Statistical Physics in the Era of Big Data

A dissertation presented by

Dashun Wang

to the Department of Physics

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Northeastern University

Boston, Massachusetts

April 26th, 2013

Statistical Physics in the Era of Big Data

by

Dashun Wang

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the College of Science of
Northeastern University
April 26th, 2013

Abstract

With the wealth of data provided by a wide range of high-throughout measurement tools and technologies, statistical physics of complex systems is entering a new phase, impacting in a meaningful fashion a wide range of fields, from cell biology to computer science to economics. In this dissertation, by applying tools and techniques developed in statistical physics, I present some of my contributions to the emerging field of Big Data in three distinct but related settings. First, we investigate long-term predictability of scientific impact. By deriving a mechanistic model for the citation dynamics of individual papers, we demonstrate that citation histories of all papers follow the same universal temporal pattern, helping us uncover the basic mechanisms that govern scientific impact. Second, we study the contextual factors that affect information spreading processes. We find that the social and organizational context significantly impacts to whom and how fast people forward information. Yet the structures within spreading processes can be well captured by a simple stochastic model, indicating surprising independence of context. Lastly, we study the mobility patterns and social interactions of mobile phone users, demonstrating the possibility of using the similarities between individual trajectories to predict social ties.

To Tian

Acknowledgements

I would like to thank my advisor, Albert-László Barabási, for accepting me into his group, providing me with his support whenever needed, and teaching me the importance of believing in my own work. I would also like to thank Prof. Alain Karma, David Lazer, and Alessandro Vespignani for serving on my thesis committee. I have benefited greatly from the many conversations with them.

This dissertation is a result of a wonderful collaborative effort. I am indebted to my coauthors and collaborators: Jim Bagrow, Vincent Blondel, Pierre Deville, Nathan Eagle, Gao Liang, Cesar Hidalgo, Ching-Yung Lin, Yu-Ru Lin, Fosca Giannotti, Luca Pappalardo, Dino Pedreschi, Roberta Sinatra, Chaoming Song, Hanghang Tong, Zhen Wen. I would especially like to thank Chaoming, who has been such a wonderful mentor, collaborator, and friend that graduate school would have not been the same without him. I would also like to thank Cesar for his help when I first joined the lab. Jim provided much help during my early stage, and I learned much from him.

I have spent two great summers at IBM Research. I am fortunate to have great mentors like Ching-Yung Lin, Hanghang Tong, and Zhen Wen. I am grateful for their support and trust throughout my internships. They were also instrumental in my job search. I would also like to thank the Watson team, especially David Ferrucci, Aditya Kalyanpur, and James Fan, for sparing their time helping me and exploring crazy ideas with me.

I am proud to be a member of CCNR, and grateful for the many past and current members of this family. They have provided numerous insightful conversations and inspirations over the past five years, especially Yong-Yeol Ahn, Sebastian Ahnert, József Baranyi, Baruch Barzel, Ginestra Bianconi, Nicholas Blumm, Liping Chi, Michele Coscia, Bruno Coutinho, Joseph De Nicolo, Zalán Forró, Jianxi Gao, Dina Ghiassian, Gourab Ghoshal, Santiago Gil, Marta Gonzales, Natali Gulbahce, Philipp Hövel, Hiroyasu Inoue, Tao Jia, Qing Jin, Maksim Kitsak, Tal Koren, Mauro Martino, Ronaldo Menezes, Deok-Sun Lee, Sune Lehmann, Yang Liu, Jörg Menche, Juyong Park, Alec Pawling, Márton Pósfai, Zehui Qu, Sabrina Rabello, Maximilian Schich, Amitabh Sharma, Filippo Simini, Georgios Tsekenis, Pu Wang, Xuezhong Zhou. You made CCNR a place I will almost surely miss. Special thanks go to the CCNR staff, especially Suzanne Aleva, Agnes Petrozcky, Trevor Gillaspy, and Sarah Weiss, who have been the indispensable force that keeps the center running.

Many friends made my life along this journey seem easy. Among them, Yiqiao Li, David Moser, and Guanqun Zhang suffered the most from my occasional complaints.

I own my deepest gratitude to my parents. Their unconditional love and support made me who I am today. My grandpa remembers everything I tell him. His love and care makes me feel blessed.

Finally, this dissertation is dedicated to my partner and wife, Tian Shen. You are the best thing that ever happened to me.

Table of Contents

Chapter 1

Introduction

Over the past few years, we have witnessed a cross-disciplinary shift. Supplied by a wide range of high-throughput tools and technologies, a wealth of data are fueling rapid advances in a number of research fields, from physics and cell biology to computer science and economics. Nowhere are these advances more apparent than in the study of social systems. Indeed, today just about everything we do is recorded and saved in a database somewhere around the globe. Every time we make a phone call, when, where, and whom we call is carefully catalogued for billing purposes by our mobile provider. So does our credit card company, who has every incentive to save as many details as they can about each swipe of our card. While we often do not realize it, we are constantly under various microscopes that transform every pieces of our life into bits and securely store them in data centers. In addition, with the development of Web 2.0, we ourselves—no matter what the occupation is—have uniformly become the most avid contributors to this data windfall. Every YouTube videos we watched were originally uploaded by our fellow web surfers. Our news feed on social media and social networking sites, like Facebook and Twitter, are generated from updates by our 'friends' on that site. Wikipedia, the largest encyclopedia, is created collaboratively by volunteers. The list goes on, and the result is a data revolution, now commonly called, Big Data.

These data offer researchers access to patterns of human behavior at a scale and level of details previously unimaginable, representing a huge opportunity for research. The fact that statistical physicists have played an important role in this inherently interdisciplinary field should be of little surprise. Indeed, with its decades of experiences in critical phenomena, statistical physics has much to offer, particularly in understanding, quantifying and modeling the dynamics and properties of a large number of individuals. Its approach to these seemingly complex problems in social systems can be simplistic and sometimes crude—by viewing humans as atoms—yet, its occasional success in doing so underlines the concept of universality in critical phenomena that the macroscopic behavior of a system is independent of microscopic details. On the other hand, while we are inundated by the Big data with its unprecedented scale and much finer resolution, we are also handed a powerful experimental apparatus to challenge our existing models, explore new tools and frameworks, and lead ourselves to areas that we could not venture before.

In this dissertation, by applying tools and techniques developed in statistical physics, I present some of my contributions to the emerging field of Big Data in three distinct but related settings, with the hope to highlight the opportunities and promises offered by Big Data. For instance, we shall see some of the first evidence of predicting social relationships based on spatiotemporal information, thanks to the availability of large-scale datasets that simultaneously capture social interactions and human movements on a national scale (Ch. 4). The predictive power was made possible by applying the correlation function in a spin system to individual trajectories. We shall see the power of combining and matching various sources of Big Data, helping us uncover the contextual factors that affect information spreading processes on both global and local scale, where a simple stochastic model yields surprising agreement with the global properties of the system (Ch. 3). We shall also see an example of agent-based modeling from first principles, in the wake of high resolution longitudinal datasets, allowing us to document a remarkable amount of regularity in a system that was perceived as noisy and unpredictable (Ch. 2).

The rest of this dissertation is organized as follows.

In Chapter 2, we present a mechanistic model (MiC) that captures long-term scientific impact. An ability to accurately assess the long-term impact of a scientific discovery has implications from science policy to individual reward. Yet, the documented lack of predictability of citation based measures frequently used to gauge impact, from impact factors to short-term citations, raises a fundamental question: is there long-term predictability in citation patterns? In this chapter we test the hypothesis that impact is a collective measure that reflects the research community's response to a discovery, hence it follows quantifiable patterns. We derive a mechanistic model for the citation dynamics of individual papers, allowing us to collapse the citation histories of papers from different journals and disciplines into a single curve, indicating that all papers follow the same universal temporal pattern. The observed patterns not only help us uncover the basic mechanisms that govern scientific impact, but also offer reliable measures of influence with potential policy implications.

Chapter 3 asks a simple question: what are the factors that affect information spreading processes? Information spreading processes are central to human interactions. Despite recent studies in online domains, little is known about factors that could affect the dissemination of a single piece of information. In this chapter we address this challenge by combining two related but distinct datasets, collected from a large scale privacy-preserving distributed social sensor system. We find that the social and organizational context significantly impacts to whom and how fast people forward information. Yet the structures within spreading processes can be well captured by a simple stochastic branching model, indicating surprising independence of context. Our results build the foundation of future predictive models of information flow and provide significant insights towards design of communication platforms.

In Chapter 4, we turn our attention to the interplay between social network and human movements. Our understanding of how individual mobility patterns shape and impact the social network is limited, but is essential for a deeper understanding of network dynamics and evolution. This question is largely unexplored, partly due to the difficulty in obtaining large-scale society-wide data that simultaneously capture the dynamical information on individual movements and social interactions. Here we address this challenge by tracking the trajectories and communication records of 6 Million mobile phone users. We find that the similarity between two individuals' movements strongly correlates with their proximity in the social network. We further investigate how the predictive power hidden in such correlations can be exploited to address a challenging problem: which new links will develop in a social network. We show that mobility measures alone yield surprising predictive power, comparable to traditional network-based measures. Furthermore, the prediction accuracy can be significantly improved by learning a supervised classifier based on combined mobility and network measures. We believe our findings on the interplay of mobility patterns and social ties offer new perspectives on not only link prediction but also network dynamics.

These three chapters (Ch. 2–4) are written in a self-contained manner, with the hope that readers who are of interest to any particular topic could "get to the meat" without having to consult the other chapters.

Chapter 2

Quantifying Long-term Scientific Impact

Of the many tangible measures of scientific impact one stands out in its frequency of use: citations [42, 28, 90, 54, 68, 117, 58, 86, 39]. The reliance on citation based measures, from the Hirsch index [54] to the g-index [38], from impact factors [42] to eigenfactors [40], and on diverse ranking based metrics [21, 87], lies in the (often debated) perception that citations offer a quantitative proxy of a discovery's importance or a scientist's standing in the research community. In this debate it is often lost the fact that our ability to foresee lasting impact based on citation patterns has well-known limitations:

(i) The *impact factor* (IF) [42], conferring a journal's historical impact to a paper, is a poor predictor of a particular paper's future citations [93]: papers published in the same journal a decade later acquire widely different number of citations, from one to thousands (Fig. 2.1A).

(ii) The *number of citations* [28] collected by a paper strongly depends on the paper's age, hence citation based comparisons favor older papers and established investigators. It also lacks predictive power: a group of papers that within a five year span collect the same number of citations are found to have widely different long-term impact (Fig. 2.1B).

(iii) Paradigm changing discoveries have notoriously limited early impact [90], precisely because the more a discovery deviates from the current paradigm, the longer it takes to be appreciated by the community [64]. Indeed, while for most papers their early and long-term citations correlate, this correlation breaks down for discoveries with most long-term citations (Fig. 2.1C). Hence, publications with exceptional long-term impact appear to be the hardest to recognize based on their early citation patterns.

(iv) Comparison of different papers is confounded by incompatible publication/citation/acknowledgement traditions of different disciplines and journals.

These limitations not only affect science policy, but also probe our understanding of complex evolving systems [24, 19, 107, 31, 17], prompting us to ask, is there long-term predictability in such short-term measures as early citation patterns? To be sure, long-term cumulative measures like the Hirsch index have documented predictable components, that can be extracted via data mining [54, 1]. Yet, given the myriad of factors involved in the recognition of a new discovery, from the work's intrinsic value to timing, chance and the publishing venue, finding regularities in the citation history of *individual papers*, the minimal carriers of a scientific discovery, remains an elusive task.

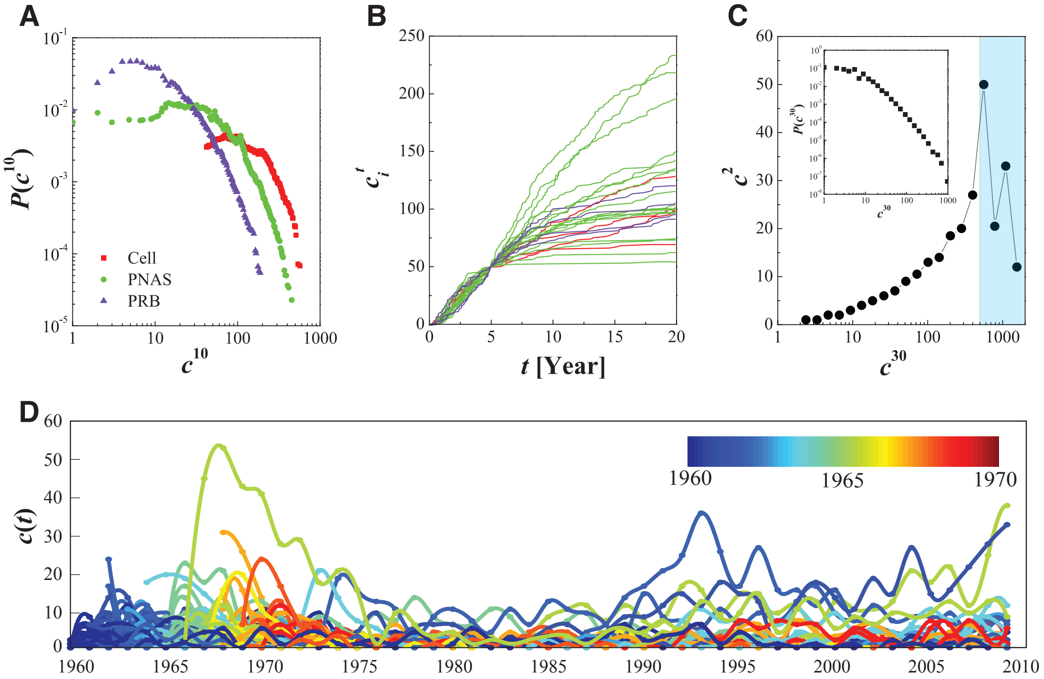


Figure 2.1: **Characterizing citation dynamics**. (A) Distribution of the cumulative citations ten years after publication (c^{10}) for all papers published in *Cell*, *PNAS*, and *Physical Review B* (*PRB*) in 1990. (B) Citation history of all papers shown in (A) that acquired 50 citations 5 years after publication, illustrating the different long-term impact despite their equal early impact. (C) Average number of citations acquired two years after publication (c^2) for papers with the same long-term impact (c^{30}), indicating that for high impact papers ($c^{30} \geq 400$, shaded area) the early citations underestimate future impact. Inset: Distribution of citations 30 years after publication (c^{30}) for PR papers published between 1950 and 1980. (D) Yearly citation $c_i(t)$ for 200 randomly selected papers published between 1960 and 1970 in the Physical Review (PR) corpus. The color code corresponds to each papers' publication year.

The difficulty in identifying reproducible patterns in citation histories is well illustrated by the citation patterns of papers extracted from the Physical Review corpus (Fig. 2.1D), consisting of 463,348 papers published between 1893 and 2010 and spanning all areas of physics [90, 69, 88]. The fat tailed nature of the citation distribution 30 years after publication indicates that while most papers are hardly cited, a few do have exceptional impact (Fig. 2.1C inset) [89, 28, 9, 90, 86]. This impact heterogeneity, coupled with widely different citation histories (Fig. 2.1D), suggests a lack of order and hence lack of predictability in citation patterns. Yet, as we show in this chapter, this lack of order in citation histories is only apparent, as citations follow widely reproducible dynamical patterns that span research fields. Quantifying these patterns allow us to derive from first principles more accurate impact measures than the currently used heuristic quantities.

2.1 Data Description

To demonstrate the practical relevance of our study, we compiled two citation datasets.

2.1.1 Physical Review Corpus

The Physical Review (PR) dataset consists of all papers published by journals within the Physical Review corpus from 1893 to 2010 (Table 2.1). The data is available by request through the APS website. The corpus is comprised of over 450,000 papers with citations. The data is unique in its longitudi-

Table 2.1: **Statistics of PR Corpus.**

Journal	Start Year	End Year	# Papers
Physical Review (Series I)	1893	1912	1,469
Physical Review	1913	1969	47,941
Reviews of Modern Physics	1929	2009	2,926
Physical Review Letters	1958	2009	95,516
Physical Review A	1970	2009	53,655
Physical Review B	1970	2009	137,999
Physical Review C	1970	2009	29,935
Physical Review D	1970	2009	56,616
Physical Review E	1993	2009	35,944
Physical Review Special Topics - Accelerators and Beams	2002	2009	1,257
Physical Review Special Topics - Physics Education Research	2005	2009	90

nal nature, spanning over 100 years. Therefore, it is ideal for understanding the long term aspects of citation histories and impact. Yet, it has two major limitations: (1) It is discipline specific, containing physics papers only. Therefore, the results obtained using this data need to be checked on data pertaining to other disciplines. (2) The data includes only internal citations. Hence, the Web of Science citations of each paper are higher than contained in this data. Such incompleteness introduces a systematic undercount of the impact of interdisciplinary papers. Hence we systematically validate our results on Web of Science data.

2.1.2 Web of Science

To correct for the limitations of the Physical Review corpus and to test the generality of our results, we also downloaded papers and citations from Web of Science database from Thomson Reuters. This dataset indexes citations using six major databases with comprehensive coverage. Thus it automatically solves limitation (2). To address limitation (1), we selected 12 journals based on their reach and impact (Table 2.2). They include general audience journals, like *Nature*, *Science*, and *Proceedings of the National Academy of Sciences* (*PNAS*); leading journals within a certain discipline, like *Cell*, *New England Journal of Medicine* (*NEJM*), *Physical Review Letters* (*PRL*); and review journals, like *Reviews of Modern Physics* (*RMP*). We downloaded all papers published by these journals in three different years (1990, 1995, 2000), and all citations collected by each of these papers until 2011.

To test the mechanism behind the temporal changes in impact factor of *Cell* and *NEJM*, we also downloaded all papers published by these journals from 1995 to 2005 and their citations. Web of Science only provides publication year of papers published prior to 1985, not publication date. This prevents us from accurately (day resolution) estimating the parameters for papers published before 1985. This limitation is corrected by the excellent longitudinality of the PR corpus.

2.2 Empirical Observations

Besides the results described in Fig. 2.1, there are rather robust characteristics for a citation system. In this section, we take PR dataset as an example and outline some major empirical observations.

Explosive Growth of Publications

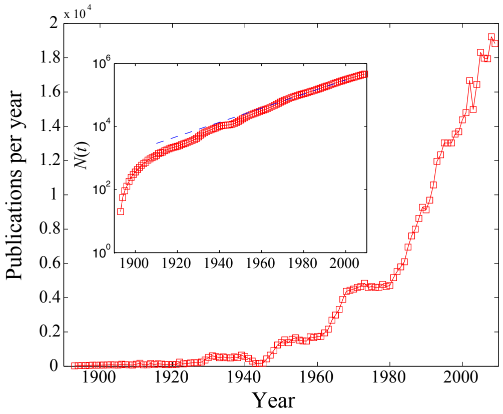


Figure 2.2: The number of papers published each year in the PR corpus. Inset: cumulative number of papers $\mathcal{N}(t)$ published up to year t.

The number of scientific publications grows exponentially, a pattern first pointed out by Price in 1963 [27]. Since then, various groups have shown that this pattern not only holds for the overall scientific enterprise, but also within each discipline [32, 101, 13]. We show in Fig. 2.2 the number of papers published in each year in the PR corpus. The inset of Fig. 2.2 gives a cumulative view, i.e., total number of papers published up to a certain year, on a log-linear scale. Figure 2.2 is in good agreement with previous findings

[13] that the number of papers published each year increases exponentially, in analogy to Moore's law describing the development of technology, indicating that

$$\mathcal{N}(t) \sim \exp(\beta t), \tag{2.1}$$

where $\beta = (17year)^{-1}$ for PR corpus. Therefore the number of papers increases by 2.73 times after 17 years, or equivalently doubles every $17 \times \ln(2) = 11.8$ years.

The 'Jump-decay' Pattern

Despite the exponential growth of the system, the average number of citations $c(t) \equiv \overline{c_i(t)}$ of all papers at time t after publication follows a distinct 'jump-decay' pattern, indicating that a paper's main impact comes in the first two years after publication and diminishes over time (Fig. 2.3A) [28].

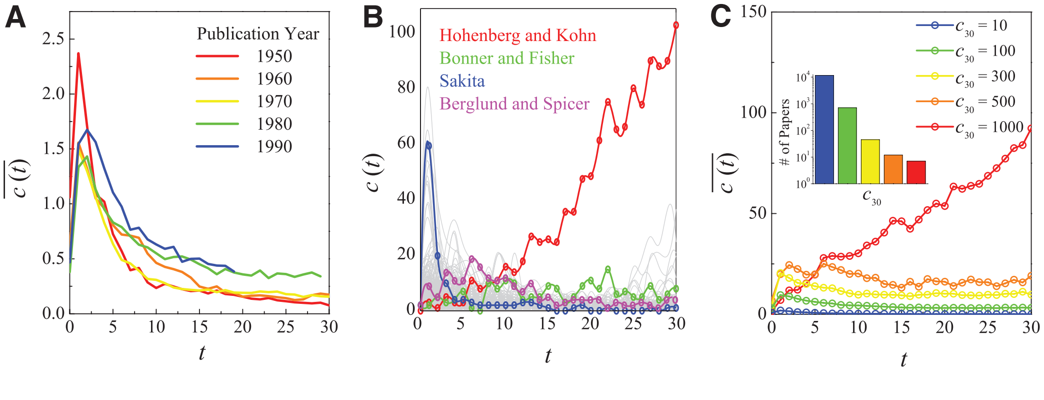


Figure 2.3: (A) Average citations for papers published in same year. (B) Citation dynamics for all papers published by PR in 1964. Four papers are highlighted for illustration. (C) Average annual citations for papers with the same cumulative citations after 30 years c^{30}.

At the same time $\overline{c_i(t)}$ hides a remarkable diversity in individual citations histories. To illustrate this we show the citation history of all papers published in 1964 (Fig. 2.3B), finding that while most obey the 'jump-decay' pattern (blue), a few reach their peak years after publication (magenta), yet others attract a constant number of citations over decades (green) and some continue to acquire an increasing number of citations even 30 years after their publication (red). To see if these differences correlate with long term impact, we grouped papers based on their total number of citations after 30 years, $c^{30} \equiv \sum_{t=1}^{30} c_i(t)$, measuring the shape of $\overline{c_i(t)}$ for each c^{30} group (Fig. 2.3C). We find that the higher is a paper's early impact (c^2), the higher is its long-term impact (c^{30}), a relationship that holds for all but the highest impact papers (Fig. 2.1C). Indeed, papers with $c^{30} \geq 1000$ have limited early impact (c^2), suggesting these exceptional long-term impact papers (about 0.01% of all) are the hardest to foresee based on their short term citation pattern.

2.3 Minimal Citation Model (MiC)

Next we turn our attention to modeling the citation dynamics of individual papers. In this section, we present a minimal citation (MiC) model, that captures all known quantifiable mechanisms that affect citation histories. We will then show that MiC, although relying on minimal assumptions, accurately captures the citation dynamics of individual papers, revealing a remarkable degree of regularity in citation histories.

2.3.1 Three Assumptions

We start by identifying three fundamental mechanisms that drive the citation history of individual papers:

A) Preferential attachment

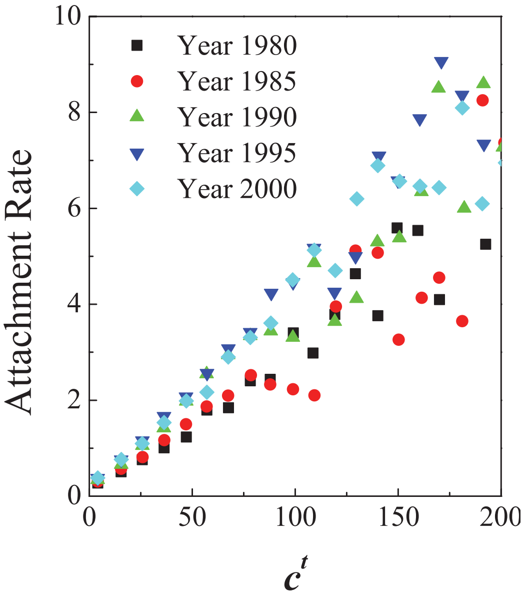


Figure 2.4: **Empirical validation of preferential attachment**. Attachment rate measures the likelihood for new papers published in different years (color coded) to cite an old paper with c^t citations. That is, for each year, c^t measures the citations of each paper before this year, and attachment rate measures the average number of times each paper with c^t citations was cited in this year. The linearity of the curves offer evidence for preferential attachment.

Preferential attachment captures the well-documented fact that highly cited papers are more visible and are more likely to be cited again than less-cited contributions [9, 90, 17, 94]. Accordingly a paper i's probability to be cited again is proportional to the total number of citations c_i the paper received previously. To document the presence of preferential attachment in our dataset, we follow the methodology reported in [90] (Fig. 2.4). Attachment rate measures the likelihood for a paper with c^t citations to get cited by a new paper. We measured this for different years (color coded). For each year, we first count the citations of each paper before this year, and then measure the number of times each paper with c^t citations was cited in this year. The linearity of the curves offer evidence for preferential attachment.

B) Aging

Aging captures the fact that new ideas are integrated in subsequent work, hence each paper's novelty fades with time [79, 30, 108]. The resulting long term decay is best described by a log-normal survival probability. This temporal relaxation function $P(\Delta t_i)$ can be measured directly from the real data. Given that a paper's citation is driven by three independent forces, that are difficult to separate from each other, we need to control the influence of these factors, isolating the temporal decay. This is similar to measuring preferential attachment from empirical data, where one keeps a constant time window and looks at the growth of degrees as a function of existing degree [90]. To achieve this we should group papers with the same fitness (η) and cumulative citations (c^t), and look at the time when they are cited again. But we do not know η beforehand. Moreover, each paper is likely characterized by different μ and σ parameters in (2.2). Therefore, by aggregating different papers, we will measure a superposition of different temporal relaxation functions. We therefore selected papers published between 1950 and 1960 in the PR corpus with fixed cumulative citations c (i.e., controlling for c, publication time and IF), and tracked the moment when their citations changed from c to $c+1$. We then measured Δt, i.e. time between their publication and when $c \to c+1$ took place.

Figure 2.5 shows both $P(\ln \Delta t \mid c)$ and $P(\Delta t \mid c)$ for fixed $c = 10$ and $c = 20$, finding that the relaxation function is best approximated by a lognormal function

$$P(\Delta t) = \frac{1}{\sqrt{2\pi}\sigma\Delta t} \exp\left(-\frac{(\ln \Delta t - \mu)^2}{2\sigma^2}\right). \tag{2.2}$$

Indeed, a lognormal distribution naturally emerges in multiplicative processes, frequently used to model the temporal relaxation function in diverse settings, from survival times after cancer diagnosis [14] and latency periods of diseases [92] to the duration of marriages [85] and length of both spoken and written conversations [114, 52]. Theoretically, we obtain a lognormal distribution if the relaxation time Δt is a product of independent, identical distributed random variables $\{x_i\}$, $\Delta t = \prod_i^n x_i$ (equivalently, $\ln \Delta t = \sum_i^n \ln x_i$ converges to a normal distribution due to the central limit theorem). Such multiplicative processes often result from independent decision processes [84, 105]. Similar mechanisms are likely at work in the case of citations: a decision to cite a paper involves balancing many different factors, from appropriateness to novelty, relevance and even citation limits, each of which

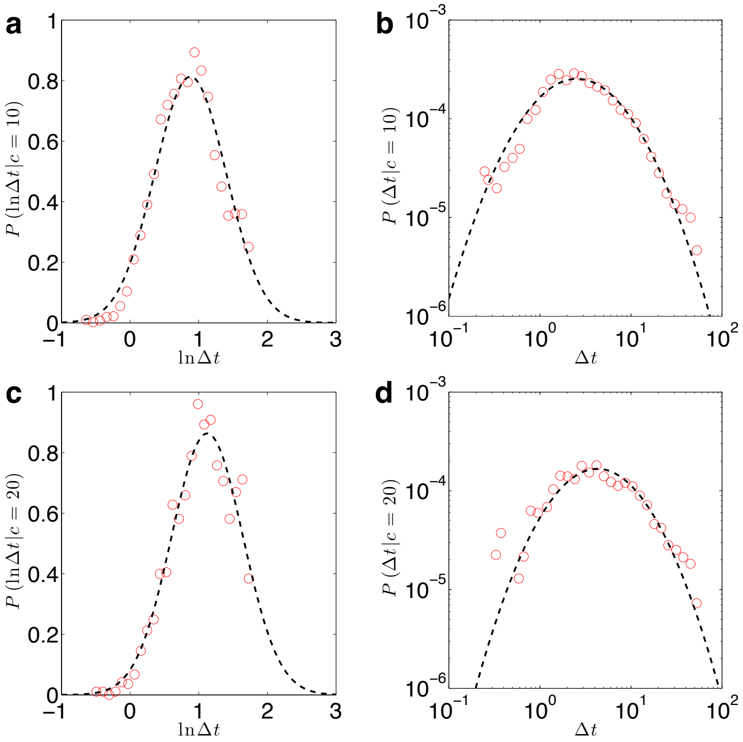


Figure 2.5: **Empirical validation of the lognormal decay (2.2)** (a) $P\,(\ln \Delta t)$ when papers change from 10 citations to 11 citations. The dashed line corresponds to the best gaussian fitting. (b) Same as (a) but for $P(\Delta t)$. Dashed line corresponds to the best lognormal fitting ($\mu = 7.85$ and $\sigma = 1.01$). Here Δt is measured in unit of years. (c) $P\,(\ln \Delta t)$ when papers change from 20 citations to 21 citations. The dashed line corresponds to the best gaussian fitting. (d) Same as (c) but for $P(\Delta t)$. The dashed line corresponds to the best lognormal fitting ($\mu = 8.29$ and $\sigma = 0.93$).

may be approximated as an independent event with random probability, resulting in a random latent waiting time. The final decision to cite requires us to satisfy all these individual conditions, best described by a multiplicative process. This argument offers an intuitive explanation for the origin of the observed lognormal relaxation time. More models that generate lognormal relaxation times are reviewed in Ref. [103].

C) Fitness

Fitness, η_i, captures the inherent differences between papers, accounting for the perceived novelty and importance of a discovery [12, 18, 31]. Novelty and importance depend on so many intangible and subjective dimensions that it is impossible to objectively quantify them all. Here we bypass the need to evaluate a paper's intrinsic value and view fitness η_i as a *collective measure capturing the community's response* to a work. As we show below, η_i can be extracted from a paper's citation history.

2.3.2 Solving the Model

In the proposed model, the total number of papers $\mathcal{N}$ grow exponentially as (2.1), and every new published paper has $m = 30$ citations to existing papers. Combining A–C, we write the probability that paper i is cited at time t after publication as

$$\Pi_i(t) \sim \eta_i c_i(t) P_i(t). \tag{2.3}$$

Hence the time evolution of the expected number of citations c_i^t satisfies

$$\frac{dc_i^t}{dN} = \frac{\Pi_i}{\sum_{i=1}^{N} \Pi_i}. \tag{2.4}$$

Combining (2.3-2.4) with Eq. (2.1) leading to $\Delta t_i = t - t_i = \beta^{-1}\ln(N/i)$, we obtain

$$\frac{dc_i}{dN} = m\frac{c_i\eta_i P_t(\beta^{-1}\ln(N/i))}{\sum_{i=1}^{N} c_i\eta_i P_t(\beta^{-1}\ln(N/i))}. \tag{2.5}$$

Assuming $c_i = m(f(\eta_i, \Delta t_i) - 1)$, we have

$$\frac{df(\eta_i, \Delta t_i)}{d\Delta t_i} = \beta\frac{\eta_i f(\eta_i, \Delta t_i) P_t(\Delta t_i)}{A}, \tag{2.6}$$

with the initial condition $f(\eta_i, 0) = 1$, where the normalization constant

$$\begin{aligned} A &\equiv \lim_{N\to\infty} N^{-1}\left\langle \sum_{i=1}^{N} \eta_i P_t(\beta^{-1}\ln(N/i)) f(\eta_i, \beta^{-1}\ln(N/i)) \right\rangle \\ &= \lim_{N\to\infty}\left\langle \int_1^N \eta_i P_t(\beta^{-1}\ln(N/i)) f(\eta_i, \beta^{-1}\ln(N/i)) d(i/N) \right\rangle \\ &= \beta\int d\eta\rho(\eta)\int_0^\infty \eta P_t(t') f(\eta, t') e^{-\beta t'} dt'. \end{aligned} \tag{2.7}$$

The solution of Eq. (2.6) is

$$f(\eta_i, \Delta t_i) = e^{\frac{\beta}{A}\eta_i \int_0^{\Delta t_i} P_t(t')dt'}, \tag{2.8}$$

thus

$$c_i^{\Delta t_i} = m\left(e^{\frac{\beta}{A}\eta_i \int_0^{\Delta t_i} P_t(t')dt'} - 1\right), \tag{2.9}$$

where the constant A can be calculated from

$$\beta\int \rho(\eta) d\eta \int_0^\infty \exp\left(-\beta t + \frac{\beta}{A}\eta\int_0^t P_t(t')dt'\right) dt = 2. \tag{2.10}$$

Plugging Eq. (2.2) into Eq. (2.9), we get

$$c_i^t = m\left(e^{\frac{\beta}{A}\eta_i\Phi\left(\frac{\ln t-\mu_i}{\sigma_i}\right)} - 1\right), \tag{2.11}$$

where $\Phi(x)$ is the cumulative normal distribution

$$\Phi(x) = \Phi(x) \equiv (2\pi)^{-1/2}\int_{-\infty}^{x} e^{-y^2/2}dy. \tag{2.12}$$

As β and A are system wide parameters, we use $\lambda_i \equiv \eta_i\beta/A$ the relative fitness for each paper i, arriving at Eq. (3) that describes the citation dynamics of paper i:

$$c_i^t = m\left(e^{\lambda_i\Phi\left(\frac{\ln t-\mu_i}{\sigma_i}\right)} - 1\right), \tag{2.13}$$

Equation (2.13) represents a minimal citation (MiC) model, that captures all known quantifiable mechanisms that affect citation histories. It predicts that the citation history of paper i is characterized by three fundamental parameters: the *relative fitness* $\lambda_i \equiv \eta_i\beta/A$, capturing a paper's importance relative to other papers; the *immediacy* μ_i, governing the time for a paper to reach its citation peak and the *longevity* σ_i, capturing the decay rate. Using the rescaled variables $\tilde{t} \equiv (\ln t - \mu_i)/\sigma_i$ and $\tilde{c} \equiv \ln(1 + c_i^t/m)/\lambda_i$, we obtain our main result,

$$\tilde{c} = \Phi(\tilde{t}), \tag{2.14}$$

predicting that each paper's citation history should follow the same universal curve $\Phi(\tilde{t})$ if rescaled with the paper-specific $(\lambda_i, \mu_i, \sigma_i)$ parameters. Given the obvious diversity of citation histories (Fig. 2.1D), this prediction is somewhat unexpected.

2.3.3 Maximum Likelihood Estimation of Model Parameters

In order to test how well MiC matches empirical data, we need to estimate the best $(\lambda_i, \mu_i, \sigma_i)$ parameters for each individual paper i given its citation history. We show in this section that this can be done by considering a non-homogeneous stochastic process, with the events corresponding to the arrival of individual citations. Imagine a stochastic process $\{x(t)\}$ where $x(t)$ represents the number of events by time t, satisfying

$$\mathrm{Prob}(x(t+h) - x(t) = 1) = \lambda_0(x, t)h + \mathcal{O}(h), \tag{2.15}$$

where $\lambda_0(x, t)$ is a time dependent rate parameter. Given an empirically observed set of N events $\{t_i\}$ within the time period $[0, T]$, where t_i indicates the moment when the paper gets cited the i^{th} time, the likelihood that a paper's citation dynamics follows the model can be evaluated by the log-likelihood function

$$\begin{aligned}\ln L &= \sum_{i=1}^{N} \ln\left(\lambda_0(i-1,t_i)\right) - \int_0^T \lambda_0(x(t),t)dt \\ &= \sum_{i=1}^{N} \ln\left(\lambda_0(i-1,t_i)\right) - \sum_{i=0}^{N} \int_{t_i}^{t_{i+1}} \lambda_0(i,t)dt.\end{aligned} \tag{2.16}$$

From Eq. (2.6), we have

$$\lambda_0(x,t) = \frac{\lambda(x+m)}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{(\ln(t)-\mu)^2}{2\sigma^2}\right) \tag{2.17}$$

and

$$\int \lambda_0(x,t)dt = \lambda(x+m)\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right). \tag{2.18}$$

Combining Eqs. (2.16) and (2.18), we find

$$\begin{aligned}\ln L = N\ln\lambda &+ \sum_{i=1}^{N}\ln(i+m-1) + \sum_{i=1}^{N}\ln\left(P(t_i)\right) \\ &- \lambda\sum_{i=0}^{N}(i+m)\left[\Phi\left(\frac{\ln(t_{i+1}-t_0)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(t_i)-\mu}{\sigma}\right)\right] \\ &= N\ln\lambda + \sum_{i=1}^{N}\ln(i+m-1) + \sum_{i=1}^{N}\ln\left(P(t_i)\right) \\ &- \lambda(N+m)\Phi\left(\frac{\ln(T)-\mu}{\sigma}\right) + \lambda\sum_{i=1}^{N}\Phi\left(\frac{\ln(t_i)-\mu}{\sigma}\right).\end{aligned} \tag{2.19}$$

After the change of variables $\hat{l} = (\ln L)/N$ and $\hat{m} = m/N$, we have

$$\begin{aligned}\hat{l} = \ln\lambda + \langle \ln((i-1)/N + \hat{m})\rangle + \left\langle \ln P(t_i) + \lambda\Phi\left(\frac{\ln(t_i) - \mu}{\sigma}\right)\right\rangle \\ - \lambda(1+\hat{m})\Phi\left(\frac{\ln(T) - \mu}{\sigma}\right).\end{aligned} \tag{2.20}$$

As the goal is to maximize $\ln L$, which is the same as maximizing $\hat{l}$ (2.20), we can obtain the set of parameters that best capture a paper's citation records $(\lambda^*, \mu^*, \sigma^*)$,

$$(\lambda^*, \mu^*, \sigma^*) = \text{arg max}_{\lambda,\mu,\sigma}\hat{l}(\lambda, \mu, \sigma), \tag{2.21}$$

or

$$\begin{aligned}\frac{\partial l(\lambda^*, \mu^*, \sigma^*)}{\partial \lambda^*} &= 0 \\ \frac{\partial l(\lambda^*, \mu^*, \sigma^*)}{\partial \mu^*} &= 0 \\ \frac{\partial l(\lambda^*, \mu^*, \sigma^*)}{\partial \sigma^*} &= 0.\end{aligned} \tag{2.22}$$

The first equation in (2.22) leads to

$$\lambda^* = \left[(1+\hat{m})\Phi\left(\frac{\ln(T) - \mu^*}{\sigma^*}\right) - \left\langle \Phi\left(\frac{\ln(t_i) - \mu^*}{\sigma^*}\right)\right\rangle\right]^{-1}. \tag{2.23}$$

and the rest two are

$$
\begin{aligned}
&\left\langle \frac{\ln(t_i) - \mu^*}{\sigma^*} - \lambda^* P_G\left(\frac{\ln(t_i) - \mu^*}{\sigma^*}\right)\right\rangle + \lambda^*(1+\hat{m})P_G\left(\frac{\ln(T) - \mu^*}{\sigma^*}\right) = 0 \\
&\left\langle \frac{\ln(t_i) - \mu^*}{\sigma^*}\left(\frac{\ln(t_i) - \mu^*}{\sigma^*} - \lambda^* P_G\left(\frac{\ln(t_i) - \mu^*}{\sigma^*}\right)\right)\right\rangle \\
&+ \lambda^*(1+\hat{m})\frac{\ln(T) - \mu^*}{\sigma^*} P_G\left(\frac{\ln(T) - \mu^*}{\sigma^*}\right) = 1,
\end{aligned}
\tag{2.24}
$$

where $P_G(x) \equiv (2\pi)^{-1/2} e^{-x^2/2}$ is the standard normal distribution.

By solving Eqs. (2.23–2.24) numerically, we obtained the parameter set $(\lambda^*, \mu^*, \sigma^*)$ for each paper based on its historical citation pattern within the time period $[0, T]$.

2.3.4 Model Validation

To test the validity of (2.14) we first determined (λ, μ, σ) for four papers selected for their widely different citation histories (Fig. 2.6A), finding that after rescaling they all collapse into a single curve (2.14) (Fig. 2.6B). The reason is explained in Fig. 2.6C: by varying λ, μ and σ, Eq. (2.13) can account for a wide range of empirically observed citation histories, from jump-decay patterns to delayed impact. Yet, to test the validity of MiC, we rescaled all papers published between 1950 and 1980 in the Physical Review corpus, finding that they all collapse into (2.14) (Fig. 2.6D). We also tested our model on all papers published in 1990 by 12 prominent journals (Table S2),

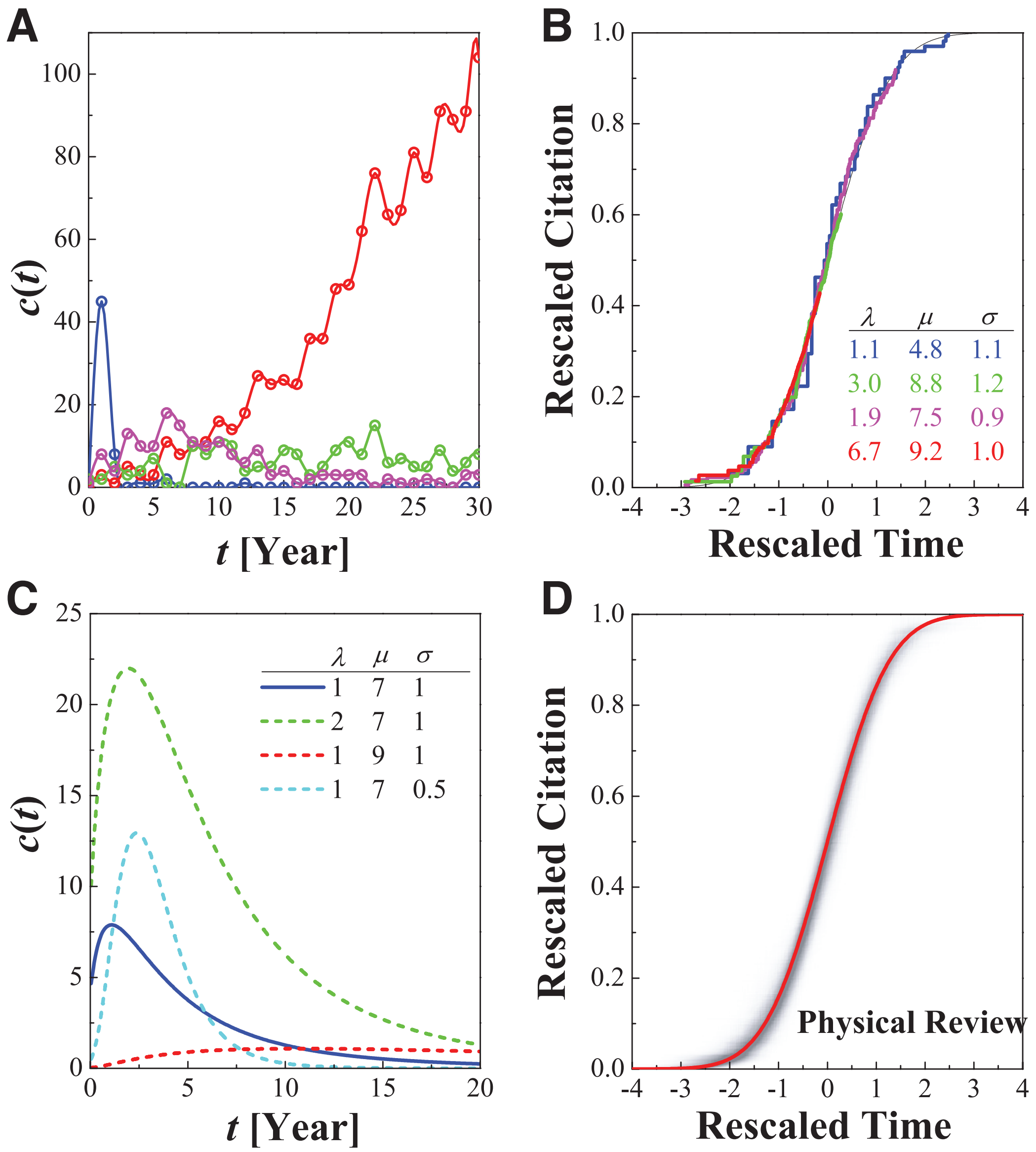


Figure 2.6: **Validating MiC**. (A) Citation history of four papers published in PR in 1964, selected for their distinct citation dynamics. (B) Data collapse for the four papers in (A) using Eq. (2.14). Legend: the (λ, μ, σ) parameters used to rescale the citation history of each paper. (C) Changes in the citation history $c(t)$ according to (2.13) after varying the (λ, μ, σ) parameters, indicating that (2.13) can account for a wide range of citation patterns. (D) Data collapse for 7,775 papers with more than 30 citations within 30 years in the PR corpus published between 1950 and 1980.

Table 2.2: **Citation statistics for 11 non-review journals and one review journal in 1990.** In line with citation items in definition of IF, we only include here reviews and articles. For 11 non-review journals, highest Λ, M and Σ are in bold faces. C^∞ is obtained by $C^\infty = m\left(e^\Lambda - 1\right)$.

Journal	Year	# Papers	Λ	M	Σ	C^∞
Cell	1990	485	**2.55**	6.99	1.23	354
NEJM	1990	330	2.54	7.34	1.24	350
Nature	1990	1,099	2.36	7.36	1.24	289
Science	1990	842	2.33	7.32	1.23	280
Neuron	1990	178	1.99	7.22	1.04	189
Lancet	1990	541	1.84	7.61	1.16	159
Gene-Dev	1990	200	1.83	7.17	1.09	157
JEM	1990	313	1.76	7.29	1.07	144
PNAS	1990	2,060	1.73	7.41	1.11	140
PRL	1990	1,633	1.61	7.78	**1.37**	121
PRB	1990	2,189	1.13	**7.93**	1.32	63
RMP	1990	18	3.95	8.09	1.62	1535

finding an excellent collapse for all (Fig. 2.7). The data collapse demonstrates that the observed differences in individual citation histories (Fig 2.1D) are rooted in variations in three measurable parameters: fitness, immediacy and longevity. Hence the diverse citation histories hide a remarkable degree of regularity, accurately captured by the MiC model (2.13)–(2.14).

Using the appropriate $(\lambda_i, \mu_i, \sigma_i)$ for each paper, the model is expected to generate a citation history that resembles the real citations of Fig. 2.1D. We show in Fig. 2.8 an example of randomly selected papers published between 1960 and 1970, finding excellent agreement between the model and empirical data.

Taken together, what this section shows is a rather encouraging signal.

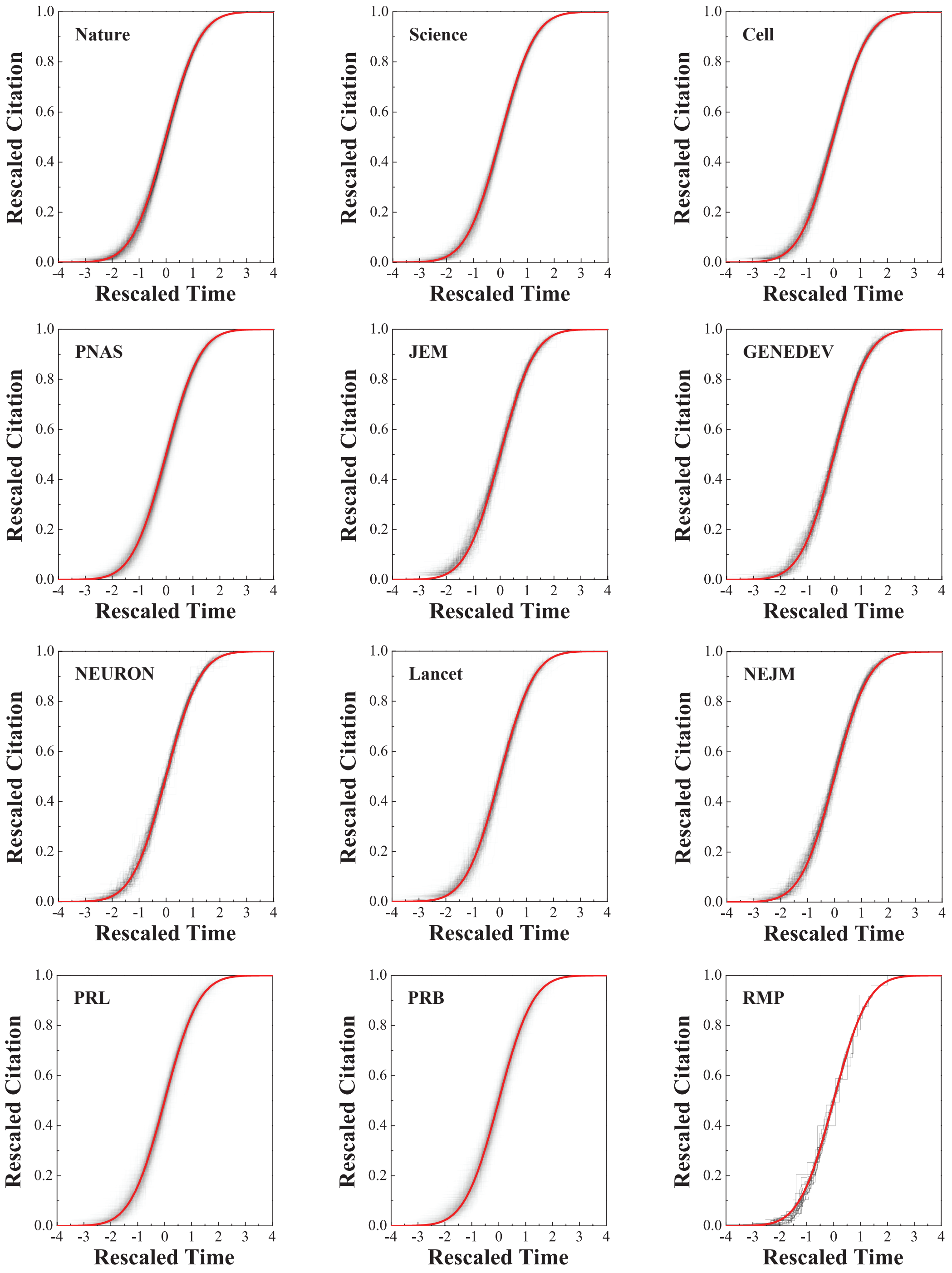


Figure 2.7: **Validating MiC for 12 journals**.

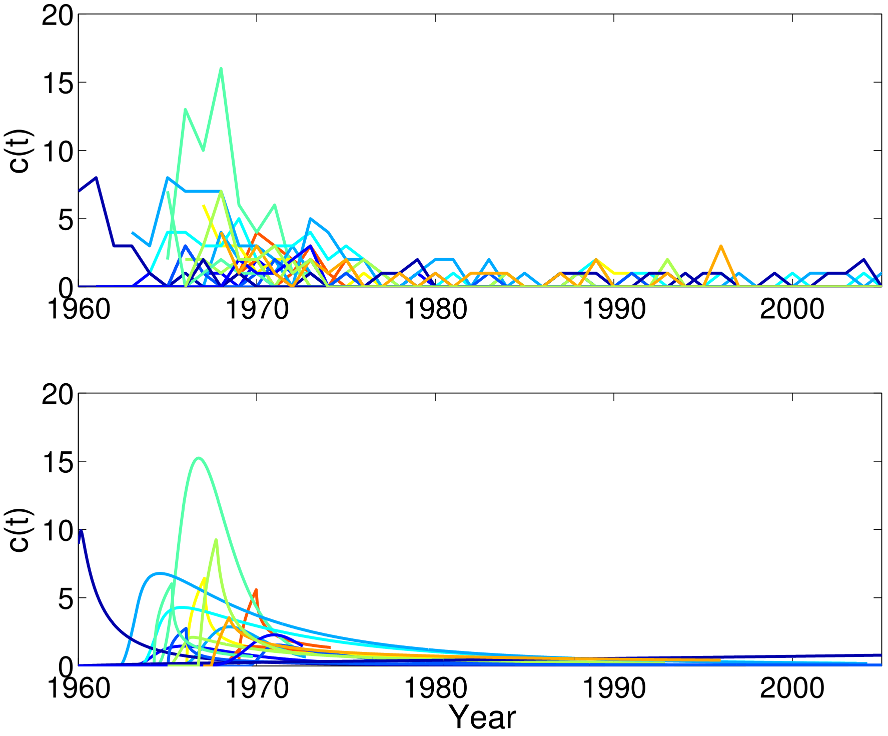


Figure 2.8: **Simulating individual citation histories.** We randomly selected two papers each year between 1960 to 1970 from the PR corpus. Their citation histories are shown on the top panel. Color code is the same as Fig. 1D, corresponding to the publication year. We estimated the set of (λ, μ, σ) parameters for each paper using the methods described in Section 2.3.3. The bottom panel shows the citation dynamics predicted by Eq. (2.13).

As noisy and unpredictable as citation histories (Fig. 2.1 and Fig. 2.3), the citation dynamics of individual papers hide a remarkable amount of regularity, that can be accurately captured by the MiC model presented here. While relying on a minimal set of ingredients that drive citation histories, MiC fits real citation dynamics remarkably well, indicating the impact dynamics that reflects the cumulative response from scientific community follows rather robust patterns, despite myriad of factors that influence them.

The rest of this chapter is organized as the following. We will present other potential models that have been or can be used to characterize citation dynamics, and discuss their strengths and limitations. Then we will spend four sections on the applications of MiC, illustrating that MiC offers us a quantitative understanding of scientific impact.

2.4 Potential Models for Citation Dynamics

The observed accuracy of the MiC model prompts us to ask whether MiC is unique in its ability to capture future citation histories. We therefore seek models that can account for the observed diversity in citation dynamics, fit citation histories, and predict future citations. While most models are not specifically designed to capture the citation dynamics of individual papers, we examine in this section some of the most relevant models and discuss their strengths and limitations. Two lines of inquiry are relevant in this context: network growth models from statistical physics built to capture citation networks, and models pertaining to diffusion of innovations in social/ecomonic sciences.

2.4.1 Network Growth Models

Scale-Free Model

The scale-free model (also known as Barabási-Albert (BA) model) is designed to reproduce the degree distribution of complex networks. Note that variants of the model were proposed by Price [28] and Simon [94]. At each step an old paper i acquires citations from a new paper with probability proportional to its current citations $\Pi_i \propto c_i^t$, a mechanism known as preferential attachment (PA).

Despite the success of the scale-free model in predicting fat-tailed citation distribution, it has difficulties capturing the citation dynamics of individual papers. Indeed, for paper i the scale-free model predicts its citation growth as [9]

$$c_i^t \sim \mathcal{N}^{1/2}, \tag{2.25}$$

where c_i^t is the cumulative number of citations paper i received, given the $\mathcal{N}$ papers in the system. This indicates that (i) all papers follow the same citation dynamics, in contrast with our observation (Fig. 1D) that each paper has a different citation history; (ii) it tells us that the citations should grow indefinitely at $\mathcal{N}^{1/2}$. If we incorporate in the model the fact that we have an exponential growth in the number of papers (see Eq. (2.1)), we find that

$$c_i^t \sim \exp(0.5\beta t), \tag{2.26}$$

indicating that the paper citations should increase exponentially over time (Fig. 2.9). Yet, the average number of citations $\overline{c_i(t)}$ of all papers i at time t after publication follows a distinct 'jump-decay' pattern (Fig. 2.3A), indicating that a paper's main impact comes during the first two years after publication and diminishes over time [28]. Therefore, both (2.25) and (2.26) represent drastic deviations from the empirical observations.

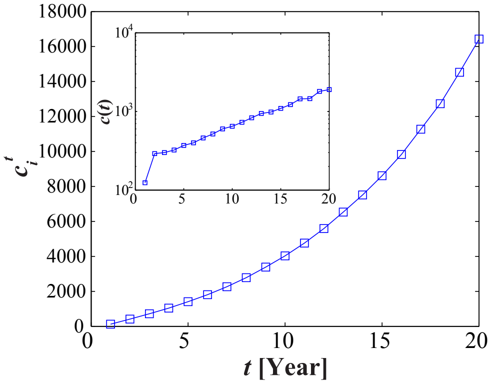


Figure 2.9: **Simulation results for the scale-free model**. We simulate a scale-free network with 100,000 nodes. Each node is associated with a time stamp such that the number of nodes in each unit time grows exponentially, following Eq. (2.1). We group together the nodes with the same time stamps, in analogy to papers published in the same year, and explore how their degrees evolve over time. Inset: the new links acquired by the selected nodes in each time step shown on a log-linear scale, demonstrating the exponential nature of the growth curve, as predicted by Eq. (2.26).

Fitness Model

In the fitness model (also known as Bianconi-Barabási (BB) model), besides the PA mechanism each paper i has an initial fitness λ_i capturing its unique likelihood to be cited in the future. That is,

$$\Pi_i \propto \lambda_i c_i^t. \tag{2.27}$$

The fitness model predicts

$$c_i^t \sim \mathcal{N}^{\alpha_i}, \tag{2.28}$$

where the exponent $\alpha_i \propto \lambda_i$, i.e. it is proportional to paper i's fitness. Given exponential growth of papers (Eq. 2.1), we find that c_i^t again increases exponentially over time, significantly deviating from the observations (Fig. 2.3A).

2.4.2 Diffusion of Innovations

The theory of diffusion of innovations aims to explain the adoption of new ideas and technologies. Although its main focus is to determine the success and failure of a product, the models often predict S-curves that are similar to the one presented in the main text. Next we explore the possibility of using diffusion S-curves to describe the citation history of individual papers.

Logistic Model

The logistic function is widely used to model population growth and product adoptions, with applications in many fields. In the context of citations one could view a paper as a new product, whose adoption leads to an increase in citations. Each paper is characterized by a different increase rate r and a total number of citations c_i^∞ that captures the differences in impact. With time, a paper's attractiveness fades, as the development along the ideas offered by the paper have been adopted by all potential adopters, hence the paper's citations approach c_i^∞. In the rate equation formalism this can be described as

$$\frac{\mathrm{d}c_i^t}{\mathrm{d}t} = r_i c_i^t \left(1 - \frac{c_i^t}{c_i^\infty}\right), \tag{2.29}$$

yielding

$$c_i^t = \frac{c_i^\infty}{1 + e^{-r_i(t-\tau_i)}}, \tag{2.30}$$

where c_i^∞, r_i and τ_i correspond to ultimate citation, longevity, and immediacy of paper i.

Bass Model

One of the most famous models in marketing and management sciences is the Bass model [10], that describes the process of new product being adopted by mass populations. The Bass model assumes the adopters of a product are influenced by two aspects: mass media and word of mouth. Hence the buyers comprise two groups. One group, the innovators as coined by Bass, is influenced only by the mass media, while the other group, the imitators, is influenced by others (word of mouth effect). Such assumptions are fairly reasonable in the context of citations. The innovators correspond to people who cite the paper spontaneously, little influenced by how many people have already cited the paper. At the same time, a paper's citations are driven by word-of-mouth diffusion (the imitators). Mathematically, this can be expressed as

$$\frac{dc_i^t}{dt} = (p + qc_i^t/c^\infty)(c^\infty - c_i^t), \tag{2.31}$$

where p characterizes “innovators”, reflecting an influence that is independent of current citations (c_i^t), and q reflects the imitation part of the model. Solving (2.31) yields

$$c_i^t = c^\infty \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}}. \tag{2.32}$$

Gompertz Model

The Gompertz model [46], named after Benjamin Gompertz, was first proposed to model human mortality. The model generates a skewed diffusion curve with long tails. In this context, early citations pave the way for new citations and drive the citation dynamics, hence the rate of research develop increases at an exponential rate. This can be formulated as

$$\frac{dc_i^t}{dt} = qc_i^t \ln(c^\infty/c_i^t), \tag{2.33}$$

yielding,

$$c_i^t = c^\infty e^{-e^{-(a+qt)}}. \tag{2.34}$$

In (2.34), a sets the displacement in c_i^t, while q characterizes the growth rate of citations.

It is worth noting that, while these three models are perhaps the most famous ones, they are far from complete to cover the list of models in diffusion of innovations. For a more comprehensive review of this body of literatures,

Table 2.3: **Modeling citation dynamics**. We identified four models that can be or have been used to fit citation histories. The table shows the corresponding rate equation and its analytical solution. In this dissertation we do not explicitly test the prediction of the Bianconi-Barabási model, as it lacks saturation for high t, hence it is unable to fit true citation histories.

Model Name	**Rate Equation**	**Solution**
MiC	$\frac{dc_i^t}{dt} \approx c_i^t \eta_i P(t)$	$c_i^t = m\left(e^{\lambda_i \Phi(\frac{\ln t - \mu_i}{\sigma_i})} - 1\right)$
Bianconi-Barabási [12]	$\frac{dc_i^t}{dt} \approx \eta_i c_i^t$	$c_i^t \sim \exp(\lambda_i t)$
Logistic [77]	$\frac{dc_i^t}{dt} = r_i c_i^t \left(1 - c_i^t / c^\infty\right)$	$c_i^t = \frac{c^\infty}{1+e^{-r_i(t-\tau_i)}}$
Bass [10]	$\frac{dc_i^t}{dt} = (p + q c_i^t / c^\infty)(c^\infty - c_i^t)$	$c_i^t = c^\infty \frac{1-e^{-(p+q)t}}{1+\frac{q}{p}e^{-(p+q)t}}$
Gompertz [46, 77]	$\frac{dc_i^t}{dt} = q c_i^t \ln(c^\infty / c_i^t)$	$c_i^t = c^\infty e^{-e^{-(a+qt)}}$

refer to [77].

2.4.3 Goodness of Fitting

The main models discussed in this section are summarized in Table 2.3. Next we evaluate the performance of these three models in describing citation dynamics, in comparison with MiC. There are two aspects we need to evaluate. One concerns with fitting. That is, how closely each model matches citation histories. This is the focus of this section. The second aspect is about the predictive power of these models, which will be discussed in more details in Section 2.5.3.

As outlined in Table 2.3, these three models, together with MiC, all have 3 parameters each, making it a fair comparison of how well each of the models fit the citation dynamics. Therefore, given a paper's citation dynamics, we can obtain the three parameters for each paper for a given model. We implemented two methods to estimate the best parameters for fitting for each the models. One is by using non-linear least square fitting, and the other one is by using Maximum Likelihood Estimation. We find these two methods have comparable performance for these three models, corroborating previous finding that in most cases these two methods perform equally well in fitting these models [77].

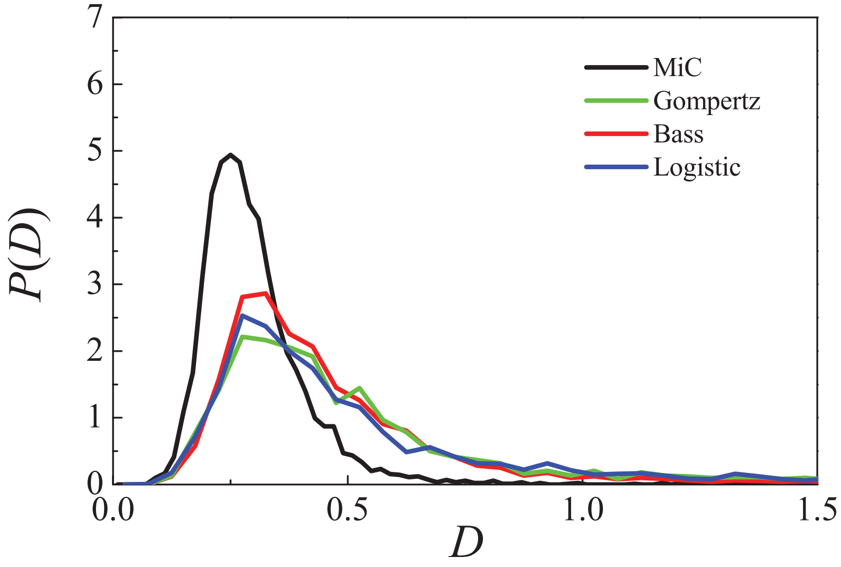


Figure 2.10: Goodness of fit using weighted Kolmogorov-Smirnov (KS) test, indicating that Eq. (2.13) offers the best fit to our testing base.

We fitted these three models to our test base of papers (all papers in the PR corpus published within the same decade (1960s) that have at least 10 citations in 5 years). To quantify how well the model fits the real data, for each paper i with total c_i^T within time period $[0, T]$ we measured the weighted KS measure

$$D_i = \max_{t\in[0,T]} \frac{|c_i^t - \tilde{c}_i^t|}{\sqrt{(1+c_i^t)(c_i^T - c_i^t + 1)}}, \tag{2.35}$$

where $\tilde{c}_i^t$ represents the citations computed by the model. A smaller D_i implies a better fit. Figure 2.10 shows the KS distribution $P(D)$ for both MiC and the three competing models, finding the best fit is offered by MiC.

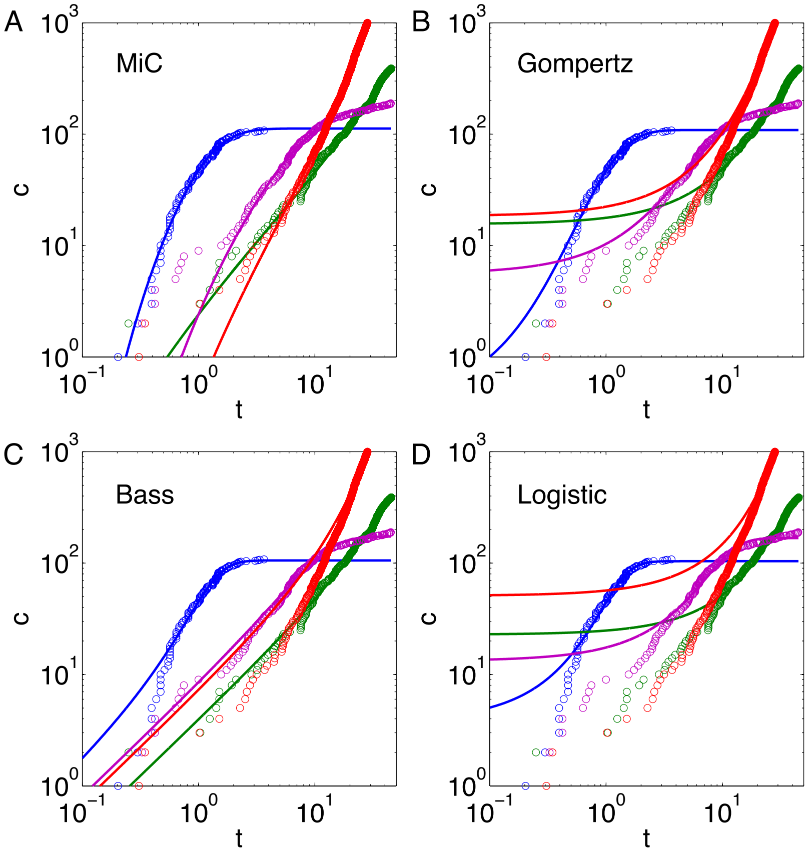


Figure 2.11: Fitting the four papers in Fig. 1G by using (A) MiC, (B) Gompertz, (C) Bass, and (D) Logistic models.

To illustrate the fitting results and compare the fit by the four model, we use the four papers shown in Fig. 2.3B and obtain the best fit (Fig. 2.11). We find, despite radical differences in citation dynamics of these four papers, MiC fit all of them consistently well. Logistic model performs the worst, mainly due to the fact that it predicts a symmetric citation curves (same growth and decay). While the Gompertz and the Bass models predict asymmetric citation pattern, they also predict an exponential (Bass) or double-exponential (Gompertz) decay of citations (Table 2.3), much faster than observed in real data. As a result, they both over-estimate the citation at small time scales (growth region), particularly for papers with low immediacy (red curve in Fig. 2.11). As Gompertz, Bass, and Logistic models predict citation tails with an exponential or faster decay, it also affects their predictive power. Refer to Section 2.5.3 for more details.

2.5 Applications

2.5.1 Parameter Distributions

The model indicates that the differences in the citation history of individual papers are encoded in the (λ, μ, σ) parameters, offering a new way to quantify a paper's impact and compare different papers through three separable factors. This is best illustrated by comparing the density functions $P(\lambda)$, $P(\mu)$ and $P(\sigma)$ for papers published in different journals in the same year (1990) (Fig. 2.12), indicating striking fitness differences among the journals.

For example $P(\lambda)$ for *PRB* is peaked at $\lambda \approx 0.5$ and is characterized by a relative paucity of high fitness publications. In contrast most *Cell* papers have high fitness, in the vicinity of 2 and 3.

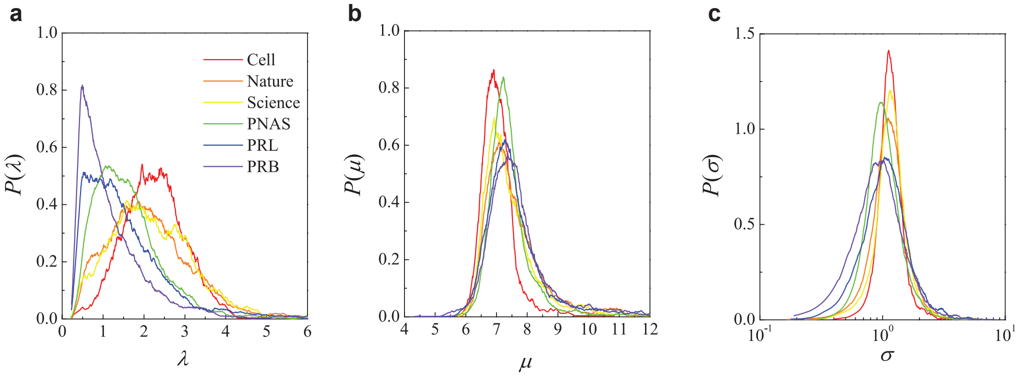


Figure 2.12: **Parameter distributions for papers published by six journals in 1990**. (a) Fitness distributions are radically different for different journals. (b) Immediacy distributions show modest differences: *Cell* has the smallest average μ among the 6 journals, while the average immediacy of *PRB* is the largest. (c) Longevity distributions of these journals are characterized by similar mean values.

We also find a modest temporal shift in fitness distributions (Fig. 2.13a). The observed $P(\mu)$ and $P(\sigma)$ distributions show a remarkable stability in time, across decades (Fig. 2.13bc). Hence in most cases the immediacy (μ) and the decay (σ) remain unchanged over decades for some journals (but in some periods they can undergo major changes, as we document for *Cell*). We also detect a weak linear correlation between λ_i and σ_i (Fig. 2.13d), indicating that papers with high fitness are more likely to have a slower decay, enhancing their long-term impact. Fitness λ and immediacy μ_i are independent for most but the high λ papers (Fig. 2.13e), suggesting that it

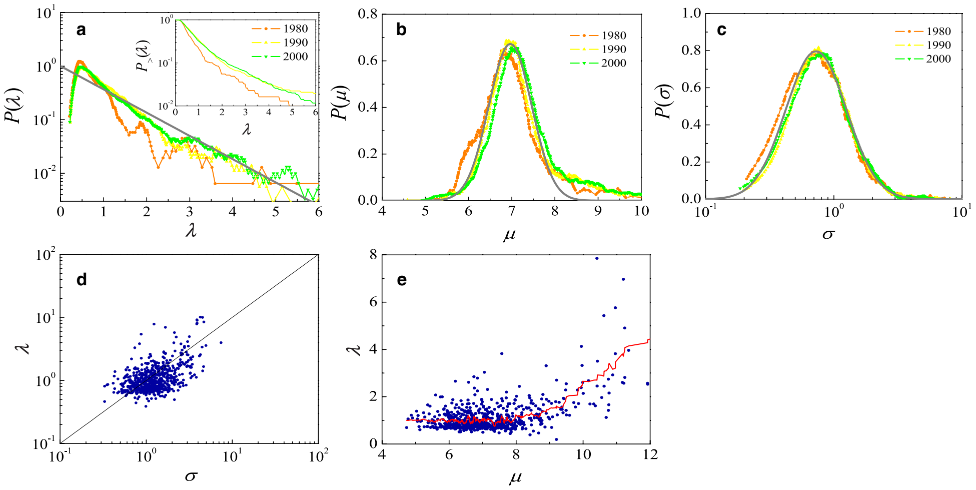


Figure 2.13: **Temporal evolution of the parameter distributions and correlations between them.** (a) Fitness distribution for papers published in different years (1980, 1990, and 2000) within the PR corpus. (b) Immediacy distributions for papers shown in (a). (c) Same, but for longevity distributions. (d) We observe a weak linear correlation between λ and σ, indicating higher fitness papers tend to have larger longevity. (e) λ and μ are largely uncorrelated, except in the large λ region, indicating that papers with very high fitness are also characterized by a delayed impact.

takes more than 20 years ($\mu > 9$) for truly influential papers (high λ) to reach their citation peak. This explains the lack of correlation between a paper's early (c^2) and long-term (c^{30}) citations for exceptionally high impact papers (Fig. 2.1C).

2.5.2 Ultimate Impact

The mechanistic nature of the model allows us to develop several fundamental measures of impact:

Ultimate impact (c^{∞}) represents the total number of citations a paper acquires during its lifetime. By taking the $t \to \infty$ limit in Eq. (2.13), we obtain

$$c_i^{\infty} = m\left(e^{\lambda_i} - 1\right), \tag{2.36}$$

a simple formula that predicts that the total number of citations acquired by a paper during its lifetime is independent of immediacy (μ) or the rate of decay (σ), and *depends only on a single parameter, the paper's relative fitness*, λ.

Impact time (T_i^*) represents the characteristic time it takes for a paper to collect the bulk of its citations. A natural measure is the time necessary for a paper to reach the geometric mean of its final citations. As m can be viewed as some sort of initial attractiveness, this is equivalent to solving for t in

$$\sqrt{m\left(m + c_i^{\infty}\right)} = m\left(e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right). \tag{2.37}$$

Considering $e^{\lambda_i \Phi(x)} \gg 1$, we can approximate the impact time T_i^* as

$$T_i^* \approx \exp(\mu_i). \tag{2.38}$$

Hence impact time is mainly determined by the immediacy parameter μ_i and is largely independent of fitness λ_i or decay σ_i.

Next we show, using the fundamental measures predicted here, the MiC model offers a journal free methodology to evaluate long term impact. To

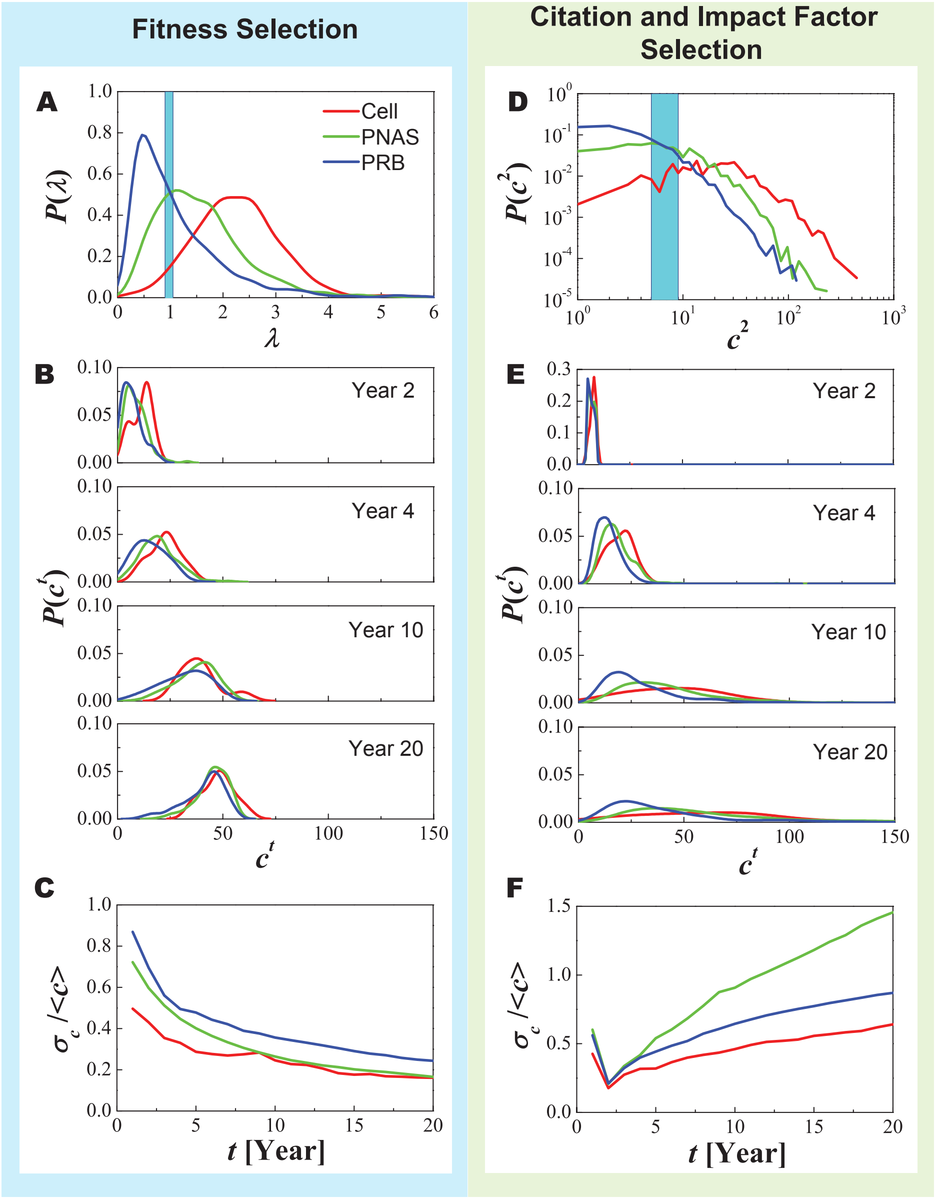


Figure 2.14: **Evaluating long-term Impact**. (A) Fitness distribution $P(\lambda)$ for papers published by *Cell*, *PNAS*, and *Physical Review B* (*PRB*) in 1990. Shaded area indicates papers in the $\lambda \approx 1$ range selected for further study. (B) Citation distributions for papers with fitness $\lambda \approx 1$ highlighted in (A) for years 2, 4, 10, and 20 after publication. (C) Time dependent relative variance of citations for papers selected in (A). (D) Citation distribution two years after publication ($P(c^2)$) for papers published by *Cell*, *PNAS*, and *PRB*. Shaded area highlights papers with $c^2 \in [5, 9]$ selected for further study. (E) Citation distributions for papers with $c^2 \in [5, 9]$ selected in (D) after 2, 4, 10, and 20 years. (F) Time dependent relative variance of citations for papers selected in (D).

illustrate this we selected three journals with widely different IFs: *Physical Review B* (*PRB*) (IF = 3.26 in 1992), *PNAS* (10.48) and *Cell* (33.62), and measured for each paper published by them the fitness λ, obtaining their distinct journal-specific $P(\lambda)$ fitness distribution (Fig. 2.14A). We then selected all papers with comparable fitness $\lambda \approx 1$, and followed their citation histories. As expected they follow different paths: *Cell* papers ran slightly ahead and *PRB* papers stay behind, resulting in distinct $P(c^T)$ distributions for years $T = 2 \div 4$. Yet, by year 20 the cumulative number of citations acquired by these papers show a remarkable convergence to each other (Fig. 2.14B), supporting our prediction that given their similar fitness λ, eventually they will have the same ultimate impact $c^\infty = 51.5$. This convergence is also supported by the decreasing $\sigma_c / \langle c \rangle$ ratio of $P(c^T)$ (Fig. 2.14C), indicating that the differences in citation counts between these papers vanish with time. In contrast, if we choose all papers with the same number of citations at year two (i.e. the same c^2, Fig. 2.14D), the citations acquired by them diverge with time and $\sigma_c / \langle c \rangle$ increases (Fig. 2.14E,F), supporting the lack of predictability in these quantities. Therefore λ and c^∞ offer a journal independent measure of a publication's long-term impact, in contrast with the lack of predictive power of c^2 and/or the IF.

2.5.3 Citation Predictions

While our primary goal is to uncover the mechanisms driving a paper's citation history, the accuracy of the MiC model raises a tantalizing question: can we use the developed framework to predict the future citations of a publication? In this section, we explore the possibility of using MiC as a prediction tool for future citations, and compare its predictive power with other models in Section 2.4.

In principle, we can use a paper's citation history up to year T_t after publication (training period T_t) to estimate the λ_i, μ_i, σ_i parameters associated with the paper and then use Eq. 2.13 to predict the paper's future citations. This process can be formalized as the following. Using the training period T_t and k_t sampling citations, we try to predict the number of citations at a future time T_p. Equation 2.13 predicts the expected increment of citations between the period $(T_t, T_p]$

$$\overline{\Delta k} = (k_t + m)\left(e^{\eta(\Phi((\ln T_p - \mu)/\sigma) - \Phi((\ln T_t - \mu)/\sigma))} - 1\right). \tag{2.39}$$

Hence, the expected citation at time T_p is

$$\overline{k}(\eta, \mu, \sigma) = (k_t + m)e^{\eta(\Phi((\ln T_p - \mu)/\sigma) - \Phi((\ln T_t - \mu)/\sigma))} - m \tag{2.40}$$

where, by assuming uniform prior distributions of (η, μ, σ), the probability of taking parameters (η, μ, σ) follows,

$$P(\eta, \mu, \sigma) \propto L = e^{\ln L(\eta, \mu, \sigma)}, \tag{2.41}$$

where the likelihood function L satisfies Eq. (2.19).

Therefore, given a citation history, we can use MiC to predict the probability for the paper to have k_p citations at the time T_p,

$$P(k_p) = \int \delta\left(\overline{k}(\eta,\mu,\sigma) - k_p\right) P(\eta,\mu,\sigma) d\eta d\mu d\sigma. \tag{2.42}$$

Here we neglect the fluctuation from the stochastic process itself, and consider only the uncertainties in parameter estimation.

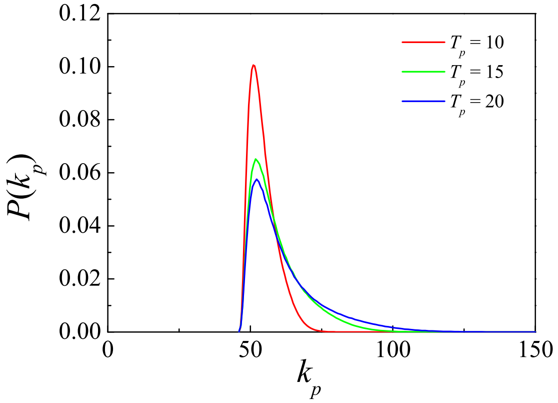


Figure 2.15: Illustrative example of $P(k_p)$ for a randomly selected paper. Different lines correspond to different testing period (T_p).

To give an intuition about (2.42), we show $P(k_p)$ for a randomly selected paper for different T_p (Fig. 2.15), illustrating the narrowly peaked nature of $P(k_p)$. Hence, the most probable future citation k_p^* can be obtained from

$$\left.\frac{dP(k_p)}{dk_p}\right|_{k_p=k_p^*} = 0, \tag{2.43}$$

and the upper/lower uncertainty can be obtained from the variance of $P(k_p)$:

$$\sigma_p^+ = \sqrt{\int_{k_p^*}^{\infty} (k_p - k_p^*)^2 P(k_p) dk_p} \tag{2.44a}$$

$$\sigma_p^- = \sqrt{\int_{k_t}^{k_p^*} (k_p - k_p^*)^2 P(k_p) dk_p} \tag{2.44b}$$

Taken together, based on an existing citation history and by combining Equations (2.43) and (2.44), we can use MiC to predict at each future time T_p, the most likely citations at the time (k_p^*) as well as the confidence range $[-\sigma_p^-, \sigma_p^+]$, represented as a citation envelope (Fig. 2.16). This is further illustrated in Fig. 2.17A, in which we show the predicted most likely citation path (red line) with the uncertainty envelope (grey area) for three papers, based on a 5 year training period. Two of the three papers fall within the envelope, for the third, however, the MiC model overestimates the future citations. Increasing the training period enhances the predictive accuracy (Fig. 2.17B).

To systematically evaluate the fraction of papers that fall within the predicted citation envelope, we measure $z_T = |c^T - k_p^*|/\sigma_p^+$, that quantifies how many standard deviations away the real citations deviate from predicted most likely citations. $z_T \leq 1$ indicates that the real citation dynamics fall within the citation envelope. If, however, $z_T > 2$, it indicates that the predicted citations exit the envelope far enough that the citations are not predicted correctly by the model. To this end, we compiled a test base of papers, consisting of all papers in the PR corpus published within the same decade

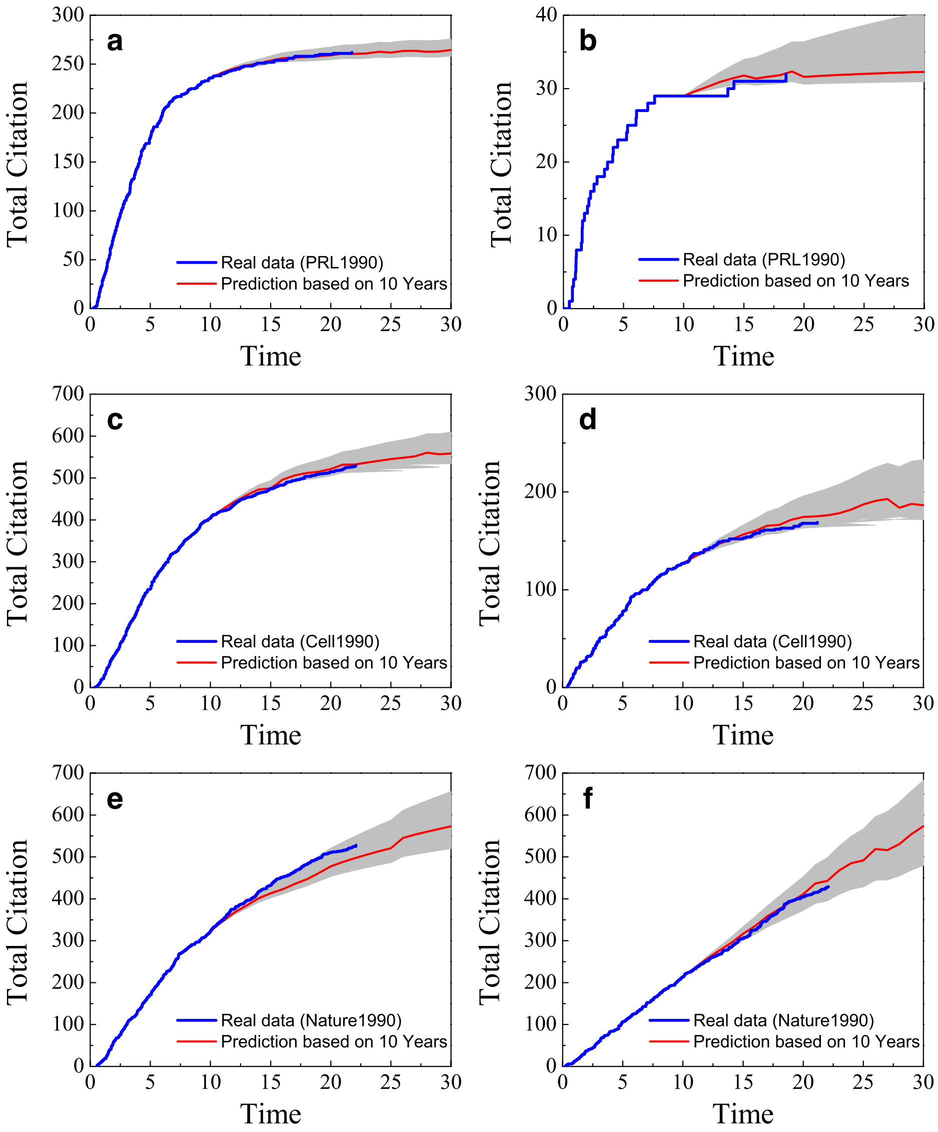


Figure 2.16: Citation predictions using MiC for 6 papers randomly selected form three different journals (*PRL*, *Cell*, and *Nature*).

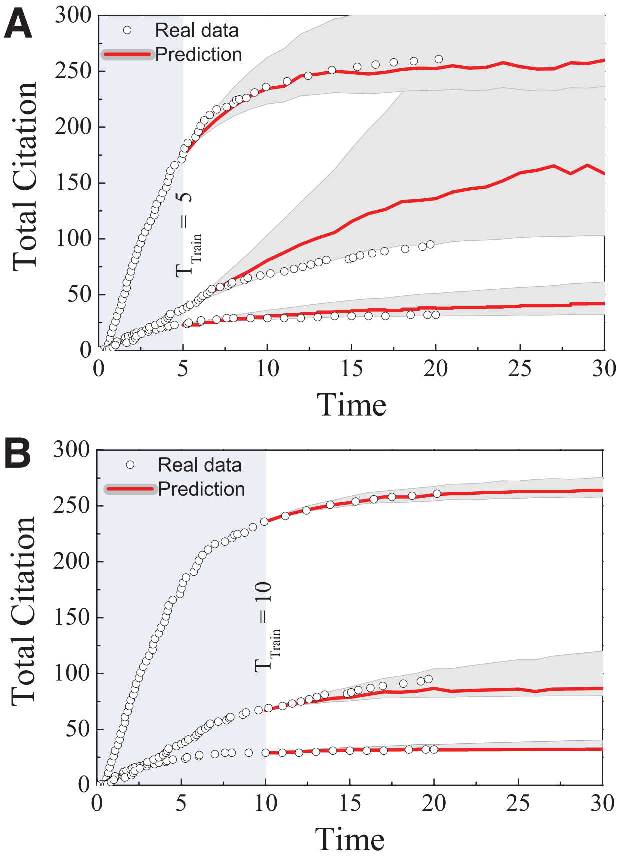


Figure 2.17: (A, B) Prediction envelope for three papers obtained using a five (A) and ten (B) years of training (shaded vertical area). The middle curve offers an example of a paper for which the prediction envelope misses the future evolution of the citations. The envelope illustrates the range for which $z \leq 1$. Comparing A and B illustrates how the increasing training period decreases the uncertainty of the prediction, resulting in a narrower envelope.

(1960s) that have at least 10 citations in 5 years (4492 papers). we measure the z_{30}-score for each paper, capturing the number of standard deviations z_{30} the real citations c^{30} deviate from the most likely citation 30 years after publication. The obtained $P(z_{30})$ distribution across all papers decays fast with z_{30} (Fig. 2.18), indicating that large z values are extremely rare. With $T_{Train} = 5$ only 6.5% of the papers leave the prediction envelope 30 years later, hence the model correctly approximates the citation range for 93.5% of papers 25 years into the future. $P(z_{30})$ distribution documents the predictive limitations of the current models. Indeed, for the Logistic, Bass and Gompertz model more than half of the papers underestimate with more than two standard deviations the true citations ($z > 2$) at year 30 (Fig. 2.18), in contrast with 6.5% for the MiC model.

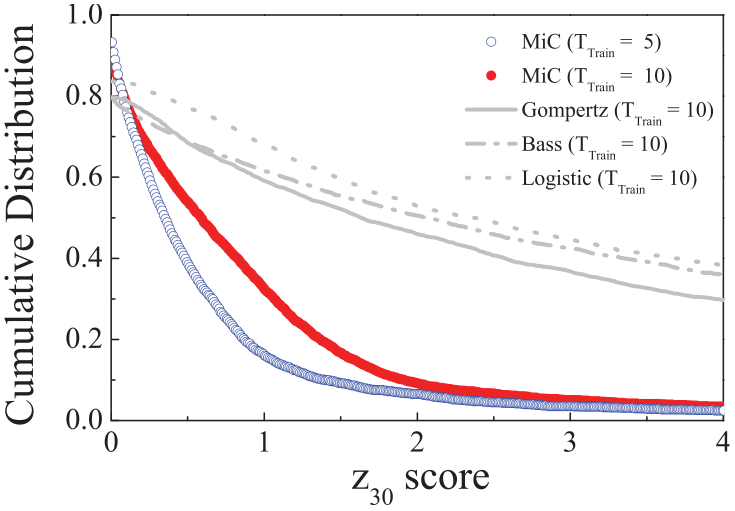


Figure 2.18: Complementary cumulative distribution of z_{30} ($P^{>}(z_{30})$), where z_{30} quantifies how many standard deviations the predicted citation history deviates from the real citation curve thirty years after publication (see also S2.6). We selected papers published in 1960s in PR corpus that acquire at least 10 citations in 5 years (4492 in total). The red curve captures predictions for 30 years after publication for $T_{Train} = 10$, indicating that for the MiC model 93.5% papers have $z_{30} \leq 2$. The blue curve relies on 5 year training. The grey curves capture the predictions of Gompertz (solid line), Bass (dash-dot line), and Logistic (dotted line) model for 30 years after publication by using 10 years as training.

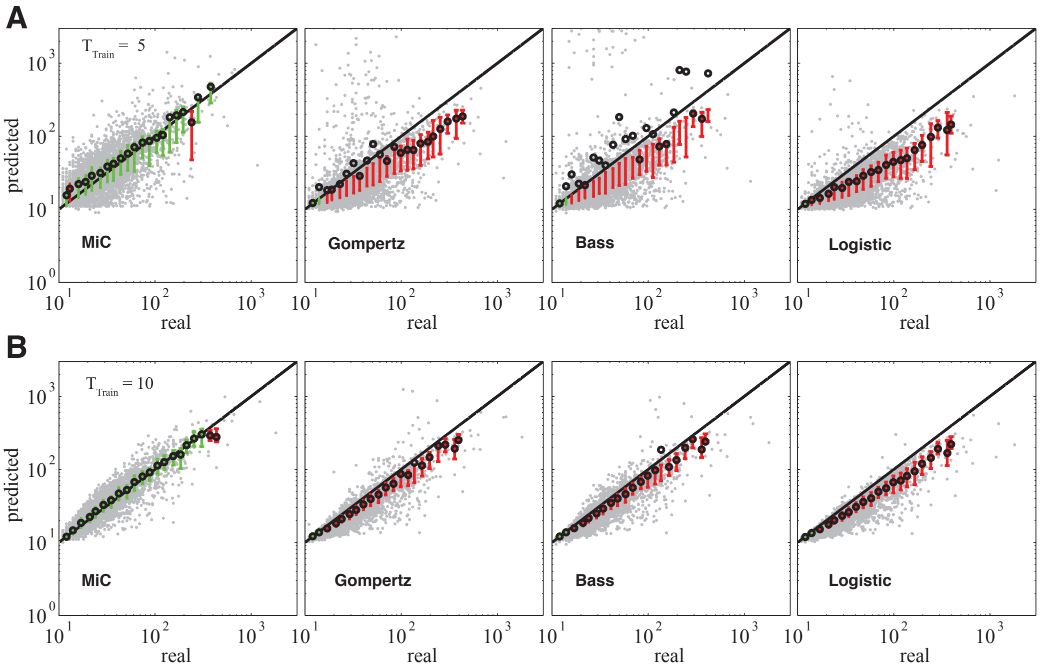


Figure 2.19: Scatter plots of predicted citations and real citations at year 30 for our test base, using as training data the citation history for the first 5 (A) or 10 (B) years. The error bars indicate prediction quartiles (25% and 75%) in each bin, and are colored green if $y = x$ lies between the two quartiles in that bin, and red otherwise. The black circles correspond to the average predicted citations in that bin.

That a large fraction of papers falling within the prediction envelop documents MiC's remarkable predictive accuracy, raising an important question: how well does the real citations match the predicted most likely citations? Indeed, while MiC correctly predicts the range of future citations for a large fraction of papers (93.5%), It is also important to know, particularly from a practical perspective, a single number for future citations instead of a citation range. An intuitive way to quantify this is through scatter plots. Hence we used a 5 and a 10 year training period to fit the parameters of each model and computed the predicted most likely citations at year 30 (Fig. 2.19). We find that independent of the training period the predictions of the Logistic, Bass and Gompertz models always lay outside the 25%–75% prediction quartiles (red bars), systematically underestimating future citations. In contrast, the prediction of Eq. (2.13) for both training periods is within the 25-75% quantiles, its accuracy visibly improving for the ten year training period (Fig. 2.19B).

2.5.4 Quantifying a Journal's Impact

The MiC model (2.13–2.14) also helps connect the impact factor (IF), the traditional measure of impact of a scientific journal, to the journal's Λ, M, and Σ parameters (the analogs of λ, μ, σ).

To quantify the impact of a journal, we count the average citations all

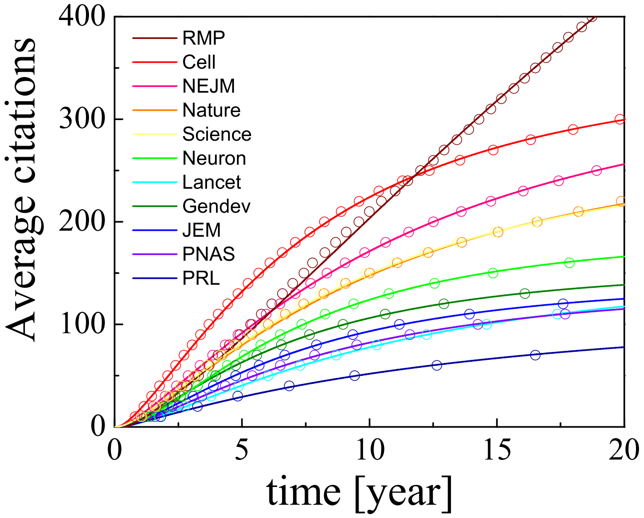


Figure 2.20: **Average cumulative citations for different journals**. The average number of citations each journal gets to all of its papers published in 1990. Circles correspond to empirically measured citations, and solid lines are based on Eq. (2.46)

papers published by the journal acquire over time,

$$C_j^t = \frac{1}{N_j} \sum_{i}^{N_j} c_i^t, \tag{2.45}$$

where N_j is the total number of papers published by journal j. Here we only consider the research papers published by each journal. This can be achieved by looking at *document type* for each paper indexed by Web of Science. Most specifically, we only consider the document types as *Review* and *Article*. We find that for some journals, like for *Physical Review Letters* (*PRL*), the vast majority of papers are in these categories. Yet for other journals, especially

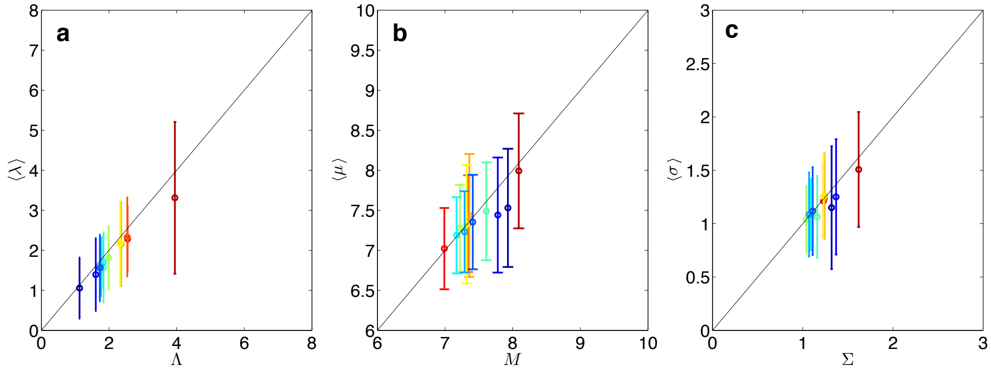


Figure 2.21: **Comparing journal parameters**. The three parameters characterizing a journal's citation history can be computed in two ways. One is to average over the parameters of individual papers ($\langle\lambda\rangle$, $\langle\mu\rangle$, $\langle\sigma\rangle$). The other is to use the average citation curve (Fig. 2.20) for a journal (Λ, M, Σ). Here we show an a reasonable agreement between the values offered by these two methods in (a) fitness (b) immediacy (c) longevity.

the ones for general audience, many are classified as "letter to editor" or "editorial". Therefore, in analogy to the "citable items" used in measuring a journal's impact factor (IF) by Journal Citation Reports, this distinction is important to understand a journal's impact.

By viewing each journal as a super paper, we find that C_j^t is also well approximated by our model (Fig. 2.20), indicating that

$$C_j^t = m\left(e^{\Lambda_j \Phi\left(\frac{\ln T - M_j}{\Sigma_j}\right)} - 1\right). \tag{2.46}$$

Therefore each journal's citations are captured by three parameters (Λ, M, Σ), in analogy with the (λ, μ, σ) parameters derived for individual papers. To check whether (Λ, M, Σ) represent the average of individual papers published by the journal, we computed for each journal shown in Fig. 2.20 the (λ, μ, σ) parameters for individual papers published in the journal, and the mean of each parameter with its journal average (Fig. 2.21). We find that $\langle\lambda\rangle$, $\langle\mu\rangle$, $\langle\sigma\rangle$ are in good agreement with Λ, M, and Σ, indicating that (Λ, M, Σ) parameters are representative for an average paper published by the journal.

Next we derive IF by using the three parameters (Λ, M, Σ) for each journal. The IF of a journal is defined as the average number of citations received per paper by that journal during the two preceding years. Let us consider for example calculating a journal's IF in 1992. In the numerator we need to measure the number of times papers published by this journal in 1990 and 1991 are cited during 1992. This includes all papers published by this journal. In the denominator, we need to normalize by the number of papers. But according to the definition from Journal Citation Reports, this normalization is for the number of "citable items" published by that journal in 1990 and 1991. In principle, the exact expression of the IF can be obtained by integrating over all papers published within the two-year time frame using their corresponding parameters. Assuming the publication date of a paper within a year does not affect its citations, we can treat all papers within a year as published on the same date. Imagine we want to calculate a journal's IF in the year T, and this journal published N_1 papers in the year $T_1 = T - 2$ and N_2 papers in $T_2 = T - 1$. Therefore, based on the definition of IF, we have

$$\mathrm{IF}(T) = \frac{\sum\limits_{i}^{N_1} c_i(T|T_1) + \sum\limits_{i}^{N_2} c_i(T|T_2)}{N_1 + N_2}, \tag{2.47}$$

where $c_i(T|T_1)$ and $c_i(T|T_2)$ are the citations in year T for paper i published in year T_1 and T_2, respectively. In Fig. 2.22ab, we compared the IFs measured by Eq. (2.47) to the reported value for *Cell* and *NEJM* (Fig. 3) between 1998-2005, finding a good agreement except for small deviations for *NEJM* in 1999 and 2000, which are likely caused by the small differences between our downloaded data and the data used by Journal Citation Report.

The proposed model allows us to calculate the journal's impact factor analytically. To do it, we substitute Eqs. (2.45) and (2.46) into (2.47), obtaining

$$\begin{aligned} \mathrm{IF}(T) &= \frac{N_1 C(T|T_1) + N_2 C(T|T_2)}{N_1 + N_2} \\ &= \frac{mN_1}{N_1+N_2}\left(e^{\Lambda(T_1)\Phi\left(\frac{M_1 - M(T_1)}{\Sigma(T_1)}\right)} - e^{\Lambda(T_1)\Phi\left(\frac{M_3 - M(T_1)}{\Sigma(T_1)}\right)}\right) \\ &+ \frac{mN_2}{N_1+N_2}\left(e^{\Lambda(T_2)\Phi\left(\frac{M_3 - M(T_2)}{\Sigma(T_2)}\right)} - e^{\Lambda(T_2)\Phi\left(\frac{M_2 - M(T_2)}{\Sigma(T_2)}\right)}\right), \end{aligned} \tag{2.48}$$

where $(\Lambda(T_1), M(T_1), \Sigma(T_1))$ and $(\Lambda(T_2), M(T_2), \Sigma(T_2))$ are the journal parameters measured at the year T_1 and T_2, respectively, and $M_1 = \ln(3\, years) = \ln(3\times 365) \approx 7.00$, $M_2 = \ln(1\, year) = \ln(365) \approx 5.90$ and $M_3 = \ln(2\, years) = \ln(2 \times 365) \approx 6.59$. Figure 2.22c documents an excellent match between Eq. (2.48) and the empirical measurement based on (2.47).

To further simplify (2.48), we assume that the changes in papers published

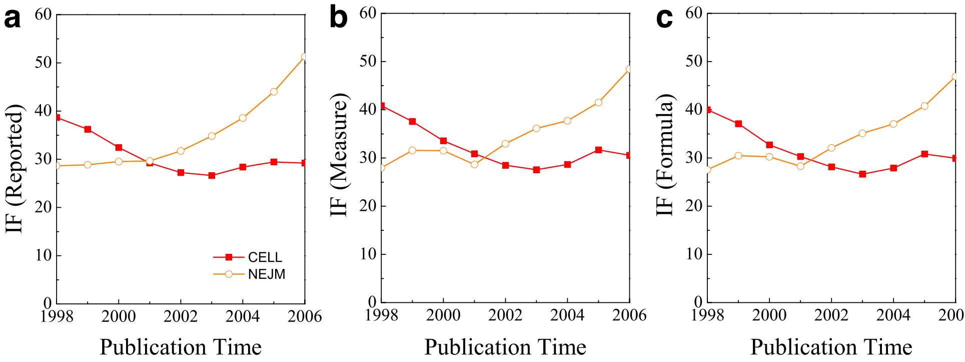


Figure 2.22: **Comparing** IF **reported by ISI with the (2.48) approximation**. (a) IF reported by ISI for *Cell* and *NEJM* from 1998 to 2006. (b) IF measured for these two journals within the time span following the definition of (2.47). (c) IF computed by plugging in the corresponding parameters in (2.48).

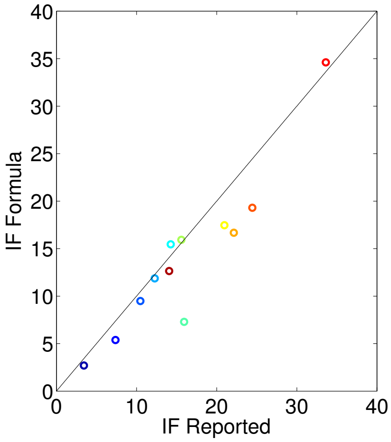


Figure 2.23: Scatter plot for journals shown in Fig. 2.20, their reported IF in 1992 and their approximated IF obtained using their parameters in 1990 in (2.49).

by a journal are small over the course of two years, in terms of both number of papers published and their citations. Under this assumption, $N_1 = N_2$

and $(\Lambda, M, \Sigma) \equiv (\Lambda(T_1), M(T_1), \Sigma(T_1)) = (\Lambda(T_2), M(T_2), \Sigma(T_2))$, Eq. (2.48) leads to

$$\mathrm{IF} \approx \frac{m}{2}\left(\exp\left[\Lambda\Phi\left(\frac{M_1 - M}{\Sigma}\right)\right] - \exp\left[\Lambda\Phi\left(\frac{M_2 - M}{\Sigma}\right)\right]\right). \tag{2.49}$$

This approximation is not able to account for the temporal evolution in IF, but allows us to compute a journal's IF using only one year of data. To see how well (2.49) approximates the reported IF, we use the citation data for the journals published in 1990 and approximate their IF in 1992. We then compare the computed IF by using (2.49) with the ones reported by ISI (Fig. 2.23). We find that despite its simplicity, the two quantities largely agree with each other for different journals, indicating that (2.49) serves as a good approximation for a journal's impact factor.

Equation (2.49) helps us understand the mechanisms that influence changes in the IF, as vividly illustrated by the evolution of *Cell* and *NEJM*: in 1998 the IFs of *Cell* and *NEJM* were 38.7 and 28.7, respectively (Fig. 2.24A). Yet over the next decade there was a remarkable reversal: *NEJM* became the first journal to reach IF = 50, while *Cell*'s IF decreased to around 30. This raises a puzzling question: has the impact of papers published by the two journals changed so dramatically? To answer this we determined Λ, M, and Σ for both journals from 1996 to 2006 (Fig. 2.24D–F). While Σ were indistinguishable (Fig. 2.24D), we find that the fitness of *NEJM* increased from $\Lambda = 2.4$ (1996) to $\Lambda = 3.33$ (2005), increasing the journal's

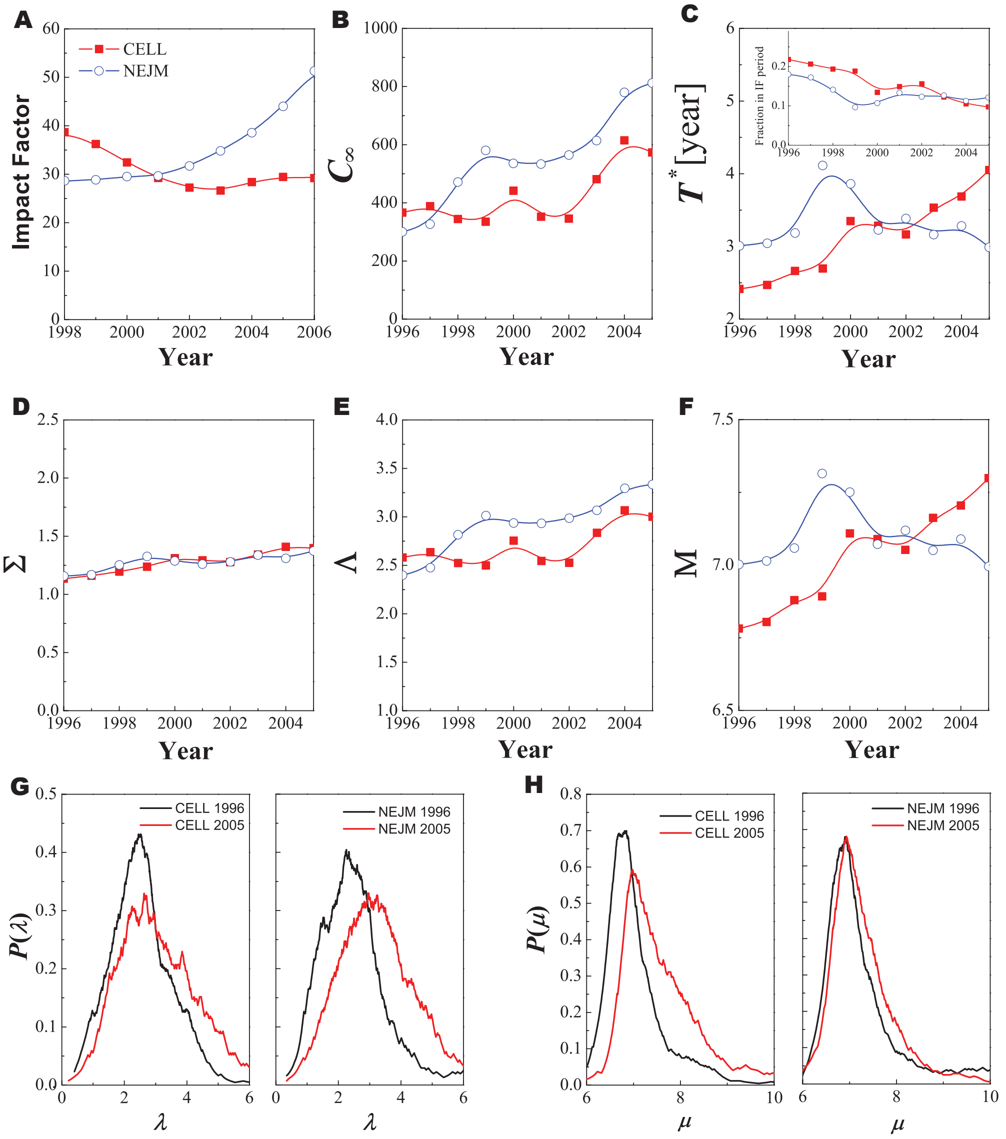


Figure 2.24: **Quantifying changes in a journal's long-term impact**. (A) Impact factor of *Cell* and *New England Journal of Medicine* (*NEJM*) reported by Thomson Reuters from 1998 to 2006. (B) Ultimate impact C^∞ (see Eq. (2.37)) of papers published by the two journals from 1996 to 2005. (C) Impact time T^* (Eq. (2.38)) of papers published by the two journals from 1996 to 2005. Inset: fraction of citations that contribute to the IF. (D–F) The measured time dependent longevity (Σ), fitness (Λ), and immediacy (M) for the two journals. (G) Fitness distribution for individual papers published by *Cell* (left) and *NEJM* (right) in 1996 (black) and 2005 (red). (H) Immediacy distributions for individual papers published by *Cell* (left) and *NEJM* (right) in 1996 (black) and 2005 (red).

ultimate impact from $C^\infty = 300$ (1996) to a remarkable $C^\infty = 812$ (2005) (Fig. 2.24B). But *Cell*'s Λ also increased in this period (Fig. 2.24E), moving its ultimate impact from $C^\infty = 366$ (1996) to 573 (2005). Yet, if both journals attracted papers with increasing long-term impact, why did *Cell*'s IF drop and *NEJM*'s grow? The answer lies in changes in the impact time $T^* = \exp(M)$: while *NEJM*'s impact time remained unchanged at $T^* \approx 3$ years, *Cell*'s T^* increased from $T^* = 2.4$ years to $T^* = 4$ years (Fig. 2.24C). Therefore, *Cell* papers have gravitated from short to long-term impact: a typical *Cell* paper gets 50% more citations than a decade ago, but fewer of the citations come within the first two years (Fig. 2.24C, inset). In contrast, with a largely unchanged T^*, *NEJM*'s increase in Λ translated into a higher IF. These conclusions are fully supported by the $P(\lambda)$ and $P(\mu)$ distributions for individual papers published by *Cell* and *NEJM* in 1996 and 2005: both journals show a clear shift to higher fitness papers (Fig. 2.24G), but while $P(\mu)$ is largely unchanged for *NEJM*, there is a clear shift to higher μ papers in *Cell* (Fig. 2.24H).

2.6 Discussions and Conclusions

The remarkable accuracy of the MiC model we documented in this chapter, both in its ability to capture the universal aspects of citation histories, as well as to predict future citations, supports our hypothesis that scientific impact is a collective phenomenon, governed by mechanisms that follow reproducible patterns [107, 20, 97]. Therefore I would anticipate that the proposed modeling framework is not limited to citations, but with appropriate adjustments will likely apply to other phenomena driven by collective processes, from patents to the popularity of twitter hash tags. Another much anticipated direction is a shift of focus from understanding the dynamics and properties of a system as a whole to the characterization of individual items within the system. The MiC model illustrates an example where a remarkable amount of regularity within individual papers is hidden behind the apparent noise and unpredictability in their citation dynamics.

At the same time, the model also has obvious limitations: it cannot account for exogenous "second acts", like the citation bump observed for superconductivity papers following the discovery of high temperature superconductivity in the 1980s, or delayed impact, like the explosion of citations to Erdős and Rényi's work four decades after their publication, following the emergence of network science [90, 24, 31, 17].

Taken together, the mechanistic understanding of citation dynamics offers a quantitative springboard to uncover the hallmarks of future impact. These questions also have major policy implications, as current measures of citation-based impact, from IF to Hirsch index [54, 1], are frequently integrated in reward procedures, the assignment of research grants, awards and even salaries and bonuses [78, 41], despite their well-known lack of predictive power. In contrast with the IF and short-term citations that lack predictive power, we find that c^{∞} offers a journal independent assessment of a paper's long term impact, with a meaningful interpretation: it captures the total number of citations a paper will ever acquire, or the discovery's ultimate impact. While additional variables combined with data mining could further enhance the demonstrated predictive power, an ultimate understanding of long-term impact will benefit from a mechanistic understanding of the factors that govern the research community's response to a discovery.

Chapter 3

Factors that Affect Information Spreading Processes

Information spreading plays an essential role in numerous human interactions, including the spread of innovations [104, 100], knowledge and information security management [49], social influence in marketing [29, 61, 70], and more. Thanks to the increasing availability of large-scale data, we have witnessed great advances in understanding how information propagates from person to person, ranging from incentivized word-of-mouth effects when recommending products [70, 56], to understanding how a single piece of information forms internet chain letters on a global scale [74].

Despite recent studies in online social networks [70, 4, 50, 72], it has been difficult to obtain detailed traces of information dissemination alongside relevant contextual data such as people's real social connections, their behavioral profiles, and job roles in organizations. Therefore, an important question is largely unanswered: *to what extent do spreading processes depend on the underlying social network and behavioral profiles of individuals*. Indeed, on one hand, information such as rumors, innovations and opinions diffuses through the underlying social networks. To whom and to how many people a user would pass such information is constrained by whom s/he connects to and how well s/he is connected in the social network, and the strength of those connections. On the other hand, the population-based heterogeneity in personal profiles coexists with complex connectivities between individuals, raising questions about to what degree the diverse profiles of individuals, from personal interests and expertise to communities and hierarchy, impact the information spreading process. Understanding the role of these features is of fundamental importance.

The lack of contextual information could change drastically, however, thanks to the pervasive use of email communications in well-documented settings, such as corporate work forces [36, 95, 37, 115, 8, 3, 62, 59]. Indeed, emails have become the most important communication method in various settings [112, 95], unveiling detailed traces of social interactions among large populations. Previous studies [112, 11] have shown that email communications serve as a good indicator of social ties. *Forwarded emails* [96], written by someone other than the sender and sent to someone who was not included in the original email, serve as an ideal proxy for the information spreading process, where the single piece of information, the original body of the email, is passed through the social network.

We compiled a new dataset by integrating two related but distinct data structures, collected from a large-scale, privacy-preserving distributed social sensor system. First, we collected two years of email communication data from $8,952$ volunteer employees within a large technology firm operating in more than 70 countries. Emails occupy the majority of information workers' time and thus provide high-quality observation of the social context, i.e., the real social connections of employees in the workplace [112, 11]. In addition to such "informal networks," we investigated the "formal networks," imposed by the corporation such as their hierarchical structure, as well as demographic data such as geography, job role, self-specified interests, performance, etc. This dataset provides us unique opportunities to study the interplay between the information spreading and its context. This issue is largely not addressed in previous studies partially due to the lack of such a multi-faceted dataset and the difficulty in matching user IDs across multiple sources.

Specifically, we investigate the impact of context on spreading processes in two levels:

- At the *microscopic* level, we are interested in the behaviors of each individual in the spreading process, e.g. to whom and how fast does a user forward information? (Sec. 3.3)

- At the *macroscopic* level, we ask what are the structural properties of the spreading processes? And what is the best model for the observed structures? (Sec. 3.4)

At the microscopic level, we find that information spreading is indeed highly dependent on social context as well as the individuals' behavioral profiles. Macroscopically, however, we find that the tree structures observed in the spreading process can be accurately captured by a simple stochastic branching model, indicating the macroscopic structures of spreading processes, i.e., to how many people a user forwards the information and the overall coverage of the information, are largely independent of context and follow a simple reproducible pattern. To the best of our knowledge, this work presents the first comprehensive analysis of the determining factors affecting information spreading processes. We believe our findings are of fundamental importance in developing prediction models for information flow, provide new insights towards the design of our social and collaborative applications, such as assisting users to disseminate information more efficiently, protecting digital information leakage, and promoting spreading strategies to achieve expected coverage.

3.1 Related Work

In this section, we review three categories of related work: studies on information spreading and cascades, social network analysis especially on emails, and virus propagation.

Information Spreading and Cascades

Various studies in online domains have been conducted to understand the structural properties of information flow. Among them, the spreading processes of specific pieces of information, including studies on internet chain letters and viral marketing, are most related to our work. Liben-Nowell and Kleinberg [74] studied information flows on a global scale using internet chain letters. They found that the structures of observed trees are narrow and deep. They proposed a probabilistic model, leveraging the structure of other social networks, to explain the deep tree-like structure. Golub and Jackson [45] then showed that the structures observed in [74] could be explained by the Galton-Watson branching model [111] combined with the selection bias of observing only the largest trees. Leskovec, Adamic and Huberman [70] studied an incentivized word-of-mouth effect by analyzing viral marketing data, focusing on the overall properties of the resulting recommendation network and its dynamics. By using data from a viral experiment of recommending newsletters, Iribarren and Moro [56] modeled the overall dynamics of information flow from individual activity patterns. There has been extensive work done in the blog domains about cascading behaviors [65, 4, 50, 72], and several models have been proposed to capture the structure of the blogosphere.

Previous work focuses on analyzing the observed properties of information flows. In contrast, the questions we are interested in this study are *Why does information spread? What are the factors that could potentially affect this process?*

Emails

Much work has focused on email communication records, from their static topological structure [36, 3] to dynamic properties [37, 8, 106, 62, 56]. These works focus on the overall structure of the social network, or on the timing of events. Recently, Karagiannis and Vojnovic [59] studied the behavioral patterns of email usage in a large-scale enterprise by looking at email replies. They examined various factors that could potentially affect email replies, focusing on pair-wise interactions and aiming to inform the design of advanced features. Our approach presents a new angle to using email data. First, we treat social networks as the backbone of the spreading processes, using the network topology to inspect the structures of information spreading and to assess models. Second, the spreading processes we study go beyond the pair-wise interactions of email replies, representing richer structural properties.

Virus Propagation

There is much literature regarding virus propagation. To name a few, Hethcote [53] studied the epidemic threshold for cliques. Briesemeister et al [15] studied the virus propagation on power law graphs by simulations. Most recent research has been devoted to real, arbitrary graphs. For example, Wang et al. [110] gave the analytic epidemic threshold for an arbitrary graph. Based on that, Tong et al [102] proposed an effective immunization strategy by approximately maximizing that threshold.

Virus propagation, although bearing some high level similarity to spreading of information, is not selective, meaning that a better connected individual in the network will infect more people. Information, on the other hand, spreads purposefully, representing a more complex behavior. Indeed, in our study we investigate the factors that affect to whom, and to how many people, a user forwards information, exploring the selective process of information spreading.

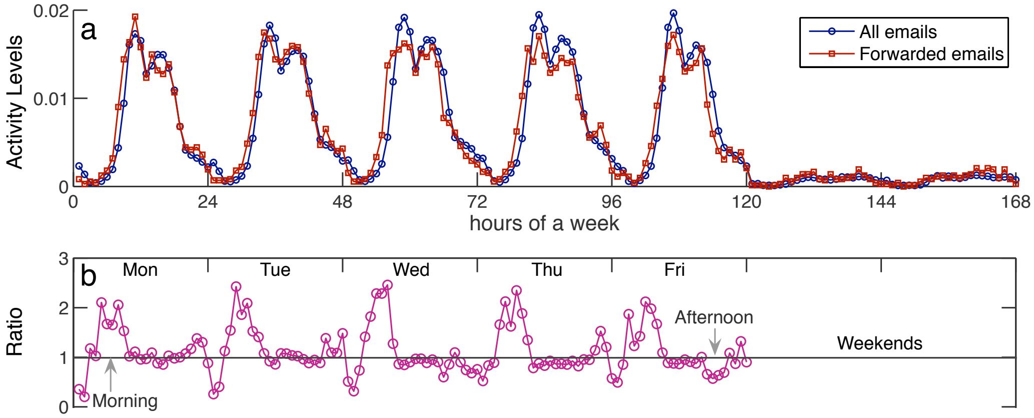


Figure 3.1: Activity levels at different times of the week for email forwarding activities and overall email traffic in (a), and ratios of the two in (b). The ratios in weekends are omitted due to the low volumes.

3.2 Preliminaries

In this section, we describe the datasets used in our study and present the basic properties of the dataset.

3.2.1 Data Description

We collected detailed electronic communication records of $8,952$ volunteer employees in more than 70 countries over a two-year period within a large global technology company with over $400,000$ employees. Each log entry specifies the sender and receiver(s) of the message, a timestamp, the subject, and the content of the body of the email. To preserve privacy[1], the email addresses of users are hashed, and the original textual content in email's body was not saved. Instead, this content is represented as a term-frequency vector containing the terms that appear in the text as well as their counts after stemming and removal of stop-words. During the two year period, we observe about 20 million emails sent by our users. For the same population, we gathered information specifying a range of personal attributes (gender, job role, departmental affiliation and report-to relation with managers, and more). We also collected detailed financial performance data for more than 10,000 consultants in the company. These consultants generate revenue by logging "billable hours". It has been found that a consultant's ability to generate revenue is an appropriate productivity measure [116]. Therefore, we measure performance of individuals with the total US dollars a consultant generated from June 2007 to July 2008. Combining the financial and communication data yields a total of 1,029 consultants for whom we have both email and financial records. The identities of participants were hashed.

As we are interested in how a specific piece of information spreads, we

[1] refer to [113] for more information about privacy related solutions

further processed our dataset using the following procedures. We started by looking for the string *Fw:* in the beginning of each email subject title. This process gives us all the emails that were forwarded. We then grouped emails with the same subject title together, reconstructing the original threads. Each forwarded thread results in an information spreading tree structure, where the single piece of information, the original body of the email, was passed from one to others. Our dataset provides us with $9,623$ such distinct threads, the starting point of our study.

3.2.2 Basic Properties

As our dataset captures communication within an enterprise, the temporal patterns and the organizational roles of individuals involved may indicate the importance of forwarded emails. We show in Fig. 3.1a the activity levels, the number of communications in each hour of the week normalized by the total number of communications in a week. We observe a clearly periodic pattern. Communication builds up in the morning and decays in the afternoon with a notable dip at noon indicating the lunch time. There are two interesting points we want to make here. First, while the activity levels of forwarded emails (red squares) follow a similar periodic pattern to the overall email traffic (blue circles), their activity levels are significantly higher than the normal email traffic on workday mornings, especially on Mondays, and lower in the afternoons especially on Fridays. This can be seen clearly in Fig. 3.1b, where we show the ratio of the two curves in Fig. 3.1a. The curve goes above 1 in the mornings, but mostly below 1 in the afternoons. This is a good indicator that forwarded emails are timely and important, representing a special class of overall email traffic. Second, access to email is limited by weekly schedules. This weekly cycle becomes important when we inspect the efficiency of information spreading in the following sections. That is, there is a time delay when forwarding an email after receiving it. For example, a delay of two days in the delivery of information, when it was received on Friday, could be due to the inability of a user to access his or her email during the weekend. Therefore, for any calculation regarding time in the following sections, we perform a check by removing the off-hours. Yet no results changed qualitatively.

In addition, we observe that 38% of the forwarded threads involve people from multiple departments. This suggests that email forwarding is an important means to facilitate cross-organization collaboration. Moreover, 43% of the emails are forwarded by managers, indicating that email forwarding is a common management tool.

3.3 Microscopic Information Spreading in Context

What factors could potentially affect the information spreading process at the microscopic level, i.e., to whom and how fast a user spreads the information? Why does some information get rapidly processed and passed on to others, while other information experiences notable delay? Or more generally, why does some information get forwarded at all in the first place?

Here we investigate several aspects of these questions. Our analysis will focus on the most fundamental building blocks of information spreading – information pathways, as illustrated in Fig. 3.2. More specifically, user A sent an email to user B at a certain time with a specific title. Then user B waited for some time and forwarded the message to user C, passing the information, the main body of the email, along via B from A to C. We refer to user A as "initiator", B as "spreader", and C as "receiver". The dissemination process can be far more complex than this simple case, as we shall see in Section 3.4, where we focus on spreading processes at the macroscopic level, exploring to what extent the overall structure of spreading processes depends on context. Yet any spreading tree structures can be reduced to a combination of such information pathways.

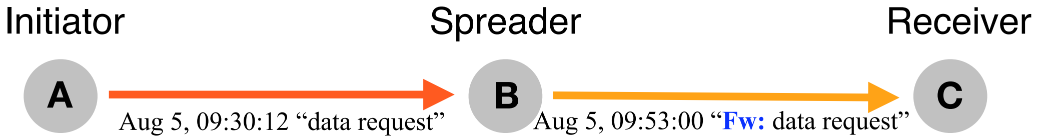

Initiator
Spreader
Receiver
A
B
C
Aug 5, 09:30:12 "data request"
Aug 5, 09:53:00 "Fw: data request"

Figure 3.2: An illustrative example of an information pathway.

Our study shows that not only does the social and organizational context affect to whom a specific piece of information is forwarded, but it also affects how fast it is forwarded. We found that information undergoes interesting re-routing processes, from weak links to strong ties, and from non-experts to experts. The efficiency of the spreading process is affected by departmental structure, but little by individual performance. These findings can guide us to build better social and collaborative environments, design applications assisting users to disseminate information more efficiently, and develop strategies to protect digital information leakage and predictive tools for recommendation systems.

3.3.1 The Underlying Social Networks

How information spreads may be influenced by the underlying social network, and understanding the interplay between the social network and spreading process is very important. First, it has a number of implications in various social systems, such as promoting new strategies in viral marketing by taking into account the effect of the network topology. Second, it plays an important role in assessing the choice of models, arguing whether a flu-like epidemiology model, which directly relies on the topology of the network, is suitable for modeling the information flow, (see Section 3.4).

We start by building a social network among our users by aggregating email communications over a one year period. We add a link between two users if there has been at least one email communication between them. The weight of the link, $w(i \to j)$, is asymmetric, defined as the number of emails sent from user i to j. As we are mostly interested in the connectivity between individuals, we focus on the static picture of the network rather than the dynamics of the network evolution.

We show in Fig. 3.3 the probability ratio of email forwarding activity[2] as a function of the weight of the links between initiators and spreaders, $w(A \to B)$, and spreaders and receivers, $w(B \to C)$, defined in the information pathways in Fig. 3.2. A positive slope would indicate that information is more likely to flow through strong ties, whereas a negative slope shows that weaker connections are more favorable for the spreading processes. Surprisingly, we observe that the information is more likely to spread initially via weak ties and then gets passed through strong connections, strong evidence of information routing by spreaders choosing social neighbors of different closeness.

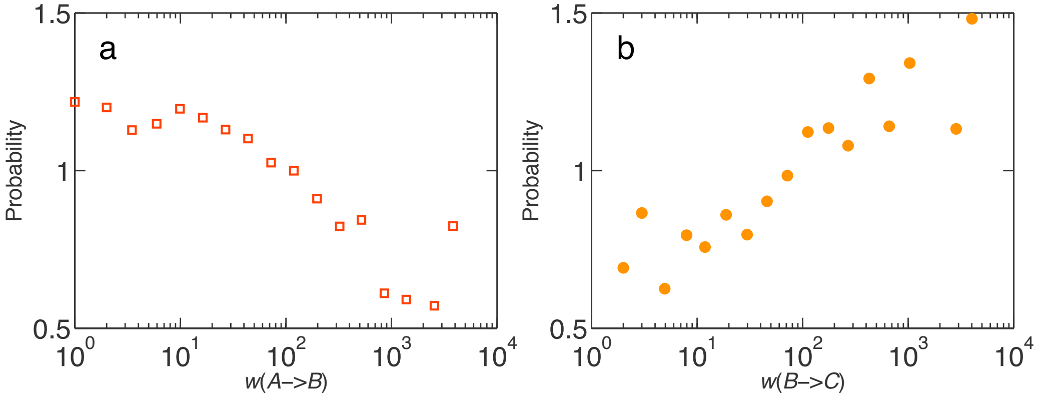


Figure 3.3: The probability ratio of email forwarding as a function of the weight of links between A and B and B and C respectively in the information pathways. Information spreading undergoes an interesting re-routing process, from weak links to strong ties.

[2] the probability ratio of email forwarding as a function of quantity q is obtained by $P^{\mathrm{Fw}}(q)/P^{\mathrm{rand}}(q)$, where $P^{\mathrm{Fw}}(q)$ is the probability of having q in forwarded emails, while $P^{\mathrm{rand}}(q)$ is the same probability for overall emails. A value equals to 1 would indicate $P^{\mathrm{Fw}}(q)$ is about what you would expect normally.

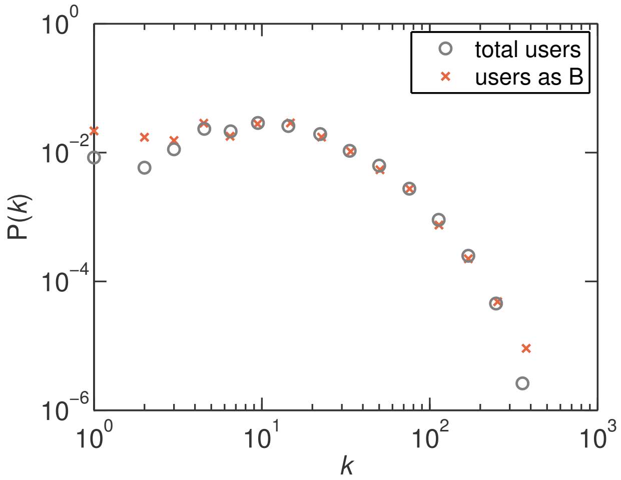


Figure 3.4: The degree distribution of the whole network and the group of the spreaders. The spreaders have comparable connectivity to randomly sampled individuals in the network.

This raises an interesting question: how well are the information spreaders connected in the network? Are they a random sample of individuals or are they a biased sample of more central social hubs? We show in Fig. 3.4 the degree distribution $P(k)$ of nodes in the whole social network as grey circles and the sample of spreaders as orange crosses. Interestingly, we find that spreaders show nearly the same distribution of connectivity as a random sample of individuals from the network.

3.3.2 Information Content and Expertise

An important question about information spreading is how the process depends on the relevance of the content of the information to the individual's expertise. Here, we explore this issue using the available message content. As mentioned previously, the content of each email in our dataset is represented as term frequencies. We build a vocabulary vector $\vec{v_i} = \langle s_1, s_2, ..., s_n \rangle$ for each user i by looking at the content of all the emails sent by i, where the length of $\vec{v_i}$ is the total number of meaningful words that have appeared in all emails for all users, thus is the same length for every user. The j-th element s_j is the score of the j-th word calculated by TF-IDF. The vector $\vec{v_i}$ will provide a measure of ranked "buzzwords" for user i, which serves as an indicator of the individual's expertise, since previous studies have shown emails are a primary form of communication within big corporations [112, 11]. Next, we build a vector $\vec{v_l}$ for each email l following the same procedure, where s_j is the TF-IDF score of the j-th word in $\vec{v_l}$. Therefore, $\vec{v_l}$ will give us a measure of the content of email l, accounting for overly common words and overly rare words. Then the similarity between the content of the information l and the individual i's expertise is defined as the cosine similarity of the two vectors, $\mathcal{S}_{i,l} = \vec{v_i} \cdot \vec{v_l} / (\|\vec{v_i}\| \, \|\vec{v_l}\|)$. We show in Fig. 3.5, how the probability ratio of information spreading changes in function of $\mathcal{S}_{i,l}$ for user i as (a) spreaders and (b) receivers, respectively. The probability ratio anti-correlates with $\mathcal{S}_{i,l}$, similarity between information content and spreaders' expertise, yet exhibits a significant positive correlation for the receivers' case. This finding offers quantitative evidence that the information undergoes a clear re-routing, demonstrating that information flows from non-experts to experts. That is, the information is more likely to be passed on by spreaders if the content is dissimilar to spreaders' expertise. It then flows to receivers who are more likely to be interested in the information.

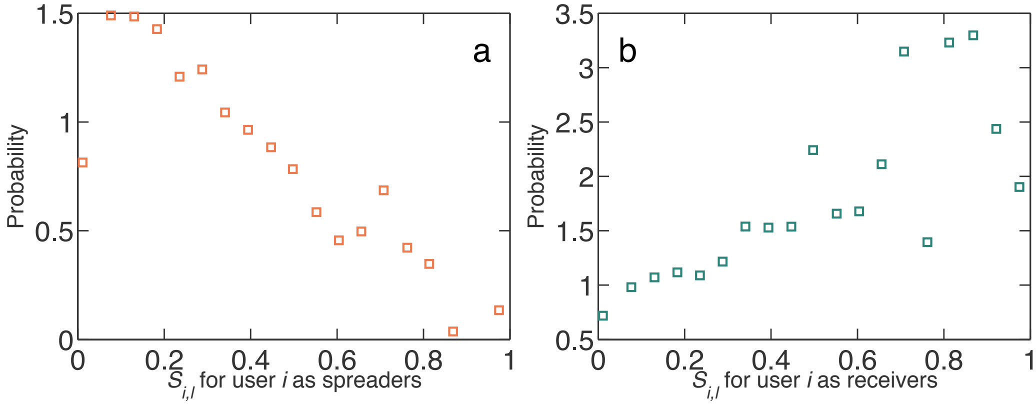


Figure 3.5: Probability ratio of information spreading changes in function of $\mathcal{S}_{i,l}$ for (a) spreaders and (b) receivers. Information spreading undergoes an interesting re-routing from non-experts to experts.

3.3.3 Organizational Context

In an enterprise, understanding how information flows within and between different departments and organizational levels is of great importance, from building a better collaborative environment to controlling information security. Here we examine the impact of organizational context in two directions: one is the influence of departmental restrictions, and the other is the organizational hierarchies.

In Fig. 3.6, we show the median time delay in information spreading for spreaders as different roles of brokerage [48]. There are in total five types of brokers. Figure 3.6 contains illustrative examples for all five: Each box represents a department, and users are from the same department if they are in the same box. If there is only one user in the department, we omit the box for brevity. Our dataset has individuals from as many as 19 departments, and the information pathways consisting of people from different departments are classified into these 5 categories. We observe that the information flows significantly faster in two cases – coordinator and gatekeeper – than the other three cases. These are the only two cases where spreaders and receivers are in the same department. The quartiles also follow similar patterns, yet are not shown in Fig. 3.6 as the waiting time follows a broad distribution [8]. Thus the bottleneck of information flow in the departmental context is to get the information out of the department. We further break down the manager and non-manager cases for each role of brokerage. We find that managers are better as a representative while non-managers are better as a liaison, but the difference between managers and non-managers is seldom large.

We now turn our attention to the impact of organizational levels. While it is intuitive to assume that users would respond faster to emails from people of higher level in the organization (e.g., the reaction of the emails are influenced by the report-to relationship), a previous study [59] on email replies revealed that the reply time does not depend on level difference. Our study shows similar results, confirming that the time delay of information appears to be independent of the hierarchy. Yet, when we look at the probability ratio of email forwarding as a function of the level difference (Fig. 3.7a) and organizational distance (Fig. 3.7b) between initiators and spreaders, we discover some non-homophily effect as opposed to the homophily effect found in [59].

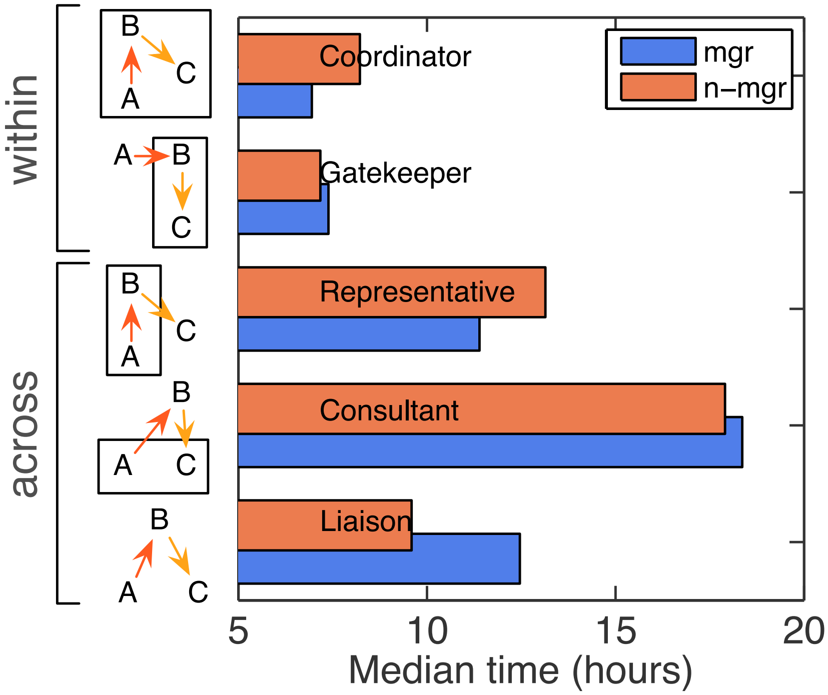


Figure 3.6: Information flow in the departmental context. Each box represents a department, and users are from the same department if they are in the same box. Information spreads faster when B and C are in the same department.

As shown in Fig. 3.7a, information is unlikely to flow between individuals in the same level compared with normal email traffic, and two extreme cases clearly stand out – either bottom up or top down the hierarchy. It tells us that, while the communication between different hierarchies does not yield a faster or slower response, it does matter when determining whether one would decide to pass on the information or not in the first place. Moreover, Fig. 3.7b further confirms the non-homophily effect that the information tends to flow between individuals at a larger distance in the formal organizational structure. This effect shows that "informal networks" and "formal networks" complement each other in information spreading.

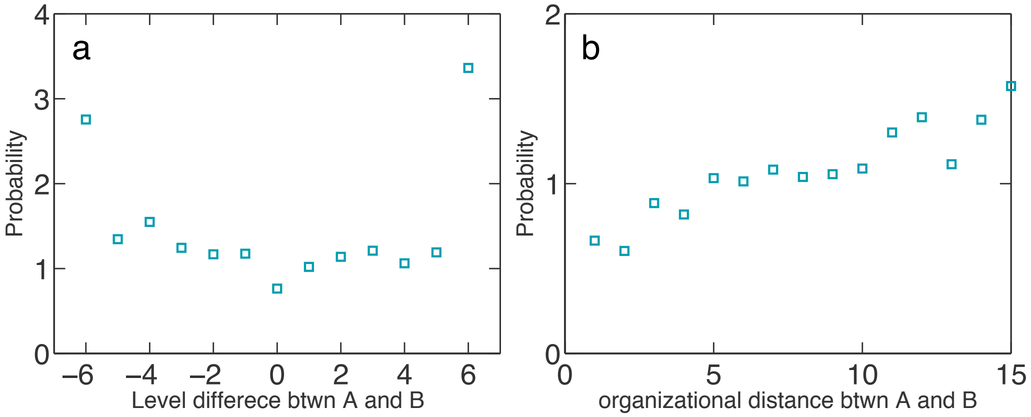


Figure 3.7: Probability ratio of email forwarding as a function of (a) hierarchical level difference and (b) organizational distance between initiators and spreaders. The information spreading exhibits some non-homophily effect.

3.3.4 Individual Characteristics

Another factor that may impact the efficiency of information spreading relates to the individual characteristics of those participating in the spreading process. Do people with different work performance behave differently in getting the word out? A natural hypothesis is that people with better performance are more efficient in spreading the information. While it is generally difficult to get a quantitative measurement of individual performance, the mentioned "billable hours" data serves this purpose. As a consultant's performance is directly related to the total revenue s/he generates, this unique data offers us an opportunity to explore for the first time how individual char-

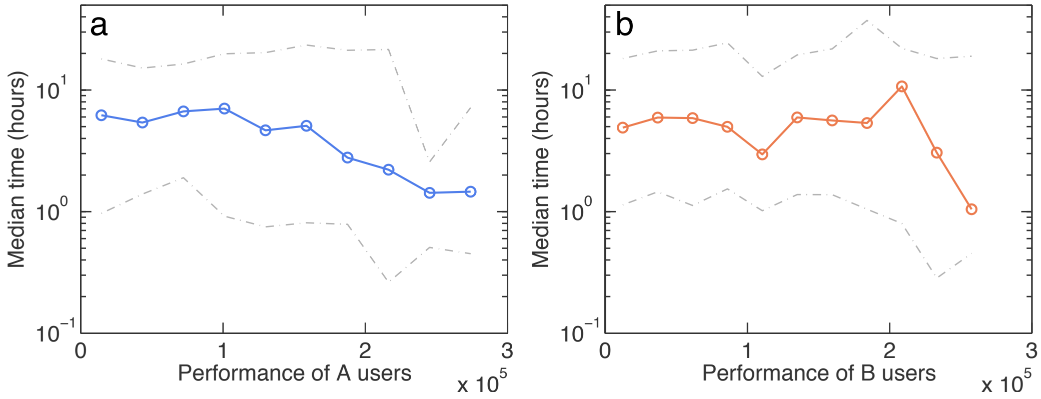


Figure 3.8: Information delay time in hours versus individual performance for (a) initiators and (b) spreaders. The efficiency of the spreading is little affected by individual performance.

acteristic affects the information spreading process. To test this hypothesis, we look at whether there is a correlation between the delay time of information spreading and the performance of the individuals. We find that the hypothesis is not supported by our data. We show in Fig. 3.8 the correlation between the median information's waiting time in hours and the performance of initiators and spreaders, respectively. The dashed grey lines show the 25% quantiles. The information's waiting time appears to be constant for both initiators and spreaders, independent of individual performance.

3.4 Macroscopic information spreading in context

We now go beyond information pathways and turn our attention at the macroscopic level, aiming to understand to what degree spreading processes rely on contextual factors. That is, to how many people a user forwards the information and the information reaches in total. Our dataset contains more than 2000 threads[3]. Each thread can be treated as a rooted tree (Fig. 3.9),where information spreads from one user to others.

We focus on two questions: (i) what are the generic properties of the tree structures within spreading processes? (ii) how much contextual information do we need to incorporate in the models in order to capture these properties? We found, in contrast to the *narrow and deep* trees observed in previous studies [74, 45], that the trees in our study are *bushy yet shallow*. The information fans out, but quickly dies out. We further demonstrated that the way information fans out, i.e., to how many people a user forwards the information, features a high degree of randomness, being independent of

[3] Since we did not have all the communications within the enterprise, we were left with a relatively small number of threads. The readers might be curious whether this sampling issue would affect our observations of the tree structures. As our upcoming stochastic model, which well captures the empirical observations, is purely based on the intrinsic media properties of email systems (i.e., number of recipients n in each email, and its distribution $P_n(n)$), we can therefore validate our results by checking the distribution of n across different datasets. To this end, we measured this quantity in other email datasets([37] and [36]). We found that all email datasets to date share the common feature that $P_n(n)$ universally follows a fat-tailed distribution, indicating that our results are robust to sampling.

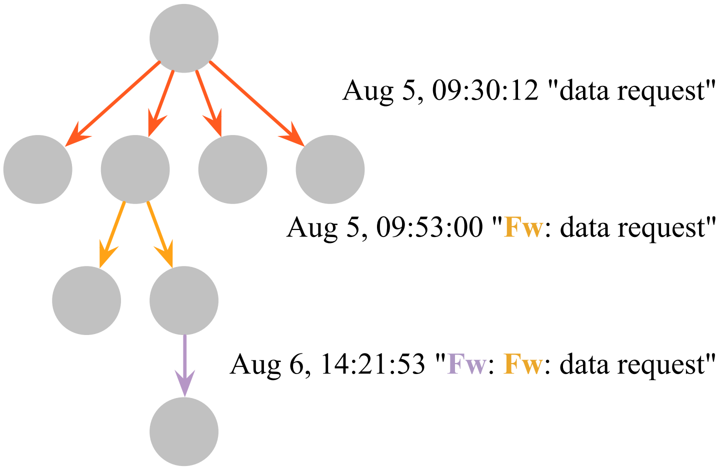

Aug 5, 09:30:12 "data request"
Aug 5, 09:53:00 "Fw: data request"
Aug 6, 14:21:53 "Fw: Fw: data request"

Figure 3.9: An illustrative example of an information spreading tree. This tree is of size 8, width 4, and depth 3.

the connectivity of spreaders in the underlying social network. The overall structural properties of spreading processes can be captured surprisingly well by a simple stochastic model, indicating that information spreading is largely independent of context at the macroscopic level.

3.4.1 Empirical Observations

In this subsection we report the main observations about structural properties of the threads. These observations build the foundation of our models. In summary, there are two interesting findings regarding the observed trees, which can not be interpreted intuitively by existing models.

- **Ultra-shallow trees**: Almost 95% of trees are of depth 2, and trees with more than 4 hops are absent.

- **Stage dependency**: The branching factor (number of children each node has) depends on the distance from the root.

Tree size, width, and depth

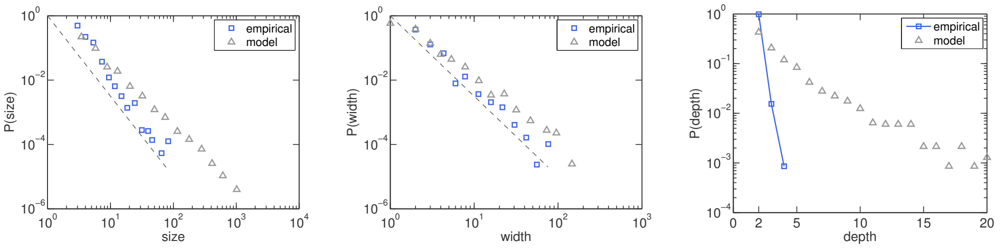


Figure 3.10: Distributions of size, width, and depth of the trees. Empirical measurements are denoted as blue squares, while the grey triangles are predictions of existing models. Dashed lines are guides for the eye, with an exponent of 2.5. The existing models overestimate the tails of the distributions.

The size, width, and depth of a tree are its three most important structural characteristics. The size of a thread is defined as the total number of people involved in the spreading process; the width of a tree is the maximum of number of nodes in each level among all levels of the tree; depth is the length of the longest path from a leaf to the root. (As our forwarding process is conditioned on the emails that were forwarded, the minimum depth of the trees is 2). The distributions of size and width both follow a power law[4], with an exponent of 2.67 and 2.53, respectively (Fig. 3.10). While the power law distribution itself is not unexpected, what is surprising is that the tails of

[4] The likelihood of power law distributions and the exponents hereafter are assessed by applying the techniques in [23].

these two distributions have similar exponents. This fact directly implies, as shown in Figure 3.11, that the size of the trees grows almost linearly with the width (a power law relation with exponent around 1). Moreover, we observe that the tree structures extracted from email forwarding activities are ultra shallow: 95% of the trees are of depth 2, and the distribution of depth decays so fast that we don't observe any tree of depth greater than 4 within 2000+ samples.

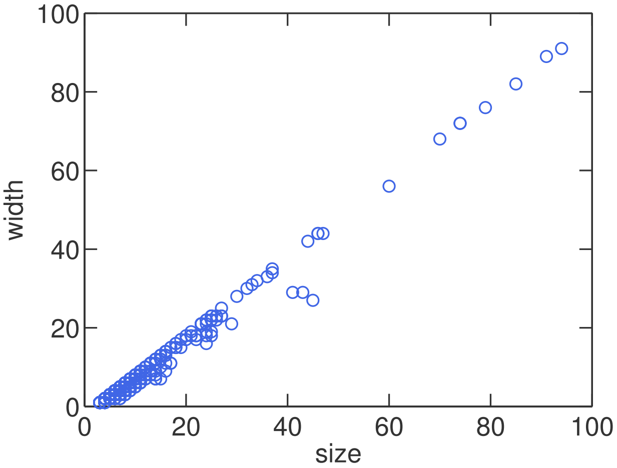


Figure 3.11: Scatter plot of the size and the width of the trees. The size of the tree grows almost linearly wrt the width of the tree.

These findings are puzzling when we apply the classical model for generating a random tree structure: a Galton-Watson branching process [111], in which each node has a random number of children κ, drawn independently according to the same distribution, denoted as $P(\kappa)$. Previous work [45] has shown that, despite the complexities of the process, this simple model fits the data quite well. We therefore follow the modeling procedure of [45] to fit our observation. We first compute the parameters $P(\kappa)$ of a Galton-Walton process by using maximum-likelihood estimation. Then we simulate the process and generate the same number of trees as empirical measurements. The distribution of size, width and depth, plotted as grey triangles in Fig. 3.11, follow a power law, with an exponent of 1.96 and 2.11. Clearly, directly applying this procedure significantly overestimates the tails of distributions, generating trees that are much bigger and deeper than observed empirically. Most prominent is the depth distribution. For trees that are in the subcritical regime, i.e., the mean μ of $P(\kappa)$ is less than 1 ($\mu < 1$), the depth distribution has an exponential tail [51]. However, the measured depth distribution decays much faster than the model prediction.

In summary, the trees we observed here are bushy yet very shallow, which implies that the information spreads efficiently, reaching out to many people then quickly dying out.

Stage dependence

The observations above raise an important question: does the information spreading process change in different stages? We therefore compute the conditional probability of κ given the distance to the root d, $P(\kappa \mid d)$, in Figure 3.12. In a Galton-Watson branching process, $P(\kappa)$ is universal across all nodes, therefore independent of d, predicting the collapse of curves in the plot. We observe that, however, the branching process does depend on the distance to the root. The power law exponent γ_0 of $P(\kappa)$ when $d = 0$ ap-

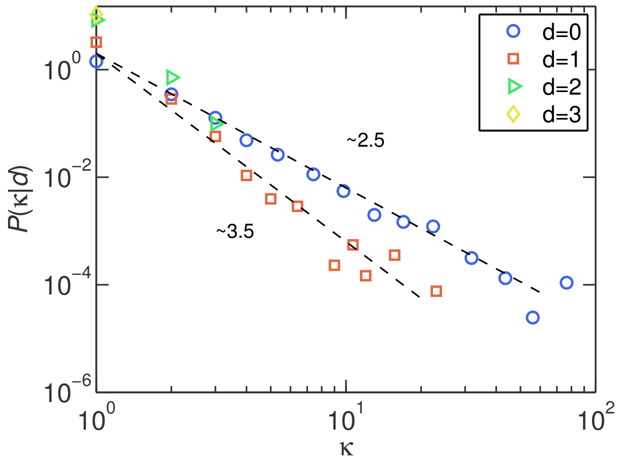


Figure 3.12: Distribution of branching factors κ conditioned on the distance to the root of the tree d. The branching process depends on the spreading stages.

proximately satisfies $\gamma_1 = \gamma_0 + 1$, where γ_1 is the exponent of $P(\kappa)$ when $d = 1$. ($\gamma_0 = 2.48$ and $\gamma_1 = 3.48$). The distribution of κ becomes steeper as we move deeper down the trees, corresponding to the stage dependence, which was also observed in a recent study [66] regarding how online conversation forms yet remained unclear why the exponent of $P(\kappa)$ changes with d, indicating that this effect is generic among different settings, and a model that could appropriately capture this feature would be of great importance in enhancing our understanding of social systems.

3.4.2 Modeling the information spreading process

What is the underlying mechanism that governs the information spreading process? Our goal here is to explore how much contextual information we

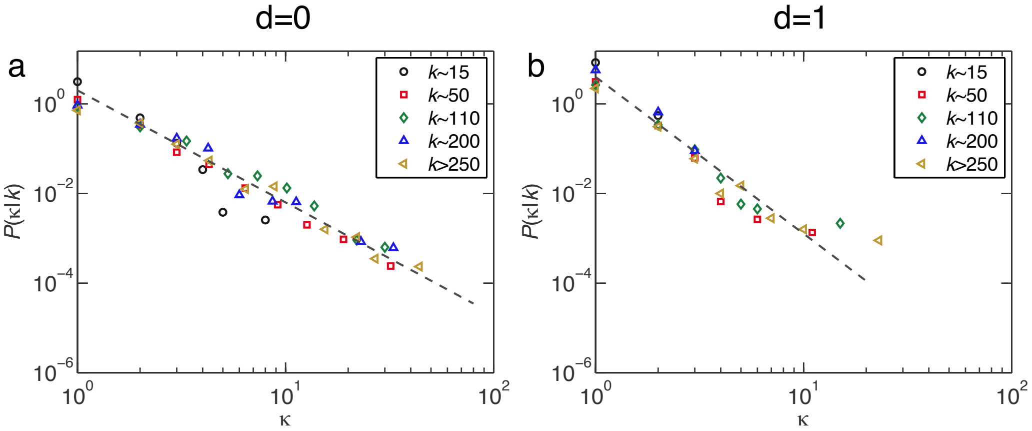


Figure 3.13: Distribution of branching factors κ conditioned on the degree of the node k. The branching factors are independent on the degree connectivity.

need to rely on to model the observed macroscopic structural properties of the trees in Sec. 3.4.1, aiming to quantify to what extent spreading processes at the macroscopic level depend on context.

The observed fat-tailed distributions of branching factors κ in Fig. 3.12 help us assess the properties of nodes in the underlying social network. As shown in Fig. 3.4, the degree distribution, $P(k)$, is also fat tailed [9, 6, 17]. Indeed, individuals are connected differently in the network. While most people have only a few connections, there are a notable number of individuals who have many social neighbors. This raises an important question: to what extent does the information spreading process depend on the underlying social network? First off, the branching factor κ for an individual in the spreading process is upper-bounded by the total number of connections s/he has. Yet beyond that, it is important to inspect whether there is a correlation between k and κ. This question has a number of important implications. In the viral marketing case for example, where the underlying social network is usually not visible, the correlation between k and κ will tell us whether it is a good marketing strategy to carefully choose the seed populations to spread an advertisement. A positive correlation suggests that it does matter who you choose to start the spreading, as social hubs would tend to send the information to more people. Yet if the correlation is not so strong, one could argue that perhaps it is not so important how one chooses the seed population. Another example comes from the difference between the spreading of information and diseases. Indeed, diseases spread from a seed to many others through networks, bearing high level similarity to the spreading of information. The models of epidemics commonly rely on infection rates, where better connected nodes infect more neighbors, corresponding to a strong correlation between k and κ. Therefore, understanding to what extent information spreading relies on the context of underlying network would quantitatively assess the difference between these two spreading processes, arguing whether the existing epidemic models are applicable to the spreading of information.

The correlation between k and κ can be examined by empirically measuring the conditional probability $P(\kappa \mid k)$. Indeed, as $P(\kappa) = \int P(\kappa \mid k)P(k)dk$, if κ is largely uncorrelated with k, $P(\kappa \mid k)$ can therefore be factored out of the integral, giving $P(\kappa) = P(\kappa \mid k)$, leading to a data collapse when plotting $P(\kappa)$ in different curves by grouping individuals of similar k. We show in Fig. 3.13 the conditional probability $P(\kappa \mid k)$ for two different stages of spreading ($d = 0$ vs. $d = 1$), respectively. Surprisingly, we observe very good collapse for different k in both figures, which indicates that there is no direct correlation between k and κ. The breadth of the dissemination of information is independent of the connectivity k of individuals. This indicates, while to whom a user forwards the information indeed depends on the underlying social network (as shown in Sec. 3.3.1), to how many people (κ) one would forward the information does not.

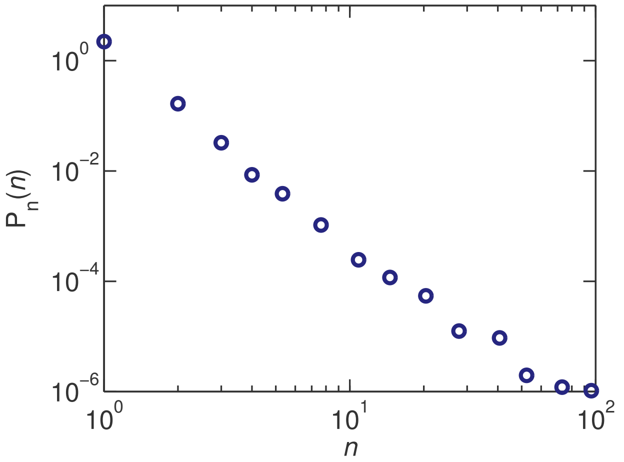


Figure 3.14: The distribution of the number of recipients for each email $P_n(n)$ is fat-tailed.

The surprising independence of node properties of the information spreading process leads us to question its dependence on the media properties of email systems. Therefore, we model the spreading processes by mimicking the way emails are sent. Indeed, an important feature of email communication, distinguishing it from other forms of communication, like cell phones,

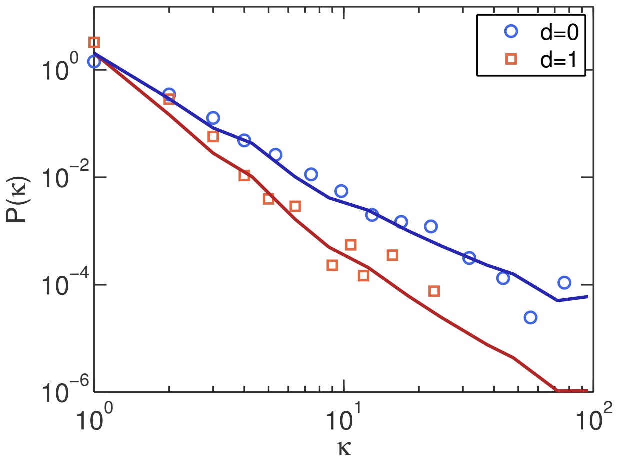


Figure 3.15: $P(\kappa)$ for $d = 0$ and $d = 1$, model prediction (solid lines) vs. experiment measure (scattered squares and circles). Our model well captures the stage dependence phenomenon of information spreading.

is the ability to send a message to multiple recipients at the same time. Therefore, the distribution of the number of recipients for all the emails being sent should follow some non-trivial form, other than $\delta(1)$ in cell phones, i.e., each phone call is made to one recipient only. Let us denote the distribution for emails system as $P_n(n)$ for now, where n represents the number

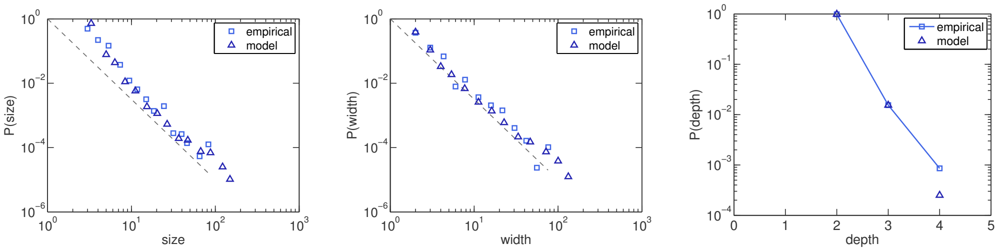


Figure 3.16: Size, width, and depth distributions of model prediction (triangles) with empirical observations (squares). The model matches well with observations. Note the last point in depth distribution is biased by empirical finite size effect, lower bounded by N^{-1}.

of recipients in each email. While some emails are forwarded, many more are not. The easiest way to look at email forwarding is to treat it as an independent decision making process, where each recipient with probability p forwards the information, or probability $1 - p$ does nothing. As email forwarding represents a small fraction of overall email traffic, p should be a small number. When a recipient decides to forward the email, s/he draws a random number from the distribution P_n to decide how many people to send the emails. So the distribution of branching factors should follow the same distribution as P_n, from which the random numbers were drawn, giving $P(\kappa) = P_n(\kappa)$. However, this should only hold for the case of $d > 0$. Indeed, as our study is focused on the emails that are forwarded, there should be an extra term for correcting this conditional probability when $d = 0$. That is, the original emails with more recipients are more likely to get forwarded, as there will be more people to make a decision whether or not to pass on the information. Following this mechanism, the distribution of branching factors at depth 0, $P(\kappa \mid d = 0)$, follows

$$\begin{aligned} P(\kappa \mid d = 0) &= A\left(1 - (1-p)^{\kappa}\right) P_n(\kappa) \\ &= A\left(1 - e^{\kappa \ln(1-p)}\right) P_n(\kappa) \end{aligned} \tag{3.1}$$

where A is the normalization factor, whereas $P(\kappa \mid d > 0)$ follows

$$P(\kappa \mid d > 0) = P_n(\kappa) \tag{3.2}$$

In the limit of $p \to 0$, to the leading power, the relationship between the scaling exponent γ_0 of $P(\kappa \mid d = 0)$, and γ_1 of $P(\kappa \mid d = 1)$, follows the simple relation $\gamma_1 = \gamma_0 + 1$ if $\kappa \ll -1/\ln(1-p) \approx 1/p$, and $\gamma_1 = \gamma_0$ if $\kappa \gg -1/\ln(1-p) \approx 1/p$.

Both parameter p and function P_n can be measured independently from our data, yielding $p = 0.012$ and a fat-tailed distribution P_n (Fig. 3.14). We can therefore simulate the distributions of size, width, and depth using these two measured parameters. The results are shown in Fig. 3.16, with observations as squares and model predictions as triangles. Surprisingly, they all match the empirical observations very well. The distributions of size and width follows a power law, with an exponent of 2.63 and 2.51, very close to the empirical observations (2.67 and 2.53). Furthermore, the observation of stage dependence could be verified analytically by plugging the parameters into eqs. (3.1) and (3.2), as plotted in blue and red lines in Fig. 3.15, respectively. It is also very well captured by the model.

The model we described above for email forwarding processes is purely stochastic and has two parameters, p and P_n, which are measured from our email dataset independently. Perhaps unexpectedly, such a simple model explains a great deal of observations. This, together with Fig. 3.13, indicates that, despite the complexity in real life, the macroscopic structures of information spreading processes are largely independent of contextual information and can be well captured and explained via simple machanisms.

3.5 Conclusions and Future Work

Applications of social systems rely on our understanding of information spreading patterns. In this work, by combining two related but distinct large scale datasets, we address the factors that govern information spreading at both microscopic and macroscopic levels. We found, microscopically, whom the information flows to indeed depends on the structure of the underlying social network, individual expertise and organizational hierarchy. The performance of individuals has little influence on the efficiency of spreading, yet departmental constraints do slow down the process. At the macroscopic level, however, although seemingly complex, the structural properties of spreading trees, i.e., to how many people a user forwards the information and the total coverage the information reaches, can be well captured by a simple stochastic branching model, indicating that the spreading process follows a random yet reproducible pattern, largely independent of context. We believe that our findings could guide users to build better social and collaborative applications, design tools and strategies to spread information more efficiently, improve information security, develop predictive tools for recommendation systems, and more.

Future directions mainly fall into two lines. The first is to develop a better prediction model for information flow. Indeed, upon understanding to whom one forwards information, when one would forward it, and to how many people, the question thereafter is can we build a better prediction model of the flows? The second direction is about the mutation of information. People sometimes add extra information or express opinions about existing information when passing along the originals to others. How does information mutate along the way? How does the mutation of information affect the patterns of spreading? These questions stand as missing chapters in our understanding of spreading processes. Indeed, with the availability of large-scale email datasets, thorough inspection of the email message contents will reveal the dynamics of information itself, which in turn can yield better predictive tools for information spreading.

Chapter 4

Connections between Human Mobility and Social Networks

4.1 Introduction

Social networks have attracted particular interest in recent years, largely because of their critical role in various applications [35, 17]. Despite the recent explosion of research in this area, the bulk of work has focused on the social space only, leaving an important question of to what extent individual mobility patterns shape and impact the social network, largely unexplored. Indeed, social links are often driven by spatial proximity, from job- and family-imposed shared programs to joint involvement in various social activities [91]. These shared social foci and face-to-face interactions, represented as overlap in individuals' trajectories, are expected to have significant impact on the structure of social networks, from the maintenance of long-lasting friendships to the formation of new links.

Our knowledge of the interplay between individual mobility and social network is limited, partly due to the difficulty in collecting large-scale data that record, simultaneously, dynamical traces of individual movements and social interactions. This situation is changing rapidly, however, thanks to the pervasive use of mobile phones. Indeed, the records of mobile communications collected by telecommunication carriers provide extensive proxy of mobility patterns and social ties, by keeping track of each phone call between any two parties and the localization in space and time of the party that initiates the call. The high penetration of mobile phones implies that such data captures a large fraction of the population of an entire country. The availability of these massive CDRs (Call Detail Record) has made possible, for instance, the empirical validation in a large-scale setting of traditional social network hypotheses such as Granovetter's strength of weak ties [83], the development of a first generation of realistic models of human mobility [47, 98] and its predictability [99]. Indeed, despite the inhomogeneous spatial resolution (the uneven reception area of mobile phone towers) and sampling rates (the timing of calls), the large volume of CDR data allows us to reconstruct many salient aspects of individual daily routines, such as the most frequently visited locations, and the time and periodicity of such visits. Therefore, these data serve as an unprecedented social microscope helping us scrutinize the mobility patterns together with social structure and the intensity of social interactions.

In this work, we follow the trajectories and communication patterns of approximately 6 Million users over three months, by using CDR data from an anonymous country, aiming to measure for any pair of users u and v:

- *How similar is the movement of* u *and* v*.* For this purpose, we introduce a series of *co-location* measures quantifying the similarity between their movement routines, prompting us to call them the *mobile homophily* between u and v.

- *How connected are* u *and* v *in the social network.* For this purpose, we adopt several well-established measures of network proximity, based on the common neighbors or the structure of the paths connecting u and v in the who-calls-whom network.

- *How intense is the interaction between* u *and* v*.* For this purpose we use the number of calls between u and v as a measure of the strength of their tie.

Our analysis offers empirical evidence that these three facets, co-location, network proximity and tie strength, are positively correlated with each other. In particular, we find that the higher the mobile homophily of u and v, the higher the chance that u and v are strongly connected in the social network, and that they have intense direct interactions. These findings uncover how the social network, made of numerous explicit who-calls-whom ties, is embedded into an underlying mobility network, made with the implicit ties dictated by the mobile homophily.

The emergence of such surprising three-fold correlation hints that it is conceivable, to some extent, to predict one of the three aspects given the other two. Indeed, we demonstrate in this study how the predictive power hidden in these correlations can be exploited to identify new ties that are about to develop in a social network. Specifically, we study the influence of co-location and mobile homophily in link prediction problems, asking: what is the performance of mobility-based measures in predicting new links, and can we predict more precisely whether two users u and v (that did not call each other in the past) will call each other in the future, by combining the measurements of their network proximity *and* mobile homophily? Our key findings are summarized as follows:

- The mobility measures on their own carry remarkably high predictive power, comparable to that of network proximity measures.

- By combining both mobility and network measures, we manage to significantly boost the predictive performance in supervised classification, detecting interesting niches of new links very precisely. For example, by considering a subset of potential links (pair of users) with high network proximity and mobile homophily, we are able to learn a decision-tree classifier with a precision of 73.5% and a recall of 66.1% on the positive class. In other words, only approximately one fourth of the predicted new links were false positives, and only one third of the actual new links were missed by the predictor.

To the best of our knowledge, this work presents the first assessment of the extent individuals' daily routines as a determinant of social ties, from empirical analysis to prediction models. With recent proliferating advances on human mobility and social networks, we believe our findings are of fundamental importance in our understanding of human behavior, provide significant insights towards not only link prediction problems but also the evolution and dynamics of networks, and could potentially impact a wide array of areas, from privacy implications to urban planning and epidemic prevention.

4.2 Mobile Phone Data

Currently the most comprehensive data that contains simultaneously both human mobility and social interactions across a large segment of the population is collected by mobile phone companies. Indeed, mobile phones are carried by their owners during their daily routines. As mobile carriers record for billing purposes the closest mobile tower each time the user uses his phone, the data capture in detail individual movements. With almost 100% penetration of mobile phones in industrial countries, the mobile phone network is the most comprehensive proxy of a large-scale social network currently in existence. We exploit in this study a massive CDR dataset of approximately 6 Million users, which, to the best of our knowledge, is the largest dataset analyzed to date containing both human trajectories and social interactions. We focused on 50k individuals selected as the most active users (identical to those that were studied in a recent publication [99]), following not only their trajectories but also their communication records during 14 successive weeks in 2007.

The resulting dataset contains around 90M communication records among the individuals, and over 10k distinct locations covering a radius of more than 1000 km. Each record, for our purposes, is represented as 4-tuple $\langle x, y, t, l \rangle$, where user x is the caller, user y is the callee, t is the time of the call, and l is the location of the tower that routed the call. The temporal granularity used in this study is the hour, justified by the finding in [47, 99, 98]. Let V denotes the set of users. For each user $x \in V$, the total number of calls initiated by x is denoted as $n(x)$. For x's i-th communication, where $1 \leq i \leq n(x)$, the time stamp, location, and the contacted user are denoted as $T_i(x)$, $L_i(x)$ and $N_i(x)$, respectively. Given a time interval between t_0 and t_1, the set of communications between pairs of users occurred within the interval is denoted as $E[t_0, t_1] \equiv \{(x, y) | x, y \in V, \exists i, 1 \leq i \leq n(x), t_0 \leq T_i(x) < t_1, N_i(x) = y\}$. In other words, we add an edge (x, y) if there has been at least one communication between x and y in the interval. Therefore, $G[t_0, t_1] \equiv \{V, E[t_0, t_1]\}$ is the resulting social network within the time interval.

To prepare for the link prediction experiments, we further separate our data into 2 parts: first 9 weeks for constructing the old network and the rest 5 weeks for the new network. For each link $e \in E$, we classify it according to its time stamp $t(e)$. $E_t \equiv \{e|e \in E, t \leq t(e) < t+1\}$ is defined as the set of edges of the resulting network after aggregating the communications in the t-th week. The "past" and "future" sets are therefore denoted as $E_{old} = \bigcup_{t=1}^{9} E_t$ and $E_{new} = \bigcup_{t=10}^{14} E_t$. In our study, we focus on nodes in the largest connected component $G_{old} = \{V_{old}, E_{old}\}$, where we observe in total $|V_{old}| = 34,034$ users and $|E_{old}| = 51,951$ links.

4.3 Network Proximity

General approaches in link prediction tasks have been focused on defining effective network based "proximity" measures, so that two nodes that are close enough on the graph but not yet connected may have a better likelihood of becoming connected in the future. As the main focus of the paper is to explore the predictive power of mobility compared and combined with topological predictors, we selected four representative quantities which have been proven to perform reasonably well in previous studies (for more details of the quantities and their performance on citation networks, see [75].)

- *Common neighbors.* The number of neighbors that nodes x and y have in common. That is, $CN(x,y) \equiv |\Gamma(x) \cap \Gamma(y)|$, where $\Gamma(x) \equiv \{y|y \in V, (x,y) \in E\}$ is the set of neighbors of x.
- *Adamic-Adar* [2]. A refinement of $CN(x,y)$ by weighting common neighbors based on their degrees, instead of simple counting. Therefore

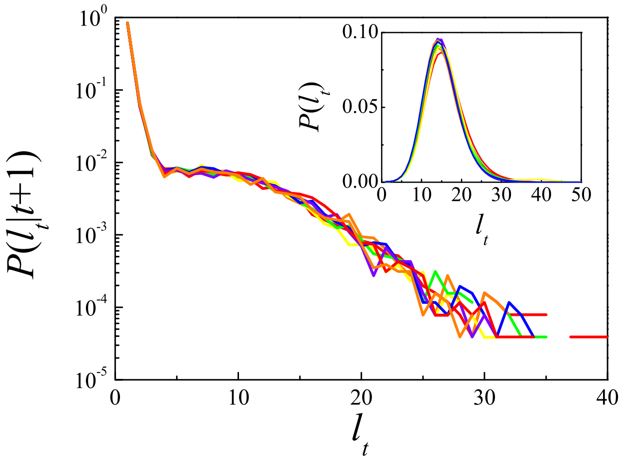


Figure 4.1: The probability density function $P(l_t|t+1)$ that a link has chemical distance l in previous week. Inset: the probability density function of chemical $P(l_t)$ for different weeks.

the contribution from hubs to common neighbors is penalized by the inverse logarithm of their degree.

$AA(x,y) \equiv \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$.

- *Jaccard's coefficient.* Defined as the size of the intersection of the neighbors of two nodes, $\Gamma(x)$ and $\Gamma(y)$, divided by the size of their union, characterizing the similarity between their sets of neighbors.

  $J(x,y) \equiv |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$.

- *Katz* [60]. Summation over all possible paths from x to y with exponential damping by length to weight short paths more heavily. $K(x,y) \equiv \sum_{l=1}^{\infty} \beta^l \cdot |paths^l_{x,y}|$, where $paths^l_{x,y}$ is the set of all paths with length l from x to y (damping factor β is typically set to 0.05.)

Most network proximity measures are related to the chemical distance on the graph, under the natural assumption that new links are more likely to occur between nodes that are within a small distance on the graph. The *chemical distance* $l(x, y|E)$ is defined as the length of the shortest path between two nodes x and y. $l(x, y|E) = 1$ implies that nodes x and y are connected, or $(x, y) \in E$. The role of chemical distance on tie formation can be tested directly by measuring the probability $P(l_t|t + 1)$ for a new link $e \equiv (x, y) \in E_{t+1}$ to have a chemical distance l_t measured at previous week t. That is, $P(l_t|t + 1) \equiv |\{e|e \equiv (x, y) \in E_{t+1}, l(x, y|E_t) = l_t\}|/|E_{t+1}|$. This distribution is shown in Fig. 4.1, different colors indicating different time windows t. We find, first of all, $P(l_t|t + 1)$ is stable over different weeks (1 through 14), indicating that the aggregation process we adopted to construct the network is robust, and that $P(l_t|t+1)$ is largely independent wrt the time windows. Second, $P(l_t|t + 1)$ decays rapidly as l_t increases, consistent with previous study [71] on other data sets. This implies that the majority of new links are between nodes within two hops from each other, i.e., nodes with common neighbors. Third, the Poisson distribution of the chemical distance for arbitrary pairs (inset of Fig. 4.1) suggests that the most probable distance for two users to form a link at random is around 12, while it is only 2 for pairs that do form new links.

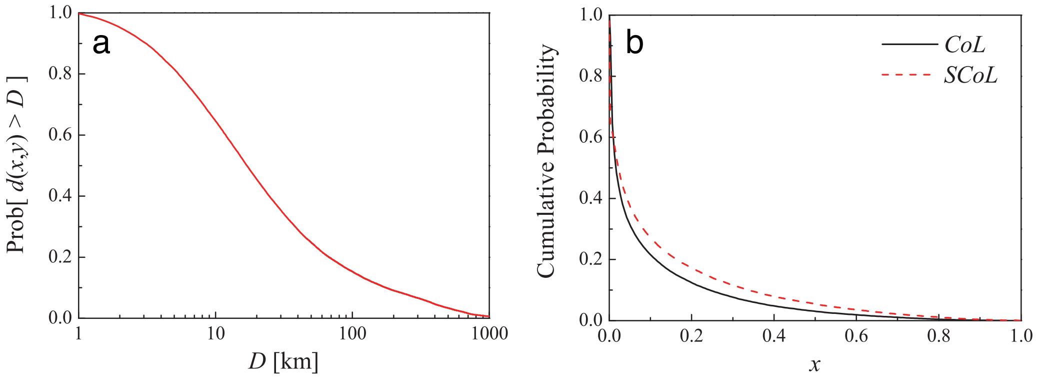


Figure 4.2: a) The probability two users i and j have distance $d(i,j) > D$. b) The probability two users i and j have Co-Location $CoL(i,j)$ (solid) and Spatial Co-Location $SCoL(i;j)$ (dashed) greater than x.

4.4 Mobile Homophily

Similar to the graph-based approaches, a natural strategy to predict new links by leveraging mobility information is to look for quantities that capture some degree of closeness in physical space between two individuals. Indeed, people who share high degree of overlap in their trajectories are expected to have a better likelihood of forming new links [91]. Therefore, we explored a series of quantities aiming to define the similarity in mobility patterns of two individuals.

- *Distance.* Let

$$ML(x) \equiv \text{argmax}_{l \in Loc} PV(x,l)$$

be the most likely location of user x, where Loc is the set of all locations

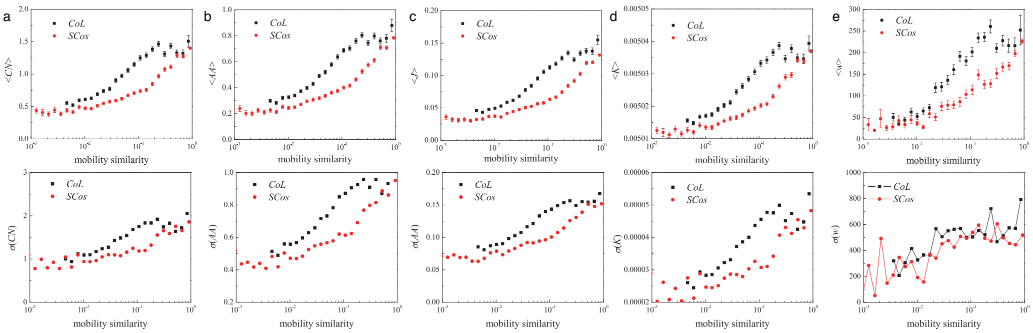


Figure 4.3: Correlations between mobility measures (CoL and $SCos$) and a) Common Neighbor, b) Adamic Adar, c) Jaccard Coefficient, d) Katz, and e) link weight. The upper panels show the mean values, whereas the lower panels show the standard deviations

(cell phone towers), and

$$PV(x,l) \equiv \sum_{i=1}^{n(x)} \delta\left(l, L_i(x)\right)/n(x)$$

is the probability that user x visits location l[1]. We define $d(x,y) \equiv \mathrm{dist}(ML(x), ML(y))$ as the distance between two users x and y, representing the physical distance between their most frequented locations.

- *Spatial Co-Location Rate*. The probability that users x and y visit at the same location, not necessarily at the same time. Assuming that the probability of visit of any two users are independent, we define:

$$SCoL(x,y) \equiv \sum_{l \in Loc} PV(x,l) \times PV(y,l)$$

[1] Here $\delta(a,b) = 1$ if $a = b$, 0 otherwise.

- *Spatial Cosine Similarity.* The cosine similarity of user x and y's trajectories, capturing how similar their visitation frequencies are, assigned by the cosine of the angle between the two vectors of number of visits at each location for x and y.

$$SCos(x, y) \equiv \sum_{l \in Loc} \frac{PV(x, l) \times PV(y, l)}{\|PV(x, l)\| \times \|PV(y, l)\|}$$

- *Weighted Spatial Cosine Similarity.* The *tf-idf* version of cosine similarity of the visitation frequencies of users x and y, where the contribution of each location l is inversely proportional to the (log of) its overall population in l. Coherent with the *tf-idf* idea in information retrieval, this measure promotes co-location in low-density areas, while penalizes co-location in populated places.

- *Co-Location Rate.* The probability for users x and y to appear at the same location during the same time frame (hour):

$$CoL \equiv \frac{\sum_{i=1}^{n(x)} \sum_{j=1}^{n(y)} \Theta\left(\Delta T - |T_i(x) - T_j(y)|\right) \delta\left(L_i(x), L_j(y)\right)}{\sum_{i=1}^{n(x)} \sum_{j=1}^{n(y)} \Theta\left(\Delta T - |T_i(x) - T_j(y)|\right)}$$

where $\Theta(x)$ is the Heaviside step function, and ΔT is set to 1 hour. This quantity takes into account the simultaneous visits of two users at the same location, i.e., both spatial and temporal proximity, normalized by the number of times they are both observed at the same time frame.

- *Weighted Co-Location Rate.* The *tf-idf* version of CoL, i.e., the probability for two users x and y to co-locate during the same hour, normalized by the (log of) population density of the co-location at that hour.

- *Extra-role Co-Location Rate.* The probability for two users x and y to co-locate during the same hour at night or weekends. As shown in [34], close proximity of two individuals during off-hours may serve as a powerful predictor for symmetric friendship.

The quantities listed above either aim at measuring the geographical closeness or the degree of trajectory overlap of two individuals, characterizing their mobile homophily. It should be noted that it is not obvious whether the spatiotemporal co-location measures, e.g., CoL, would yield better estimates of the probability of face-to-face interactions than spatial only measures, e.g., $SCoL$. Indeed, on one hand, CoL quantifies the co-presence of two users in the same place around the same moments, corresponding to a high likelihood of meeting face-to-face. Yet there are circumstances where two users do co-locate but are not captured by the data if any one of them did not place any phone calls. And this latter case is captured to some extent by $SCoL$, as the necessary condition for two individuals to meet is the spatial overlap of their trajectories.

We now explore the distributions of the various measures over the linked pairs of individuals $(x, y) \in E_{old}$. In Fig. 4.2a we show the complementary

Table 4.1: Pearson Coefficients

	CoL	$Scos$	CN	J	AA	K	w	dML
CoL	1	0.76286	0.25359	0.19618	0.2251	0.18952	0.14521	-0.17894
$Scos$	0.76286	1	0.30789	0.25657	0.28679	0.24933	0.14402	-0.24938
CN	0.25359	0.30789	1	0.82384	0.88147	0.81108	0.11348	-0.10136
J	0.19618	0.25657	0.82384	1	0.94437	0.99939	0.05989	-0.098562
AA	0.2251	0.28679	0.88147	0.94437	1	0.93806	0.086881	-0.10126
K	0.18952	0.24933	0.81108	0.99939	0.93806	1	0.053842	-0.095631
w	0.14521	0.14402	0.11348	0.05989	0.086881	0.053842	1	-0.029339
dML	-0.17894	-0.24938	-0.10136	-0.098562	-0.10126	-0.095631	-0.029339	1

cumulative distribution function (CCDF) of geographical distances $d(x, y)$. We find that $d(x, y)$ follows a fat-tailed distribution, consistent with previous studies [67, 63, 73], meaning that while most friends live close to each other, there are also friends who are far apart. The CCDF plots of CoL and $SCoL$ are shown in Fig. 4.2b as solid and dashed line, respectively. $SCoL$ measures the probability for two users to appear at the same location, capturing, spatially, the degree of trajectory overlapping. CoL quantifies the probability of appearing at the same place around the same time, characterizing the spatio-temporal overlap of trajectories. We find that “friends” typically do co-locate, in that most pairs $(x, y) \in E_{old}$ exhibit non-zero spatial or spatio-temporal overlap in their trajectories, and such overlap decays very fast.

4.5 Correlation between mobile homophily and network proximity

We explore a series of connections between similarity in individual mobility patterns and social proximity in the call graph, by measuring the correlation between the proposed mobility and network quantities, using again the edges in G_{old}. We also consider the strength of the ties in the network, quantified by the number of calls placed between any two users (during the first 9 weeks of our observation period.) In Fig. 4.3, we plot the mean values and the standard deviations of Common neighbors, Adamic-Adar, Jaccard's coefficient, Katz, and the strength of social ties for different values of Co-Location and Spatial Cosine Similarity, discretized by logarithmic binning. We find that the quantities that characterize the proximity in the social graph systematically correlate with mobility measures. The more similar two users' mobility patterns are, the higher is the chance that they have close proximity in the social network, as well as the higher is the intensity of their interactions. Furthermore, Fig. 4.4 demonstrates that the geographical distance between two individuals decays logarithmically with mobility measures. We omit the plots where the network proximity measures and the tie strength are on the x-axis, due to space limitations, but we observe a qualitatively similar trend in all cases. The Pearson coefficients of each pairs of variables are reported in Table 4.1. It is interesting to observe that tie strength, although conceived as a network measure, is more strongly correlated with mobile homophily than with network proximity measures.

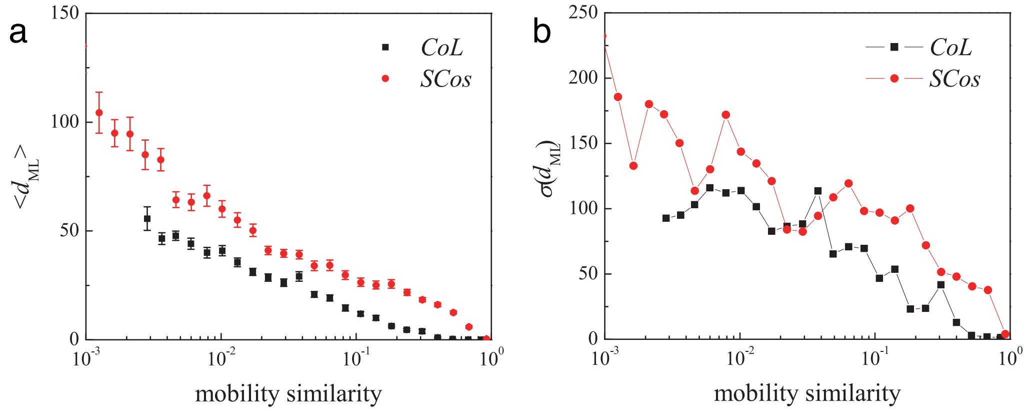


Figure 4.4: Correlations between mobility measures (CoL and $SCos$) and distance between two individuals. (mean values in a and standard deviations in b)

Taken together, our results indicate that mobile homophily, network proximity and tie strength strongly correlate with each other. This fact implies that mobile homophily is a viable alternate candidate to predict network structures, and motivates the investigation of a novel approach to link prediction that takes into account both mobility and network measures. Moreover, we find that the standard deviation for the correlation plots are not small, hinting that there are extra degrees of freedom which allow us to further improve our predictive power by using supervised classification methods combining the mobility and network dimensions together.

4.6 Link Prediction

4.6.1 Design of the link prediction experiment

We now study the link prediction problem in the context of our mobile social network. Link prediction is a classification problem, aimed at detecting, among all possible pairs of users that did not call each other in the past, those that will communicate in the future. We define a *potential link* any pair of users (u, v) such that $(u, v) \notin E_{old}$, i.e., users u and v did not call each other from week 1 through 9, and a *new link* any potential link (u, v) such that $(u, v) \in E_{new}$, i.e., users u and v did not call each other from week 1 through week 9, but did call each other (at least once) from week 10 through week 14. Finally, we define a *missing link* any potential link which is not a new link, i.e., a pair of users that did not call each other in the entire period from week 1 through week 14. For any potential link (u, v), let $NL(u, v)$ be a binary variable with value 1 if (u, v) is a new link, and 0 if (u, v) is a missing link.

In this setting, link prediction is formalized as a binary classification problem over the set of all potential links, where the class label is specified by the NL variable, and the predictive variables are the network and mobility quantities introduced in Sec. 4.3 and 4.4, measured over the first period from week 1 through week 9. According to this formulation, we aim to predict whether a potential link becomes a new link in the "future" based on the observation of its "past" network connectedness and co-location.

Our dataset consists of $n = 34,034$ users and $m = 51,951$ links, resulting in $(n(n-1)/2) - m = 579,087,610$ potential links. Yet the actual new links are only $12,484$ – about 2 new links every 10^5 potential links! The significant number of potential links creates obvious computational challenges, both in terms of memory and time. Moreover, the huge disproportion between new links and missing links implies an extreme unbalance between the positive and negative class, which makes the classification task prohibitive. To cope with both difficulties, we followed two complementary strategies for selecting subsets of potential links: *i*) *progressive sampling*: we consider increasingly large samples of missing links, up to some manageable size, and *ii*) *links with common neighbors*: we concentrate on the interesting case of pairs of nodes that are two hops away in the network, i.e., nodes with common neighbors, and consider the entire population of potential links between such nodes. We report below the results obtained in our link prediction analysis in both cases.

Another dimension of our study is the kind of classification used. Adhering to the machine learning terminology [76], we consider both *unsupervised* and *supervised* link prediction:

- The unsupervised method, originally proposed in [75], consists in ranking the set of potential links using one of the available network or mobility quantities, and then classifying as new links the k top-ranked potential links, where k is the expected number of new links (as measured in the dataset.) The rest are classified as missing links.

	predicted class = 0	predicted class = 1
actual class = 0	TN (true neg.)	FP (false pos.)
actual class = 1	FN (false neg.)	TP (true pos.)

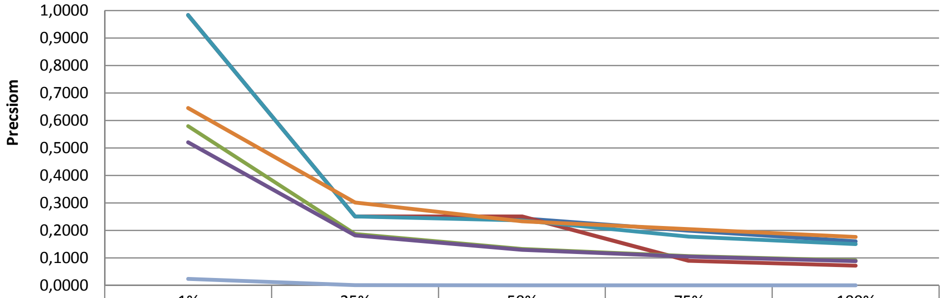
Table 4.2: Confusion matrix of a binary classifier

- The supervised method consists in learning a classifier, e.g., a decision tree, using a training set of new links and missing links, and then classifying each pair as a new or missing link according to the class assigned by the learned classifier.

Different unsupervised classifiers are obtained by considering the various network and mobility measures, and different supervised classifiers are obtained by considering different combinations of the same quantities as predictive variables. We systematically constructed the complete repertoire of classifiers, based either on network quantities, or mobility quantities, or the combination of the two; we then compared their quality and predictive power. To this extent, we put particular attention on the metric used to assess a classifier, given that simple accuracy (over either the training or test set) is a misleading measure for classifiers learned over highly unbalanced datasets. Indeed, recall that in our case the trivial classifier that labels each potential link as missing has a 99.998% accuracy. The real challenge in link prediction is achieving high *precision* and *recall* over positive cases (new links), defined in terms of the confusion matrix of a classifier (see Table. 4.2): $precision = \frac{TP}{TP+FP}$, and $recall = \frac{TP}{TP+FN}$. Traditionally, precision and recall are combined into their harmonic mean, the F-measure. However, we put more emphasis on precision, as the most challenging task is to classify some potential links as new links with high probability, even at the price of a non negligible number of false negatives. We also use lift and gain charts to compare the precision of the various classifiers over the percentiles of the examined test cases.

4.6.2 Progressive sampling of missing links

In our first set of experiments we created various unsupervised and supervised classifiers over the complete dataset of positive cases, i.e., 12,484 new links, augmented with up to 51M negative cases of missing links. We assess the precision achieved by each classifier when used with all 12,484 new links and increasing fractions of missing links, i.e., to 1%, 25%, 50%, 75% and 100% of the total 51M missing links sampled. Figure 4.5 summarizes our findings for unsupervised classifiers. For each network/mobility quantity Q and each dataset with increasing samples of missing links, we rank the potential links in the dataset for decreasing values of Q, and the top ranked 12,484 links are predicted as new links. Each line in Fig. 4.5 describes how the precision for different quantities decays with the size of missing links. On the positive side, all unsupervised classifiers are significantly better than random guessing, and the decay of their precision tends to stabilize. Nevertheless, as these 51M links are only about 10% of the total missing links, we conclude that all quantities exhibit modest predictive power. The most surprising finding is that the co-location measures have a comparable precision to network measures: slightly worse than best network predictors (Katz, Adamic-Adar), but better than Common Neighbors. Moreover, mobility measures have a slower decay than network measures over increasing negative sample size. The observation that the two classes of measures have approximately similar predictive power offer further evidence that social connectedness is strongly correlated with mobile homophily.



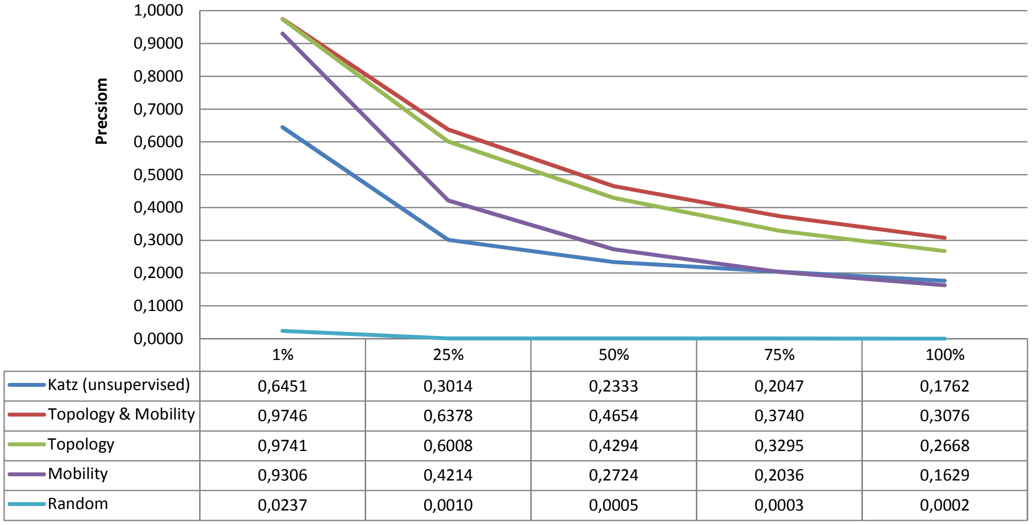
	1%	25%	50%	75%	100%
Adamic Adar	0,9841	0,2507	0,2441	0,1988	0,1602
Common Neighbors	0,9829	0,2507	0,2507	0,0895	0,0715
Cosine Colocation	0,5794	0,1871	0,1325	0,1069	0,0906
ST Colocation	0,5203	0,1817	0,1295	0,1049	0,0884
Jaccard	0,9833	0,2507	0,2363	0,1777	0,1505
Katz	0,6451	0,3014	0,2333	0,2047	0,1762
Random	0,0237	0,0010	0,0005	0,0003	0,0002

Figure 4.5: Precision of unsupervised classifiers over increasing fractions of missing links (1%, 25%, 50%, 75% and 100% of the total 51M missing links sampled). Ranking is obtained using the various network and mobility measure; precision refers to the fraction of new links among the top-ranked 12,484 potential links; the precision of the random classifier is shown as baseline.

Figure 4.6 illustrates the supervised case: we consider the best classifiers obtained using network and mobility measures, both in isolation and combined together. Once again, we consider negative samples of increasing size, up to 51M missing links, and measure the decay of precision as in the unsupervised case. We considered a vast repertoire of classification algorithms (decision trees, random forests, SVM, logistic regression) under diverse parameter settings, and report in the chart the most robust classifiers, evaluated with cross validation, with strongest evidence against overfitting. In the chart we also compare the precision of the supervised methods with that of the best unsupervised predictor (Katz). We observe that the precision of the supervised classifiers is about double of their unsupervised counterpart, and mobility measures once again achieve comparable predictive powers to the traditional network measures. The best precision, around 30% in the 51M case, is obtained using the network and mobility measures combined together. Therefore, using network measures in combination with co-location measures yields a sensible improvement. Indeed, the probability of correctly predicting a new link is 1500+ times larger than random guessing.

Potential links with common neighbors

To get better insight, we concentrate on the nodes that are two hops away from each other in E_{old}, i.e., all potential links (u, v) of mobile users in our complete network such that u and v have at least one common neighbor during the first two months. The motivations behind this approach are two-fold. First, most new links that do form belong to this category (Fig. 4.1), and we hope to boost our prediction models by focusing on this most promising set of links. Second, by focusing on these links, the total number of potential links becomes computationally manageable, which enables us to assess the asymptotic behavior of prediction accuracy.



	1%	25%	50%	75%	100%
Katz (unsupervised)	0,6451	0,3014	0,2333	0,2047	0,1762
Topology & Mobility	0,9746	0,6378	0,4654	0,3740	0,3076
Topology	0,9741	0,6008	0,4294	0,3295	0,2668
Mobility	0,9306	0,4214	0,2724	0,2036	0,1629
Random	0,0237	0,0010	0,0005	0,0003	0,0002

Figure 4.6: Precision of the best supervised classifiers found over increasing fractions of missing links (1%, 25%, 50%, 75% and 100% of the total 51M missing links sampled), using only network measures, only mobility measures, and combination of both. Precision of best unsupervised classifier (K) and random classifier is shown as baseline.

There are 266,750 potential links in this case, of which 3,130 (1.17%) formed a new link. Note that, different from the previous case, we now consider the entire population of missing links. We study the precision of the unsupervised and supervised methods in this case. In the unsupervised case, the precision for the different measures is computed by considering the fraction of new links in the top-ranked 3,130 cases in the list ordered by the precision of each measure in descending order:

As we now have a complete set of negative cases, we corroborate our findings in Sec. 4.6.2 that mobility measures indeed yield remarkably high predictive power in the unsupervised setting, comparable to network measures in

Measure	Precision
Katz	9.1%
Adamic-Adar	7.8%
Spatial Cosine Similarity	5.6%
Weighted Spatial Cosine Similarity	5.6%
Extra-role Co-Location Rate	5.1%
Weighted Co-Location Rate	5.1%
Common Neighbors	5.1%
Co-Location Rate	5.0%
Jaccard	3.0%

Table 4.3: Precision in unsupervised learning for different features

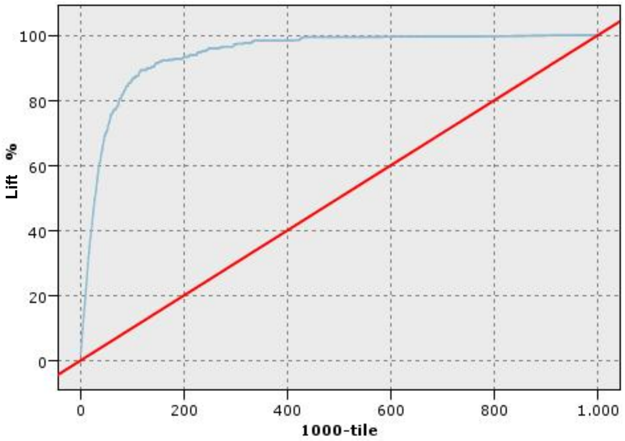


Figure 4.7: Lift chart of the best decision tree found in the dataset of potential links with common neighbors; the x-axis represents the percentiles of the potential links in the test set ranked by decreasing probability of being new links, as specified by the learned classifier; a point (x, y) in the blue curve represents the fact that y% of the actual new links are found when considering the top-ranked x% potential links predicted as positive. The red straight line is the lift of the random classifier. In our classifier, more than 85% new links are found considering only the 10% most probable positive potential links.

the link prediction literature. Furthermore, various mobility measures have very similar performance, indicating these measures all adequately capture the similarity in mobility patterns.

In the supervised case, after systematic, yet heuristic, exploration of a

	pred. class = 0	pred. class = 1
actual class = 0	6.627	82
actual class = 1	117	228

Table 4.4: Classification result in supervised learning

large space of classification methods with different parameters, we construct a decision-tree using Quinlan's C4.5 classification over the combined network and mobility measures, with cross validation to control over-fitting, applied to the subset of potential links with common neighbors under the further constraint $AA > 0.5$ and $SCoL > 0.7$. Our tree has the following confusion matrix over an independent test set (1 = new link), implying a precision of 73.5% and a recall of 66.1%.

Both precision and recall are one order of magnitude larger than all previous figures. The lift chart (Fig. 4.7) for this classifier shows how, e.g., 86.4% of new links are found by considering only the top 10% positive cases, as ranked by the classifier in descending order of their probability of being new links. Interestingly, we find that the classifier obtained with the procedure discussed above, but using network measures only, has precision 36.2% and recall 6.1%, suggesting that the combination of topology and mobility measures is crucial to achieve high precision and recall. In other words, learning a supervised classifier based on combined network and mobility measures significantly boosts the precision and recall of predicted new links. The price to pay is that we need to focus on a niche of promising potential links with high AA and $Scos$ coefficients, concentrating on a relatively small number of candidates, yet for those we gain a very high probability of guessing the correct new links. While stressing the use of specific classification techniques, e.g., ad-hoc link prediction methods optimized for highly-unbalanced data, such as HPLP [76, 22], to achieve better precision is beyond our goals here, it is indeed an interesting open question for future research.

4.7 Related work

In this section, we review three categories of related work: studies on human mobility patterns, link prediction in social networks, and interplay between physical space and network structure.

4.7.1 Human Mobility

In the past few years, the availability of large-scale datasets, such as mobile-phone records and global-positioning-system (GPS) data, has offered researchers from various disciplines access to detailed patterns of human behavior, greatly enhancing our understanding of human mobility.

From statistical physics perspective, significant efforts have been made to understand the patterns of human mobility. Brockmann et al. [16] tested human movements using half a million dollar bills, finding that the dispersal of bills is best modeled by continuous-time random walk (CTRW) models. González et al. [47] then showed that each individual is characterized by a time-independent travel distance and a significant probability to re-visit previous locations, by using mobile phone data of $100,000$ individuals. Song et al. [98] then proposed a statistically self-consistent microscopic model for individual human mobility. Researchers have also found individuals' daily routines are highly predictable, by using principal component analysis [33] and measuring mobility entropy [99].

From data mining perspective, there have been a number of studies mining frequent patterns on human movements. General approaches are based on frequent patterns and association rules, and build predictive models for future locations. To name a few, Morzy used a modified version of Apriori [81] and Prefixspan [82] algorithms to generate association rules. Jeung et al. [57] developed a hybrid approach by combining predefined motion functions with the movement patterns of the object, extracted by a modified version of the Apriori algorithm. Yavas et al. [118] predicted user movements in a mobile computing system. Furthermore, Giannotti et al. [43, 44] developed trajectory pattern mining, and applied it to predict the next location at a certain level of accuracy by using GPS data [80].

4.7.2 Link prediction in social networks

Link prediction has attracted much interest in recent years after the seminal work of Liben-Nowell and Kleinberg [75]. It is a significant challenge in machine learning due to the inherent extreme disproportion of positive and negative cases. Existing approaches have focused on defining various proximity measures on network topology, to serve as predictors of new links in both supervised [5, 109, 76, 55] and unsupervised [75] frameworks. Most of the empirical analyses are based on co-authorship networks, and the domain-dependent features developed in certain studies (see, e.g., [5]) are tailored to this particular data set. The supervised high-performance link prediction method HPLP in [76, 55] has also been applied to a large phone dataset, using only network proximity measures.

The fundamental difference of our study from this literature is that we focus on the impact of human mobility, an intrinsic property of human behavior, on link prediction. Indeed, we have designed a broad range of mobile homophily measures and explored their power in predicting new links. Our research is orthogonal to the above line of research, in the sense that any general link prediction method can be used in combination with our mobility features, e.g., the machine learning techniques for extremely unbalanced classes.

4.7.3 Interplay between physical space and network structure

Although it is in general difficult to obtain data that contain simultaneously the geographical and network information, there have been a few interesting attempts to assess the interplay between the two. For example, there is empirical evidence [67, 63, 73] showing that the probability of forming a social tie decays with distance as a power law. Based on this fact, Backstrom, et al. [7] introduced an algorithm that predicts the location of an individual. A few recent studies focused either on small populations of volunteers, whose whereabouts and social ties were monitored at fine detail using ad-hoc smart-phone applications [34] and location-sharing services [26], or on large but specific online communities such as Flickr [25]. Although none of these data could provide a society-wide picture of either social interactions or individuals' daily routines, these studies indeed indicate that the strong correlation between physical space and network structures emerges in many diverse settings.

4.8 Conclusions and future work

Recent advances on human mobility and social networks have turned the fundamental question, to what extent do individual mobility patterns shape and impact the social network, into a crucial missing chapter in our understanding of human behavior. In this work, by following daily trajectories and communication records of 6 Million mobile phone subscribers, we address this problem for the first time, through both empirical analysis and predictive models. We find the similarity between individuals' movements, their social connectedness and the strength of interactions between them are strongly correlated with each other. Human mobility could indeed serve as a good predictor for the formation of new links, yielding comparable predictive power to traditional network-based measures. Furthermore, by combining both mobility and network measures, we show that the prediction accuracy can be significantly improved in supervised learning.

We believe our findings on the interplay of mobility patterns and social ties offer new perspectives on not only link prediction but also network dynamics. At the same time, they also have important privacy implications. Indeed, the surprising power of mobility patterns in predicting social ties indicates potential information leakage from individuals' movements to their friendship relations, posing a new challenge in privacy protection. Furthermore, we believe our results could impact a wide array of phenomena driven by human movements and social networks, from urban planning to epidemic prevention.

The results presented in this paper also open up many interesting directions for future research. The first is to search for improvement in link prediction tasks by judiciously mixing mobility and network measures. For example, we find that adding co-location measures into Adamic-Adar could yield a precision of 9.6% in unsupervised classification, overtaking any traditional measures listed in the paper. While exhaustively searching for such quantities is beyond our goals here, further work in this direction would be very important. Another interesting direction is to look at the inverse problem with respect to this work. Indeed, upon uncovering the strong correlations between mobility similarity and social connectedness and predicting links based on mobility patterns, the question thereafter is can we gain more insights about individuals' whereabouts by leveraging our knowledge of their social ties and activity patterns? In sum, the increasing availability of mobile phone data and the emergence of location-based social networking websites has the power to revolutionize our understanding of the interplay between mobility and social networks, making this field particularly fallow for new results.

Chapter 5

Conclusions

The technology, together with continuous growth of the Web and the development of Web 2.0, has inundated us with a remarkable amount of information. These data have the potential to fundamentally transform our understanding in a large variety of fields. In a much simplified view, with its scale and reach, Big Data is like a newly invented telescope that would allow us to look at the stars that we could not see before, offering us ample opportunities to do things that are otherwise impossible. With its finer resolution and higher precision, Big Data is also like a superb experimental apparatus that would allow us to examine existing theory and to articulate and exhaust new models, offering us a chance to refine and improve our current understanding and systems.

Among the questions outlined in this dissertations, many of them can now be answered because of the availability of appropriate datasets. For example, the question about the interplay between human mobility and social networks inevitably requires a large-scale dataset that could simultaneously capture the information about both aspects over a massive population. Mobile phone data serve nicely for this purpose. To probe the question of what determines information dissemination, we need to compile a comprehensive dataset that captures the spread of information as well as the context where the spreading processes woven. However, while data seems to be the key to starting these projects, the realization of these ideas was only made possible through the thinking and tools developed in statistical physics. The products are pleasant surprises with oversimplifications. Indeed, given the myriad of factors involved in the recognition of a new discovery, from the work's intrinsic value to timing, chance and the publishing venue, it is hard to imagine the level of regularity and predictability we ended up documenting in this dissertation. Similarly, while our decision in disseminating a specific piece of information likely depends on a range of factors from its content to audiences, the scale and reach of the spreading processes can be captured by a simple stochastic model. These results support our hypothesis, that macroscopic properties of a large-scale system follow generic and reproducible patterns, independent of microscopic details.

As I type in the last paragraph of this dissertation with my laptop, I cannot help but think back to the dawn of modern computing, when the smallest computer takes up a whole room by itself in a rather cumbersome way. This left me with a glimpse of thought—as if our second chapter had not taught us enough about the difficulties in predicting the future impact at an early stage—If one day future historians look back to this data revolution, what would they say about Big Data?

Bibliography

[1] D.E. Acuna, S. Allesina, and K.P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[2] L.A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[3] Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187 – 203, 2005.

[4] Eytan Adar and Lada A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, pages 207–214, 2005.

[5] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[6] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.

[7] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, pages 61–70, 2010.

[8] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

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

[10] F.M. Bass. Comments on "a new product growth for model consumer durables the bass mode". *Management science*, 50(12):1833–1840, 2004.

[11] Nancy K. Baym, Yan Bing Zhang, and Mei-Chen Lin. Social Interactions Across Media. *New Media & Society*, 6(3):299–318, 2004.

[12] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54:436, 2001.

[13] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.P. Bouchaud, and A.L. Barabási. Dynamics of ranking processes in complex systems. *Physical Review Letters*, 109(12):128701, 2012.

[14] John W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):pp. 15–53, 1949.

[15] Linda Briesemeister, Patric Lincoln, and Philip Porras. Epidemic profiles and defense of scale-free networks. *WORM 2003*, Oct. 27 2003.

[16] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[17] G. Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007.

[18] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702, 2002.

[19] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591, 2009.

[20] F. Cecconi, M. Marsili, J.R. Banavar, and A. Maritan. Diffusion, peer pressure, and tailed distributions. *Physical Review Letters*, 89(8):88102, 2002.

[21] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with googles pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[22] David A. Cieslak and Nitesh V. Chawla. Learning decision trees for unbalanced data. In *ECML/PKDD*, pages 241–256, 2008.

[23] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[24] R. Cohen and S. Havlin. *Complex networks: structure, robustness and function*. Cambridge University Press, 2010.

[25] D.J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436, 2010.

[26] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 119–128, New York, NY, USA, 2010. ACM.

[27] D.J. de Solla Price. *Little Science, Big Science... and Beyond*. Columbia University, 1963.

[28] D.J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[29] Pedro Domingos and Matthew Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[30] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks with aging of sites. *Physical Review E*, 62(2):1842, 2000.

[31] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford, 2003.

[32] D.T. Durack. The weight of medical knowledge. *New England Journal of Medicine*, 298(14):773–775, 1978.

[33] N. Eagle and A.S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

[34] N. Eagle, A.S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274, 2009.

[35] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[36] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks. *Phys. Rev. E*, 66(3):035103, Sep 2002.

[37] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004.

[38] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[39] J.A. Evans and J.G. Foster. Metaknowledge. *Science*, 331(6018):721–725, 2011.

[40] A. Fersht. The most influential journals: Impact factor and eigenfactor. *Proceedings of the National Academy of Sciences*, 106(17):6883–6884, 2009.

[41] I. Fuyuno and D. Cyranoski. Cash for papers: Putting a premium on publication. *Nature*, 441(7095):792–792, 2006.

[42] E. Garfield. The history and meaning of the journal impact factor. *JAMA: the journal of the American Medical Association*, 295(1):90–93, 2006.

[43] Fosca Giannotti, Mirco Nanni, and Dino Pedreschi. Efficient mining of temporally annotated sequences. In *SDM*, 2006.

[44] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339, 2007.

[45] B. Golub and M.O. Jackson. Using selection bias to explain the observed structure of Internet diffusions. *Proceedings of the National Academy of Sciences*, 107(24):10833, 2010.

[46] Benjamin Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583, 1825.

[47] M.C. González, C.A. Hidalgo, and A.L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[48] R.V. Gould and R.M. Fernandez. Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, 19(1989):89–126, 1989.

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

[50] Daniel Gruhl, Ramanathan V. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.

[51] T.E. Harris. *The theory of branching processes*. Dover Pubns, 2002.

[52] G. Herdan. The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, 45(1-2):222–228, 1958.

[53] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599–653, 2000.

[54] J.E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569, 2005.

[55] Z. Huang, X. Li, and H. Chen. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 141–142. ACM, 2005.

[56] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.*, 103(3):038702, Jul 2009.

[57] Hoyoung Jeung, Qing Liu, Heng Tao Shen, and Xiaofang Zhou. A hybrid prediction model for moving objects. In *ICDE*, pages 70–79, 2008.

[58] B.F. Jones, S. Wuchty, and B. Uzzi. Multi-university research teams: shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[59] T. Karagiannis and M. Vojnovic. Behavioral profiles for advanced email features. In *Proc. of WWW*, pages 711–720, 2009.

[60] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[61] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[62] Gueorgi Kossinets and Duncan J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90, 2006.

[63] G. Krings, F. Calabrese, C. Ratti, and V.D. Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:L07003, 2009.

[64] T.S. Kuhn. *The structure of scientific revolutions.* University of Chicago press, 1996.

[65] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *World Wide Web*, pages 159–178, 2005.

[66] Ravi Kumar, Mohammad Mahdian, and Mary McGlohon. Dynamics of conversations. In *KDD*, pages 553–562, 2010.

[67] R. Lambiotte, V.D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[68] S. Lehmann, A.D. Jackson, and B.E. Lautrup. Measures for measures. *Nature*, 444(7122):1003–1004, 2006.

[69] S. Lehmann, B. Lautrup, and AD Jackson. Citation networks in high energy physics. *Physical Review E*, 68(2):026113, 2003.

[70] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. In *ACM/EC*, pages 228–237, 2006.

[71] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD*, pages 462–470, 2008.

[72] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie S. Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.

[73] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences*, 102(33):11623, 2005.

[74] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[75] David Liben-Nowell and Jon M. Kleinberg. The link prediction problem for social networks. In *CIKM*, pages 556–559, 2003.

[76] Ryan Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. New perspectives and methods in link prediction. In *KDD*, pages 243–252, 2010.

[77] V. Mahajan, E. Muller, and F.M. Bass. New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing*, pages 1–26, 1990.

[78] A. Mazloumian, Y.H. Eom, D. Helbing, S. Lozano, and S. Fortunato. How citation boosts promote scientific paradigm shifts and Nobel prizes. *PloS one*, 6(5):e18975, 2011.

[79] M. Medo, G. Cimini, and S. Gualdi. Temporal effects in the growth of networks. *Physical Review Letters*, 107(23):238701, 2011.

[80] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *KDD*, pages 637–646, 2009.

[81] Mikolaj Morzy. Prediction of moving object location based on frequent trajectories. In *ISCIS*, pages 583–592, 2006.

[82] Mikolaj Morzy. Mining frequent trajectories of moving objects for location prediction. In *MLDM*, pages 667–680, 2007.

[83] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

[84] R. Plamondon. A kinematic theory of rapid human movements. *Biological cybernetics*, 72(4):295–307, 1995.

[85] FW Preston. Pseudo-lognormal distributions. *Ecology*, pages 355–364, 1981.

[86] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[87] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80(5):056103, 2009.

[88] F. Radicchi, S. Fortunato, and A. Vespignani. Citation networks. *Models of Science Dynamics*, pages 233–257, 2012.

[89] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

[90] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58:49, 2005.

[91] M.T. Rivera, S.B. Soderstrom, and B. Uzzi. Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology*, 36:91–115, 2010.

[92] Philip E. Sartwell. The distribution of incubation periods of infectious disease. *American Journal of Epidemiology*, 51(3):310–318, 1950.

[93] P.O. Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(7079):498, 1997.

[94] H.A. Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.

[95] Hilary Smith, Yvonne Rogers, and Mark Brady. Managing one's social network: Does age make a difference. In *In: Proc. Interact 2003, IOS*, pages 551–558. Press, 2003.

[96] Marc A. Smith, Jeff Ubois, and Benjamin M. Gross. Forward thinking. In *CEAS*, 2005.

[97] R.V. Solé, R. Ferrer-Cancho, J.M. Montoya, and S. Valverde. Selection, tinkering, and emergence in complex networks. *Complexity*, 8(1):20–33, 2003.

[98] C. Song, T. Koren, P. Wang, and A.L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 2010.

[99] C. Song, Z. Qu, N. Blumm, and A.L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018, 2010.

[100] David Strang and Sarah A. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24(1):265–290, 1998.

[101] W.H. Suh, K.S. Suslick, G.D. Stucky, and Y.H. Suh. Nanotechnology, nanotoxicology, and neuroscience. *Progress in neurobiology*, 87(3):133–170, 2009.

[102] Hanghang Tong, B. Aditya Prakash, Charalampos Tsourakakis, Tina Eliassi-Rad, Christos Faloutsos, and Duen Horng "Polo" Chau. On the vulnerability of large graphs. In *ICDM*, 2010.

[103] R. Ulrich and J. Miller. Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, 1993.

[104] Thomas W. Valente. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2:163–164, 1996.

[105] G.J.P. VanBreukelen. Parallel information processing models compatible with lognormally distributed response times. *Journal of Mathematical Psychology*, 39(4):396–399, 1995.

[106] A. Vázquez, J.G. Oliveira, Z. Dezsö, K.I. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):36127, 2006.

[107] T. Vicsek and A. Zafeiris. Collective motion. *Physics Reports*, 517:71–140, 2012.

[108] D. Walker, H. Xie, K.K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.

[109] Chao Wang, Venu Satuluri, and Srinivasan Parthasarathy. Local probabilistic models for link prediction. In *ICDM*, pages 322–331, 2007.

[110] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. *SRDS*, 2003.

[111] H.W. Watson and F. Galton. On the probability of the extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland*, pages 138–144, 1875.

[112] Barry Wellman. *The Internet in Everyday Life (The Information Age)*. Blackwell Publishers, December 2002.

[113] Zhen Wen and Ching-Yung Lin. On the quality of inferring interests from social neighbors. In *KDD*, pages 373–382, 2010.

[114] CB Williams. A note on the statistical analysls of sentence-length as a criterion of literary style. *Biometrika*, 31(3-4):356–361, 1940.

[115] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327 – 335, 2004.

[116] L. Wu, C.-Y. Lin, S. Aral, and E. Brynjolfsson. Value of social network – a large-scale analysis on network structure impact to financial revenue of information technology consultants. In *The Winter Conference on Business Intelligence*, 2009.

[117] S. Wuchty, B.F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[118] Gökhan Yavas, Dimitrios Katsaros, Özgür Ulusoy, and Yannis Manolopoulos. A data mining approach for location prediction in mobile environments. *Data Knowl. Eng.*, 54(2):121–146, 2005.