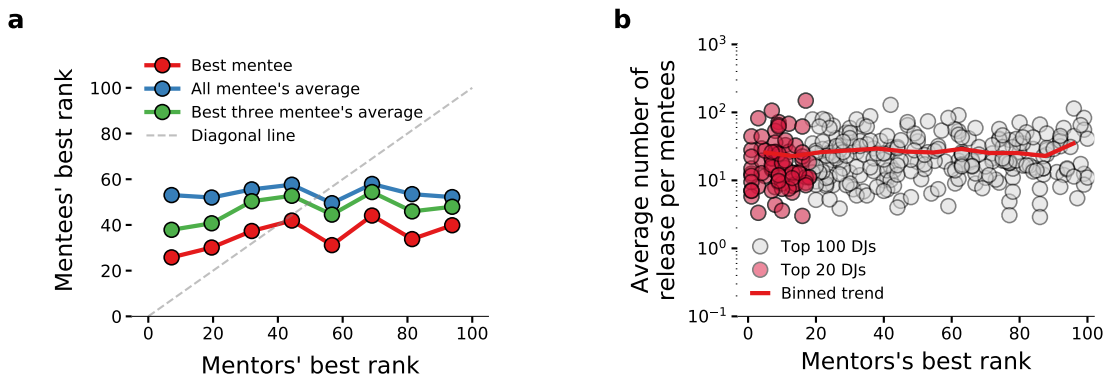


Figure 4.10. Mentors in electronic music.

a, Average best rank of mentees as a function of mentors' best rank. Mentees correspond to three groups: the best mentee of a mentor (red line, Spearman rank correlation $r_s \approx 0.199$), the average best rank of its the best three mentees (green line, $r_s \approx 0.192$), and the average best rank of all its mentees (blue line, $r_s \approx 0.035$). The diagonal line illustrates the case where mentees reach similar best ranks as their mentors. **b**, Number of releases normalized by the number of mentees for mentor DJs, expressing the frequency of their mentoring activities, and measured as a function of the mentor's best rank. Top 20 DJs are highlighted by red (Spearman rank correlation $r_s \approx 0.04$) and the rest by grey ($r_s \approx -0.12$).

In addition, comparing the number of mentees each DJ has, relative to the number of releases they produce, I measured a correlation as low as $r_s \approx 0.04$ for the top 20 DJs (Figure 4.10b). The fact that most mentees are mentored by top DJs is thus only due to top DJs being more productive. Therefore, seemingly, star DJs do not carry an extra 'star-mentor' effect; all DJs seem to follow the

same pattern and simply release more music when they collaborate more, which includes co-releases with new artists. A cumulative advantage process may be in effect to help top DJs by keeping their top positions, since the more successful DJs are, the more resources they have access to, which leads to higher chances of recruiting new mentees, as well as producing new releases.

Inspired by previous work on the effect of mentorship in scientific careers [51, 255], I further extended my analysis on the effects and the characteristics of mentoring on DJ success. It turned out that the most likely scenario is that mentees work together with the same mentor twice during their careers, as shown in Figure 4.11a. This finding is in close relation to the results on academic careers by Li et al. [255], which emphasizes the importance of rare, more precisely, “one-off” collaborations in academia. Testing the statistical significance of this difference of one vs. two-times collaboration between academia and electronic music is beyond the scope of this dissertation. However, a possible suggestion for more thorough analyses lay along the lines of the different production rates of the two fields.

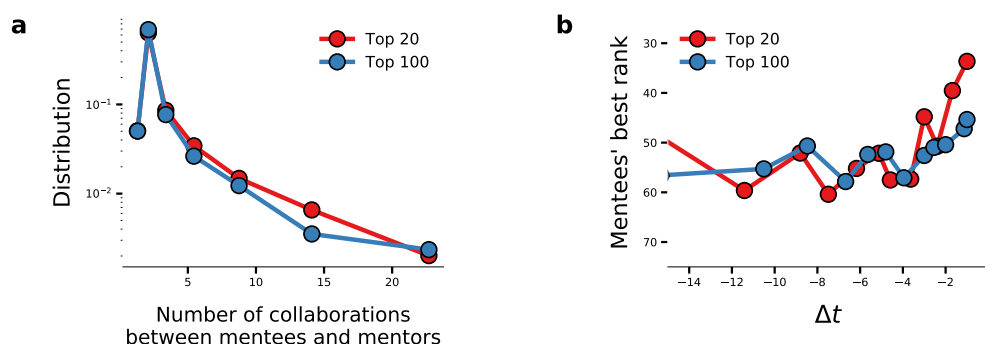


Figure 4.11. Mentee-mentor relationships.

a, Binned distribution of the number of times mentored DJs worked with the various mentors they had. **b**, Mentees' average best rank as a function of the entry-time difference Δt (experience gap) between every possible mentor and mentee pair.

Finally, I measured the effects of the experience gap between mentors and mentees on the mentees' success. I defined this experience gap as the top 100-entry-time difference between each pair of mentor and mentee who worked together:

$$\Delta t = t_{\text{mentor}} - t_{\text{mentee}}. \quad (4.4)$$

The distribution of Δt visualized in Figure 4.11b highlights that the best rank of mentees is insensitive to the experience gap for $\Delta t < -4$. I confirmed this by measuring the Spearman rank correlation between the entry-year difference

of mentors and mentees and the mentees' best rank for $\Delta t < -4$, which has a value of ~ 0.030 for top 20 mentors and ~ 0.027 for all the DJs. However, for $\Delta t > -5$, the graph shows a slightly stronger correlation of ~ 0.275 for top 20 DJs and ~ 0.121 when all the DJs are included. These results imply that only smaller entry-time (i.e., experience) differences boost the success of mentees, which may be rather related to the effects of joining trending communities than be the direct influence of the mentors themselves.

Taken together, the presented results point out that mentorship plays a vital role in the rise of new DJs and the growth of their prolific environment. Still, mentorship alone is not enough to explain the emergence of superstars. Such events seem to depend on (as of now) unknown mechanisms that cannot be inferred solely by an analysis of the music co-release network.

4.6 Discussion

Electronic music, one of the most popular music genres, has evolved into a complex ecosystem, with DJs and producers releasing and collaborating across multiple subgenres over the past two decades. Here I have investigated the temporal evolution of this field, focusing on how to pinpoint and distinguish a long-standing elite from the rest of the electronic music artists. I have also proposed potential mechanisms that could lead to the differences between elite musicians and less successful artists. First, I connected the dynamics of the top 100 ranking list of DJs to their underlying co-release activities to infer major driving factors of success. I reported that the historical top 100 splits into two distinct regimes in terms of the stability of their dynamics, showing the existence of a persistent elite in the DJ world. From collaboration patterns, I showed that those superstars who have reached the No. 1 position usually tend to lead segregated communities, which rise, peak, and fall separately over time, often representing changes in genres. I also saw that a major social force driving these communities is mentorship since new DJs usually join the top 100 after co-releasing music with already established artists. DJs who have been mentored before typically achieve significantly higher success, yet their chances of overcoming their mentors are slim. I observed that while star DJs exist, star mentors do not: the success of mentors has little influence on the expected success of their mentees.

While my results highlight interesting and major patterns in the growing ecosystem of electronic music artists, they have some limitations as well. The top 100 ranking of DJs reflects the opinion of a particular segment of electronic music fans, mostly limited to online platforms. Live shows and festivals, also a major platform of electronic music, are disregarded. This shortcoming may be

alleviated by incorporating data from social media and other music providing platforms (to have a less biased picture of the online landscape), or by using information about live shows, ticket and record sales, to connect this work with offline behavior. Another major question is how well these findings can be generalized to other genres. Are the observed phenomena particular to electronic music, or do rock, pop, and other musical genres follow similar trends? Since various rankings exist for other genres, such as in *Billboard Magazine* [95], and collaboration and co-release data are also available (for instance, on Discogs), most of this analysis could be replicable and may be tested soon.

Possible venues of related future research include an understanding of the differences between the trajectories of those who never make it to the top 100 against those who do and the analysis of the early-career patterns of these two groups. As a step further from descriptive analysis, an interesting direction is the development of predictive models that capture not only the next top 100 or No. 1 DJ's identities but also the next new entries: people who are already out there with the potential for becoming the stars of the next generation. An even more pressing issue is gender bias – the fraction of female DJs is shockingly low in the top 100, and this analysis could be used to track down the roots of this pronounced gender gap. I would also suggest to study musical features and extract various descriptors of the audio data itself, as well as combine the collaboration network with co-follow networks extracted from multiple social media outlets.

In this chapter and the corresponding publication, I have proposed a first attempt to understand the emergence of success in electronic music by obtaining quantitative findings on the existence and behavior of an exclusive elite of star DJs and producers. These results not only give insights into an interesting and vividly dynamic social system but also offer a good starting point for further research and policy suggestions. These include directions such as how to make electronic music more inclusive and less biased, help junior artists to be less exposed to long-standing stars, and take steps towards more merit- (and less business-based) spaces for artistic creativity.

CHAPTER 5

SUCCESS EVOLUTION OF URBAN VENUES

5.1 Introduction

In this chapter, I study the dynamics of the success of urban venues, such as restaurants, parks, and museums. To capture their popularity, I rely on large-scale data covering the area of Greater London from the location-based online social network application Foursquare City Guide [16]. I show that the aggregated overall popularity measures (for instance, the total number of visitors or likes a venue receives) are misleading because the venues' success trajectories can take distinctively different shapes over time, resembling substantially different success-outcomes. This implies that their success shows more complex patterns than previously thought and that even established venues with a similar popularity level can have completely different future perspectives. In particular, I find that venues can take six different popularity trajectories: they can either monotonously *rise* or *fall* over time, follow a *rise & fall*, or a *fall & rise* arc, show early *ephemeral* success or become *underdogs*.

Next, I map out the possible factors that could influence which shape a venue's trajectory will take alongside the following three dimensions. First, I elaborate on what different places have to offer to their visitors, such as gastronomical and restorative experiences [256, 257], or the possibilities of social

This chapter is based on the working paper "Six shapes of success", where Milan Janosov conducted the experiments and analyzed the results. All authors, Milan Janosov, Luca Maria Aiello, and Daniele Quercia conceived the experiments and contributed to writing the manuscript.

encounters [258–261]. In addition, I include both the effects of the pricing [262] of the venues and their uniqueness comparing to their competitors [263]. Second, I analyze the location of the venues and their impact on venue success. This includes the density, the popularity, the availability of different amenities, and the various functional roles of the different neighborhoods such as residential and metropolitan areas [264–269]. Finally, I cover several dimensions of the people who flow through these venues. Visitors can influence the success of urban spaces in many ways, such as the regularity of their visits, their financial allowances, and economic status, or their social status [270–277]. Taken together, I describe each urban venue by these three types of attributes: *i*) the venues’ characteristics (*what*); *ii*) the neighborhood and location of the venues (*where*); and *iii*) the clientele profile of these urban spaces (*who*).

Finally, I use machine learning models to capture the driving factors of urban success. To that end, first, I use binary classifiers to compare successful and unsuccessful venues. Then I build classifiers to capture the essential features predicting which of the six success shapes a given venue would take. My analysis reveals that the most predictive group of features in each case belong to the *who* category highlighting the importance of the urban venues’ social embeddedness.

5.2 Data

5.2.1 Data description

Foursquare City is a popular location-based online social network platform with $\sim 110\text{M}$ venues visited by more than $\sim 50\text{M}$ registered users worldwide [16, 278]. Foursquare provides a public API (<https://developer.foursquare.com>) to access their endpoints, which I used to collect the information about all the 178,321 venues and the 243,487 users that visited them in Greater London.

The application offers multiple features and actions for its users: they can like venues, write tips about them, upload multiple pictures by tagging the particular locations, and checking-in at venues. In addition, users can connect to other users and follow their activities – building an underlying social network and making the social fabric of the different cities visible. These user activities are also reflected in the level of venues in the aggregated number of user likes, tips, check-ins, and photos each venue receives over time, providing a natural way of measuring popularity on Foursquare.

The presented data collecting and processing procedure resulted in a database comprising information about both venues, users, and their neigh-

borhoods as follows.

Venue data. The basic information Foursquare provides about venues are:

- Geographical coordinates (latitude and longitude);
- Price range;
- Taxonomical category (e.g. Event, Food, Nightlife, etc.); and
- Time-stamped user interactions (likes, tips, check-ins, picture uploads).

User data. Foursquare provides different information-fields about its users as well. From those, I collected the profile of every Foursquare user who interacted with any of the venues at least once (e.g., liked or uploaded a picture). The available fields on each user profile are the following:

- Home location;
- Gender;
- Full list of time- and location-tagged activities; and
- List of friends.

To have a broader picture of the underlying social system of the Foursquare users, I extended my data collection to not only the users mentioned above but to all their friends as well and collected the same metadata about them, too.

Neighborhood data. I extended my database with neighborhood-level information from other sources:

- The number of inhabitants for the city of London [279]; and
- The Index of Multiple Deprivation (IMD score) that reflects the economic conditions of the neighborhoods [280].

I introduce these measures in detail in Subsection 5.4.

5.2.2 Data cleaning

In this analysis, I wanted to capture the popularity of the venues from the perspective of the local visitors; therefore, I had to filter out those users who were not residents in the city of London. However, only 28,327 of the 243,487 users listed London in their *homeCity* attribute, 74,217 of the users were specifically from other cities, and 140,943 users didn't specify their home location at all. Moreover, only 2,487 out of the 28,327 Londoners shared their exact coordinates for their home location, which I used as ground-truth information for inferring the home coordinates of the rest of the users for the upcoming feature engineering processes.

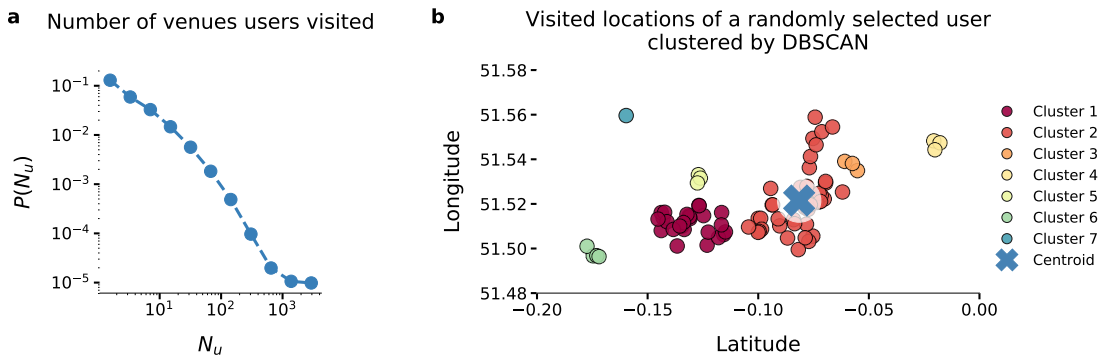


Figure 5.1. Visitors' venue distribution and the DBSCAN method.

a, Shows the probability distribution of the number of venues each user visited (N_u). **b**, The panel illustrates the DBSCAN algorithm [281] on an example of a user's visited venues, where the algorithm found seven distinct spatial clusters marked by different colors. The largest turned out to be Cluster 1, and the estimated home location of the user was the centroid of it marked by a blue cross.

While users typically visit multiple venues (Figure 5.1a), only 2.8k Londoners shared their home coordinates, illustrating that users usually don't show Foursquare activities at home. Therefore, to infer the home location coordinates of the remaining users, I followed a quantitative approximative approach inspired by previous works by Cheng et al. [270] and Noulas et al. [261].

First, I had to decide whether those 140,943 users with unknown home cities are from London or not. Second, if they were from London, I had to approximate their home coordinates. For these, I used the DBSCAN spatial clustering algorithm and its Scikit Learn implementation [281, 282] to detect the spatial clusters of the visited venues of the users. Next, I picked the largest cluster for each user and considered that to be the home area of those particular users. As the final step, I assumed the home location (most likely location on average) to be the centroid of that cluster, and kept those users whose centroid fall within

the boundaries of London. The DBSCAN method is illustrated in Figure 5.1b by using the set of locations a randomly chosen user visited.

DBSCAN has two parameters given a distance metric, which I chose to be Euclidian for the sake of simplicity: *eps* describing the maximum distance between two points so that they can be considered to be in the neighborhood of each other, and *min* representing the number of points in a neighborhood for a point to be considered as a core point, including the point itself. To pick the most suitable parametrization of the DBSCAN model, I tested various values of *eps* in the range of (0.0005, 0.2) and the values of *min* in the range of (2,10) and estimated the home location of the 2,487 users for which I had ground-truth information. I found the error between the estimated and ground-truth coordinates, measured as the Euclidian distance between these two points, to be minimal at *eps* = 0.02 and *min* = 3, and without setting any thresholds to the minimum number of venues a user visited. Later, I used these settings approximate the home location of the rest of the users. With this particular parametrization of the DBSCAN model, I managed to associate spatial coordinates to all the Londoners, which represent the most likely spatial location of the individuals, and arriving to a set of 101k users with London-based home coordinates.

5.2.3 Measuring urban success

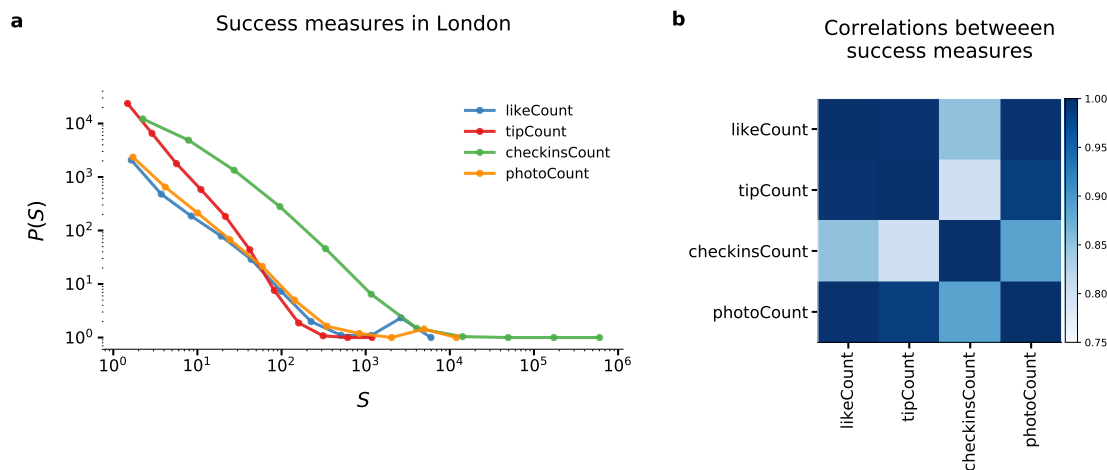


Figure 5.2. Measures of urban success.

a, The figure shows the distribution of the different success measures S (number of likes, check-ins, tips, and photos) on Foursquare for the city of London. **b**, The matrix shows the pairwise correlations between the four available success measures for London.

Foursquare provides several ways to measure the popularity of the urban venues present on their platform based on user activities [260]. These are the

number of users that checked-in at certain venues, the number of tips given about the venues, the number of photos uploaded, and the number of likes received. These measures, similarly to other measures such as the citation and play count distributions in the previous chapters, are based on the response of a wider audience, and their distributions follow heavy-tailed functions (Figure 5.2a). Besides, since all these quantities depend on the typical number of users that interact with a venue, they are all highly correlated (Figure 5.2b) with an average correlation of 0.919.

While the statistics in Figure 5.2a shows aggregated information, the dataset I used includes timestamps on each activity at a sub-second resolution, allowing us to reconstruct the popularity time series of the different venues over time. I carried out the present analysis by using like count as the success measure (unless noted otherwise); however, as the correlations of Figure 5.2b suggests, this particular choice should not influence the key findings.

5.3 Success trajectories

In this chapter, I aim to better understand the temporal properties of success, namely, the shapes of the popularity *trajectories* of the different venues (similarly to the age-curves of creative careers). To analyze that, first, I defined and transformed the venues' *trajectories* as popularity time series. Then, I used machine learning techniques to extract and quantify the different shapes of urban success.

5.3.1 Measuring and transforming success trajectories

Venue success, e.g., the number of likes a venue i received at time t , is a variable denoted by $S_i(t)$ in the interval of time $t \in [0, T]$. While the temporal resolution of my data is on the scale of seconds, this resolution does not directly provide information on the much slower life-cycles of urban venues. Therefore I binned the time dimension into time slots of 6 months, which yields up to 11 points on my dataset's duration of 5.5 years. In addition, I restricted the analysis to established venues that have been present for at least 2.5 years (5 semesters) in the studied dataset, which resulted in 21,758 venues. To this end, I denoted the semester-binned temporal popularity by $s_i(t)$. Next, to rule out the temporal changes in the popularity of the platform itself from the popularity of the venues, I normalized each time series by the average popularity of all the venues in each time-bin, accounting for the inflation of the popularity measure

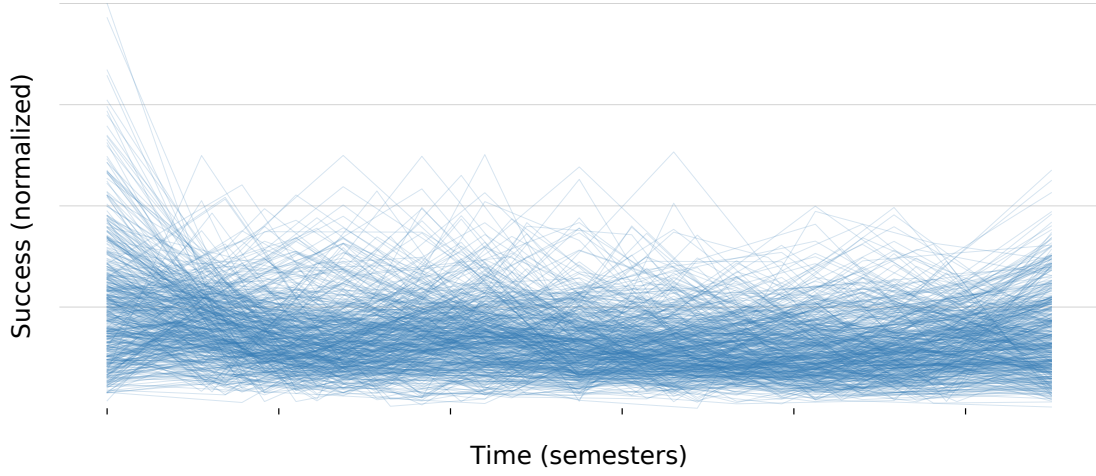


Figure 5.3. Success trajectories of urban venues.

The figure shows the transformed, discretized and normalized popularity trajectories of 500 randomly picked venues with shaded lines.

as the platform gained and lost popularity over time:

$$s_i(t) := \frac{s_i(t)}{\sum_{t=0}^T s_i(t)}. \quad (5.1)$$

Finally, as I intended to capture general trends across all kinds of venues disregarding their raw popularity, I had to ensure that their popularity trajectories are comparable by re-scaling each trajectory's values by its mean:

$$s_i(t) := \frac{s_i(t)}{\frac{1}{|V|} \cdot \sum_{j \in V} s_j(t)}. \quad (5.2)$$

After conducting these initial transformations, the venues' success trajectories show little regularities at first (Figure 5.3). In the following subsection, I am going to use machine learning techniques to show how to extract an underlying, hidden structure from these seemingly random time series.

5.3.2 Clustering success trajectories

Here I introduce the clustering of the 21,758 senior venues of London based on their normalized popularity trajectories ($s_i(t)$) in an unsupervised fashion to see whether particular trends emerge.

First, I had to find a way to measure the similarity (distance) between the popularity time series. For this, I used Dynamic Time Warping (DTW) [283], a

widely used method in signal processing and time series clustering [284]. The DTW method finds the best temporal alignment between two time series of arbitrary lengths and estimates the lowest distance between those two.

Building on this distance metric, I used hierarchical clustering [285, 286], an agglomerative clustering technique that initially assigns each time series to a cluster of its own, and then combines them hierarchically by merging the closest pairs of clusters based on their DTW distance at every iterative step until they all belong to one giant cluster. The closeness of the clusters can be captured in different ways. Here I used the complete-linkage variant of this algorithm that measures the distance d between two clusters C_a and C_b as the similarity of their most dissimilar members [287]:

$$d(C_a, C_b) = \max_{s_i \in C_a, s_j \in C_b} DTW(s_i, s_j). \quad (5.3)$$

Compared to other linkage methods, the complete-linkage variant typically results in more compact clusters [288, 289].

The most important shortcoming of hierarchical clustering is the difficulty of estimating the number of clusters that naturally emerge from the data [290]. To resolve this problem, I used the so-called gap statistics [291], which determines the quality of clustering by comparing the instances' homogeneity within the measured clusters against randomly assigned cluster identifiers. The gap statistics indicated ten as the optimal number of groups. However, upon visual inspection, it became apparent that while the success time series within clusters were very homogeneous, some pairs of clusters contained time series with very similar trajectories. Therefore, I merged the ten clusters into six higher-level clusters in an algorithmic fashion (corresponding to the six distinct shapes) using the following definitions:

(1) Ephemeral and (2) Underdog. Clusters whose vast majority of success trajectories ($> 75\%$) are mostly flat except for a sharp initial drop (Ephemeral) or a final peak (Underdog). Those curves are characterized by high differences between *i*) the first and last values in their popularity time series and *ii*) the variance σ of the values over the first and last thirds of the time series. For the Ephemeral group, this is expressed as: $s_i(t_{start}) \gg s_i(t_{end}) \wedge \sigma(s_i(t_{start}), \dots, s_i(\frac{1}{3} \cdot t_{end})) \gg \sigma(s_i(\frac{2}{3} \cdot t_{end}), \dots, s_i(t_{end}))$. In addition, I used the same formula for underdogs but by replacing \gg with \ll to capture the increasing trend.

(3) Rise and (4) Fall. Venues that's success trajectories are dominated ($> 75\%$) by steadily increasing (rise) or decreasing (fall) trends. To uncover these linear relationships, I used a linear regression to model the time slots of a series (t_{start}

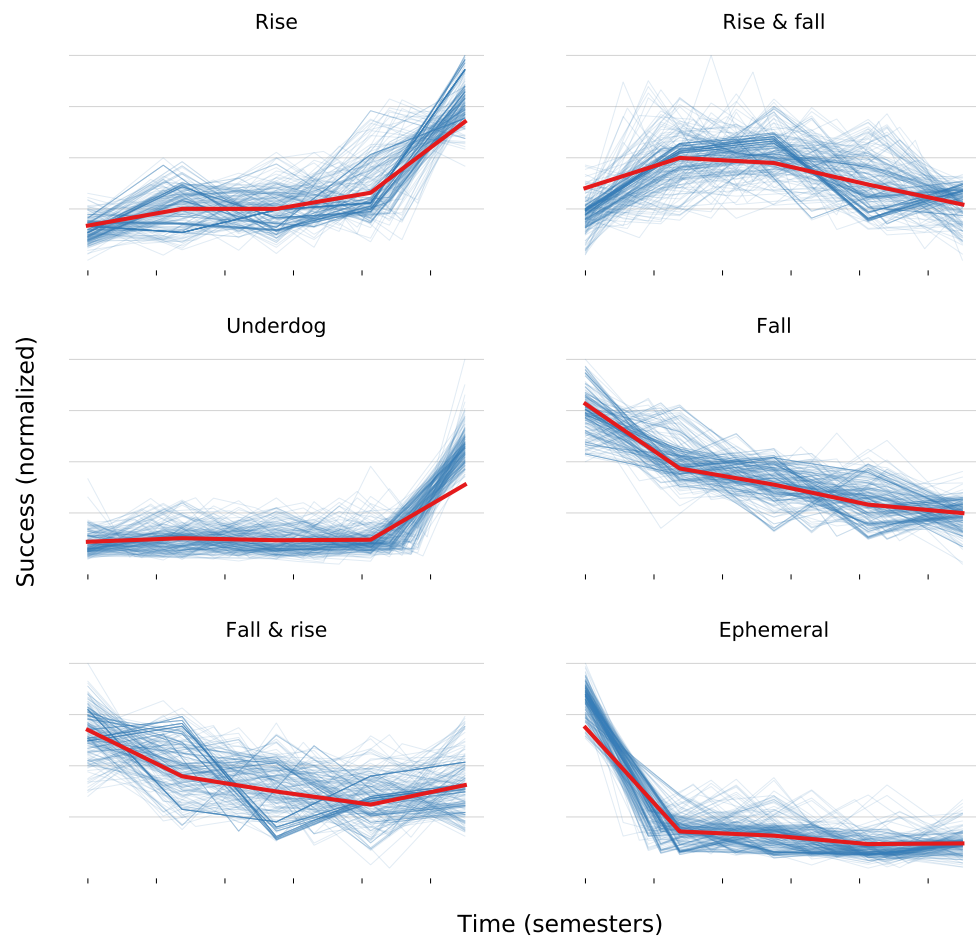


Figure 5.4. Clustered success trajectories.

The figure shows the six typical success trajectory clusters by visualizing 200 random examples from each. The centroids of the clusters are marked by the red line, while the individual venues' popularity time series are shown in blue.

to t_{end}) as a function of the success values at those times ($s_i(t_{start})$ to $s_i(t_{end})$), namely: $t \sim \alpha + \beta s_i(t)$. When the resulting R^2 was higher than 0.75, then due to the strong increasing linear relationship between time and success I classified the venue into the Rise/Fall category based on the sign of the regression line's slope.

(5) Rise & fall and (6) Fall & rise. Clusters of venues whose vast majority of success trajectories ($> 75\%$) can be fitted by a U-shaped curve [292] (Fall & rise) or an inverse U-shaped curve (Rise & fall) with $R^2 > 0.75$.

With these definitions, I managed to classify all the relevant venues into one of these six shapes, which are visualized in Figure 5.4.

5.4 Describing urban spaces

My analysis reveals that the success trajectories of urban venues follow six distinct shapes. From this, the question arises: what decides which of these six shapes is a certain venue's trajectory going to take? What are the most important features deciding whether a venue will be an all-time classic or disappears shortly after it opened? To answer these questions, I explored three different directions on potential driving factors. These three families of features capture the nature of the venues (*what*), their location and neighborhood (*where*), and the people who visit them (*who*). Most of these features can directly be derived from the data provided by Foursquare, while in some cases, I also included external sources of data.

5.4.1 Features describing the venues' characteristics (*what*)

Venue category. The top-level category of the venues in the Foursquare taxonomy, which takes a categorical value from Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. This feature gives basic information on what the different venues have to offer to their visitors.

Price category. This is the price tier of the venue, on a scale from 1 to 4 (least to most expensive), as determined by Foursquare (<https://developer.foursquare.com/docs/api/venues/managed>). A venue's price has effects on its popularity in several ways. First, the more affordable a venue is, the larger the pool of potential customers is. Second, the price can easily have a priming effect

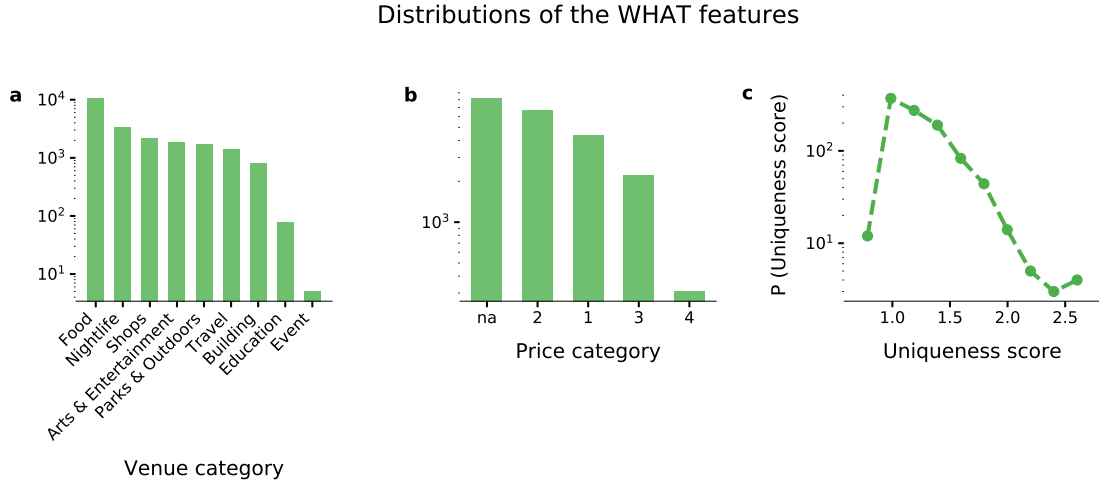


Figure 5.5. Urban features describing the venues' characteristics (*what*).
a-b, The histograms of the senior Londoner venues' categories and price ranges. **c**, The distribution of the categories' uniqueness score of these venues.

on how people judge their experience as well by putting venues into a different niche [262].

Uniqueness score. A necessary condition for commercial competitive advantage is heterogeneity [263]: a venue's offer must differ from competing businesses in the same area. To capture these aspects, I defined the uniqueness score of a venue i in an area a in the following way. First I calculated the average Euclidean distance between the location of i and the locations of all the venues in a that belong to the same category of i , and then normalized by the average distance between i and of all the other venues within that area, to account for venue density:

$$uniqueness_i = \frac{\langle d(i, j | \text{category}_i = \text{category}_j) \rangle_j}{\langle d_{i,j} \rangle_j}. \quad (5.4)$$

The distributions of these features are shown in Figure 5.5.

5.4.2 Features describing the venues' neighborhoods (*where*)

The location of a venue impacts its success opportunities for various reasons, such as population density and economic development. For instance, it has been shown before that the composition of Foursquare venues in an area is predictive of its economic deprivation [293]. To quantify these effects, first, I defined the venue areas using the official geographic partitioning provided by

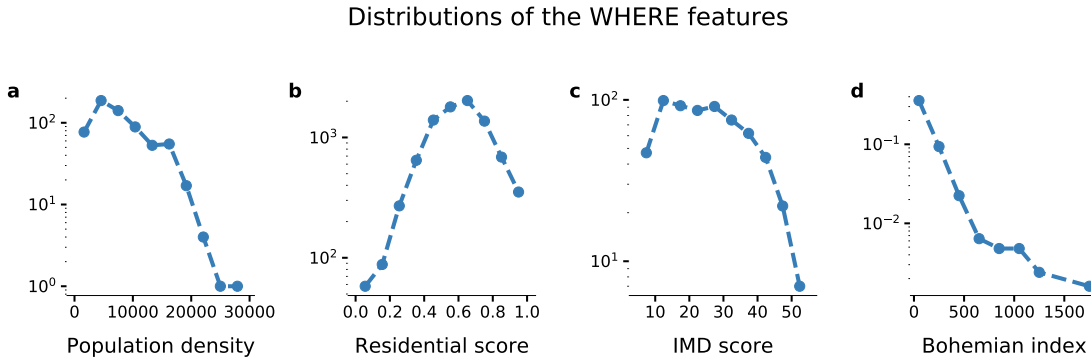


Figure 5.6. Urban features describing the venues' neighborhoods (*where*).

The panels show the probability distributions of the measures characterizing urban neighborhoods (on the level of Wards) and individual venues' locational attributes for the city of London.

the census of the UK. In the case of the city of London, these units are called Wards, 638 areas defined by the Office for National Statistics [279] and designed to contain about 13k residents each on average. To account for these mentioned dimensions of neighborhoods, I used the following measures:

Population density. Population density of urban areas fosters the flow of mobility and information as well as interpersonal interactions and increasing opportunities for businesses [265, 294]. Therefore, population density is a key component for the development of different functional areas [266]. To measure the population density, I gathered the number of residents and the surface of the area of each Ward from the census data of London [280], and computed population density as the number of inhabitants per square kilometer.

Residential score. Due to its significantly different role for not being a social place, I decided to handle residential areas separately. To assess the extent to which a venue i is in a residential area, I computed the average distance of i from other Foursquare venues under the category *Residence* within the same area i is located in:

$$residential_i = \langle d(i, j | \text{category}_j = \text{residential}) \rangle_j. \quad (5.5)$$

IMD score. I relied on the UK Index of Multiple Deprivation [280] (IMD), available at Ward level for London. The IMD score is a composite measure of deprivation across several domains, such as education, barriers to housing, crime, employment, and access to healthcare. I collected this index for the year

2015, which falls about in the middle of the period captured by my Foursquare data.

Bohemian index. The Bohemian index [268, 269] uses occupation data to estimate the bohemian population of a given area [295]. It is computed as the fraction of people employed in arts, entertainment, and recreation over the total population of residents.

The distributions of these measures are shown in Figure 5.6.

5.4.3 Features describing the venues' visitors (*who*)

The success of venues depends just as much on where they are located, and what they are offering to their visitors, as on the people who decide to visit them. To capture these social aspects, I extracted several features describing the clientele-profile of the Londoner venues and aggregated these measures to the level of the venues themselves.

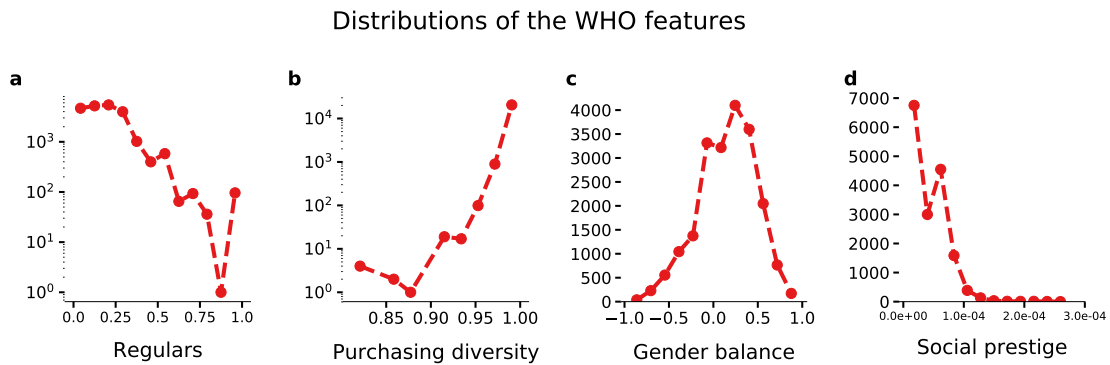


Figure 5.7. Urban features describing the venues' visitors (*who*).

The panels show the probability distributions of the measures characterizing the venues' visitors.

Regulars. It has been reported before that a steady influx of regular visitors secures a consistent level of popularity over time [270, 271]. Therefore, I estimated the ratio of the returning visitors by the frequency of their photo-uploads (since this the most typical repeated action). I computed the regularity score as the ratio between the number of users who uploaded photos of the venue i in at least two distinct days ($|U_i^{pics2+}|$) and the total number of users who uploaded

photos of the venue ($|U_i^{pics}|$):

$$regulars_i = \frac{|U_i^{pics2+}|}{|U_i^{pics}|}. \quad (5.6)$$

Purchasing diversity. It also has been shown that liveliness of the urban fabric benefits from mixing people of varied social and economic extractions [272, 273]. To this end, I measured the purchasing diversity of the visitors of the different venues in the following way. For every user u , I computed the probability density function of the price ranges of the venues u interacted with through likes, tips, or photos. As a result, u is described with a four-dimensional *user price vector* $A_u = [\alpha_{u,1}, \alpha_{u,2}, \alpha_{u,3}, \alpha_{u,4}]$, $\sum_j \alpha_{u,j} = 1$, whose entries represent the probability of user u visiting a venue of price range from 1 (cheapest) to 4 (priciest). Given a target venue i , I obtained a *venue price vector* C_i by summing the price vectors of all the users who interacted with i at least once (U_i):

$$C_{i,k} = \sum_{u \in U_i} \alpha_{u,k}, \quad (5.7)$$

and by normalizing it I obtained a probability distribution:

$$\bar{C}_{i,k} = \frac{C_{i,k}}{\sum_j C_{i,j}}. \quad (5.8)$$

From this, I defined the diversity of purchasing power of venue i 's visitors by computing the Shannon entropy of C_i :

$$diversity_i = - \sum_k \bar{C}_{i,k} \cdot \log \bar{C}_{i,k}. \quad (5.9)$$

Gender balance. Since the advantages of bringing people together from different walks of life are good predictors of the socio-economic development of the neighborhoods [296], I decided to include the most basic meta-information I have about Foursquare users: their gender. I estimated the gender balance of venue i 's visitors by using the previously introduced measure of *average gender* [297]:

$$gender_i = \frac{1}{|U_i|} \sum_{u \in U_i} g_{i,u}, \quad (5.10)$$

where $|U_i|$ denotes the number of visitors of venue i , and $g_{i,u} = \{-1, 1\}$ encodes the gender (female, male) of visitor u at venue i . When $gender_i$ is negative, the visitors are more predominantly female.

Social prestige. The *centrality* of the actors on social networks [277] is a notion of paramount importance to model such phenomena, as it captures “central” actors who play a key role in gathering and relaying resources such as attention or information. From this, I assumed that places that attract high-centrality people get higher chances of increasing their popularity by leveraging on the network of these central individuals.

To measure the individuals’ network centrality, I collected information about all the users U_0 that have interacted with venues in London and augmented this set of users by their friends on the Foursquare social graph G . I then retained only the subset $U_{res} \subset U$ of users who are resident in the city of London and restricted the graph of interactions to the social links among them ($G_{res} = \{(u, v) | u \in U_{res} \wedge v \in U_{res}\}$). Using the friendship graph G_{res} , where each user is embedded in the physical space via its predicted home coordinates, I computed the distance-weighted PageRank centrality [117] of all visitors of venue i to model their social prestige. The network G_{res} is visualized on Figure 5.8. Finally, I defined the social prestige of venue i as the average centrality score across all visitors of venue i :

$$socialprestige_i = \frac{\sum_{u \in U_i} PageRank(G, u)}{|U_i|}. \quad (5.11)$$

The distributions of these features are shown in Figure 5.7.

5.4.4 Features summary

Table 5.1 summarizes the features described in detail in this subsection covering three different domains of urban venues along the question of *what* are these venues offering, *where* are they located, and *who* are their typical visitors.

Category	Measure	Short definition
What	Venue category	The primary category of the venue.
	Price category	The price category of the venue.
	Uniqueness score	The relative density of the venues’ category within its neighborhood.
Where	Population density	Number of inhabitants per square kilometre.
	Residential score	Average distance of residential venues.
	IMD Score	Score for living environment conditions.
	Bohemian index	Fraction of people employed in arts, entertainment and recreation.
Who	Regulars	Fraction of returning visitors.
	Purchasing diversity	Diversity in purchasing power of the venues’ visitors.
	Gender balance	The relative ratio of the number of people of different genders.
	Social prestige	Average pagerank of the venues’ customers from their social network.

Table 5.1. Summary of urban features

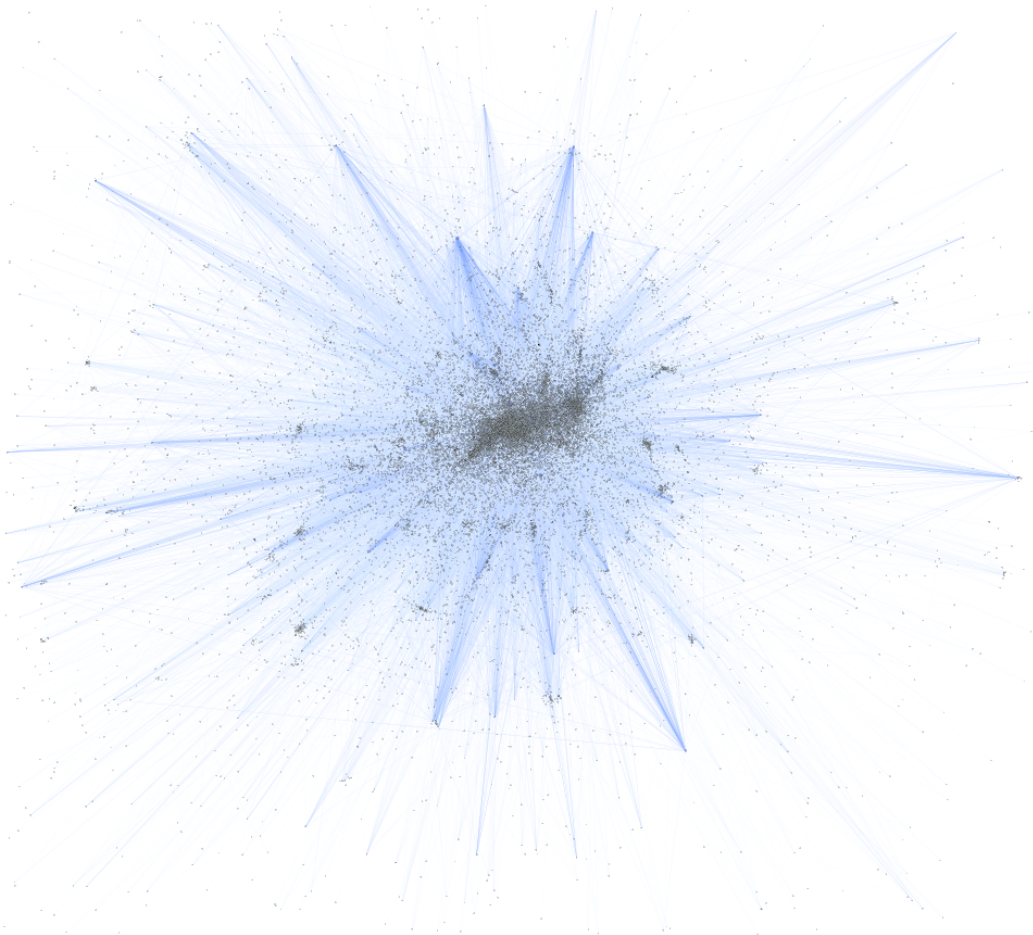


Figure 5.8. The social graph of Foursquare users resident in London.

The nodes of the network represent the different Foursquare users that have given or predicted home coordinates within London. The visualization layout is placing the nodes on a plane based on their estimated home coordinates (latitude and longitude), while the color of the nodes shows the values of their PageRank centralities, the nodes' sizes are proportional to their degree (number of connections a user has).

5.5 Predicting venue success

In this section, I use the extracted features to first classify venues as successful and unsuccessful ones in a binary way, and then show how the different features can predict the shape of the venues' success trajectories.

5.5.1 Binary success classification

I first built a predictive model to differentiate between successful or unsuccessful venues. For this, I considered a venues' cumulative success, calculated as the total number of likes it received during the observational period T :

$$S_i = \sum_{t \in [0, T]} S_i(t), \quad (5.12)$$

based on the terminology of Subsection 5.3.1. I repeated the same binarization with check-in, and tip counts as well to generalize my findings.

Next, I split the venues into quartiles based on their S_i values' place in the global $P(S_i)$ distribution and discriminated between venues that fall into the top quartile of the success distribution (successful, positive examples) and those falling in the bottom quartile (unsuccessful, negative examples). This formulation effectively prunes the middle quartiles and makes it possible to focus on the classification of venues whose success pattern is distinct from the average case. By this procedure, I also ensured a perfectly balanced ratio of positive to negative examples in each class.

For the prediction, after trying several simpler models (e.g. decision tree, random forest), I trained an XGBoost [298, 299] classifier, a classifier that proved itself most effective in a wide range of classification and regression tasks [300]. XGBoost is based on tree boosting: an ensemble of 'weak' decision trees—called *estimators*, that combined yield very accurate predictions. I ran XGBoost with: 250 estimators; a maximum depth of the individual decision trees of 5; a learning rate of 0.1 (i.e., the rate at which the influence of old trees is reduced in successive iterations of the model to prevent overfitting), which was proved to lead to good generalization error [301]; and a subsample size of 0.8 (fraction of training data that is randomly subsampled before growing each tree).

Next I used this XGBoost classifier with the *what*, *where*, and *who* features summarized in Table 5.1. As the XGBoost is robust against scaling, the only pre-processing I carried out is the one-hot-encoding of the only categorical feature, venue category. During the classification, I measured the performance of the model through average accuracy ($accuracy = \frac{\# \text{ correctly classified venues}}{\# \text{ total venues}}$) over 10-fold cross-validation, which is an appropriate measure since I have a balanced

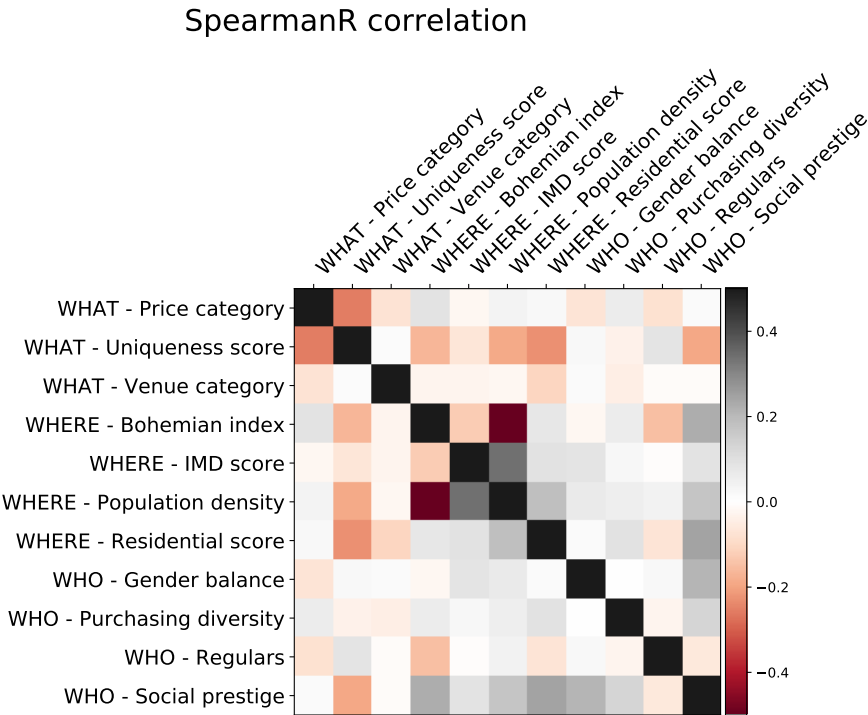


Figure 5.9. Correlations between urban features.

The correlation matrix shows the relationship between the derived urban features (Table ??).

sample and no particular distinction between the two classes. In addition, I tested the co-linearity of the features and found that they show minor redundancies: a cross-correlation analysis shows most signals are orthogonal (Figure 5.9), with slight (either negative or positive) correlations between features belonging to the same feature groups.

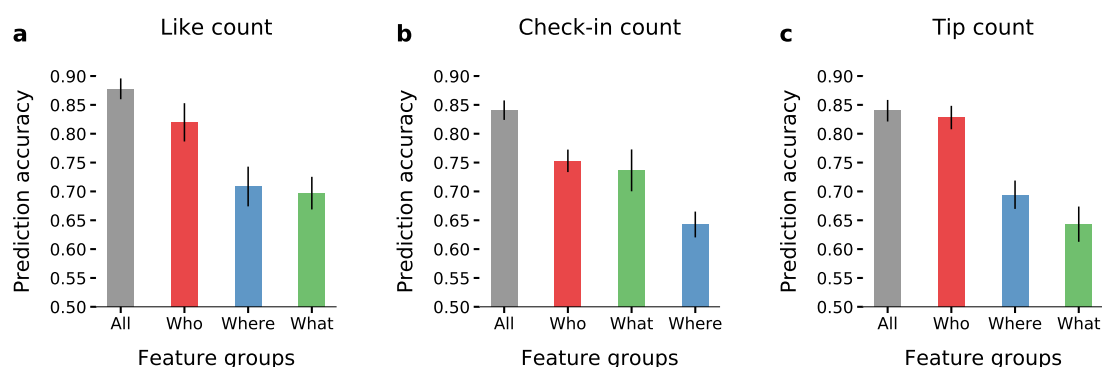


Figure 5.10. Binarized success prediction results.

The figure shows the prediction accuracies of the binary classification of venue success where success is measured as **a**, the number of likes **b**, the number of check-ins **c**, and the number of tips the venues received.

The prediction results, summarized in Figure 5.10, indicate that it is possible to precisely tell apart successful and unsuccessful venues: on average, the model classifies correctly $\sim 85\%$ of the instances when relying on all the features, regardless of the success measure of choice (e.g., like count, check-in count, tip count). The performance drops considerably when training the model on only the *what* or *where* features. More interestingly, a model trained on the *who* features only suffers only a slight performance drop compared to the full model.

I investigated the practical functionality of the XGBoost: the estimation of the importance of the different features when predicting the number of likes, check-ins, and tips each venue received. The ten most predictive features are visualized in Figure 5.11. The results show that indeed the most predictive features concern the clientele of the urban venue. First and foremost, their purchasing power and the diversity of it. Another major predictor is the fraction of regular visitors, which is a good signal on how well the successful venues manage to build up their pool of visitors and keep their engagement high. In addition, in the case of check-in counts, which is probably associated with more relaxed activities, such as sitting down for dinner or lunch, are very typical at venues serving food. On the other hand, the location features show only moderate predictive power, among which, the Bohemian-index performs the best.

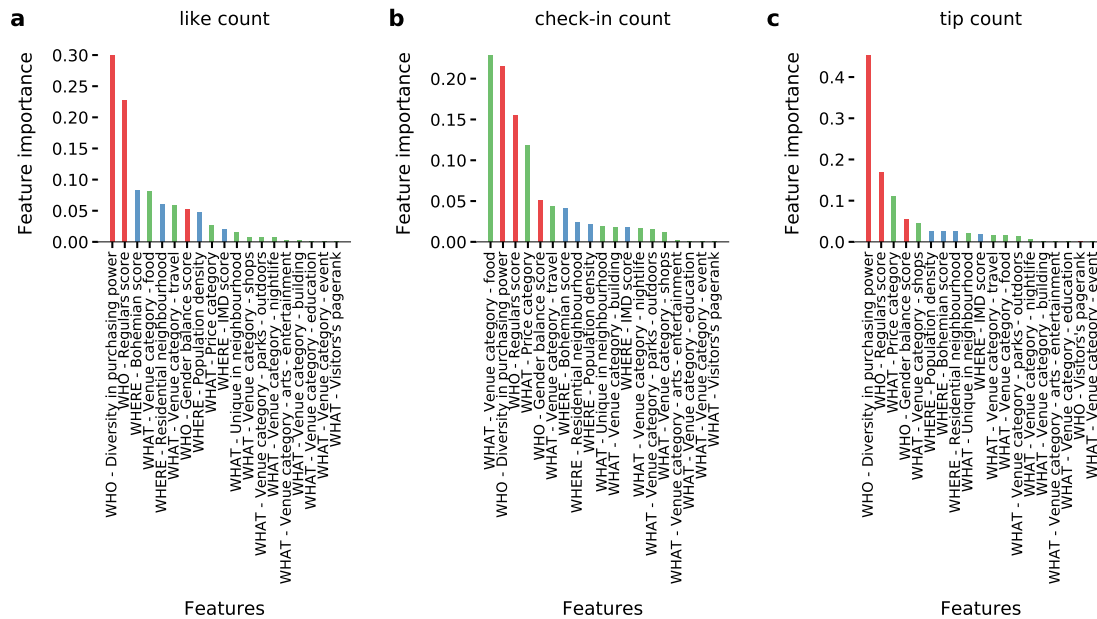


Figure 5.11. Feature importances in binary prediction.

The figure shows the relative importance of each feature in the case of binary success classification and different measures of success.

This suggests that the most successful areas are the trending, artistic, so-called “hipsterish” parts of London.

5.5.2 Six-shape prediction

	6 classes	5 classes	4 classes
Number of venues	900	4290	7128
Prediction accuracy	0.287	0.239	0.267
Random accuracy	$1/6$	$1/5$	$1/4$
Prediction over random	72%	19%	7%

Table 5.2. Six-shape prediction accuracies.

The number of venues, the prediction accuracies and the random baseline values for the cases of using all the six shapes, and when I drop the first and second least frequent ones.

In this subsection, I show predictive results on how the previously introduced urban features may signal the different success trajectories each venue will follow over time. I compared several cases on shape-prediction as the number of different venues in different classes was not equally distributed, and I aimed to make balanced predictions. Therefore I used all the six shapes,

then dropped the least common, and then the two least common shapes and repeated the predictions.

To carry out this prediction, I used a similar setup to the binary classification of Subsection 5.5.1 with the XGBoost classifier, with the difference of using multiple class labels. Namely, I used 250 estimators and 5-fold cross-validation. In addition, I carried out a grid-search on the parameter range of *max_depth* : [4,5] , *learning_rate* : [0.1, 0.2, 0.3], and. *subsample* : [0.7, 0.8, 0.9]. I summarize the prediction accuracies in Table 5.2, showing that surprisingly I got the most accurate model with the largest number of classes. While the reported accuracy scores are not high on an absolute scale, this is the first study ever conducting such measurements, and my best results still outperform the random cases by more than 70%.

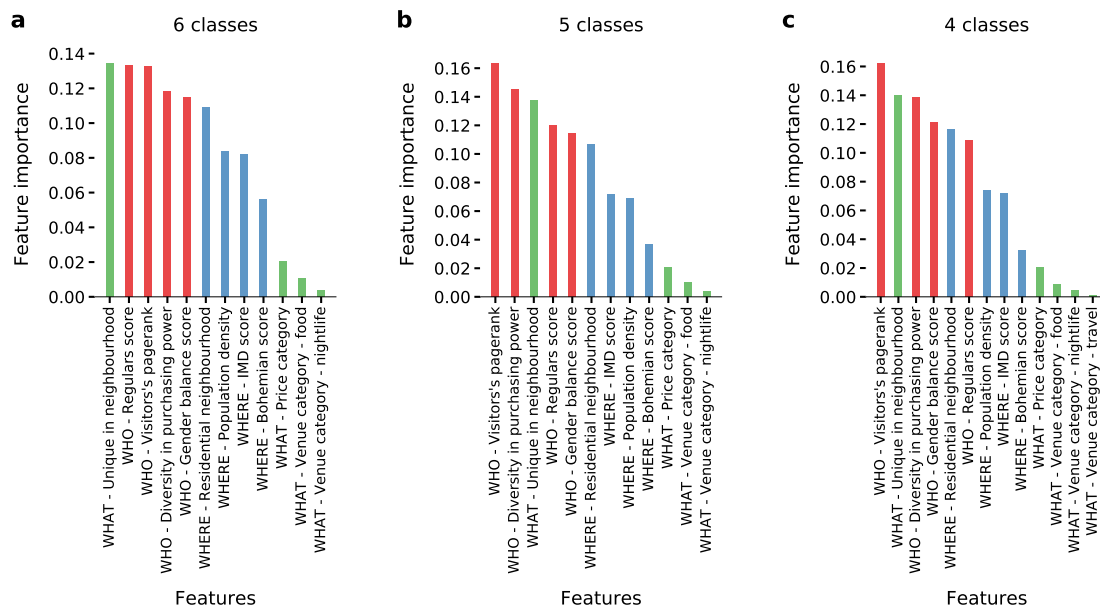


Figure 5.12. Six-shape success prediction results.

The figure shows the feature importances in case of the success-shape prediction where **a**, I used all the six classes, or down-sampled the data set to **b**, five **c**, or four classes.

By studying the relative feature importances in the success-shape prediction, my results highlight the particular role of the social capital of the venues even further. On the one hand, figure 5.12 shows that in each classification setup, there were only either one *what* or *where* features in the top five. For instance, the venues' uniqueness, and the residential score, fitting into previous theories [258–261, 263]. On the other hand, all the rest of the top five features were along the social dimension with relative feature importance scores higher than 10%. In particular, a venue's social prestige seems to play the most important

role in deciding which trajectory its success takes, pinpointing the importance of the social fabric each venues' visitors are embedded into. On the contrary, category-related features, for instance, the venue price range or the food category, turn out to be a poor predictor with importance lower than 2%. Finally, all the features describing the neighborhood, such as the Bohemian-index or the IMDB score average at the importance of around 5 – 7%.

5.6 Discussion

In this chapter, I built on the location-based online social service Foursquare City to understand the dynamics of the successful evolution of urban venues of various kinds. First, the temporal clustering of the normalized popularity trajectories (success shapes) of the urban venues revealed that urban success is not binary, but it has six distinct shapes: venues' popularity time series can simply rise or fall, rise & fall or fall & rise, show ephemeral success or turn out to be underdogs.

Second, I collected different urban features based on the literature that can characterize urban venues alongside three dimensions. I named these dimensions as *i) what*, describing the characteristics of the venues themselves (e.g., venue category or price range); *ii) where*, summarizing the typical features of the neighborhood of the venues (e.g., economic development); *iii) and who*, a group of features representing the clientele and social embeddedness of the venues (e.g., fraction of regular visitors).

Third, I conducted a machine-learning-based predictive analysis to estimate the importance and predictability of urban success. I showed that when I consider success to be binary, then whether a venue is going to be successful or not can be predicted with a relative accuracy of $\sim 85\%$. Moreover, my analysis showed that the most predictive feature group is the *who* features describing the social environment of the venues. Furthermore, my predictions on the six shapes of success with an accuracy 70% higher than the random case reveal that the importance of the social features is just as pronounced when it comes to predicting the dynamics of the success as in the binary case.

In conclusion, my work revealed that urban success has an unexpected property: the temporal evolution of the popularity of urban venues follows six distinct shapes: the rise, fall, rise & fall, fall & rise, ephemeral, and underdog curves. By deriving relevant features describing urban spaces along the *where*, *what*, and *who* dimensions, I also built machine learning models predicting venue success. This modeling approach revealed that both the binarized success and the temporal success trajectory of the urban venues mostly depend on the venues' clientele, implying that the key to success is to become unique

social melting pots.

CHAPTER 6

CONCLUSION

In this thesis, I studied several aspects of the evolution of success on five different domains of life: science, music, film, literature, and urban spaces, split across three chapters corresponding to three publications and working papers.

First, I built on an existing modeling approach called the Q-model to decompose creative impact in science, music, film, and literature. By combining the Q-model with classical test theory, I proposed a framework to compare the role of randomness on 28 creative professions, such as film directors, pop musicians, and mathematicians. I compared the temporal evolution of the impact and the network position of creative individuals and found that there is an even split between those individuals for whom network peaks first and success follows, and those whose success peaks first and their network centrality follows. However, my analysis showed that these two substantitally different networking behaviors do not result in significant differences in expected success.

Second, I studied the dynamics of the Top 100 ranking of the most popular DJs and identified a clear threshold that separates a long-standing elite from the rest of the DJ world. I found that these all-time star DJs are typically centering different clusters of artists. My analysis showed that these clusters rise and fall distinctively over time, reflecting changes in musical trends. Moreover, I proposed a mentorship definition that captures the role of senior artists in building the communities around them. This mentorship turned out to give clear advantages for the mentees, but limit their chances of becoming all-time stars at the same time.

Third, my analysis of the evolution of the popularity of urban venues revealed that urban success has six distinct shapes: urban venues' popularity trajectories can (1) rise or (2) fall over time, can follow (3) rise & fall or (4) fall & rise arcs, or can be either (5) ephemeral or (6) underdog. To further elaborate

on these six shapes, I derived several features describing venues based on their profile, their neighborhoods' properties, and different features of their clientele. Surprisingly, the best predictors of urban success turned out to be the social features describing the urban venues' visitors, including their social prestige estimated by their network properties.

The three papers covered in this dissertation all contribute to the literature in unique ways in themselves and offer several ways for direct practical applications from scientific policymaking to urban planning. For instance, the first contribution of Chapter 3 is to show that the Q -model holds for different fields other than the already studied scientific ones. This serves as further proof of the validity and practical usefulness of the Q -parameter capturing success with the prospect of applications on several fields as a universal success metric. Furthermore, since the results presented in Chapter 3 highlight that the influence of random fluctuations in creative impact is comparable to the magnitude of the individuals' contribution, my findings alert the readers on the shortcomings of using exclusively citation-based metrics for research impact evaluation.

My work in Chapter 4 on the emergence and decay of DJ-communities not only uncovered the existence of a small elite but also highlighted a possible mechanism responsible for this, which has important practical implications. As now a data-driven understanding of some basic rules of the DJ world is clear, new doors open to design a better mentoring system. This could include finding better ways of including and leaving more space for underserved individuals. Another important direction would be tracking down biases in a pragmatic data-driven fashion, for instance, the striking gender gap in electronic music. In a broader sense, the tools and ideas I presented about mentorship do not strictly apply to electronic music or rely on any specific attributes of that profession. They could easily be adapted to study the benefits and limitations of mentorship and success in other commercial environments, from organizational development to human resources.

The results presented in Chapter 5 showed a surprising effect of social networks on urban spaces: they seem to be the key for venue success. This has interesting implications on both urban planning and marketing strategies in general since, apparently, building up a coherent social network matters more than the location and the profile of urban venues. In addition, this draws urban planners' attention to the importance of creating unique social melting pots rather than extravagant but less human-friendly environments, also suggesting future collaborations between the fields of social psychology and urban planning & informatics.

In conclusion, this thesis deepened our understanding of the evolution of success in five different domains by presenting unique and new scientific re-

sults. On the one hand, I show several examples of the temporal aspects of success, such as the relation of creative individuals' big hits and the success shapes of urban spaces. On the other, I relate success to the underlying social networks from collaboration networks to mentorship, and the effect of social factors on urban success, highlighting a universal role of the network effects in success. Furthermore, as this brief concluding chapter also outlines, my findings not only have academic importance and interest but are directly related to real-life problems and offer practical solutions. I truly believe that the work presented in this thesis is a unique interdisciplinary contribution to the scientific literature, in particular at the interface of network science, data science, and computational social science.

BIBLIOGRAPHY

- [1] Pais, A.: *Subtle Is the Lord: The Science and the Life of Albert Einstein*. Oxford University Press USA, New York City, New York, US (1982)
- [2] Thomson Reuters Corporation: *Web of Science* (2018). <https://webofknowledge.com>
- [3] Simonton, D.K.: *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, Cambridge, UK (2004)
- [4] Meyers, J.: *George Orwell*. Routledge, Abingdon-on-Thames, UK (2002)
- [5] Abra, J.: Do the muses dwell in Elysium? Death as a motive for creativity. *Creativity Research Journal* **8**(3), 205–217 (1995)
- [6] Simonton, D.K.: Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review* **104**(1), 66 (1997)
- [7] Lindauer, M.S.: The youthful rise, early fall, and short span of creativity. In: *Aging, Creativity and Art*, pp. 45–57. Springer, New York City, New York, US (2003)
- [8] Adams, C.W.: The age at which scientists do their best work. *Isis* **36**(3/4), 166–169 (1946)
- [9] Lehman, H.C.: *Age and Achievement*, foreword by Lewis M. Terman. Princeton University Press for the American Philosophical Society (1953)
- [10] Campbell, D.T.: Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review* **67**(6), 380 (1960)

- [11] Simonton, D.K.: Creative productivity and age: A mathematical model based on a two-step cognitive process. *Developmental Review* **4**(1), 77–111 (1984)
- [12] Simonton, D.K.: Age and outstanding achievement: What do we know after a century of research? *Psychological Bulletin* **104**(2), 251 (1988)
- [13] Feist, G.J.: The past and future of the psychology of science. *Review of General Psychology* **10**(2), 92–97 (2006)
- [14] Amazon.com, Inc.: Internet Movie Database (2017). www.imdb.com
- [15] Zink Media: Discogs music release database (2017). www.discogs.com
- [16] Foursquare Labs: Foursquare City Guide (2019). <https://foursquare.com>
- [17] Light, R.P., Polley, D.E., Börner, K.: Open data and open code for big science of science studies. *Scientometrics* **101**(2), 1535–1551 (2014)
- [18] Google LLC: Google Scholar (2018). <https://scholar.google.com/>
- [19] Hartnett, J.: Discogs. com. *The Charleston Advisor* **16**(4), 26–33 (2015)
- [20] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., *et al.*: Computational social science. *Science* **323**(5915), 721–723 (2009)
- [21] Clauset, A., Larremore, D.B., Sinatra, R.: Data-driven predictions in the science of science. *Science* **355**(6324), 477–480 (2017)
- [22] Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., Stanley, H.E.: The science of science: From the perspective of complex systems. *Physics Reports* **714**, 1–73 (2017)
- [23] Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., *et al.*: Science of science. *Science* **359**(6379), 0185 (2018)
- [24] Garfield, E., Merton, R.K.: *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities* vol. 8. Wiley New York, Hoboken, New Jersey, US (1979)
- [25] Hirsch, J.E.: An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences* **102**(46), 16569–16572 (2005)

- [26] Maslov, S., Redner, S.: Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience* **28**(44), 11103–11105 (2008)
- [27] Wang, D., Song, C., Barabási, A.-L.: Quantifying long-term scientific impact. *Science* **342**(6154), 127–132 (2013)
- [28] Zuccala, A.A., Verleysen, F.T., Cornacchia, R., Engels, T.C.: Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Journal of Information Management* **67**(3), 320–336 (2015)
- [29] Petersen, A.M., Pan, R.K., Pammolli, F., Fortunato, S.: Methods to account for citation inflation in research evaluation. *Research Policy* **48**(7), 1855–1865 (2019)
- [30] Hoonlor, A., Szymanski, B.K., Zaki, M.J.: Trends in computer science research. *Communications of the ACM* **56**(10), 74–83 (2013)
- [31] Sinatra, R., Deville, P., Szell, M., Wang, D., Barabási, A.-L.: A century of physics. *Nature Physics* **11**(10), 791 (2015)
- [32] Bonaventura, M., Latora, V., Nicosia, V., Panzarasa, P.: The advantages of interdisciplinarity in modern science. *arXiv preprint arXiv:1712.07910* (2017)
- [33] Battiston, F., Musciotto, F., Wang, D., Barabási, A.-L., Szell, M., Sinatra, R.: Taking census of physics. *Nature Reviews Physics* **1**(1), 89 (2019)
- [34] Barabási, A.-L.: *The Formula: The Universal Laws of Success*. Little, Brown and Company, Boston, Massachusetts, United States (2018)
- [35] Yin, Y., Wang, Y., Evans, J.A., Wang, D.: Quantifying the dynamics of failure across science, startups and security. *Nature* **575**(7781), 190–194 (2019)
- [36] Williams, O.E., Lacasa, L., Latora, V.: Quantifying and predicting success in show business. *Nature Communications* **10**(1), 1–8 (2019)
- [37] Yucesoy, B., Barabási, A.-L.: Untangling performance from success. *EPJ Data Science* **5**(1), 17 (2016)
- [38] Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V.D., Barabási, A.-L.: Career on the move: Geography, stratification, and scientific impact. *Scientific Reports* **4**, 4770 (2014)

- [39] Clauset, A., Arbesman, S., Larremore, D.B.: Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* **1**(1), 1400005 (2015)
- [40] Jia, T., Wang, D., Szymanski, B.K.: Quantifying patterns of research-interest evolution. *Nature Human Behaviour* **1**(4), 0078 (2017)
- [41] Shen, H.-W., Barabási, A.-L.: Collective credit allocation in science. *Proceedings of the National Academy of Sciences* **111**(34), 12325–12330 (2014)
- [42] Jones, B.F.: Age and great invention. *The Review of Economics and Statistics* **92**(1), 1–14 (2010)
- [43] Sinatra, R., Wang, D., Deville, P., Song, C., Barabási, A.-L.: Quantifying the evolution of individual scientific impact. *Science* **354**(6312), 5239 (2016)
- [44] Liu, L., Wang, Y., Sinatra, R., Giles, C.L., Song, C., Wang, D.: Hot streaks in artistic, cultural, and scientific careers. *Nature* **559**(7714), 396 (2018)
- [45] Li, J., Yin, Y., Fortunato, S., Wang, D.: Nobel laureates are almost the same as us. *Nature Reviews Physics* **1**(5), 301 (2019)
- [46] Stewart, J.: The distribution of talent. *Marilyn Zurmuehlen Working Papers in Art Education* **2**(1), 21–22 (1983)
- [47] Mauboussin, M.: Untangling skill and luck: How to think about outcomes—past, present, and future. Legg Mason Capital Management (2010)
- [48] Mauboussin, M.J.: *The Success Equation: Untangling Skill and Luck in Business, Sports, and Investing*. Harvard Business Press, Brighton, Massachusetts, US (2012)
- [49] Burt, R.S.: Structural holes versus network closure as social capital. In: *Social Capital*, pp. 31–56. Routledge, Abingdon-on-Thames, UK (2017)
- [50] Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., Schweitzer, F.: Predicting scientific success based on coauthorship networks. *EPJ Data Science* **3**(1), 9 (2014)
- [51] Petersen, A.M.: Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences* **112**(34), 4671–4680 (2015)

- [52] Gogoglou, A., Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: A scientist's impact over time: The predictive power of clustering with peers. In: Proceedings of the 20th International Database Engineering & Applications Symposium, pp. 334–339 (2016). ACM
- [53] Vedres, B.: Forbidden triads and creative success in jazz: The Miles Davis factor. *Applied Network Science* **2**(1), 31 (2017)
- [54] Wachs, J., Daróczy, B., Hannák, A., Páll, K., Riedl, C.: And now for something completely different: Visual novelty in an online network of designers. In: Proceedings of the 10th ACM Conference on Web Science, pp. 163–172 (2018). ACM
- [55] Vedres, B., Cserpes, T.: Forbidden triads and innovation. *Social Science Research Network* (2019)
- [56] Guimera, R., Uzzi, B., Spiro, J., Amaral, L.A.N.: Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**(5722), 697–702 (2005)
- [57] Uzzi, B.: A social network's changing statistical properties and the quality of human innovation. *Journal of Physics A: Mathematical and Theoretical* **41**(22), 224023 (2008)
- [58] Bel, R., Smirnov, V., Wait, A.: On Broadway and sports: How to make a winning team. *Organizational Economics Proceedings* **2**(1) (2014)
- [59] Tröster, C., Mehra, A., van Knippenberg, D.: Structuring for team success: The interactive effects of network structure and cultural diversity on team potency and performance. *Organizational Behavior and Human Decision Processes* **124**(2), 245–255 (2014)
- [60] Bonaventura, M., Ciotti, V., Panzarasa, P., Liverani, S., Lacasa, L., Latora, V.: Predicting success in the worldwide start-up network. *Scientific Reports* **10**(1), 1–6 (2020)
- [61] Dennis, W.: Bibliographies of eminent scientists. *The Scientific Monthly* **79**(3), 180–183 (1954)
- [62] Dennis, W.: Predicting scientific productivity in later maturity from records of earlier decades. *Journal of Gerontology* (1954)
- [63] Simonton, D.K.: Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry*, 309–328 (1999)

- [64] Dennis, W.: Creative productivity between the ages of 20 and 80 years. *Journal of Gerontology* **21**(1), 1–8 (1966)
- [65] Simonton, D.K.: Career landmarks in science: Individual differences and interdisciplinary contrasts. *Developmental Psychology* **27**(1), 119 (1991)
- [66] Dennis, W.: Age and productivity among scientists. *Science* (1956)
- [67] Lehman, H.C.: Reply to Dennis' critique of age and achievement. *Journal of Gerontology* **11**(3), 333–337 (1956)
- [68] Feist, G.J.: Quantity, quality, and depth of research as influences on scientific eminence: Is quantity most important? *Creativity Research Journal* **10**(4), 325–335 (1997)
- [69] Beaver, D., Rosen, R.: Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics* **1**(3), 231–245 (1979)
- [70] Bayer, A.E., Smart, J.C.: Career publication patterns and collaborative "styles" in American academic science. *The Journal of Higher Education* **62**(6), 613–636 (1991)
- [71] Hagstrom, W.O.: Traditional and modern forms of scientific teamwork. *Administrative Science Quarterly*, 241–263 (1964)
- [72] Lee, S., Bozeman, B.: The impact of research collaboration on scientific productivity. *Social Studies of Science* **35**(5), 673–702 (2005)
- [73] Gordon, M.: A critical reassessment of inferred relations between multiple authorship, scientific collaboration, the production of papers and their acceptance for publication. *Scientometrics* **2**(3), 193–201 (1980)
- [74] Kram, K.E.: Improving the mentoring process. *Training & Development Journal* (1985)
- [75] Kram, K.E., Isabella, L.A.: Mentoring alternatives: The role of peer relationships in career development. *Academy of Management Journal* **28**(1), 110–132 (1985)
- [76] Higgins, M.C., Kram, K.E.: Reconceptualizing mentoring at work: A developmental network perspective. *Academy of Management Review* **26**(2), 264–288 (2001)

- [77] Simonton, D.K.: Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin* **129**(4), 475 (2003)
- [78] Merton, R.K.: The Matthew effect in science: The reward and communication systems of science are considered. *Science* **159**(3810), 56–63 (1968)
- [79] Petersen, A.M., Jung, W.-S., Yang, J.-S., Stanley, H.E.: Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences* **108**(1), 18–23 (2011)
- [80] Larivière, V., Gingras, Y.: The impact factor's Matthew effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology* **61**(2), 424–427 (2010)
- [81] Perc, M.: The Matthew effect in empirical data. *Journal of The Royal Society Interface* **11**(98), 20140378 (2014)
- [82] Radicchi, F.: In science “there is no bad publicity”: Papers criticized in comments have high scientific impact. *Scientific Reports* **2**, 815 (2012)
- [83] Catalini, C., Lacetera, N., Oettl, A.: The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences* **112**(45), 13823–13826 (2015)
- [84] Bornmann, L., Daniel, H.-D.: Does the h-index for ranking of scientists really work? *Scientometrics* **65**(3), 391–392 (2005)
- [85] Vanclay, J.K.: Bias in the journal impact factor. *Scientometrics* **78**(1), 3–12 (2009)
- [86] Jacso, P.: Grim tales about the impact factor and the h-index in the web of science and the journal citation reports databases: reflections on Vanclay's criticism. *Scientometrics* **92**(2), 325–354 (2012)
- [87] Kaur, J., Radicchi, F., Menczer, F.: Universality of scholarly impact metrics. *Journal of Informetrics* **7**(4), 924–932 (2013)
- [88] Piwowar, H.: Altmetrics: Value all research products. *Nature* **493**(7431), 159 (2013)
- [89] Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C.R.: Do altmetrics work? Twitter and ten other social web services. *PLOS ONE* **8**(5), 64841 (2013)

- [90] Kousha, K., Thelwall, M., Abdoli, M.: Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology* **68**(8), 2004–2016 (2017)
- [91] Szell, M., Ma, Y., Sinatra, R.: A Nobel opportunity for interdisciplinarity. *Nature Physics* **14**(11), 1075 (2018)
- [92] King, D.A.: The scientific impact of nations. *Nature* **430**(6997), 311 (2004)
- [93] Salganik, M.J., Dodds, P.S., Watts, D.J.: Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**(5762), 854–856 (2006)
- [94] Thelwall, M., Kousha, K.: Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology* **68**(4), 972–983 (2017)
- [95] Segall, L.: Billboard Magazine. <http://www.billboard.com>
- [96] Flôres Jr, R.G., Ginsburgh, V.A.: The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician* **45**, 97–104 (1996)
- [97] Radicchi, F., Fortunato, S., Markines, B., Vespignani, A.: Diffusion of scientific credits and the ranking of scientists. *Physical Review E* **80**(5), 056103 (2009)
- [98] Blumm, N., Ghoshal, G., Forró, Z., Schich, M., Bianconi, G., Bouchaud, J.-P., Barabási, A.-L.: Dynamics of ranking processes in complex systems. *Physical Review Letters* **109**(12), 128701 (2012)
- [99] Thrust Publishing Ltd: DJMag Top 100 (2019). <https://djmag.com/top-100-djs>
- [100] CSRankings: Computer Science Rankings (2019). <https://csrankings.org>
- [101] Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., Rafols, I.: Bibliometrics: The Leiden Manifesto for research metrics. *Nature News* **520**(7548), 429 (2015)
- [102] Leydesdorff, L., Milojevic, S.: Scientometrics. arXiv preprint arXiv:1208.4566 (2012)
- [103] Lyall, C., Bruce, A., Marsden, W., Meagher, L.: The role of funding agencies in creating interdisciplinary knowledge. *Science and Public Policy* **40**(1), 62–71 (2013)

- [104] Szell, M., Sinatra, R.: Research funding goes to rich clubs. *Proceedings of the National Academy of Sciences* **112**(48), 14749–14750 (2015)
- [105] Bromham, L., Dinnage, R., Hua, X.: Interdisciplinary research has consistently lower funding success. *Nature* **534**(7609), 684 (2016)
- [106] Murray, D.L., Morris, D., Lavoie, C., Leavitt, P.R., MacIsaac, H., Masson, M.E., Villard, M.-A.: Bias in research grant evaluation has dire consequences for small universities. *PLOS ONE* **11**(6), 0155876 (2016)
- [107] Shen, Z., Yang, L., Pei, J., Li, M., Wu, C., Bao, J., Wei, T., Di, Z., Rousseau, R., Wu, J.: Interrelations among scientific fields and their relative influences revealed by an input–output analysis. *Journal of Informetrics* **10**(1), 82–97 (2016)
- [108] Erdős, P., Rényi, A.: On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**(1), 17–60 (1960)
- [109] Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684), 440 (1998)
- [110] Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)
- [111] Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**(1), 47 (2002)
- [112] Newman, M.E.: Assortative mixing in networks. *Physical Review Letters* **89**(20), 208701 (2002)
- [113] Newman, M.E.: The structure and function of complex networks. *SIAM Review* **45**(2), 167–256 (2003)
- [114] Barabási, A.-L.: *Linked: The new science of networks*. AAPT (2003)
- [115] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics. *Physics Reports* **424**(4-5), 175–308 (2006)
- [116] Barabási, A.-L., *et al.*: *Network Science*. Cambridge University Press, Cambridge, UK (2016)
- [117] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)

- [118] Mei, Q., Guo, J., Radev, D.: Divrank: The interplay of prestige and diversity in information networks. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1009–1018 (2010). ACM
- [119] Jomsri, P., Sanguansintukul, S., Choochaiwattana, W.: CiteRank: Combination similarity and static ranking with research paper searching. *International Journal of Internet Technology and Secured Transactions* **3**(2), 161–177 (2011)
- [120] Su, C., Pan, Y., Zhen, Y., Ma, Z., Yuan, J., Guo, H., Yu, Z., Ma, C., Wu, Y.: PrestigeRank: A new evaluation method for papers and journals. *Journal of Informetrics* **5**(1), 1–13 (2011)
- [121] Mariani, M.S., Medo, M., Zhang, Y.-C.: Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics* **10**(4), 1207–1223 (2016)
- [122] Mariani, M.S., Medo, M., Lafond, F.: Early identification of important patents through network centrality. *arXiv preprint arXiv:1710.09182* (2017)
- [123] Spitz, A., Horvát, E.-Á.: Measuring long-term impact based on network centrality: Unraveling cinematic citations. *PLOS ONE* **9**(10), 108857 (2014)
- [124] Redner, S.: How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* **4**(2), 131–134 (1998)
- [125] Lehmann, S., Lautrup, B., Jackson, A.D.: Citation networks in high energy physics. *Physical Review E* **68**(2), 026113 (2003)
- [126] Redner, S.: Citation characteristics from 110 years of *Physical Review*. *Physics Today Online* **58**(6), 49–54 (2005)
- [127] Radicchi, F., Fortunato, S., Castellano, C.: Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* **105**(45), 17268–17272 (2008)
- [128] Peterson, G.J., Pressé, S., Dill, K.A.: Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences* **107**(37), 16023–16027 (2010)

- [129] Radicchi, F., Castellano, C.: Rescaling citations of publications in physics. *Physical Review E* **83**(4), 046116 (2011)
- [130] Newman, M.E.: The first-mover advantage in scientific publication. *Europhysics Letters* **86**(6), 68001 (2009)
- [131] Newman, M.E.: Clustering and preferential attachment in growing networks. *Physical Review E* **64**(2), 025102 (2001)
- [132] Yucesoy, B., Wang, X., Huang, J., Barabási, A.-L.: Success in books: A big data approach to bestsellers. *EPJ Data Science* **7**(1), 7 (2018)
- [133] Mendel, G.: Versuche uber pflanzen-hybriden. *Verhandlungen des naturforschenden Vereins in Brünn* **4**, 3–47 (1866)
- [134] Garfield, E.: Premature discovery or delayed recognition-why. *Current Contents* (21), 5–10 (1980)
- [135] Van Raan, A.F.: Sleeping beauties in science. *Scientometrics* **59**(3), 467–472 (2004)
- [136] Ke, Q., Ferrara, E., Radicchi, F., Flammini, A.: Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences* **112**(24), 7426–7431 (2015)
- [137] Ye, F.Y., Bornmann, L.: “smart girls” versus “sleeping beauties” in the sciences: The identification of instant and delayed recognition by using the citation angle. *Journal of the Association for Information Science and Technology* **69**(3), 359–367 (2018)
- [138] CBS Interactive: LastFM (2019). <https://www.last.fm>
- [139] Rossetti, G., Milli, L., Giannotti, F., Pedreschi, D.: Forecasting success via early adoptions analysis: A data-driven study. *PLOS ONE* **12**(12), 0189096 (2017)
- [140] Batista, P.D., Campitelli, M.G., Kinouchi, O.: Is it possible to compare researchers with different scientific interests? *Scientometrics* **68**(1), 179–189 (2006)
- [141] Castellano, C., Radicchi, F.: On the fairness of using relative indicators for comparing citation performance in different disciplines. *Archivum Immunologiae et Therapiae Experimentalis* **57**(2), 85 (2009)
- [142] Price, D.J.D.S.: Networks of scientific papers. *Science*, 510–515 (1965)

- [143] Boyack, K.W., Klavans, R., Börner, K.: Mapping the backbone of science. *Scientometrics* **64**(3), 351–374 (2005)
- [144] Porter, A., Rafols, I.: Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* **81**(3), 719–745 (2009)
- [145] APS: American Physical Society (2019). www.aps.org
- [146] Herrera, M., Roberts, D.C., Gulbahce, N.: Mapping the evolution of scientific fields. *PLOS ONE* **5**(5), 10355 (2010)
- [147] Perc, M.: Self-organization of progress across the century of physics. *Scientific Reports* **3**, 1720 (2013)
- [148] Uzzi, B., Mukherjee, S., Stringer, M., Jones, B.: Atypical combinations and scientific impact. *Science* **342**(6157), 468–472 (2013)
- [149] Mukherjee, S., Romero, D.M., Jones, B., Uzzi, B.: The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science Advances* **3**(4), 1601315 (2017)
- [150] Lotka, A.J.: The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* **16**(12), 317–323 (1926)
- [151] Dorogovtsev, S.N., Mendes, J.F.: Ranking scientists. *Nature Physics* **11**(11), 882 (2015)
- [152] Wildgaard, L., Schneider, J.W., Larsen, B.: A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics* **101**(1), 125–158 (2014)
- [153] Acuna, D.E., Allesina, S., Kording, K.P.: Future impact: Predicting scientific success. *Nature* **489**(7415), 201 (2012)
- [154] Wang, M., Yu, G., Yu, D.: Mining typical features for highly cited papers. *Scientometrics* **87**(3), 695–706 (2011)
- [155] Penner, O., Pan, R.K., Petersen, A.M., Kaski, K., Fortunato, S.: On the predictability of future impact in science. *Scientific Reports* **3**, 3052 (2013)
- [156] Chakraborty, T., Nandi, S.: Universal trajectories of scientific success. *Knowledge and Information Systems* **54**(2), 487–509 (2018)

- [157] Arakelyan, S., Morstatter, F., Martin, M., Ferrara, E., Galstyan, A.: Mining and forecasting career trajectories of music artists. In: Proceedings of the 29th on Hypertext and Social Media, pp. 11–19 (2018). ACM
- [158] Biondo, A.P., Rapisarda, A., et al.: Talent vs Luck: The role of randomness in success and failure. arXiv preprint arXiv:1802.07068 (2018)
- [159] Jones, B.F., Weinberg, B.A.: Age dynamics in scientific creativity. Proceedings of the National Academy of Sciences **108**(47), 18910–18914 (2011)
- [160] Wardil, L., Hauert, C.: Cooperation and coauthorship in scientific publishing. Physical Review E **91**(1), 012825 (2015)
- [161] Wuchty, S., Jones, B.F., Uzzi, B.: The increasing dominance of teams in production of knowledge. Science **316**(5827), 1036–1039 (2007)
- [162] Waltman, L.: An empirical analysis of the use of alphabetical authorship in scientific publishing. Journal of Informetrics **6**(4), 700–711 (2012)
- [163] Oppenheim, C.: Fractional counting of multiauthored publications. Journal of the American Society for Information Science **49**(5), 482–482 (1998)
- [164] Kim, J., Kim, J.: Rethinking the comparison of coauthorship credit allocation schemes. Journal of Informetrics **9**(3), 667–673 (2015)
- [165] Granovetter, M.S.: The strength of weak ties. Social Networks, 347–367 (1977)
- [166] Pan, R.K., Saramäki, J.: The strength of strong ties in scientific collaboration networks. Europhysics Letters **97**(1), 18007 (2012)
- [167] Feld, S.L.: Why your friends have more friends than you do. American Journal of Sociology **96**(6), 1464–1477 (1991)
- [168] Eom, Y.-H., Jo, H.-H.: Generalized friendship paradox in complex networks: The case of scientific collaboration. Scientific Reports **4**, 4603 (2014)
- [169] Milojević, S.: Principles of scientific research team formation and evolution. Proceedings of the National Academy of Sciences **111**(11), 3984–3989 (2014)
- [170] Price, D.d.S., Beaver, D.: Collaboration in an invisible college. American Psychologist **21**(11), 1011–1018 (1966)

- [171] Klug, M., Bagrow, J.P.: Understanding the group dynamics and success of teams. *Royal Society Open Science* **3**(4), 160007 (2016)
- [172] Zuckerman, H.: Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review*, 391–403 (1967)
- [173] Scandura, T.A.: Mentorship and career mobility: An empirical investigation. *Journal of Organizational Behavior* **13**(2), 169–174 (1992)
- [174] Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J., Song, M.: Standing on the shoulders of giants. *Journal of Informetrics* **11**(1), 307–323 (2017)
- [175] Malmgren, R.D., Ottino, J.M., Amaral, L.A.N.: The role of mentorship in protégé performance. *Nature* **465**(7298), 622 (2010)
- [176] Sekara, V., Deville, P., Ahnert, S.E., Barabási, A.-L., Sinatra, R., Lehmann, S.: The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences* **115**(50), 12603–12607 (2018)
- [177] Petersen, A.M., Fortunato, S., Pan, R.K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H.E., Pammolli, F.: Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* **111**(43), 15316–15321 (2014)
- [178] Newman, M.E.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E* **64**(1), 016132 (2001)
- [179] Newman, M.E.: Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E* **64**(1), 016131 (2001)
- [180] Newman, M.E.: Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences* **101**(suppl 1), 5200–5205 (2004)
- [181] Travers, J., Milgram, S.: An experimental study of the small world problem. *Social Networks*, 179–197 (1977)
- [182] Freeman, L.C.: A set of measures of centrality based on betweenness. *Sociometry*, 35–41 (1977)
- [183] Fan, Y., Li, M., Chen, J., Gao, L., Di, Z., Wu, J.: Network of econophysicists: A weighted network to investigate the development of econophysics. *International Journal of Modern Physics B* **18**(17n19), 2505–2511 (2004)

- [184] Hou, H., Kretschmer, H., Liu, Z.: The structure of scientific collaboration networks in scientometrics. *Scientometrics* **75**(2), 189–202 (2007)
- [185] Fatt, C., Ujum, E., Ratnavelu, K.: The structure of collaboration in the Journal of Finance. *Scientometrics* **85**(3), 849–860 (2010)
- [186] Leifeld, P., Wankmüller, S., Berger, V.T., Ingold, K., Steiner, C.: Collaboration patterns in the german political science co-authorship network. *PLOS ONE* **12**(4), 0174671 (2017)
- [187] Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**(12), 7821–7826 (2002)
- [188] Gleiser, P.M., Danon, L.: Community structure in jazz. *Advances in Complex Systems* **6**(04), 565–573 (2003)
- [189] TiVo Corporation: AllMusic (2019). <https://www.allmusic.com>
- [190] Park, J., Celma, O., Koppenberger, M., Cano, P., Buldú, J.M.: The social network of contemporary popular musicians. *International Journal of Bifurcation and Chaos* **17**(07), 2281–2288 (2007)
- [191] Park, D., Bae, A., Schich, M., Park, J.: Topology and evolution of the network of western classical music composers. *EPJ Data Science* **4**(1), 2 (2015)
- [192] Burke, J., Rygaard, R., Yellin-Flaherty, Z.: Clam!: Inferring genres in the Discogs collaboration network (2014)
- [193] Uzzi, B., Spiro, J.: Collaboration and creativity: The small world problem. *American Journal of Sociology* **111**(2), 447–504 (2005)
- [194] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. *Science* **298**(5594), 824–827 (2002)
- [195] Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., Hütt, M.-T.: Motifs in co-authorship networks and their relation to the impact of scientific publications. *The European Physical Journal B* **84**(4), 535–540 (2011)
- [196] Burt, R.S., *et al.*: Brokerage and Closure: An Introduction to Social Capital. Oxford University Press, Oxford, UK (2005)

- [197] Budner, P., Grahl, J.: Collaboration networks in the music industry. arXiv preprint arXiv:1611.00377 (2016)
- [198] Auber, D., Chiricota, Y., Jourdan, F., Melançon, G.: Multiscale visualization of small world networks. In: IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714), pp. 75–81 (2003). IEEE
- [199] Seidman, S.B.: Network structure and minimum degree. *Social Networks* **5**(3), 269–287 (1983)
- [200] Anderson, R.M., Anderson, B., May, R.M.: *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, UK (1992)
- [201] Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F.F.: K-core organization of complex networks. *Physical Review Letters* **96**(4), 040601 (2006)
- [202] Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A.: Identification of influential spreaders in complex networks. *Nature Physics* **6**(11), 888 (2010)
- [203] Cattani, G., Ferriani, S.: A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the Hollywood film industry. *Organization Science* **19**(6), 824–844 (2008)
- [204] Juhász, S., Tóth, G., Lengyel, B.: Brokering the core with periphery–collaboration networks and individual success in the Hungarian film industry. *Social Science Research Network*, 3217131 (2018)
- [205] Luukkonen, T., Persson, O., Sivertsen, G.: Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values* **17**(1), 101–126 (1992)
- [206] Yan, E., Ding, Y.: Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology* **60**(10), 2107–2118 (2009)
- [207] Uddin, S., Hossain, L., Rasmussen, K.: Network effects on scientific collaborations. *PLOS ONE* **8**(2), 57546 (2013)
- [208] Biscaro, C., Giupponi, C.: Co-authorship and bibliographic coupling network effects on citations. *PLOS ONE* **9**(6), 99502 (2014)

- [209] Servia-Rodríguez, S., Noulas, A., Mascolo, C., Fernández-Vilas, A., Díaz-Redondo, R.P.: The evolution of your success lies at the centre of your co-authorship network. *PLOS ONE* **10**(3), 0114302 (2015)
- [210] Jadidi, M., Karimi, F., Lietz, H., Wagner, C.: Gender disparities in science? Dropout, productivity, collaborations and success of male and female computer scientists. *Advances in Complex Systems* **21**(03n04), 1750011 (2018)
- [211] Fraiberger, S.P., Sinatra, R., Resch, M., Riedl, C., Barabási, A.-L.: Quantifying reputation and success in art. *Science* **362**(6416), 825–829 (2018)
- [212] Janosov, M., Battiston, F., Sinatra, R.: Success and luck in creative careers. *arXiv preprint arXiv:1909.07956* (2019)
- [213] Pluchino, A., Biondo, A., Rapisarda, A.: Talent vs Luck: The role of randomness in success and failure. *arXiv preprint arXiv:1802.07068* (2018)
- [214] Lee, Y.-N., Walsh, J.P., Wang, J.: Creativity in scientific teams: Unpacking novelty and impact. *Research Policy* **44**(3), 684–697 (2015)
- [215] Zagovora, O., Weller, K., Janosov, M., Wagner, C., Peters, I.: What increases (social) media attention: Research impact, author prominence or title attractiveness? *Proceedings of the 23rd International Conference on Science and Technology Indicators*, 1182–1190 (2018)
- [216] Galton, F.: Hereditary genius. 1869. *Natural Inheritance* (1889)
- [217] Flugel, J.C., West, D.J.: A hundred years of psychology. (1964)
- [218] Pluchino, A., Burgio, G., Rapisarda, A., Biondo, A.E., Pulvirenti, A., Ferro, A., Giorgino, T.: Exploring the role of interdisciplinarity in physics: Success, talent and luck. *PLOS ONE* **14**(6), 0218793 (2019)
- [219] Crocker, L., Algina, J.: *Introduction to Classical and Modern Test Theory*. Education Resources Information Center, U.S. Department of Education, (1986)
- [220] Lord, F.M.: A strong true-score theory, with applications. *Psychometrika* **30**(3), 239–270 (1965)
- [221] Metascore values: Metacritic database using expert's evaluations (2017). <http://www.metacritic.com>

- [222] Amazon.com, Inc.: Goodreads book database and social network site (2017). www.goodreads.com
- [223] Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In: *Evolutionary Computation, 1996., Proceedings of IEEE International Conference On*, pp. 312–317 (1996). IEEE
- [224] Vásárhelyi, G., Virág, C., Somorjai, G., Nepusz, T., Eiben, A.E., Vicsek, T.: Optimized flocking of autonomous drones in confined environments. *Science Robotics* **3**(20), 3536 (2018)
- [225] Kristof, W.: Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika* **39**(4), 491–499 (1974)
- [226] Kline, T.: *Psychological Testing: A Practical Approach to Design and Evaluation*. SAGE Publishing, Thousand Oaks, California, US (2005)
- [227] Kean, J., Reilly, J.: Item response theory. *Handbook for Clinical Research: Design, Statistics and Implementation*, 195–198 (2014)
- [228] Allen, M.J., Yen, W.M.: *Introduction to Measurement Theory*. Waveland Press, Long Grove, Illinois, US (2001)
- [229] Figg, W.D., Dunn, L., Liewehr, D.J., Steinberg, S.M., Thurman, P.W., Barrett, J.C., Birkinshaw, J.: Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* **26**(6), 759–767 (2006)
- [230] Hsu, J.-W., Huang, D.-W.: Correlation between impact and collaboration. *Scientometrics* **86**(2), 317–324 (2011)
- [231] Janosov, M., Musciotto, F., Battiston, F., Iñiguez, G.: Elites, communities and the limited benefits of mentorship in electronic music. *arXiv preprint arXiv:1908.10968* (2019)
- [232] Burt, R.S.: Structural holes and good ideas. *American Journal of Sociology* **110**(2), 349–399 (2004)
- [233] DJMag: DJ Magazine (2019). <https://djmag.com>
- [234] Simonton, D.K.: The swan-song phenomenon: Last-works effects for 172 classical composers. *Psychology and Aging* **4**(1), 42 (1989)

- [235] Simonton, D.K.: Emergence and realization of genius: The lives and works of 120 classical composers. *Journal of Personality and Social Psychology* **61**(5), 829 (1991)
- [236] Ericsson, K.A., *et al.*: The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge Handbook of Expertise and Expert Performance* **38**, 685–705 (2006)
- [237] Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., *et al.*: Life in the network: The coming age of computational social science. *Science* **323**(5915), 721 (2009)
- [238] Kirschbaum, C.: Trajectory types across network positions: Jazz evolution from 1930 to 1969. In: *Knowledge and Networks vol. 2*, pp. 143–167. Springer, New York City, New York, US (2017)
- [239] Klimek, P., Kreuzbauer, R., Thurner, S.: Fashion and art cycles are driven by counter-dominance signals of elite competition: Quantitative evidence from music styles. *Journal of The Royal Society Interface* **16**(151), 20180731 (2019)
- [240] Youngblood, M.: Cultural transmission modes of music sampling traditions remain stable despite delocalization in the digital age. *PLOS ONE* **14**(2), 0211860 (2019)
- [241] Park, M., Thom, J., Mennicken, S., Cramer, H., Macy, M.: Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nature Human Behaviour* **3**(3), 230 (2019)
- [242] AMF: Amsterdam Music Festival (2019). <https://amf-festival.com>
- [243] Cocho, G., Flores, J., Gershenson, C., Pineda, C., Sánchez, S.: Rank diversity of languages: Generic behavior in computational linguistics. *PLOS ONE* **10**(4), 0121898 (2015)
- [244] Morales, J.A., Sánchez, S., Flores, J., Pineda, C., Gershenson, C., Cocho, G., Zizumbo, J., Rodríguez, R.F., Iñiguez, G.: Generic temporal features of performance rankings in sports and games. *EPJ Data Science* **5**(1), 33 (2016)
- [245] Sánchez, S., Cocho, G., Flores, J., Gershenson, C., Iñiguez, G., Pineda, C.: Trajectory stability in the traveling salesman problem. *Complexity* **2018** (2018)

- [246] Morales, J.A., Colman, E., Sánchez, S., Sánchez-Puig, F., Pineda, C., Iñiguez, G., Cocho, G., Flores, J., Gershenson, C.: Rank dynamics of word usage at multiple scales. arXiv preprint arXiv:1802.07258 (2018)
- [247] Various sources: Crowdsourced collection of DJMag Top100 rankings (2019). https://vk.com/topic-4286148_28357370
- [248] Various sources: Crowdsourced collection of DJMag Top100 rankings (2019). https://edm.fandom.com/wiki/DJ_Mag_Top_100_DJs
- [249] Aiello, L.M., Schifanella, R., Redi, M., Svetlichnaya, S., Liu, F., Osindero, S.: Beautiful and damned. combined effect of content quality and social ties on user engagement. *IEEE Transactions on Knowledge and Data Engineering* **29**(12), 2682–2695 (2017)
- [250] Coscia, M., Neffke, F.M.: Network backboning with noisy data. In: *Proceedings Of The Eleventh International Conference On Data Engineering*, pp. 425–436 (2017). IEEE
- [251] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), 10008 (2008)
- [252] John D. Luerssen: Rolling Stone interview with Afrojack (2013). <https://www.rollingstone.com/music/music-news/afrojack-pays-tribute-to-mentor-david-guetta-53734/>
- [253] Armada Music: Demo submission website of the record label Armada Music (2019). <https://demodrop.armadamusic.com/>
- [254] Hexagon Record Label: Demo submission website of the record label Hexagon (2019). <https://www.hexagonhq.com/demo>
- [255] Li, W., Aste, T., Caccioli, F., Livan, G.: Early coauthorship with top scientists predicts success in academic careers. *Nature Communications* **10**(1), 1–9 (2019)
- [256] Korpela, K., Hartig, T.: Restorative qualities of favorite places. *Journal of Environmental Psychology* **16**(3), 221–233 (1996)
- [257] Korpela, K.M., Hartig, T., Kaiser, F.G., Fuhrer, U.: Restorative experience and self-regulation in favorite places. *Environment and Behavior* **33**(4), 572–589 (2001)

- [258] Oldenburg, R., Brissett, D.: The third place. *Qualitative Sociology* **5**(4), 265–284 (1982)
- [259] Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in Foursquare. *ICWSM* **11**(70-573), 2 (2011)
- [260] Li, Y., Steiner, M., Wang, L., Zhang, Z.-L., Bao, J.: Exploring venue popularity in Foursquare. In: 2013 Proceedings IEEE INFOCOM, pp. 3357–3362 (2013). IEEE
- [261] Noulas, A., Mascolo, C., Frias-Martinez, E.: Exploiting Foursquare and cellular data to infer user activity in urban environments. In: 2013 IEEE 14th International Conference on Mobile Data Management, vol. 1, pp. 167–176 (2013). IEEE
- [262] Herr, P.M.: Priming price: Prior knowledge and context effects. *Journal of Consumer Research* **16**(1), 67–75 (1989)
- [263] Peteraf, M.A.: The cornerstones of competitive advantage: A resource-based view. *Strategic Management Journal* **14**(3), 179–191 (1993)
- [264] Glaeser, E.L., Kolko, J., Saiz, A.: Consumer city. *Journal of Economic Geography* **1**(1), 27–50 (2001)
- [265] Bettencourt, L.M., Lobo, J., Strumsky, D.: Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy* **36**(1), 107–120 (2007)
- [266] Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., Batty, M.: Constructing cities, deconstructing scaling laws. *Journal of The Royal Society Interface* **12**(102), 20140745 (2015)
- [267] Lee, S.Y., Florida, R., Acs, Z.: Creativity and entrepreneurship: A regional analysis of new firm formation. *Regional Studies* **38**(8), 879–891 (2004)
- [268] Florida, R.: Bohemia and economic geography. *Journal of Economic Geography* **2**(1), 55–71 (2002)
- [269] Clifton, N.: The “creative class” in the UK: An initial analysis. *Geografiska Annaler: Series B, Human Geography* **90**(1), 63–82 (2008)
- [270] Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. *ICWSM* **2011**, 81–88 (2011)

- [271] Noulas, A., Scellato, S., Lathia, N., Mascolo, C.: A random walk around the city: New venue recommendation in location-based social networks. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 144–153 (2012). IEEE
- [272] Jacobs, J.: *The Death and Life of Great American Cities*. Vintage, New York City, New York, United States (1961)
- [273] Florida, R.: *The Rise of the Creative Class* vol. 9. Basic books New York, New York City, New York, US (2002)
- [274] Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The role of social networks in information diffusion. In: Proceedings of the 21st International Conference on World Wide Web, pp. 519–528 (2012). ACM
- [275] Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in on-line social networks: A survey. *ACM Sigmod Record* **42**(2), 17–28 (2013)
- [276] Centola, D., Baronchelli, A.: The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences* **112**(7), 1989–1994 (2015)
- [277] Borgatti, S.P.: Centrality and network flow. *Social Networks* **27**(1), 55–71 (2005)
- [278] Foursquare Labs: about Foursquare (2018). <https://foursquare.com/about>
- [279] Office for National Statistics: Census Geography (2011). <https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography>
- [280] London Datastore: Indices of Deprivation (2015). <https://data.london.gov.uk/dataset/indices-of-deprivation-2015>
- [281] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.*: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
- [282] <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>: sklearn.cluster.DBSCAN (2019)
- [283] Müller, M.: Dynamic time warping. *Information Retrieval for Music and Motion*, 69–84 (2007)

- [284] Liao, T.W.: Clustering of time series data—a survey. *Pattern Recognition* **38**(11), 1857–1874 (2005)
- [285] Johnson, S.C.: Hierarchical clustering schemes. *Psychometrika* **32**(3), 241–254 (1967)
- [286] Jones, E., Oliphant, T., Peterson, P.: {SciPy}: Open source scientific tools for {Python} (2014)
- [287] Day, W.H., Edelsbrunner, H.: Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* **1**(1), 7–24 (1984)
- [288] Bayne, C.K., Beauchamp, J.J., Begovich, C.L., Kane, V.E.: Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition* **12**(2), 51–62 (1980)
- [289] Barreto, S., Ferreira, C., Paixao, J., Santos, B.S.: Using clustering analysis in a capacitated location-routing problem. *European Journal of Operational Research* **179**(3), 968–977 (2007)
- [290] Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis* vol. 344. John Wiley & Sons, Hoboken, New Jersey, US (2009)
- [291] Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423 (2001)
- [292] Lind, J.T., Mehlum, H.: With or without U? The appropriate test for a u-shaped relationship. *Oxford Bulletin of Economics and Statistics* **72**(1), 109–118 (2010)
- [293] Quercia, D., Saez, D.: Mining urban deprivation from Foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing* **13**(2), 30–36 (2014)
- [294] Hristova, D., Aiello, L.M., Quercia, D.: The new urban success: How culture pays. *Frontiers in Physics* **6**, 27 (2018)
- [295] Quattrone, G., Proserpio, D., Quercia, D., Capra, L., Musolesi, M.: Who benefits from the sharing economy of Airbnb? In: *Proceedings of the 25th International Conference on World Wide Web*, pp. 1385–1394 (2016). International World Wide Web Conferences Steering Committee

- [296] Hristova, D., Williams, M.J., Musolesi, M., Panzarasa, P., Mascolo, C.: Measuring urban social diversity using interconnected geo-social networks. In: Proceedings of the 25th International Conference on World Wide Web, pp. 21–30 (2016). International World Wide Web Conferences Steering Committee
- [297] Palchykov, V., Kaski, K., Kertész, J., Barabási, A.-L., Dunbar, R.I.: Sex differences in intimate relationships. *Scientific Reports* **2**, 370 (2012)
- [298] John Lu, Z.: The elements of statistical learning: Data mining, inference, and prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**(3), 693–694 (2010)
- [299] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). ACM
- [300] Nielsen, D.: Tree boosting with XGBoost-why does XGBoost win “every” machine learning competition? Master’s thesis, NTNU (2016)
- [301] Friedman, J.H.: Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38**(4), 367–378 (2002)

