Computational Social Sciences

Bruno Gonçalves
Nicola Perra *Editors*

Social Phenomena

From Data Analysis to Models

Springer

Computational Social Sciences

A series of authored and edited monographs that utilize quantitative and computational methods to model, analyze, and interpret large-scale social phenomena. Titles within the series contain methods and practices that test and develop theories of complex social processes through bottom-up modeling of social interactions. Of particular interest is the study of the co-evolution of modern communication technology and social behavior and norms, in connection with emerging issues such as trust, risk, security, and privacy in novel socio-technical environments.

Computational Social Sciences is explicitly transdisciplinary: quantitative methods from fields such as dynamical systems, artificial intelligence, network theory, agent-based modeling, and statistical mechanics are invoked and combined with state-of-the-art mining and analysis of large data sets to help us understand social agents, their interactions on and offline, and the effect of these interactions at the macro level. Topics include, but are not limited to social networks and media, dynamics of opinions, cultures and conflicts, socio-technical co-evolution, and social psychology. Computational Social Sciences will also publish monographs and selected edited contributions from specialized conferences and workshops specifically aimed at communicating new findings to a large transdisciplinary audience. A fundamental goal of the series is to provide a single forum within which commonalities and differences in the workings of this field may be discerned, hence leading to deeper insight and understanding.

More information about this series at http://www.springer.com/series/11784

Bruno Gonçalves • Nicola Perra
Editors

Social Phenomena

From Data Analysis to Models

Springer

Editors

Bruno Gonçalves
Centre de Physique Théorique
Aix-Marseille Université
Campus de Luminy, Case 907
Marseille, France

Nicola Perra
MoBS Lab
Northeastern University
Boston, MA, USA

Computational Social Sciences
ISBN 978-3-319-14010-0 ISBN 978-3-319-14011-7 (eBook)
DOI 10.1007/978-3-319-14011-7

Library of Congress Control Number: 2015939174

Springer Cham Heidelberg New York Dordrecht London

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

To Duygu Balcan,
Forever in our memory

Foreword

An unprepared reader could be easily fooled by this book's title. While it hints at being a classic social science book, readers will quickly discover that the 12 chapters are in many cases more similar to Computer Science or Physics articles, and the authors of the chapters are not just social scientists, but rather interdisciplinary teams with a strong representation of physicists, computer scientists, and applied mathematicians. Indeed, this book is about Social Systems and Social Phenomena, but the approach followed is the one that has emerged in the last 10 years at the convergence of complex systems, networks, big data, and social sciences. This emerging field of research has been given the name of "computational social science" in a farseeing paper by Lazer and coworkers in 2009.[1]

I have worked in the area of complex systems for about two decades, and I can witness the huge fascination that social science has always had on the community of complex systems researchers. On the other hand, social phenomena can be seen in many cases as self-organizing systems with many degrees of freedom that develop collective behavior, and exhibit non-trivial emergent phenomena. All these features are the quintessential summary of a complex system, and it is no wonder that complex systems scientists have used their mathematical and computational tools to approach social science questions such as the emergence of consensus, social opinion dynamics, conflicts, and cooperation. However, although they provided powerful conceptual metaphors, these approaches often have suffered from being oversimplified and not grounded on actual data. Very often social scientists could not help but critique those attempts by saying that complex systems scientists were looking at society as an array of ordered magnetic spins, a view that was too simplistic by many accounts.

In the last decade, however, the research landscape has been redefined by the big data revolution. It is not just that an increasing number of socio-economic data have been made readily available by the progressive digitalization of our world. The advent of mobile and pervasive technologies, the Web, and the myriad of digital

[1] D. Lazer et al., Computational Social Science, Science 323, pp. 721–723 (2009).

social networks have triggered an unprecedented avalanche of social behavioral data ranging from human mobility and social interaction to the very real-time monitoring of conversation topics, memes, and information consumption. Data on mobile device usage allow the measure of people mobility, even in remote regions of the world. Phone call records show us patterns of social interactions that can be integrated with a multitude of social networks and microblogging data. The spread of memes and information over these networks can be monitored in real time on a planetary scale. Finally, new pervasive technologies are capable of gathering data down to the level of face-to-face interactions for thousands of people at once.

By rushing into this "Data El Dorado" scientists have thus been able to understand the complex networks underlying social interactions and to analyze the dynamics of social phenomena. Instead of a simple array of individuals, models are now informed by the intricate and large-scale connectivity patterns encapsulated in the theory of complex networks. Size does matter, and having a high quality dataset for thousands or millions of individuals has triggered the search for statistical patterns, ordering principles, and generative mechanisms that could be used to achieve greater realisms in the modeling of social phenomena. Nowadays computational social science has definitely moved from toy/conceptual models to data-driven approaches that can be validated quantitatively. From the spread of emerging infectious diseases and crime rates to road traffic and crowd movement, computational approaches are now achieving quantitative success, both for scenario analysis and in real-time forecasts.

The research activity emerging from ever-increasing data availability, novel computational tools and methods, and the rich conceptual framework provided by complex networks and systems science provides an exciting understanding of a variety of socio-technical systems. It is also promising to be truly disruptive in the way we act on and manage those systems and in the development of new interactive and adaptive information and communication technologies. The research landscape in this area is, however, fast paced and scattered across different areas. The many scientific contributions of recent years are dispersed across different disciplinary journals and conference proceedings. This book is one of the first editorial attempts at providing a coherent presentation of recent areas of investigation that range from human mobility to online interactions and the financial market. The book editors, Bruno Gonçalves and Nicola Perra, have been able to assemble a fantastic number of contributing authors who are among the scientific leaders at the forefront of the research activity presented here. Each chapter provides a clear and rigorous introduction to the incredible advances witnessed in this research field in the last 10 years. The final result is a book that delivers, for an entire research field, a coherent presentation of the workflow that we could simply summarize as "from data to knowledge". This book will certainly be an important contribution to the field—one from which many more advances of our future understanding of socio-technical systems will be built.

Boston, MA, USA
March 2015

Alessandro Vespignani

Preface

The story of this book is one that spans the better part of a decade. Even though we are both physicists by training, our interest in the study of social behavior goes back many years. It was thus natural that, in 2009, when we found ourselves working in the same group in Bloomington, Indiana, we would work together in the data-driven study of Human Behavior. These were the early days of the big data revolution when Twitter was practically unknown in the research community and new datasets were appearing almost every day.

A few years before, Barabási had drawn the first wave of attention to human behavior with a series of papers on what he called "Human Dynamics" that focused on the study of the impact that a broad-tailed inter-event time distribution can have on some dynamical process. This ramp-up in attention culminated in 2009 when Science published a position paper by some of the leading physicists, economists, and social scientists: A call to arms to combine large-scale datasets with new computational and analytical tools under the umbrella of "Computational Social Science".

Our collaboration started with a work on the empirical validation of Dunbar's number using Twitter data and continued on to cover the effect that behavioral changes can have on epidemic spreading and on how broadly distributed activity patterns influence the structure of social networks.

In 2012, after we had both left Bloomington behind, we jointly organized the first edition of the Computational Approaches to Social Modeling (ChASM) workshop collocated with the International Conference on Computational Science (ICCS) with the explicit goal of bridging the chasm between the social and physical sciences and bringing together practitioners and theorists from Computational, Physical, and Social Sciences to exchange ideas and techniques useful to the study of human behavior. In 2014, ChASM celebrated its third edition as a workshop of the ACM Web Science conference back where it all started, in Bloomington, Indiana. Preparations for the next edition are ongoing.

In parallel with ChASM we also organized several editions of a "Special Topic" session in the American Physical Society annual March Meeting. Here the goal was to help diffuse the idea of studying social behavior to an audience of "traditional" physicists. It was during the runup to one of these sessions that we were contacted by Chris Coughlin, a Physics & Complex Systems Editor at Springer, with the invitation to organize this edited volume.

With this book the goal is to showcase what some of the leading researchers, from fields of study as different as Social, Computational, and Physical Science, are doing in this important subject. We tried to give the authors as much freedom as possible while still preserving the unified view and touching on what we consider to be the most interesting developments in this field.

For this opportunity we are truly thankful to the Springer editors and, in particular, to all the authors who have agreed to participate in this work.

Marseille, France Bruno Gonçalves
Boston, MA, USA Nicola Perra
March 2015

Contents

Part II Social Behavior Under Stress

Chapter 1
Introduction

Bruno Gonçalves and Nicola Perra

When it was first conceived by Tim Berners-Lee in 1990, the World Wide Web (WWW) [1] was intended as a way for the publication and sharing of information among researchers at CERN. The original WWW browser allowed users to both browse and edit pages but the full vision of a network where anyone could be a producer and publisher of content didn't come to fruition until almost a decade later. Before the DotCom boom and the arrival of wikis, blogs, etc. was possible, a whole global infrastructure had first to be built. Routers and service providers to route traffic, Web browsers to allow users to access pages provided by Web servers, caching and billing protocols to improve performance and allow for the development of commercial enterprises, among many others.

A direct consequence of these advances was the inadvertent generation of unprecedented quantities of information documenting what pages are accessed by whom, who buys what product, who emails whom, and about every other activity occurring online. Originally collected for logging, billing, and debugging purposes, it would not be long before this kind of data attracted the attention of companies and researchers as a means to better understand their users and research subjects. Neither the potential nor the challenges that posed by this untapped wealth of information went unnoticed for long.

The Big Data revolution [2] that followed, and is still ongoing, is poised to change not only the way online systems work but also how we study Human Behavior on a large scale. Indeed, hiding within the mountain of data is not only information on

B. Gonçalves (✉)
Aix Marseille Université, Université de Toulon, CNRS, CPT, UMR 7332,
13288 Marseille, France
e-mail: bgoncalves@gmail.com

N. Perra
Northeastern University, Boston, MA, USA
e-mail: n.perra@neu.edu

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_1

how people use the system but also on how individuals communicate and interact. An email server not only records when a specific email was sent, but also who sent it and who it was addressed to, if there was a reply, etc. Search engines store all the queries submitted associating them with users' information as IP, location, gender, age, and other personal information, when available. A cell phone company must keep track not only of who made the call and who received it, but also the date and time it was made, to which cell tower those two users were connected to and how long the call lasted. Wikipedia records which account or IP address edited which page and what changes were made.

Using this so-called metadata, much progress was done in the study of social interactions, but it wouldn't be until the recent rise of full fledged online social systems that are specifically designed to facilitate social interaction and discussions (like Facebook, Twitter, or Google+), large scale collaboration (such as GitHub, Wikipedia or OpenStreetMaps), online gaming worlds (of which World of Warcraft and Eve Online are perhaps the most famous examples) or even dating (Match.com, OkCupid, etc.) that we would be able to start having a more complete view of the functioning of society.

Furthermore, the miniaturization of sensors and electronic devices has made increasingly easier the realization of tools to record duration, frequency, and other features also of offline contacts. Such devices, based on Bluetooth, WiFi, and RFID technologies, allow for the first time probing, at scale, face-to-face interactions in many different settings ranging from schools and hospitals to museums and conferences.

With the advent of these systems, it became possible for the first time to observe many aspects of social behavior that had never been amenable to large scale analysis. Each different system was created with a specific goal in mind and the choices made during the design process limit the kind of phenomena that can be analyzed. Research in this area takes advantage of a veritable bounty of different datasets, but with particular emphasis on Online Queries, Twitter, Cell phones, Bibliographic, and Offline Interactions databases, due to their intrinsic richness. Below we highlight some of the characteristics, advantages, and limitations of these types of data sources.

1.1 Online Queries

The short history of the WWW is signed by the release of Google in 1998. The company revolutionized search engines making simple and effective users' exploration for information. Indeed, with the exponential growth of webpages retrieving content was becoming increasingly difficult. The first search engines used natural language processing techniques to assess the relevance of webpages to specific queries. Google's founders realized that considering just the properties of single pages neglecting the structure of the network where they were embedded was not the optimal strategy. Starting from this observation they introduced the

PageRank [3]. The algorithm measures the relevance/importance of a webpage considering the relevance/importance of the webpages linked to it. The PageRank constituted a real paradigm shift in information retrieval, and clearly showed the importance of going beyond the local properties of nodes (webpages) when dealing with complex networks. Thanks to Google, and many advances that followed, we are now able to browses about 60 trillion of pages within few clicks.

Current estimates consider that about half of the population of the planet is active online [4]. Although the coverage is still far from being homogenous across the globe, users come from many different backgrounds, languages, and age groups resulting in a wide range of interests behind online activities. Among these, the use of search engines is one of the most common. Indeed despite the final goal, people accessing the WWW, are likely to start their sessions with a query to Google, Yahoo, or Bing. Online searches are expression of interests for specific products, events, or topics. While implications of this simple observation run deep in within our digital society, here we focus just on those associated with the study of social phenomena. In particular, increases in the volume of queries associated with specific keywords are driven by external (exogenous) or internal (endogenous) events. Examples are the spreading of infectious diseases, elections, social protests, online movements, and trends in financial markets. Online queries can be considered as proxies for such events, and their study allows near real time analyses at an unprecedented scale/resolution.

Online searches data come also with several important limitations. The data are proprietary and cannot be shared for privacy and financial concerns. Researches have access just to aggregated indicators subject to several constraints as the lack of any information about the users, or the ability to compare the relative interest of large number of keywords. The data is typically available just for very popular queries. This might limit the possibility of monitoring the unfolding of new trends or topics. Finally, search engines are dynamic entities. They are constantly changed to achieve better performances and to be more user friendly. Some features, as for example the auto-completion, modify the way we access or explore information. These modifications and their effects in our behaviors should be considered when studying societal phenomena through the lens of search engines. Furthermore, comparing trends in different period of time might introduce strong biases. Unfortunately, the lack of transparence in the data collection and post processing makes these crucial steps often impossible.

1.2 Twitter

Twitter is perhaps the most widely studied online social network. Twitter was designed to be a broadcast system so that one person could easily send a message to thousands or even millions of others. Given this asymmetry between content producers and consumers, it makes sense to have directional connections with individuals electing to follow someone who may or may not follow them back and that any content produced is considered public by default. Anyone who follows, say, Alice, will automatically receive all the content produced by Alice. By following Alice, Bob is explicitly declaring an interest in what Alice says and the more followers Alice has the more famous she is, providing a lens through which to observe the evolution of popularity and the rise of celebrities.

Twitter was originally conceived to be used through SMS, which led to an intrinsic limitation on the amount of text that can be included in one single tweet. SMS are limited to 160 characters and Twitter reserved the first 20 characters for the id of the user, resulting in the now famous 140 character limit. Users immediately started to try to find ways to work around this limitation using abbreviations and hashtags. Hashtags then took on a life of their own and became, perhaps, the defining feature of Twitter and a fixture of social systems, eventually being adopted by Facebook, Google+, and many others. Hashtags mark the topic under discussion and can be freely adopted by any user. Studying how they rise and fall in popularity allows us to analyze what are the broad topics under discussion at a given point in time.

As the system grew and users became more engaged with it, some mechanism to forward information a user received from the individuals he followed to his followers became necessary. Informally, users adopted a convention to quote one another while giving full credit to the original poster, a process that became known as ReTweet. Through the analysis of retweets we are able to observe how information spreads through social connections.

Despite the original formulation as a broadcasting system, the social component is becoming increasingly more important and conventions for mentioning and replying to other users have also been adopted. As a result one can observe how actual conversations occur between two or more Twitter users.

The most recent development has occurred with the widespread adoption of geocoding. Twitter has always allowed users to declare in their profile where they lived. As GPS enabled smart phones reached the market some Twitter clients started updating the users location field with the GPS coordinates provided by the cellphone whenever they tweeted. Twitter eventually modified its infrastructure to allow GPS information to be associated with individual tweets instead of just the users, allowing us to track where the user is located whenever he tweeted from a smartphone. This provides yet another layer to the phenomena that can be studied through Twitter. A conceptual illustration of the different types of interactions occurring in Twitter can be seen in Fig. 1.1.

As with any new tool, Twitter has, along with its many possibilities, also some severe limitations. Twitter users are tendentially younger and wealthier than the general population [5]. The use of GPS enabled smartphones is biased towards richer populations who tend to travel more. It also remains to be conclusively demonstrated that we interact with others online similarly to how we do offline, but the wealth of results obtained using this type of datasets and that corroborate or agree with results obtained with more traditional social science approaches points in that direction. All of these limitations pose challenges that must be addressed.

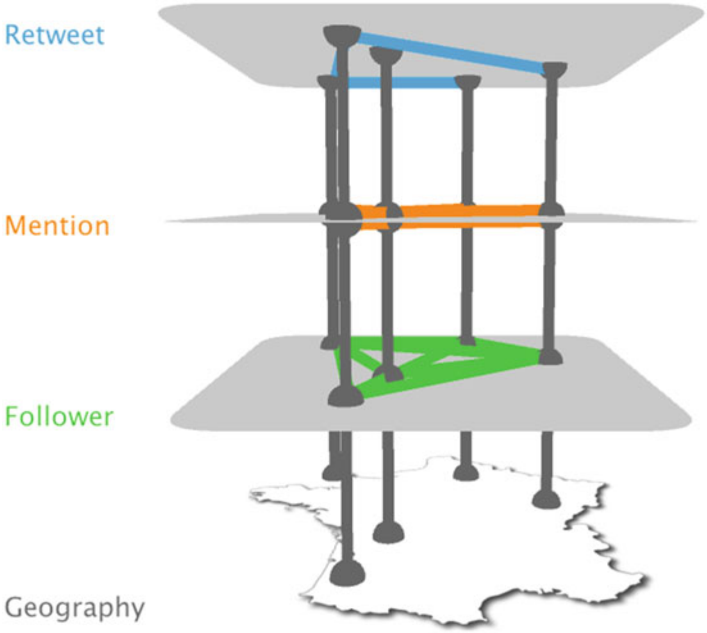


Fig. 1.1 The different layers of Twitter. As example we consider users located in some cities of France

1.3 Cell Phones

While Twitter use is limited to a specific subset of the general population, cell phones have quickly gone from a niche technology reserved to the rich and famous to the default mode of communication for the vast majority of the world population with market penetration, in some countries, surpassing 100 % or more than one cell phone number per person.

Cell phones are becoming increasingly sophisticated and sensor rich significantly increasing the range of large scale population measurements that are possible. While some studies rely on the use of custom made applications aimed at specific smartphone models, the most successful efforts have been done in collaboration with cell phone operators using only Call Detail Records. CDRs are collected by mobile carriers for billing and legal purposes and include information on any action that the user performs on their device that implies the use of the network (phone calls, SMS, MMS, or internet access). Call duration, origin and destination are recorded along side the date and time and the physical location of the user can be inferred by triangulating from the position of the cell phone towers that are within range of the device.

Cell phone data of this kind provides the widest possible view on the social interactions of an entire population so it is much less sensitive to the limitations

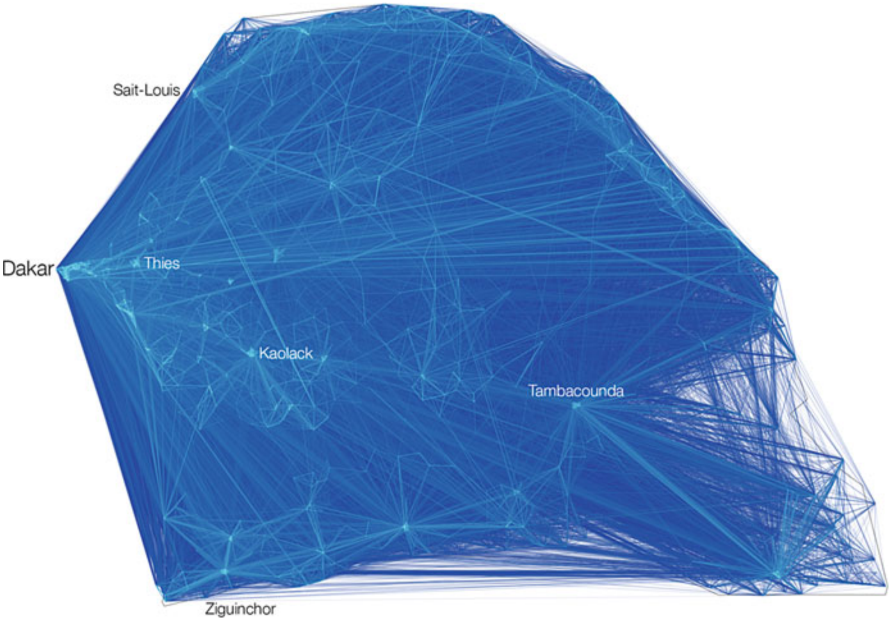


Fig. 1.2 Phone call network during one day in Senegal

mentioned above for Twitter. The biggest limitation to their use is one of privacy. While in the case of Twitter all activity is considered public, users are much more privacy conscious about their cell phone activity and cell phone service providers are afraid of the potential ramifications of privacy breaches. This has severely limited the use of this wealth of behavioral data to researchers inside or in close collaboration with mobile operators. Notably, Orange recognizes the potential of cell phone data and actively tries to overcome the privacy limitations with their Data for Development (D4D) challenges.[1] For each challenge they release anonymized call and mobility datasets for their entire user base in one developing country (Ivory Coast in 2012 and Senegal in 2014) to researchers that submit a proposal on how to use this data to help foster the development of that country. The 2014 edition is still ongoing but the 2012 one resulted in several dozen original articles being published with various approaches on how to use this data. In Fig. 1.2 we plot the phone call network for a single day in Senegal based on the D4D dataset. Each node is a cell phone tower and the color of edges between towers indicates the strength of the connection with lighter colors representing stronger connections. It is easy to see how the density of cell phone towers follows the population distribution making major cities such as Dakar in the central West Coast clearly identifiable.

[1] http://www.d4d.orange.com.

Another important limitation of cell phone datasets is that, although it contains information about the timing and frequency of communication, nothing is known about the content. This makes it impossible to use this kind of data to observe which topics are popular at the societal level or to directly track information diffusion. Also, location information is limited by the distribution of cell phone towers that closely follows the population distribution, with high concentrations and precision in urban areas and much lower levels of service in more rural areas.

1.4 Bibliographic Databases

As a society relies on efficient means of communication, such as cell phones and transportation, to function and prosper, Science relies on the publication of peer-reviewed manuscripts as a way of diffusing its latest findings and foster the debate about which directions to follow. Each manuscript, in addition to its scientific content, includes also information about who the authors are, which institution they work for and what were their sources of inspiration in the form of a list of references (Fig. 1.3).

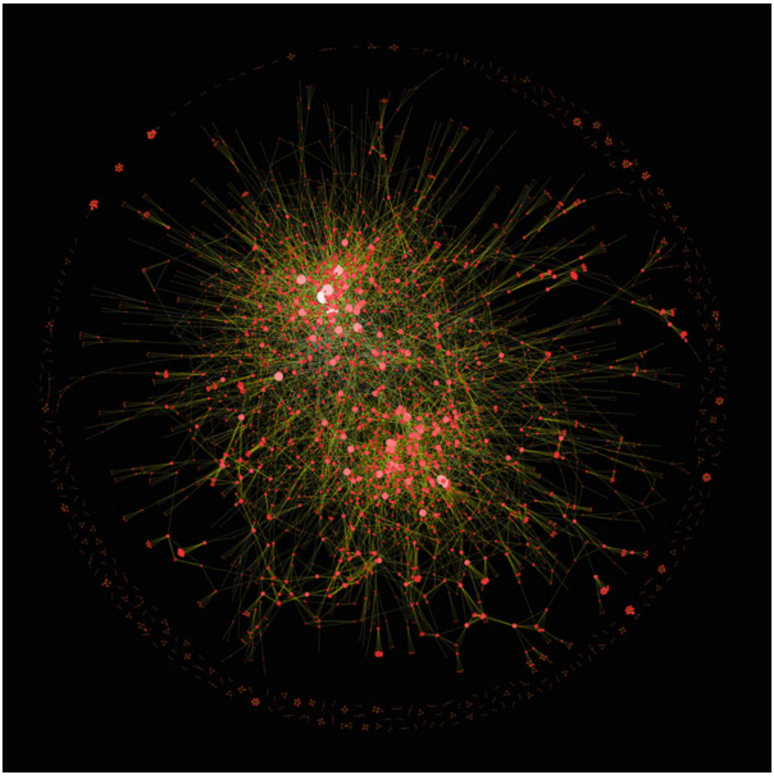


Fig. 1.3 PRL collaboration network for the 1974–2004 period, from [6]

With the evolution of science in the last centuries, the number of scientific journals and conferences has only increased and so has the number of scientific production in the form of papers. In order to allow individual scientists to navigate the ensuing sea of information large scale databases were collated containing information on several million manuscripts, who the authors were and who cited whom. These databases provide an unprecedented view on the scientific enterprise and the longest running dataset on large scale collaborative work towards a common goal. The ebb and flow of scientific collaborations has shown that as fields evolve and become more complex, the number of authors and references per manuscript has steadily increased and the citation network has documented how ideas generated in one field or area of expertise eventually reach out to influence researchers in completely separate fields.

Recent works have also focused on using this type of information as a basis to try to develop quantitative and fair measures of scientific productivity, influence and merit. Are the most important scientists those who generate more papers or those who receive more citations. How to account for varying numbers of researchers in different fields? How to recognize papers that will prove to be influential ahead of time?

While such databases provide an extremely detailed view and often full historical coverage of a given field or family of journals they tend to be limited by the fact that they cover only a limited subset of the full range of scientific production. For example, Thompson Reuters Web of Science[2] offers perhaps the most complete coverage of scientific journals, but has extremely limited coverage of the type of peer reviewed conference proceedings that are common in Computer Science and related fields. On the other hand, Scopus,[3] the largest bibliometric database, has a much more complete coverage of conferences but a more limited coverage of journals. Finally, Google Scholar[4] the most herculean effort to offer full coverage suffers from the fact that it is limited to web accessible sources resulting in a limited coverage of older, historical, issues.

1.5 Offline Interactions

As clear from the previous sections the digital revolution is providing a wealth of datasets to probe and explore human dynamics and social phenomena. Some more than others, i.e. phone calls and geolocalized mention networks on Twitter,

[2] http://thomsonreuters.com/thomson-reuters-web-of-science/.

[3] http://www.scopus.com/.

[4] http://scholar.google.com.

can be used also as proxies of actual, offline, interactions. The basic assumption behind these approaches is that phone calls, or discussion on Twitter are different expression of an underlying network of social ties. However, some features of each interaction type can be driven by the particular design of the medium used. This observation is particularly important when studying dynamical processes unfolding on networks structures as the spreading of infectious diseases. Indeed, viruses can spread just through the direct physical contacts between susceptible and infected individuals.

The collection of real data of human contacts has been traditionally done through questionnaires or surveys. While this collection method provides a rich set of information, it suffers from well-known limitations. Examples are excessive costs, difficulties in finding participants, and several biases associated with self-reporting procedures. Gathering data about face-to-face interactions using more direct and unobtrusive approaches become then of particular importance also to measure independently the quality of indirect sources as Twitter, phone calls, and surveys.

The development of tools able to accomplish this goal has been hampered by technological and other practical issues for many years. Interestingly, the digital revolution has lifted such limitations, making increasingly easier the cost-effective production of very small and portable sensor able to measure proximity. Indeed, we have now the possibility of creating inexpensive wearable tools, based on a range of technologies as Bluetooth, WiFi, and RFID, able to monitor and record face-to-face as well as other interactions. Remarkably, such sensors succeed in defining and recording objectively close contacts, accessing also to short encounters. However, there are still a set of important limitations. The data collection is typically done in closed and controlled settings. This might introduce biases in individuals behavior. Furthermore, due to experimental challenges the group of individuals under study is still relative small.

1.6 Structure of the Book

The remainder of the book is divided into two parts. In Part I, "Human Behavior Under Normal Conditions," we focus on characterizing the daily behavior of individuals going about their daily lives. In Part II, "Social Behavior Under Stress," we analyze instead how individuals act under extraordinary circumstances such as War, Epidemics, or Crime. Our aim is that by considering both sides of the same coin we are able to summarize current state-of-the-art research and start taking the first steps towards a more general understanding of Human Behavior.

We start in Chap. 2 by studying large scale cell phone datasets to analyze human mobility. Mobility is a fundamental aspect of our daily lives. We travel on vacation, commute from home to work, go visit friends and relatives in nearby neighborhoods or distant cities or even in order to participate in social and sport events. Understanding how we move over the course of a day is fundamental to help improve the infrastructure and organization of our cities. The wide availability of mobile devices facilitates the observation, in real time, of where people are, where they are going, and where the mobility bottlenecks are. An understanding of which is fundamental if we are to optimize our transportation systems and improve the efficiency of our cities. Furthermore, the quantitate characterization of human mobility at different scales is instrumental to model processes driven by our movements as, for example, the spreading of infectious diseases. Surprisingly, the authors find that the overwhelming majority of individuals is both predictable and unique. Most of our time is spent at home, at work or in between, which makes it easy to predict where a given person will be at a point in time, but the exact location of these places and how we reach them is fundamentally unique and personal.

In Chap. 3 we move on to the study of Human face-to-face interactions. Here the authors use specially crafted sensors to measure real world face-to-face interactions. Their devices detect when one individual is facing another in close proximity for an extended period. With this rich dataset they are able to characterize in detail face-to-face interactions and present a new methodology to identify mesoscopic structures of the ensuing patterns. As close proximity is a fundamental requirement for the spreading of infectious diseases the authors also consider how the empirical patterns observed impact the spreading of diseases and lay the groundwork for a research agenda in this fascinating and practically unexplored area.

After covering mobility and face-to-face interactions we are in a perfect position to move on to the study of epidemics, one of the most prominent driving forces of human history. In Chap. 4 the authors present a review summary of epidemic modeling approaches. Starting from the simplest of mathematical models, the entire formalism necessary to understand state-of-the-art epidemic models is developed with a strong focus on recent advancements. In particular, two realistic data driven models are analyzed in detail, GLEaM and FLuTe, that while starting from completely different levels of approximation have gradually converged towards being able to tackle common goal of large scale forecasting of epidemics. The authors finalize with an overview of digital epidemiology, an emerging branch of modeling approaches that stems from the big data revolution.

Chapter 5 continues the analyses of the possibilities of big data by considering applications to the study of financial markets. Investing decisions are made individually but as investors research online leave traces containing valuable information about future stock movements of a given company. The authors demonstrate that peaks in online search activity predate large market movements, giving credence to this idea and demonstrating the feasibility of using online activity to study and predict offline behaviors.

In Chap. 6 we continue the analysis of the online world by studying the mutual influence between information flows and social connections. Following a review of the literature on online interactions a longitudinal case study of Yahoo! Meme is presented. The authors analyze the complete history of the system studying how individual user behavior impacts the structure of the network and vice versa. Interestingly, the authors found that combining the dynamics occurring on the network with the dynamics of the network is crucial to reproduce empirical observations.

Chapter 7 looks in more depth into the question of individual user behavior and the factors that motivate it. What factors determine online collaboration and what leads perfect strangers to dedicate large fractions of their free time to help each other by contributing content to online communities? Why are such behaviors more common online than in our everyday lives? The authors tackle these questions by a combination of empirical studies and modeling efforts in order to identify the role played by the various factors.

We close Part I of the book by continuing, in Chap. 8, the discussion of human collaboration through the study of bibliographic databases. The authors provide a review of the most relevant recent results in field of bibliometric with a special focus on the statistical description of citation distributions and citation dynamics. Importantly, the authors discuss methods to rescale citation distributions across fields allowing the observation and characterization of their universal features. Furthermore, a framework to predict the future impact of a publication based on its behavior in the first years of publication is presented.

In Part II we move on from the study of Human Behavior under normal conditions and consider instead how we behave under extraordinary circumstances. We initiate the discussion in Chap. 9 where we modify the epidemic models introduced in Chap. 4 to take into account behavioral changes induced by risk perception during the course of an epidemic. A game theoretical approach is used to show how individual defensive behaviors can actually have a negative impact over the course of epidemic leading to a re-emergence of the disease.

In Chap. 10 our analyses move on from the consideration of the consequences or motivations of individual behavior to focusing instead on detecting specific patterns of behavior. The authors apply techniques from social network analysis to the study of cell phone networks with the aim of uncovering criminal behavior or illicit activities and introduce *LogViewer*, a computational framework developed with the goal of helping criminal investigators in the field perform these analyses without the added burden of having to be social network analysis experts. Several use-cases based on real-world criminal investigations are also discussed.

Chapter 11 takes one step further and considers global terrorism. A wide set of data sources covering the complete range of geographical scales is used to analyze the common patterns underlying asymmetric conflicts where terrorists, rebels, revolutionaries or freedom fighters are drawn to fight a larger and more conventional force. Despite all the differences between the conflicts considered several common patterns emerge pointing towards universal human behaviors in asymmetrical struggles. A generative model is proposed that is able to reproduce the patterns observed using a minimal set of physically motivated parameters.

Finally, in Chap. 12 we move definitely away from individual behavior and consider instead crowd behavior. The author presents a perspective on the literature on the use of social media like Twitter to analyze crowd behavior. Different aspects are considered, with a special emphasis on practical applications towards event detection and prediction. Implications and challenges are considered and future research directions are proposed.

References

1. Berners-Lee, T., Fischetti, M., & Foreword By-Dertouzos, M. L. (2000). *Weaving the web: The original design and ultimate destiny of the World Wide Web by its inventor*. New York: HarperInformation.
2. Lynch, C. (2008). Big data: How do your data grow? *Nature, 455*(7209), 28–29.
3. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1), 107–117.
4. Internet Live Stats. (2015). Internet users - http://www.internetlivestats.com/internet-users/.
5. Duggan, M., Ellison, M., Lampe, N. B., Lenhart, C., & Madden, M. (2015). Social media update 2014. Technical report, Pew Research Center.
6. Perra, N., Gonçalves, B., Pastor-Satorras, R., & Vespignani, A. (2012). Activity driven modeling of time varying networks. *Nature Scientific Reports, 2*, 469.

Part I
Social Behavior Under Normal Conditions

Chapter 2
Modeling and Understanding Intrinsic Characteristics of Human Mobility

Jameson L. Toole, Yves-Alexandre de Montjoye, Marta C. González, and Alex (Sandy) Pentland

Abstract Humans are intrinsically social creatures and our mobility is central to understanding how our societies grow and function. Movement allows us to congregate with our peers, access things we need, and exchange information. Human mobility has huge impacts on topics like urban and transportation planning, social and biologic spreading, and economic outcomes. So far, modeling these processes has been hindered by a lack of data. This is radically changing with the rise of ubiquitous devices. In this chapter, we discuss recent progress deriving insights from the massive, high resolution data sets collected from mobile phone and other devices. We begin with individual mobility, where empirical evidence and statistical models have shown important intrinsic and universal characteristics about our movement: we, as human, are fundamentally slow to explore new places, relatively predictable, and mostly unique. We then explore methods of modeling aggregate movement of people from place to place and discuss how these estimates can be used to understand and optimize transportation infrastructure. Finally, we highlight applications of these findings to the dynamics of disease spread, social networks, and economic outcomes.

J.L. Toole (✉)
Engineering Systems Division, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: jamesontoole@gmail.com

Y.-A. de Montjoye • A. (Sandy) Pentland
Media Lab, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: yva@mit.edu; pentland@mit.edu

M.C. Gonzáles
Department of Civil and Environmental Engineering, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: bgoncalves@gmail.com

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7_2

2.1 Introduction

Mobility has been a steering force for much of human history. The movement of peoples has determined the dynamics of numerous social and biological processes from tribal mixing and population genetics to the creation of nation-states and the very definition of our living areas and identities. Urban and transportation planners, for example, have long been interested in the flow of vehicles, pedestrians, or goods from place to place.

With more than half of the world's population is now living in urban areas,[1] understanding how these systems work and how we can improve the lives of people using them is more important than ever. Insights from models informed by novel data sources can identify critical points in road infrastructure, optimize public services such as busses or subways, or study how urban form influences its function. Epidemiologists are also relying heavily on models of human movement to predict and prevent disease outbreaks [1, 2] as global air travel makes it possible for viruses to quickly jump continents and dense urban spaces facilitate human-to-human contagion. This has made understanding human movement a crucial part of controlling recent disease outbreaks.[2] Finally, social scientists are increasingly interested in understanding how mobility impacts a number of social processes such as how information spreads from person to person in offices and cafes across the world. These interactions have been theorized to impact crime rates, social mobility, and economic growth [3, 4] and understanding their dynamics may improve how we live, work, and play.

The growing need to understand and model human mobility has driven a large body of research seeking to answer basic questions. However, the lack of reliable and accessible data sources of individual mobility has greatly slowed progress testing and verifying these theories and models. Data on human mobility has thus far been collected through pen and paper surveys that are prohibitively expensive to administer and are plagued by small and potentially biased sample sizes. Digital surveys, though more convenient, still require active participation and often rely on self-reporting [5]. Despite the development of statistical methods to carefully treat this data [6–8] new, cheaper, and larger data sources are needed to push our understanding of human mobility efforts further.

The evolution of technology over the past decade has given rise to ubiquitous mobile computing, a revolution that allows billions of individuals to access people, goods, and services through "smart" devices such as cellular phones. The penetration of these devices is astounding. The six billion mobile phones currently in use triples the number of internet users and boast penetration rates above 100 % in the developed word, e.g. 104 % in the United States and 128 % in Europe.[3] Even

[1] United Nations Department of Economic and Social Affairs—World Urbanization Prospects—2014 Update. http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf.

[2] http://www.worldpop.org.uk/ebola/.

[3] GSMA European Mobile Industry Observatory 2011 http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf.

in developing countries, penetration rates are of 89 %[4] and growing fast. These devices and the applications that run on them passively record the actions of their users including social behavior and information on location[5] with high spatial and temporal resolution. Cellular antennas, wifi access points, and GPS receivers are used to measure the geographic position of users to within a few hundred meters or less. While the collection, storage, and analysis of this data presents very real and important privacy concerns [9, 10], it also offers an unprecedented opportunity for researchers to quantify human behavior at large-scale. With billions of data points captured on millions of users each day, new research into computational social science [11] has begun to augment and sometimes replace sparse, traditional data sources, helping to answer old questions and raise new ones.

In this chapter, we present an overview of mobility research in the current data rich environment. We describe a variety of new data sources and detail the new models and analytic techniques they have inspired. We start by exploring research on individuals that emphasizes important intrinsic and universal characteristics about our movement: we are slow to explore, we are relatively predictable, and we are mostly unique. We then discuss efforts to add context and semantic meaning to these movements. Finally, we review research that models aggregates of human movements such as the flow of people from place to place. Throughout and at the end of this chapter, we point out applications of this research to areas such as congestion management, economic growth, or the spreading of both information and disease.

2.2 New Data Sources

Traditional data sources for human mobility range from census estimates of daily commutes to travel diaries filled out by individuals. These surveys are generally expensive to administer and participate in as they require intensive manual data encoding. To extract high-resolution data, individuals are often asked to recall large amounts of information on when, where, and how they have traveled making them prone to mistakes and biases. These challenges make it hard for surveys to cover more than a day or week at a time or to include more than a small portion of the population (typically less than 1 %).

Mobile phones, however, with their high penetration rates, represent a fantastic sensor for human behavior. A large fraction of location data from mobile phones are currently in the form of call detail records (CDRs) collected by carriers when users perform actions on their devices that make use of the telecommunications

[4] ITU. (2013) ICT Facts and Figures http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf.

[5] Lookout (2010) Introducing the App Genome Project https://blog.lookout.com/blog/2010/07/27/introducing-the-app-genome-project/.

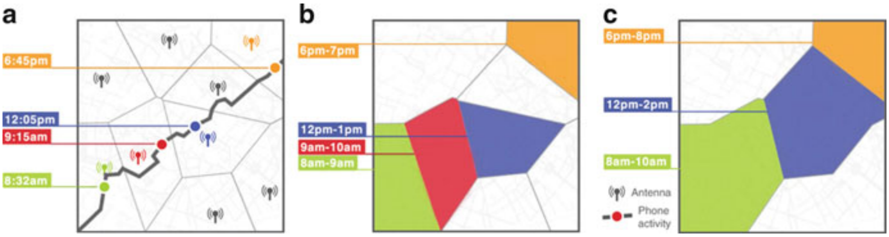


Fig. 2.1 Mobile phones are increasingly being used to collect high-resolution mobility data. This figure from de Montjoye et al. [9] depicts (**a**) a sequence of calling events made by a user at different locations. (**b**) These events are localized to the area served by the closest mobile phone tower to the use and (**c**) can be aggregated into individual specific neighborhoods where a user is likely to be found at different times of the day or week

network. The location of each device at the time a call, text, or data request is registered (Fig. 2.1) is recorded by carriers for billing, network performance, and legal purposes. Locations are inferred either by observing the tower through which the phone is connected or by triangulation with nearby towers. With the increasing use of mobile phones, each individual generates tens to hundreds of these digital breadcrumbs on a daily basis and this number is only increasing. Through specific agreements or through open-data challenges [12], location data on millions of users is readily available to researchers and has been used extensively to augment and sometimes replace traditional travel surveys. This data now forms the core of numerous new mobility studies and models some of which we describe below.

Though generally less common than CDRs, applications running on smartphones may access even more precise estimates of a user's position. A variety of these sensors, from GPS to wifi, can pinpoint the location of a device to within just a few meters and can record data every few minutes [13]. Similarly, protocols such as bluetooth and NFC (near field communication) allow devices to discover and connect to one another within a few meter radius, creating ad hoc sensor and social proximity networks [14]. Some of these applications and underlying social-networks explicitly add crucial context to mobility data. Foursquare invites users to "check-in" at specific places and establishments, Twitter will automatically geotag tweets with precise coordinates from where they were sent, and the Future Mobility survey app passively maintains an activity diary [5] requiring little input from users.

Infrastructure and public services have also become much smarter and now collect data on their usage to improve and help plan operations. Toll booths automatically count and track cars and this data has helped create accurate and real-time traffic estimates used by mapping and navigation services to provide better routing information. Subways, streetcars, and busses use electronic fare systems that record when millions of users enter and exit transportation systems to help better predict demand. In addition to smarter public infrastructure, the ecosystem created by digital devices has given birth to entirely new transportation services such as Hubway, the Boston bike rental service, that collects data of every bike ride and has even released some publicly[6] or Uber, an on-demand car service, that uses historical usage data to balance the time a user has to wait for a car to arrive and the time drivers spend without clients. Finally, on-board devices and real-time data feeds from automatic vehicle location (AVL) systems power applications such as NextBus to track the location of thousands of busses and subways across the world to display and predict when the next bus will arrive. While smart infrastructure comes with its own privacy challenges [15],[7] vehicle and public transport data offer additional information to urban planners and mobility modelers to better understand these systems.

Finally, most practical mobility models need to properly account for geography such as mountains and rivers, transportation infrastructure such as bridges and highways, differences in density between urban and rural areas, and numerous other factors. Thankfully, the digitization of maps has led to an explosion of geographic data layers. Geographic information systems (GIS) have improved dramatically while falling data storage prices have made it possible for small and large cities to offer their public mapping data to citizens in an online, machine readable format. The U.S. Census Bureau's TIGERline program, San Francisco's OpenSF, and New York City's PLUTO data warehouse are just a few sources that offer huge repositories of publicly accessible geographic data on everything from building footprints and the location of individual trees in a city. Open- and crowd-sourced initiatives like OpenStreetMap allow anyone in the world to contribute and download high-resolution digital maps of roads, buildings, subways, and more, even in developing areas that may not have institutional resources to create them. Private efforts such as Google Maps and MapBox offer high-resolution satellite imagery, route planning, or point of interest information through free or low cost APIs. Put together, these resources provide a digital map of the world that serves as a rich backdrop on which to study human mobility and the infrastructure built to facilitate it.

Put together, new sources from CDRs to public transport data, from mobile phone applications to AVLs generate a data with size and richness prohibitively expensive to match via traditional methods. Collected passively and without any effort from the user, this data is often more robust to manipulation by conscious or unconscious biases and provides a signal that is difficult to fake. While we are convinced of the potential of this data, it is always important to remember that it is not without pitfalls. It would be illusory to think that all of the old biases or hidden variables would simply disappear because the data is large. In some cases, data is only recorded when an individual interacts with a device which may bias when samples are taken [16]. Similarly it is important to keep in mind that even if it covers a significant fraction of the population this data might not be representative. Finally, these data generally come stripped of context. We do not know why an individual

[6] Hubway Data Visualization Challenge (2012) http://hubwaydatachallenge.org/.

[7] New York taxi details can be extracted from anonymized data, researchers say (2014) http://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn.

has chosen to move or what they will be doing there. For these reasons, sampling and robust statistical methods are still–maybe more than ever[8]—needed to use this data to augment our current understanding of human mobility while still providing robust conclusions. We now discuss a number of studies that aim to do just this.

2.3 Individual Mobility Models

Understanding mobility at an individual level entails collecting and analyzing sets of times, places, and semantic attributes about how and why users travel between them. For example, on a typical morning one may wake up at home, walk to a local coffee shop on the way to the bus that takes them to work. After work they may go to the grocery store or meet a friend for dinner before returning home only to repeat the process the next day. The goal of modeling this mobility is to understand the underlying patterns of individuals using new high resolution data. Models can be used to plan infrastructure or public transport. Furthermore, models provide insights into the underlying nature of human behavior helping us understanding how we are slow to explore, relatively predictable, and mostly unique.

Early modeling work draws a great amount of inspiration from statistical physics, with numerous efforts making parallels with human mobility and random walk or diffusion processes. One of the used data from the crowd-sourced "Where's George" project. Named after George Washington, whose head appears on the $1 bill, the project stamped bills asking volunteers to enter the geographic location and serial number of the bills in order to build a travel history of various banknotes. As bills are primarily carried by people when traveling from store to store, a note's movement serves as a proxy for human movement. Modeling the bills trajectories as continuous random walks, Brockmann et al. found that their movement appears to follow a Lévy flight process [20]. This process is characterized by subsequent steps whose angular direction is uniformly distributed, but whose step-lengths follow a fat-tailed distribution. While small jumps are most probable, bills have a significant probability of making long jumps from time to time. These findings are aligned with observations that humans tend to make many short trips in a familiar area, but also take longer journey's now and then.

In 2008, Gonzalez et al. [17] showed that the movement of these bills does not tell the whole story. Using a CDRs dataset of more than $100,000$ users over a 6-month period in a European country (Fig. 2.2a), they showed that the step-length distribution for the entire population was better approximated by a truncated power-law $P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\Delta r/\kappa)$ with exponent $\beta = 1.79$ and cutoff distances between 80 km and 400 km. This suggests that Lévy flights are only a good approximation of individual's mobility for short distances. To understand the

[8] Flowing data—Where People Run in Major Cities http://flowingdata.com/2014/02/05/where-people-run/.

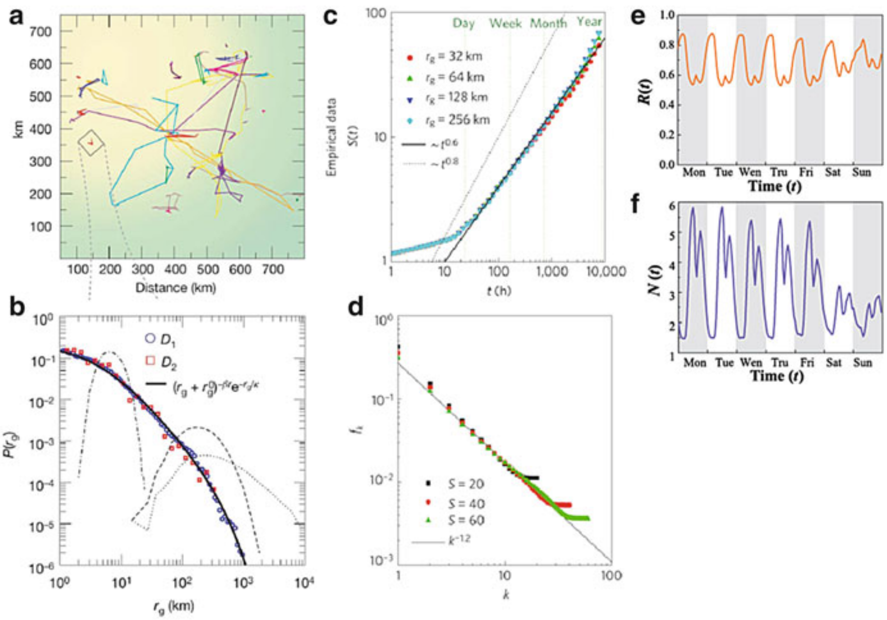


Fig. 2.2 (**a**) Individual mobility trajectories are passively collected from mobile devices [17]. (**b**) Measuring the distribution of radius of gyrations, r_g within a population of 100,000 users in a European country reveals considerable heterogeneity in typical travel distance of individuals. Moreover, this distribution cannot be explained by modeling each individual's movement as realizations of a single Lévy flight process [17]. (**c**) and (**d**) Show the slower than linear growth in new locations visited over time $S(t)$ and that the probability a location is visited next is inversely proportional to the frequency it has been visited in the past [18]. (**e**) This preferential return contributes to strikingly high predictability $R(t)$ over time while (**f**) the number of unique locations visited in any given hour is highly periodic and corresponds to the sleep-wake cycles of individuals [19]

mechanism that gives rise to this distribution, the authors borrowed a quantity from polymer physics known as the "radius of gyration" r_g:

$$r_g(t) = \sqrt{\frac{1}{N(t)} \sum_{i=1}^{N(t)} (\mathbf{r} - \mathbf{r_{cm}})^2}, \tag{2.1}$$

where N(t) are the number of observed locations and r_{cm} is the mean location of the user during the observation period. In essence, the radius of gyration is a measurement of the characteristic distance an individual travels during an observation period t. The authors then showed that the distribution of r_g in the population is itself well approximated by a truncated power-law with $r_g^0 = 5.8$ km, $\beta_{r_g} = 1.65$, and a cutoff of $\kappa = 350$ km (Fig. 2.2b). Simulations suggest that the step-length distribution of the entire population is produced by the convolution of heterogeneous Lévy flight processes, each with a different characteristic jump size determined by an individual's radius of gyration. Put differently, each person's mobility can be approximated by a Lévy flight process up to trips of some individual characteristic distance r_g. After this distance, however, the probability of long trips drops far faster than would be expected from a traditional Lévy flight.

Further investigation by the authors revealed the source of this behavior: the idiosyncrasy of human movements. Unlike random processes, humans are creature of habits and tend to return to previously visited locations such as home or work. The nature of these returns was also found to follow a very particular pattern. An individual returns to a previously visited location with a probability proportional to that location's rank $P(L) \sim 1/L$ amongst all the places he or she visits. These non-random, predictable return visits are unaccounted for in random walk and Lévy flight models and have been shown to be at the heart of deviations of observed behavior from random processes. Additional studies [21] have found similar patterns in both other CDRs datasets and Foursquare or Twitter check-ins.

Subsequent work by Song et al. [18] further studied how individual-specific locations need to be taken into account in mobility models. Using a similar CDR dataset, the authors showed three important characteristics of human behavior. First, the number of unique locations visited by individuals $S(t)$ scales sub-linearly with time $S(t) \sim t^{\mu}$ where $\mu = 0.6$ (Fig. 2.2c). Second, the probability an individual returning to a previously visited locations scales with the inverse of the rank of that location $P(L) \sim L^{-\zeta}$ where $\zeta = 1.2$ (Fig. 2.2d), a phenomena labeled as "preferential return." Third, the mean displacement, Δr, of an individual from a given starting point shows slower than logarithmic growth, demonstrating the extremely slow diffusion of humans in space. In essence, these findings pinpoint the dampening of explorative human movement overtime. Long jumps are observed so infrequently that they do not affect the average displacement of individuals. The authors then propose a new model of human mobility to capture these three characteristics. The model is as follows: starting at time t, an individual will make a trip at some future time Δt drawn from a fat-tailed probability distribution measured from CDRs. With probability $\rho S^{-\gamma}$, the individual travels to a new, never-before visited location some distance Δr away, where Δr is drawn from the fat-tailed distribution characterized in the previous model. With probability $1 - \rho S^{-\gamma}$ an individual returns to a previously visited location according to the inverse rank equation.

These early models do not attempt to recover periodic aspects of movement (e.g., daily commuting) or semantic meaning of visits (e.g., to visit a friend or go shopping), or attempt to do so. They do, however, emphasize important statistical and scaling properties of human mobility and often successfully reproduce them. Taken together, these models show that human explore slowly space, returning more often than not to known places and with less long steps than predicted by a power-law distribution.

Approaching the problem from the perspective of machine and statistical learning, another set of models has uncovered and explored another facet of human mobility: how predictable we are. In [19], Song et al. used information theory metrics on CDRs to show the theoretical upper-bound on predictability using three entropy measures: the entropy S, the random entropy S^{rand}, and the uncorrelated entropy S^{unc}. They then use their empirical distributions to derive an upper bound on a user's predictability ($\prod^{\max}$, $\prod^{\mathrm{rand}}$, and $\prod^{\mathrm{unc}}$). On average, the potential predictability of an individual's movement is an astounding 93 % and no user displayed a potential predictability of less than 80 %. To further quantify predictability, the author introduced two new metrics. They defined regularity $R(t)$ as the probability a user is found at their most visited location during a given hour t, along with the number of unique locations visited during a typical hour of the week $N(t)$ (Fig. 2.2e, f). Both show strong periodicity and regularity. These quantities have since been measured in different data sets in different cities and countries and have been shown to be consistent among them [21].

While the previous study provided a theoretic upper bound on the predictability of an individual, a number of statistical learning techniques have been developed to make predictions in the traditional sense. Early work in the area, predating even analytic computations, used Markov models and information on underlying transportation networks to predict transitions between mobile phone towers within cities. These models have been used to improve quality of service of wireless networks through proper resource allocation [22–25]. Later work incorporated various trajectory estimation and Kalman filtering algorithms to predict movements in small spaces such as college campuses [26, 27].

Temporal periodicity was used by Cho et al. [28] in their Periodic Mobility Model and social behavior incorporated in the Period Social and Mobility Model. These approaches derive the probability distribution of a user to be at any given location at a given time from previous location data. The latter also account for the location history of social contacts. The authors used these models to estimate that as much as 30 % of our trips may be taken for social purposes. Multivariate nonlinear time series forecasting produced similar results [29, 30] predicting where an individual will be either in the next few hours or at a given time of a typical day. These models, however, are all focused on predicting the geographic position of individuals at different times and do not attempt to understand what individuals may be doing there or any other semantics of place.

Though acquiring semantic information about mobility is more difficult than simply measuring geographic coordinates, it provides a much richer abstraction to study behavior. In one of the first studies to mine the behavior of college students using mobile phones, Eagle et al. [31] gave a few hundred students smart phones that recorded not only locations, but also asked users to label each place with its function such as home or work. Applying principal component analysis to these abstract movements from semantic place to semantic place (as opposed to geographic movements alone), the authors found that an individual's behavior could be represented as a linear combination of just a few "eigenbehaviors." These eigenbehaviors are temporal vectors whose components represent activities such as being at home or being at work. They can be used to predict future behaviors, perform long range forecasts of mobility, and label social interactions [14, 32]. The price paid for such detailed predictions, however, is the need for semantic information about locations. Geographic positions need to be tagged with attributes such as home or work in order for them to be grouped and compared across individuals.

Another approach to studying more abstract measurements of individual location information comes from recent work by Schneider et al. [33]. The authors introduced "mobility motifs" by examining abstract trip chains over the course of a day. A daily mobility motif is defined a set of locations and a particular of visits. More formally, these motifs constitute directed networks where nodes are locations and edges are trips from one location to another. For example, the motif of an individual whose only trips in a day are to and from work will consist of two nodes with a two directed edges (one in both directions). Counting motifs in mobility data from both CDRs and traditional travel surveys, they found that, on average, individuals visit three different places in a given day. They then construct all possible daily motifs for a given number of locations n and compute their frequencies. Shockingly, while there exists over 1 million ways for a user to travel between 6 or fewer locations, 90 % of people use one of just 17 motifs and nearly a quarter follow the simple two location commute motif introduced earlier (Fig. 2.3a). The authors found similar results in travel survey data and introduced a simple Markov model for daily mobility patterns which reproduces empirical results.

It is tempting to hypothesize that high theoretical and practical predictability results from high levels of similarity between individuals in a region. Perhaps the pace of life, full of mono-centric downtowns, or the structure of transportation systems funnel users to the same places and route choices. de Montjoye et al. [9] explored this hypothesis and found that, while predictable, an individual's movement patterns are highly unique. The authors introduced "unicity" $\mathcal{E}_p$ as the fraction of traces uniquely defined by a random set of p spatiotemporal points where a trace T is a set of spatiotemporal points, each containing a location and

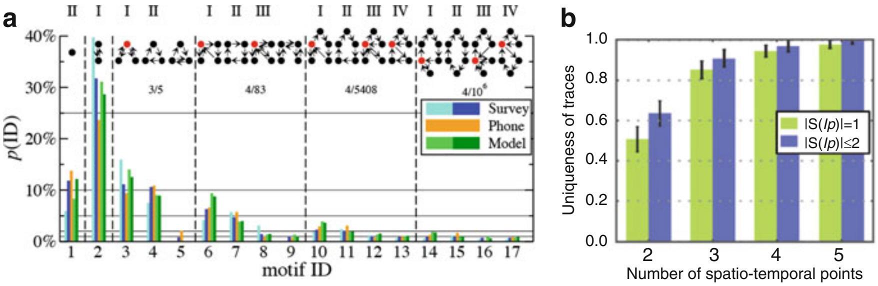


Fig. 2.3 (**a**) Removing geographic coordinates from locations and only focusing on a set of unique places and the directed travel between them, mobility motifs reveal that the daily routines of people are remarkably similar. Despite over 1 million unique ways to travel between 6 or fewer points, just 17 motifs are used by 90 % of the population. Moreover, the frequency of their appearance in CDR data matches very closely with more traditional survey methods [33]. (**b**) Despite this similarity and predictability, our movement displays a high degree of unicity. Just four spatiotemporal points is enough to differentiate a user from 95 % of all others individuals [9]

a timestamp. A trace is said to be uniquely defined by a set of points I_p if it is the only trace that matches I_p in the entire dataset. Applying this measure to a CDR dataset on 1.5 million users, the authors found that just four spatiotemporal points is enough to uniquely identify 95 % of all users (Fig. 2.3b). The authors further study unicity when the data is coarsened spatially or temporally. They found $\mathscr{E} \sim (v * h)^{\beta}$ unicity decrease as a power function with the spatial (v) and temporal resolution of the data (h) and that $\beta \sim -p/100$. Taken together, these equations show that unicity decreases slowly with the spatial and temporal resolution of the data and that this decrease is easily compensated by the number of points p. High uniqueness in human mobility traces exists across many spatiotemporal scales. These results not only raise many questions about the privacy of massive, passively collected metadata datasets, but also highlight an incredibly interesting nuance of human mobility: though individuals are predictable, they are also unique.

Merging concepts of predictability and unicity, work by Sun et al. [34] used temporal encounter networks to study repeated co-locations between passengers using data from bus passengers in Singapore. Temporal encounter networks were constructed by connecting individuals if they rode the same bus at the same time. An average individual encountered roughly 50 people per trip and these trips were highly periodic, occurring at intervals associated with working hours as well as daily and weekly trips. A pair of individuals who encountered each other tended to meet an average of 2.5 times over the course of a week. The distribution of time between encounters reveals strong periodicity, with passengers riding the same bus to work in the morning riding the same home, or riding the same bus at the same time each morning. This finding illustrates the idiosyncrasies of human mobility. Not only we visit a few places very during the day, but we also do so at the same times and by the same routes. Amazingly, though both of these results suggest that our unicity should be low, the previous work shows us that this is not the case.

In summary, new data sources have allowed researchers to show that, over weeks and months, human movement is characterized by slow exploration, preferential return to previous visited places, exploration of daily motifs, and predictable uniqueness. These regularities have been used to develop algorithms capable of predicting movement with high degrees of accuracy and have been shown to mediate other important processes such as social behavior and disease spread. Individual mobility patterns, however, are not the only level of granularity of interest to researchers, city planners, or epidemiologist. Aggregate movement can be either derived from individual level model or modeled as an emergent, personified phenomena. In the next section, we discuss works and models which aim at describing and modeling aggregate movement and flows of many individuals from place to place.

2.4 Aggregate Mobility

Aggregated mobility is used for planning urban spaces, optimizing transportation networks, studying the spread of ideas or disease, and much more. Perhaps the largest component in these models are origin-destination matrices that store the number of people traveling from any location to any other at different times or by different means. Like many complex systems, aggregate behavior is often more than the sum of individual parts and can be modeled separately. Additional layers of complexity are also needed to account for and sometimes explain individual choice of mode of transportation or route as described by the "four step model" [8, 37].

Like their individual-focused counterparts, many of these aggregate models are inspired by physical processes. Some of the earliest techniques for estimating origin-destination matrices are gravity models which have been used to model flows on multiple scales, from intra-city to international [8, 38]. Borrowed directly from Newton's law of gravitation, the number of trips T_{ij} taken from place i to place j is modeled as a function of the population of each place m_i and m_j and some function of the distance between them $f(r_{ij})$. The intuition is that the population of a place, its mass, is responsible for generating and attracting trips and thus the total flux between the two places should be proportional to the product of the two masses while the distance between them mitigates the strength of this connection. In the fully parameterized version of this model, an exponent is applied to the population at the origin and destination $T_{ij} = a\frac{m_i^{\alpha} m_j^{\beta}}{f(r_{ij})}$ to account for hidden variables that may be specific to local regions or populations. While the classical gravity model from physics is recovered by setting $\alpha = \beta = 1$, and $f(r_{ij}) = r_{ij}^2$, these parameters are generally calibrated for specific application using survey data.

Gravity models, however, are not without limitations. First, they rely on a large number of parameters to be estimated from sparse survey data which often leads to overfitting and, second, they fail to account for opportunities that exist between the two masses of people. The latter fault results in the same flow of people being estimated between two locations whether there is an entire city or an empty desert between them. Intuitively, one would expect that trips between places would be affected by the intervening opportunities to complete a journey. These shortcomings lead Simini et al. to develop the radiation model [35]. Again borrowing from physics (this time radiation and absorption), they imagined individuals being emitted from a place at a rate proportional to its population and absorbed by other locations at a rate proportional to the population there. In this model, the probability that an emitted person arrives at any particular place is a function of their probability of not being absorbed before getting there. The model is as follows: $T_{ij} = T_i \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}$, where T_i is total number of trips originating from location i and s_{ij} is the population within a disc centered on location i with a radius equal to the distance between i and j. The radiation model does not directly depend on the distance between the two places, taking instead into account the opportunities in-between them (Fig. 2.4a). Unlike the gravity model, the radiation model is parameterless and requires only data on populations to estimate flow. The authors showed that despite its lack of parameters, the radiation model provides better estimates of origin-destination flows than the gravity model for areas the size of counties or larger.

Yang et al. adapted Simini's radiation model to correct for distortions caused at different scales [39]. They showed the original radiation model's lower accuracy in urban environment is due to the relatively uniform density and small distances

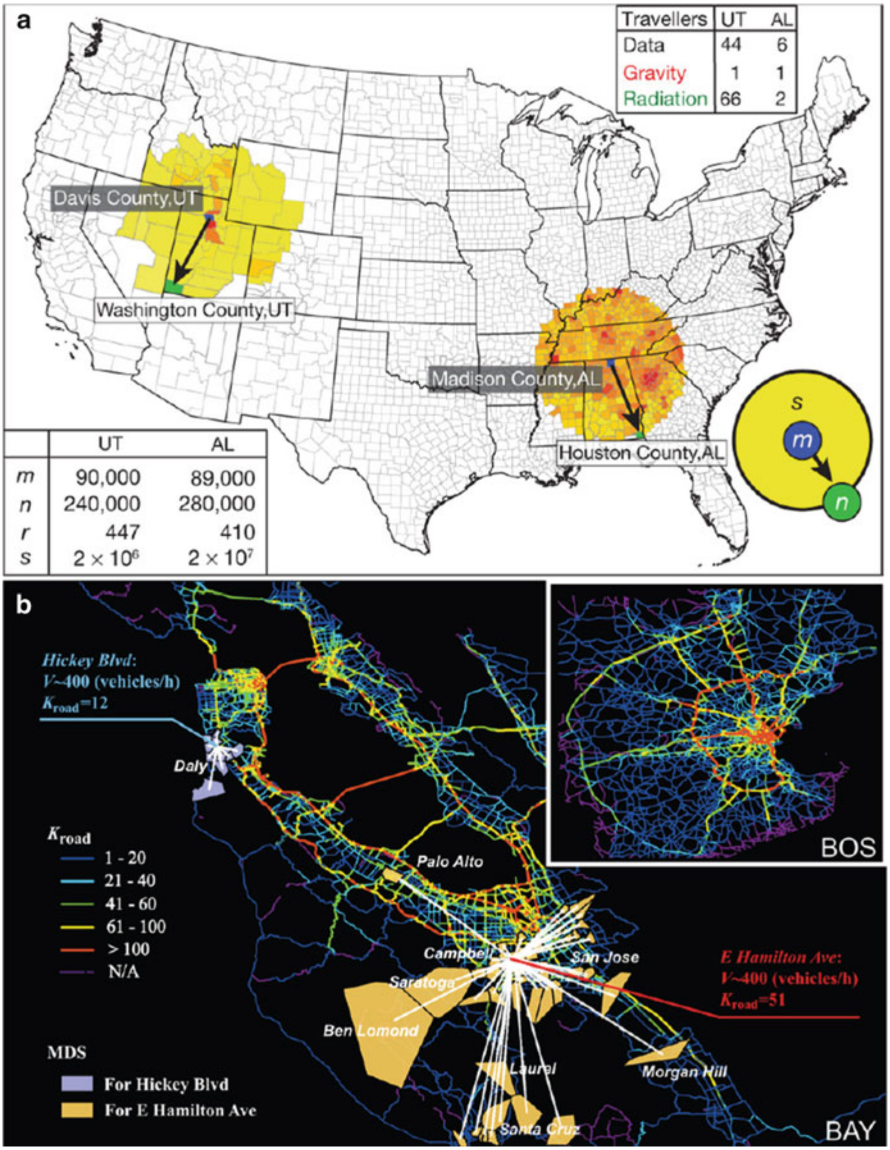


Fig. 2.4 (**a**) The radiation model accounts for intervening opportunities, producing more accurate estimates of flows between two places than more traditional gravity models [35]. (**b**) Routing millions of trips measured from CDR data to real road networks makes it possible to measure the importance of a road based on how many different locations contribute traffic to it, K_{road}. Understanding how transportation systems perform under different loads presents new opportunities to solve problems related to congestion and make infrastructure more efficient [36]

that characterize cities. In dense urban areas, distances are all relatively short and an individual may choose to visit a particular location due to hedonic attributes regardless of whether it is convenient to get to or not. Yang et al. subsequently introduced a scaling parameter α in the function describing the conditional probability an individual is absorbed at a location. This single parameter was enough to correct for these distortions and to provide a model that works on any length scale. Moreover, the authors suggested that for urban areas, the density of points of interest (POIs) such as restaurants and businesses is a better predictor of the absorption of a place than its population. Iqbal et al. [40] have demonstrated an improved way to extract valid, empirical OD matrices from CDR data to validate the model.

Finally, activity-based approaches [6] model user intent more explicitly. They hypothesize that all trips are made to fulfill certain needs or desires of an individual. Travel and survey diaries are used to identify those needs for different segments of the population and how they are typically fulfilled. This knowledge can then be used by the model given the demographics of individuals and environmental factors. These models are closely related to agent-based models simulating the behavior of city residents and rely heavily on the idea of economic utility.

From a practical perspective, city planners need to know not only how many people will go from point A to point B at a certain time of the day but also the mode of transportation and route choice of these individuals. For example, we would like to predict which route they will take so that we can properly estimate the stress placed on transportation systems and potentially optimize performance. Models of route choice typically assume that individual rationally chose the path from A to B that minimize some cost function such as total travel time or distance. Paths can be computed on a road network using shortest path algorithms such as the traditional Dijkstra algorithm or A-Star, an extension that enjoy better performance thanks to heuristics. Other informations such as speed limits can also be taken into account to estimate free flow travel times.

More advanced models are needed to account for the impact of congestion as drivers rarely encounter completely empty freeways. Iterative traffic assignment algorithms model congestion endogenously [41]. Trips are first split into segments containing only a fraction of total flow between two points. Trips in each segment are then routed along shortest paths independently of all other trips in that segment keeping counts of how many trips were assigned to each road. The travel times are then adjusted according to a volume delay function that accounts for the current congestion on a road where congestion is computed as the ratio between the volume of traffic assigned to the segment and the capacity of the road (referred to as volume-over-capacity). Trips in the next segment are then routed using updated costs until all flow has been accounted for. In this way, as roads become more congested and the travel time increases, drivers in later iterations are assigned to different, less congested routes. Values of total volume on each road, congestion, and travel times can then be validated against traffic counters, speed sensors, or data from vehicle fleets like taxis and busses but also smartphones such as in the Mobile Millennium project [42–45].

Wang et al. [36] further explored the use of CDRs as input for these iterative algorithms to estimate traffic volume and congestion. After correcting for differences in market share and vehicle usage rates, they measure trips by counting consecutive phone calls of individuals as they move through the city to generate flow estimates that were then routed. Using this approach, Wang et al. show the distribution of traffic volume and congestion to be well approximated by an exponential mixture model. This model depends on the number of major and minor roadways in a cities network. Using the same approach, the authors describe the usage patterns of drivers by a bipartite usage graph connecting locations in the city to roads used by those travelers (Fig. 2.4b). Roads can be defined by the number of locations that contribute traffic to them while place can be described by the roads used to visit them. The "function" of a road can then be classified by comparing its typological to its behavioral importance. For example, a bridge may be topologically important because it is the only way to cross a river, but a main street may be behaviorally important because it attracts motorists from many different neighborhoods. Using these measures, researchers were able to devise congestion reduction strategies that target the 2 % of neighborhoods where trip reduction will have the largest network wide effect. They found this smart reduction strategy is three to six times as effective as a random trip reduction strategy. Further work used this analysis to predict traffic jams [46, 47].

Private cars, however, are not the only mode of transportation studied. Using smartphones and AVL data, researchers have been mapping the routes followed by public transport and even privately owned mini-buses in the developing countries [48–50]. Similarly, data on air travel has been increasingly available to study aggregated mobilities between cities for applications in epidemiology (see below).

2.5 Human Behavior and Mobility

While of obvious interest to travelers, urban planners and transportation engineers, people's movement strongly impacts other areas. Though by no means an exhaustive list, we highlight three areas here: social behavior, disease and information spread, and economic outcomes. Many of these dynamics are discussed in greater detail in further sections of this volume.

2.5.1 Mobility and Disease Spread

Human movement via cars, trains, or planes has always been a major vector in the propagation of diseases. Consequently, the human mobility data and models discussed so far have increasingly been used to study the propagation of diseases. For example, CDR data has been used to map mobility patterns in Kenya helping researchers in their fight against Malaria [1, 54]. More recently, CDR and other data

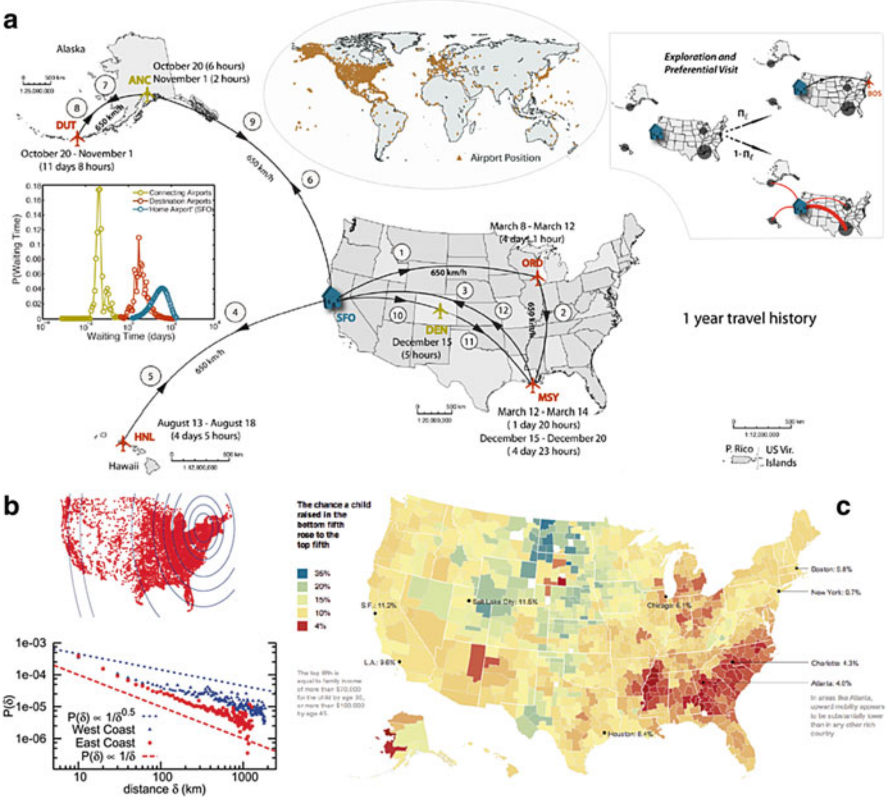


Fig. 2.5 (**a**) Global air travel has dramatically increased the speed at which diseases can spread from city to city and continent to continent [51]. (**b**) Mobility also impacts social behavior as we are far more likely to be friends with someone who lives nearby than far away [52]. (**c**) Mobility and the access it provides has strong correlations with economic outcomes. Children have dramatically different chances at upward economic mobility in certain places of the United States than others [53]

from West-Africa has been used to model regional transportation patterns to help control the spread of Ebola.[9] Finally, air travel data has become central to the study of global epidemics when planes allow an individual to travel between nearly any two points on the globe in a matter of hours. The global airline network therefore often determines how potent an epidemic could be and its likely path across the globe [2, 51, 55, 56] (Fig. 2.5a).

[9] Cell-Phone Data Might Help Predict Ebola's Spread (2014) http://www.technologyreview.com/news/530296/cell-phone-data-might-help-predict-ebolas-spread/

2.5.2 *Mobility and Social Behavior*

Intent is a crucial element of human mobility and movement is often a means to a social end. Despite new communications technologies making it easier than ever to connect across vast distances, face-to-face interactions still play an important role in social behavior whether it is the employees of a company commuting to a central workplaces or friends meeting at a restaurant on a weekend. The link between social contacts and mobility has becoming increasingly prominent in research as mobility data is often collected through mobile phones or location-based social networks.

Using data from an online social-network, Liben-Nowell showed the probability of being friends with another individual to decrease at a rate inversely proportional to the distance between them suggesting a gravity model of the form discussed above [52] (Fig. 2.5b). Subsequent work verified Liben-Nowell findings in other social networks [57, 58] while Toole et al. [59] showed the importance of taking into account geography when studying social-networks and how information spreads through them. Moreover, geographic characteristics can be used to predict the social fluxes between places [60]. Conversely, social contacts are very useful in predicting where an individual would travel next [28, 29, 61] and Cho et al. find that while 50–70 % of mobility can be explained as periodic behavior, another 10–30 % are related to social interactions.

Models such as the one proposed by Grabowicz et al. [58] have subsequently been developed to incorporate this dynamic and evolve both social networks and mobility simultaneously. The authors incorporate social interactions by having individuals travel in a continuous 2D space where an individual travel's is determined by the location of their contacts and use location as a determinant of new social tie creation. The model is as follows: with probability p_v, an individual moves to the location of a friend, and, with probability $1-p_v$, they choose a random point to visit some distance Δr away. But, while social ties impact mobility, mobility can also impact social ties. Upon arriving at a new location, the individual can thus choose to form social ties with other individuals within a radius with probability p or random individuals anywhere in the space with probability p_c, a free parameter. Although simple, this model is able to reproduce many empirical relationships found in social and mobility data.

2.5.3 *Mobility and Economic Outcomes*

Mobility not only provides people with social opportunities, it also provides economic ones. Economists and other social scientists have developed numerous theories on the role of face-to-face interactions in socio-economic outcomes and economic growth. In-person meetings are thought to unlock human capital, making us productive [62, 63]. For example, jobs in dense cities tend to pay higher wages than the same jobs in more rural areas even after controlling for factors such as age and education [64] in part due to productivity and creativity gains made possible by the rich face-to-face interactions that close spatial proximity facilitates. Universal urban scaling laws have been repeatedly found showing that societal attributes from the number of patents to average walking speed scales with population and theoretic models have been proposed that suggest density is at the heart of these relationships [3, 4, 65]. While density is one way of propagating these benefits, increased mobility is another. Poorer residents of cities have been, for example, shown to have better job prospects and higher chances of retaining jobs when given a personal car instead of being constrained by public transit [66]. Finally, Chetty et al. [53] found strong correlations between intergenerational economic mobility and variables related to the commuting times and spatial segregation of people (Fig. 2.5c). While we are only beginning to explore these relationships, early returns suggest that mobility is a critical component of many economic systems.

2.6 Conclusion

In this chapter, we reviewed a number of ways new data sources are expanding our understanding of human mobility. Applying methods from statistical physics, machine learning, and traditional transportation modeling, reproducible characteristics of human movement become visible. We explore slowly [17, 18], we are highly predictable [19, 29], and we are mostly unique [9]. Models of aggregate flows of people from place to place have also found success with analogies to statistical physics validated by new data sources [35]. More accurate measurements of city-wide traffic have made it easier than ever to assess the performances of transportation systems and devise strategies to improve them [36]. Valuable in their own rights, these insights have informed our understanding of other social phenomena as well, leading to more accurate models of disease spread, social interactions, and economic outcomes. As cities become home to millions of people each year, the insights gained from these new data are critical in making them more sustainable, safer, and better places to live.

References

1. Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., et al. (2012, October). Quantifying the impact of human mobility on malaria. *Science, 338*(6104), 267–270.
2. Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2006, February). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America, 103*(7), 2015–2020.
3. Bettencourt, L. M. A. (2013). The origins of scaling in cities. *Science, 340*, 1438–1441.
4. Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., & Pentland, A. (2013) Urban characteristics attributable to density-driven tie formation. *Nature Communications, 4*, 1961.

5. Cottrill, C. D. A., Pereira, F. C. A., Zhao, F. A., Dias, I. F. B., Lim, H. B. C., Ben-Akiva, M. E. D., et al. (2013) Future mobility survey. *Transportation Research Record, 2354*, 59–67.
6. Ben-Akiva, M. E., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand*. Cambridge: MIT Press.
7. Hall, R. W. (Ed.) (1999). *Handbook of transportation science*. International series in operations research & management science (Vol. 23). Boston: Springer.
8. de Dios Ortúzar, J., & Willumsen, L. G. (2011). *Modelling transport*. Chichester: Wiley.
9. de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Nature Scientific Reports, 3*, 1376.
10. de Montjoye, Y.-A., Shmueli, E., Wang, S. S., & Pentland, A. S. (2014). OpenPDS: Protecting the privacy of metadata through SafeAnswers. *PLoS ONE, 9*, e98790.
11. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721–723.
12. de Montjoye, Y. A., Smoreda, Z., Trinquart, R., Ziemlicki, C., & Blondel, V. D. (2014, July). D4D-Senegal: The second mobile phone data for development challenge.
13. Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. In *Pervasive and mobile computing* (Vol. 7, pp. 643–659).
14. Eagle, N., & Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology, 63*, 1057–1066.
15. Kosta, E., Graux, H., & Dumortier, J. (2014). Collection and storage of personal data: A critical view on current practices in the transportation sector. In *Privacy technologies and policy SE - 10* (Vol. 8319, pp. 157–176).
16. Ranjan, G., Zang, H., Zhang, Z.-L., & Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review, 16*(3), 33–44. http://dl.acm.org/citation.cfm?id=2412101.
17. González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782.
18. Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010, September). Modelling the scaling properties of human mobility. *Nature Physics, 6*(10), 818–823
19. Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010) Limits of predictability in human mobility. *Science, 327*(5968), 1018–1021.
20. Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature, 439*, 462–465.
21. Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. In *ICWSM* (pp. 81–88).
22. Kim, H. S. (2003, January). QoS provisioning in cellular networks based on mobility prediction techniques. *IEEE communications magazine, 41*(1), 86–92.
23. Liu, T., Bahl, P., Chlamtac, I. (1998). Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks. *IEEE Journal on Selected Areas in Communications, 16*, 922–935.
24. Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., & Toledo, S. (2009). VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems - SenSys '09* (pp. 85–98).
25. Krumm, J., Horvitz, E., Dourish, P., & Friday, A. (2006). Predestination: Inferring destinations from partial trajectories. *UbiComp 2006: Ubiquitous Computing, 4206*, 243–260.
26. Minkyong, K., Kotz, D., & Songkuk, K. (2006). Extracting a mobility model from real user traces. In *Proceedings - IEEE INFOCOM*.
27. Lee, K., Hong, S., Kim, S. J., Rhee, I., & Chong, S. (2009). SLAW: a new mobility model for human walks. In *IEEE INFOCOM 2009*.
28. Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 11* (p. 1082). New York: ACM Press.

29. De Domenico, M. (2012). Interdependence and predictability of human mobility and social interactions. *Journal Pervasive and Mobile Computing, 9*, 798–807.
30. Scellato, S., Musolesi, M., Mascolo, C., Latora, V., & Campbell, A. T. (2011). NextPlace: A spatio-temporal prediction framework for pervasive systems. In *Pervasive computing,* Lecture notes in computer science (Vol. 6696, pp. 152–169). Heidelberg: Springer.
31. Eagle, N., & Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing, 10*, 255–268.
32. Sadilek, A., & Krumm, J. (2012). Far out: Predicting long-term human mobility. In *AAAI* (pp. 814–820).
33. Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society, Interface the Royal Society, 10*(84), 20130246.
34. Sun, L., Axhausen, K. W., Lee, D.-H., & Huang, X. (2013, August). Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences, 110*(34), 13774–13779.
35. Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature, 484*(7392), 8–12.
36. Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012, January). Understanding road usage patterns in urban areas. *Scientific Reports, 2*, 1001.
37. McNally, M. G. (2008, November). *The four step model*. Irvine: Center for Activity Systems Analysis.
38. Hansen, W. G. (1959, May). How accessibility shapes land use. *Journal of the American Institute of Planners, 25*(2), 73–76.
39. Yang, Y., Herrera, C., Eagle, N., & González, M. C. (2014, January). Limits of predictability in commuting flows in the absence of data for calibration. *Scientific Reports, 4*, 5662.
40. Iqbal, Md. S., Choudhury, C. F., Wang, P., & González, M. C. (2014, March). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies, 40*, 63–74.
41. Spiess, H. (1990, May). Technical note-conical volume-delay functions. *Transportation Science, 24*(2), 153–158.
42. Samaranayake, S., Blandin, S., & Bayen, A. (2011). Learning the dependency structure of highway networks for traffic forecast. In *Proceedings of the IEEE conference on decision and control* (pp. 5983–5988).
43. Herring, R., Nasr, T. A., Khalek, A. A., & Bayen, A. (2010). Using mobile phones to forecast arterial traffic through statistical learning. *Electrical Engineering, 59*, 1–22.
44. Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies, 18*, 568–583.
45. Jariyasunant, J. (2012). *Improving traveler information and collecting behavior data with smartphones*. PhD thesis.
46. Wang, J., Mao, Y., Li, J., Xiong, Z., & Wang, W.-X. (2015). Predictability of road traffic and congestion in urban areas. *PLoS One, 10*(4), e0121825. doi:10.1371/journal.pone.0121825. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0121825.
47. Wang, P., Liu, L., Li, X., Li, G., & González, M. C. (2014, January). Empirical study of long-range connections in a road network offers new ingredient for navigation optimization models. *New Journal of Physics, 16*(1), 013012.
48. Lorenzo, G. D., Sbodio, M. L., Calabrese, F., Berlingerio, M., Nair, R., & Pinelli, F. (2014, January). AllAboard. In *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14* (pp. 335–340). New York: ACM Press.
49. Ching, A. M. L. (2012). *A user-flocksourced bus intelligence system for Dhaka*. Diss. Massachusetts Institute of Technology.

50. Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., & Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences 111*(37), 13290–13294. http://www.pnas.org/content/111/37/13290.short.
51. Nicolaides, C., Cueto-Felgueroso, L., González, M. C., & Juanes, R. (2012). A metric of influential spreading during contagion dynamics through the air transportation network. *PLoS ONE,* 7, e40961.
52. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*(33), 11623–11628.
53. Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). *Where is the land of opportunity? The geography of intergenerational mobility in the United States*. No. w19843. National Bureau of Economic Research. http://qje.oxfordjournals.org/content/early/2014/10/16/qje.qju022.full#cited-by.
54. Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013, April). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society, Interface / the Royal Society, 10*(81), 20120986.
55. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. L., & Vespignani, A. (2009, December). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America, 106*(51), 21484–21489.
56. Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011, January). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports, 1*, 62.
57. Backstrom, L., Sun, E., & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (pp. 61–70).
58. Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., & Eguiluz, V. M. (2013). Entangling mobility and interactions in social media. *PLoS One, 9*(3), e92196. http://dx.plos.org/10.1371/journal.pone.0092196.
59. Toole, J. L., Cha, M., & González, M. C. (2012). Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS ONE,* 7(1), e29528.
60. Herrera-Yagüe, C., Schneider, C. M., Smoreda, Z., Couronné, T., Zufiria, P. J., & González, M. C. (2014). The elliptic model for communication fluxes. *Journal of Statistical Mechanics: Theory and Experiment, 2014*(4), P04022.
61. van den Berg, P., Arentze, T. A., & Timmermans, H. J. P. (2010). Size and composition of ego-centered social networks and their effect on geographic distance and contact frequency. *Transportation Research Record, 2135*, 1–9.
62. Kim, S. (1989). Labor specialization and the extent of the market. *Journal of Political Economy, 97*, 692–705.
63. Freedman, M. L. (2008). Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *Journal of Urban Economics, 64*, 590–600.
64. Yankow, J. J. (2006). Why do cities pay more? An empirical examination of some competing theories of the urban wage premium. *Journal of Urban Economics, 60*, 139–161.
65. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007, April). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America, 104*(17), 7301–7306.
66. Gurley, T., & Bruce, D. (2005). The effects of car access on employment outcomes for welfare recipients. *Journal of Urban Economics, 58*, 250–272.

Chapter 3
Face-to-Face Interactions

Alain Barrat and Ciro Cattuto

Abstract Face-to-face interactions of humans play a crucial role in their social relationships as well as in the potential transmission of infectious diseases. Here we discuss recent research efforts and advances concerning the measure, analysis and modelling of such interactions measured using strategies ranging from surveys to decentralised infrastructures based on wearable sensors. We present a number of empirical characteristics of face-to-face interaction patterns and novel techniques aimed at uncovering mesoscopic structures in these patterns. We also mention recent modelling efforts and conclude with some open questions and challenges.

3.1 Introduction

Our modern interconnected societies make many channels available for communications and social interactions, such as phone calls, email, virtual conferences, micromessaging, or online social networks. Despite this wealth of alternatives, direct face-to-face interactions between individuals remain an essential element of human behaviour and of human societies. Mining and analysing face-to-face interaction patterns between individuals therefore has a clear impact towards the fundamental knowledge and understanding of human behaviour and social networks. Most crucially, contact patterns among individuals play an important role in determining the potential transmission routes of infectious diseases, in particular of respiratory pathogens. An accurate description of these patterns represents therefore a crucial tool for identifying contagion pathways, for informing models

A. Barrat (✉)
Aix Marseille Université, Université de Toulon, CNRS, CPT, UMR 7332,
13288 Marseille, France

Data Science Laboratory, ISI Foundation, Torino, Italy
e-mail: alain.barrat@cpt.univ-mrs.fr

C. Cattuto
Data Science Laboratory, ISI Foundation, Torino, Italy
e-mail: ciro.cattuto@isi.it

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_3

of epidemic spread, and for the design and evaluation of control measures such as the targeting of specific groups of individuals with appropriate prevention strategies or interventions.

Empirical data describing direct interactions between individuals are however by nature difficult to gather as, contrarily to online interactions, phone calls or electronic communications, they do not leave any digital trace. Various data collection strategies have therefore been used, in particular in the epidemiological context and at different scales: surveys and diaries, synthetic population models, and, thanks to the increase in the availability and use of novel technologies, wearable sensors (see [1] for a review).

Here we first briefly review the measurement strategies and some of their advantages and intrinsic limitations (Sect. 3.2). We then discuss in Sects. 3.3 and 3.4 a number of empirical characteristics of face-to-face interactions as obtained by recent projects using wearable sensors, and review in Sect. 3.5 some recent attempts at modelling these processes. We conclude in Sect. 3.6 by presenting a number of open questions.

3.2 Proxies of Face-to-Face Interactions and Measurement Strategies

Face-to-face interactions between individuals occur in a variety of contexts and situations, contributing to phenomena as diverse as social coordination, information propagation, disease spread and more. Gathering data and understanding the patterns of these direct contacts is therefore of interest to fields of research ranging from the fundamental understanding of human behaviour to the epidemiology of transmissible diseases, and many efforts have been devoted to these tasks. We refer the reader to [1] for a recent review of the methods and technologies that have been used in various projects and provide here only a brief discussion on the range of methods available and non-exhaustive references to the corresponding research efforts.

A commonly employed method consists in asking individuals about their contacts, using surveys and diaries. Volunteer participants are asked to record their social interactions during a certain time period, for instance on a specific day, or on consecutive days. While social science studies can be interested in all such interactions (face-to-face, by email or by phone), the strongest focus on obtaining information on face-to-face interactions has emerged in epidemiological studies of infectious diseases, as such direct interactions are considered as relevant for transmission events. Many efforts have therefore been deployed to use contact diaries under various forms (using both paper and web-based questionnaires), either in specific contexts ranging from hospitals to schools or among the general population [2–9], and sometimes at very large scale with thousands of respondents [5, 7, 8]. Surveys have both advantages and limitations. One of the main advantages is that well-studied questionnaires allow to gather information not only on the existence of contacts but also on additional characteristics, such as their context (home, work, travel), estimates of their durations, existence of repeated contacts with the same individual, or even the distance from home at which the contacts take place [8]. Metadata such as the age, gender and occupation of the respondent can also be correlated with his/her contact numbers and durations. Questionnaires can even ask to specify for each contact if it involved physical contact and distinguish periods of well-being and illness of the respondent [10]. Surveys have also important limitations. First, questionnaires are costly and it is notoriously difficult to recruit participants [8, 9]. Second, self-reporting procedures entail biases that are difficult to estimate [4, 11, 12], as participants might not recall all their contacts or might make incorrect estimates of their durations. As surveys give access to ego-networks, the fraction of triangles in contact networks is also difficult to estimate and typically relies on each individual estimating if two of his/her contacts have themselves been in contact [8]. Finally, subtleties in questionnaire design might also influence the results, as discussed in [8]: for instance, the distribution of the number of reported contacts varies significantly whether individuals have to report the name of each contact or not.

Alternative approaches to the use of surveys have emerged in the recent years, giving usually access to proxies of face-to-face interactions. For instance, the availability of large-scale computing facilities and of detailed socio-demographic data have made it possible to recreate in silica synthetic populations at the scale of a whole city or country. These synthetic populations are typically used to generate contact networks to simulate the spread of infectious diseases [13–15]. Interestingly, the contact patterns obtained within such synthetic populations have been shown to match those obtained in large-scale surveys [14, 15].

Another approach takes advantage of the development of various types of sensors which can in particular measure the proximity of other similar devices, using technologies ranging from Bluetooth, WiFi or RFID [16–24]. Depending on the range of the signals used, such methods might yield information only on proximity at a range that might not imply face-to-face interaction (e.g., Bluetooth signals between devices can typically be received through a wall), or can be tuned to specifically detect close-range face-to-face proximity [19, 21, 22]. We will here mostly report on results obtained with the latter technology. Wearable sensors are nowadays simple to use and come at reasonable costs. They also afford an objective definition of contact and can report even short encounters. Their main limitation comes from the fact that they do not register contacts with individuals not participating to the data collection (and therefore not wearing any sensor) and therefore provide data on the contacts among a closed population. Sampling issues can thus arise if not all the members of the population of interest agree to wear the sensors [22].

Given the respective advantages and limitations of methods based on surveys and wearable sensors, a comparison of data collected by both types of methods in a given population is of great interest. To our knowledge, only one such study has been performed to date, namely in a high school context, showing in particular that many contacts registered by sensors are not reported in surveys, especially for short contacts, while long contacts are better reported [12]. More such studies in various contexts would be highly welcome in the future.

3.3 Face-to-Face Interactions as Temporal Networks

One of the advantages of decentralised sensing infrastructures based on wearable sensors is that they not only yield information on the existence of a face-to-face interaction between two individuals but also give access to the starting and ending time of each such interaction, with a certain time resolution (typically of the order of 20 s to a couple of minutes). The collected data can therefore adequately be represented as a time-varying social network of contacts within the monitored community, i.e., an instance of a "temporal network" [25].

The amount of activity, quantified as the number of observed face-to-face contacts in a given time-window, varies substantially over time and can be very different in different contexts. For instance, children in a primary school interact much more than adults in offices. Despite these differences, some generic statistical properties of the temporal networks of human interactions have emerged through the various data collection efforts. First, the time intervals between successive contacts are broadly distributed, spanning several orders of magnitude: most intercontact durations are short, but very long durations are also observed, and no characteristic timescale emerges [16, 19, 22, 26–28]. This bursty behaviour is a well-known feature of human dynamics and has been observed in a variety of systems driven by human actions [29]. Moreover, the distributions of the durations of single contacts are also broad, spanning several orders of magnitude, and their functional form displays a remarkable robustness across contexts [22, 28], measurement periods, and measurement methods [30], as illustrated in Fig. 3.1.

Overall, temporal networks of face-to-face contacts between individuals exhibit strongly heterogeneous dynamics, with robust statistical features. This implies two important facts for modelers, in particular when dealing with processes depending on contact durations between individuals, such as epidemic spreading. First, the broadness of the distributions means that taking into account only average contact durations and assuming that all contacts are equivalent might be a too coarse representation of the reality. Indeed, different contacts might yield very different transmission probabilities: many contacts are very short and correspond to a small transmission probability, but some are much longer than others and could therefore play a crucial role in disease dynamics, Second, the robustness of the distributions found in different contexts means that these distributions can be assumed to depend negligibly on the specifics of the situation being modeled and thus directly plugged into the models.

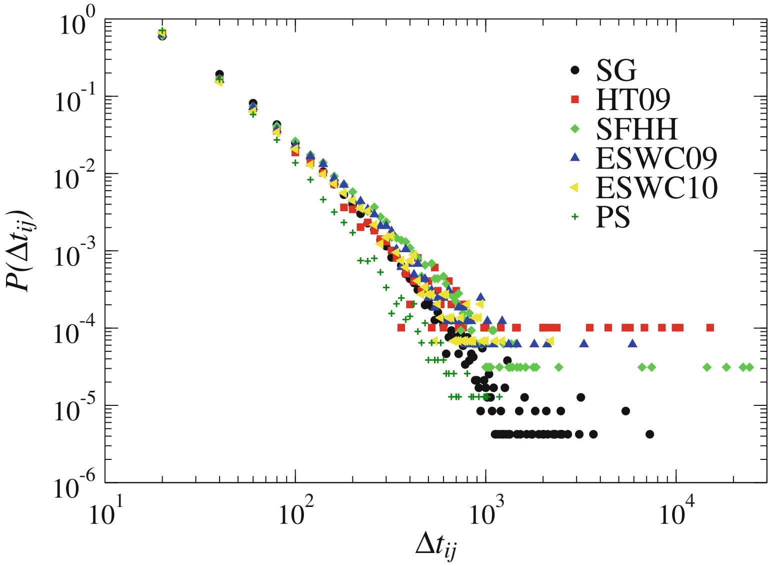


Fig. 3.1 Distributions of the face-to-face contact durations measured in different environments ranging from a museum (SG) to a school (PS) and several scientific conferences

3.4 Structures and Structure Discovery

3.4.1 Structures in Aggregated Data

3.4.1.1 Contact Networks

It is often useful to aggregate the temporal network of contacts between individuals over a given time window, in order to obtain static summaries of the contact sequence. In the obtained aggregated network, each node represents an individual, and a link between two nodes *i* and *j* denotes the fact that the corresponding individuals have been in contact at least once during the time window under consideration. Each such link is weighted by a summary of the temporal contact activity that took place between *i* and *j*, such as the number of contact events or the cumulative duration *w* of the contact events between the corresponding individuals.

The time window considered for aggregation can range from the finest time resolution of the recording system (that can be of the order of seconds or minutes) up to the entire duration of the data collection (e.g., days or weeks). In the case of surveys, the detailed temporal information on the contacts' timing is often not available, and the natural aggregation time scale is 1 day. Surveys thus typically give access to daily aggregated networks. Overall, different aggregation levels typically provide complementary views of the network dynamics at different scales.

The obtained aggregated networks unveil important information about the contact patterns of the population under study. A first characterisation is provided by the statistical distributions of nodes' degrees: in a contact network, the degree of a node (individual) is given by the number of distinct individuals with whom that individual has been in contact. In datasets collected through wearable sensors, the observed degree distributions are typically narrow, with an exponential decay at large degrees and characteristic average values that depend on the particular context [30, 31]. Interestingly, contact data obtained through surveys can lead either to narrow or broad degree distributions, as discussed in [8], and the result might be influenced by the way in which the survey is designed. When individuals are asked to report a precise list of persons encountered during the day, the obtained degree distributions are typically narrow [4, 6, 7], and the data of [8] actually show a good agreement with sensor data. However, if the respondents have in addition the possibility to report encounters with "groups" of individuals without specifying the identity of each group member, the distribution becomes broad [8].

While the number of distinct individuals met is certainly important when discussing behavioural patterns of humans, the durations and cumulated durations of face-to-face contacts also carry crucial information, in particular with respect to social or epidemiological contexts. The distributions of links' weights is thus a very relevant characteristic of these networks. Such distributions have been found to be broad in many different datasets collected either through sensors [19, 22, 24, 28] or surveys [8]: most pairs of interacting individuals have been in face-to-face proximity for a short total amount of time, but some cumulated contact durations are very long. No characteristic interaction timescale can be naturally defined, except for obvious temporal cutoffs due to the finite duration of the measurements. Strikingly, and similarly to the case of the durations of single contacts, recent studies have shown a strong robustness of the functional shape of these distributions in different contexts and even different data collection methods [22, 28, 30]. The empirically found distributions seem therefore to be a robust property of human behaviour and can be used directly for modelling purposes in various contexts.

While statistical distributions of node and link features display a strong robustness, the detailed structures of aggregated networks of contacts are much more diverse depending on the context. For instance, aggregated networks of interactions during a typical day at a small conference are rather "compact" with a close-knit structure [31], as participants are typically engaged in interacting with known individuals as well as in meeting new persons. Networks of contacts among children in a primary school or students in a high school display on the contrary a strong community structure, shown in Fig. 3.2, as a consequence of the grouping of individuals in classes [30, 33]. A similar structure has been observed in an office building, where workers from the same department have more contacts than with workers from other departments, even during lunch hours [34]. In hospitals, different structures emerge due to the different roles of individuals: as shown in [35, 36], nurses tend to form a rather dense group of nodes in the aggregated network. The network of links involving patients and caregivers has, on the other hand, a particular structure linking each patient to a specific caregiver, with very few links among caregivers or among patients (see [35, 37] for graphical illustrations).

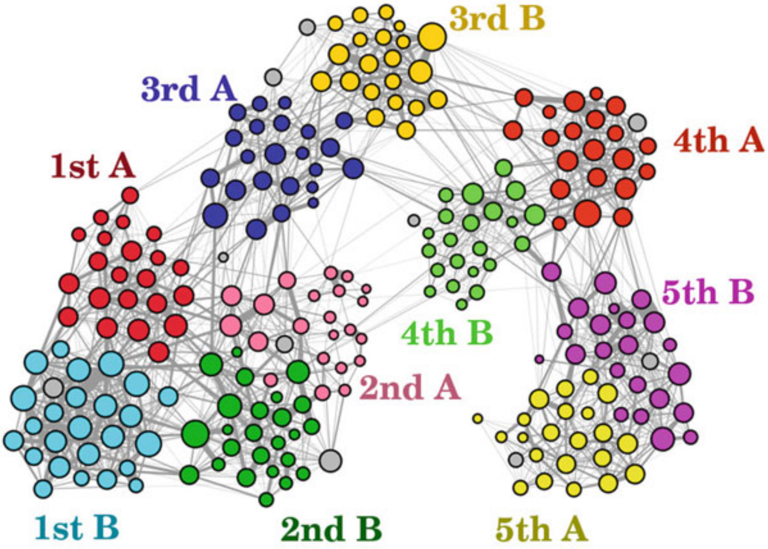


Fig. 3.2 Primary school contact network, aggregated over 1 day. Only links that correspond to cumulated face-to-face proximity in excess of 5 min are shown. The *color* of nodes indicates the grade and class of students. *Grey* nodes represent teachers. The network layout was generated by using the force atlas graph layout implementation available in Gephi [32]

3.4.1.2 Contact Matrices

It is often convenient to go one step further in the aggregation of contact data when the population under study is structured, i.e., when individuals can be classified according to specific characteristics or role (e.g., according to their age class or professional activity). A convenient summary of their contact patterns is then provided by contact matrices whose elements give the average number (or duration) of the contacts that individuals in one given class have with individuals of another class. Such a representation can be used at different scales: to describe, e.g., the contact patterns between individuals having different roles in a hospital ward (e.g., nurses, doctors, patients) [35, 36], or between children or students of different classes in schools [30, 33], but also to account for the mixing patterns between individuals of different age classes in the population of a country, as obtained by surveys [5].

Of note, the use of contact matrices for modeling contact patterns relies on a set of restricted homogeneous mixing assumptions within each class and on the representativeness of the average mixing behaviour between classes. Such an approach neglects the strong fluctuations observed in the distributions of the numbers and durations of contacts between two individuals of given classes [5, 6, 28, 38]. It neglects also the fact that contact networks are typically sparse, and that the density of links connecting individuals in given classes depends on the specific classes and is sometimes very small: many pairs of individuals never have any contact.

In order to provide a data representation that is not as specific as a high-resolution temporal network of contacts but does not discard relevant heterogeneities in the contact patterns, the contact matrix of distributions (CMD) has been introduced in [38]. This representation, instead of considering only the average of the contact time between individuals of specific classes, considers the whole distribution of contact times, typically fitted by a negative binomial distribution. Similarly to the customary contact matrices, the CMD is not an individual-based representation, and does not retain the detailed structure of the empirical contact network. It thus keeps the simplicity of a contact matrix formulation by grouping the individuals into role classes, but takes into account the heterogeneity of contact durations between individuals and the sparseness of the contact network. Such a representation is useful for designing interventions as it can suggest easily generalizable strategies that target specific classes of individuals [38].

3.4.1.3 Different Types of Contacts

Let us finally note that we have here mostly discussed aggregated networks of contacts as registered by wearable sensors in different contexts. Contacts are then gathered only between individuals participating to the data collection, and within the considered environment. Individuals however have contacts in different situations, ranging from home to workplaces and transportation means. In this respect, surveys can help understand and quantify how contacts depend on context. For instance, the large-scale survey analysed in [8] highlights how contact time decreases with age and how contacts involving touch tend be of longer duration. It also shows that home contacts account for the majority of contact hours, while work corresponds to more numerous but shorter contacts. Different occupations correspond also to different average daily numbers of contacts. Finally, the survey answers show that the time in contact decreases when the distance from home increases [8].

3.4.2 Longitudinal Structures

Human activity and contact patterns are highly non-stationary. In particular, the number of contacts among a given population varies strongly in time, obeying typical circadian rhythms and possibly modulated by the unfolding of scheduled activities [22]. It is therefore important to assess how statistical properties of contacts are impacted by and possibly coupled with these activity variations. Moreover, high-resolution datasets on contacts between individuals are typically gathered during few days or weeks in a certain context, and assessing the long-term stability of the data characteristics across different periods is also crucial.

3.4.2.1 Short-Term Stability

Despite the strong variations in activity, i.e., in the numbers of registered contacts, the main statistical properties of the contacts have been empirically shown to be stable [22, 28]. In particular, the contact duration distributions measured over different time windows coincide, as well as the structure of contact matrices across different workdays [30, 34, 36]. In fact, even the activity timelines can be remarkably stable across days when they depend on schedules either externally imposed as in schools [30] or driven by the organisation of work as in hospitals [36], as illustrated in Fig. 3.3.

On the other hand, surveys have shown that important differences between contact matrices describing contact patterns in the population are observed between work and non-work days [6, 8, 39], as well as, for a given individual, between periods of well-being and periods of illness [10].

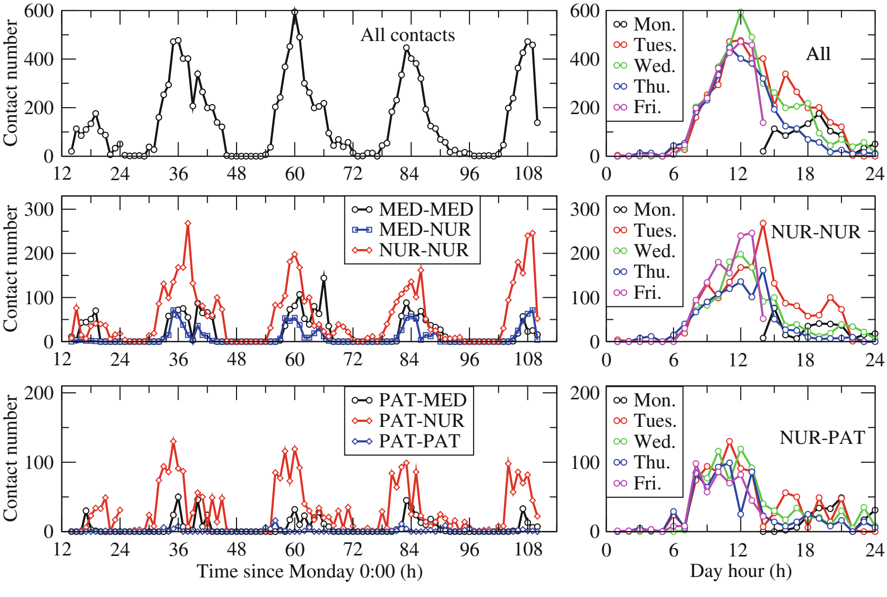


Fig. 3.3 Number of contacts per 1-h periods in a hospital ward. *Top row*: global number of contacts. *Middle and bottom rows*: number of contacts involving patients, healthcare workers and medical doctors. The *left* plots give the number of contacts as a function of the time since the start of the week (Monday, 0:00 AM). The *right* plots display the number of contacts of several types in each day, as a function of the hour of the day, to show the similarity of the curves in different days. Abbreviations: *NUR* paramedical staff (nurses and nurses' aides), *PAT* patient, *MED* medical doctor. From [36]

3.4.2.2 Long-Term Stability

Few datasets afford a comparison between contact patterns observed in a given context or in similar contexts during different periods. In particular, the comparison of the data gathered in different hospital wards [23, 35, 36] shows the robustness of stylised facts such as the central role of nurses and the small number of contacts between patients. Very few studies report and compare high-resolution contact networks measured in the same context at different points in time. Fournet and Barrat [30] compares contact data gathered in the same high school in two different years, and reports a very strong qualitative and quantitative similarity between contact matrices for different years.

3.4.3 Mesoscopic Structures and Latent Factor Analysis

In the previous sections we discussed the short-term and long-term stability of some statistical distributions of interest. The aggregation over time or over node attributes that is required to compute such distributions projects away many specificities, structures, and correlations of the original data. Depending on the problem at hand, these aggregated representations may overlook or confound important structural features of the network.

For example, a node or group of nodes may belong to different communities at different points in time: aggregating the network over time will artificially merge the communities and create a cluster that does not represent the network at any point in time. Similarly, groups of nodes may exist that share similar activity patterns over time. This is a common occurrence in environments such as schools, where an externally imposed schedule of social activities (e.g., class and lunch breaks) drives and constrains the interactions that are possible at a given time. In this case, temporal aggregation of the network may retain the topology of interactions but loses the information on correlated activity patterns, which may play an important role for, e.g., epidemic processes unfolding over the temporal network [40]. In general, correlated topological and temporal features of the network may give rise to structures that are neither local features of individual nodes or edges nor global structures, such as, for instance, a suitably defined network backbone. Hence, in the following we will refer to these structures as “mesoscopic structures”. It is important to remark that meso-scale structures are not limited to the (possibly hierarchical) community structure of the network: communities are usually defined as cohesive clusters, whereas the structures under study may also comprise two-mode communities [41], groups of links with correlated activity patterns, and more.

Detecting mesoscopic structures in high-resolution social network data is an outstanding challenge that calls for principled approaches and efficient computational techniques. Recent work focuses on extending well-known community detection techniques to the case of temporal networks. A common approach is to detect communities in static networks snapshots obtained by aggregating the temporal network over consecutive time intervals. The changes of the community structure over time are then analysed to relate communities found at different times and track their evolution. Simple approaches to mine the temporal community structure of a system are based on a continuity assumption for the (static) community structure detected at successive time intervals [42–44]. These approaches may prove useful in specific cases, but fail in the presence of discontinuous activity patterns, abrupt community formation or dissolution, and in general they cannot deal with temporal correlations over extended periods of time. Instead of separately treating the community structure and the temporal evolution of the network, some studies [45–47] pioneered global approaches to the problem of community detection in temporal networks.

More recently, we have investigated the use of techniques for latent factor analysis to simultaneously identify mesoscopic network structures and track their activity over time, without the assumption that the sought structures should be cohesive clusters. The starting point for this analysis is a mathematical representation of time-varying network data that treats topology and time on an equal footing: A temporal network can be naturally represented as a time-ordered sequence of adjacency matrices, each describing the state of the network at a discrete point in time. The adjacency matrices can be combined into a three-way tensor $\mathscr{T} \in \mathbb{R}^{N \times N \times S}$, where N is the number of nodes of the network and S the number of network snapshots. The tensor $\mathscr{T}$ encodes the entire information about the temporal network and has been recognized as a convenient representation both for multi-layer networks and temporal networks [48].

Once the network and its evolution are represented in a tensor form, we can use a variety of methods from data mining and machine learning to identify latent structures. We focused on tensor decomposition techniques that were developed in diverse domains like signal processing, psychometrics and brain science [49, 50]. In particular, we investigated the use of non-negative tensor factorization [50, 51] because, like non-negative matrix factorization [52], it is recognized as a powerful tool for learning parts-based representations. The basic idea is to approximate the tensor $\mathscr{T}$ by a sum of products of lower-dimensional factors, each of which can be interpreted in terms of groups of nodes and temporal activity patterns. Formally, $\mathscr{T}$ can be approximated by a sum $\tilde{\mathscr{T}}$ of rank-1 tensors:

$$\tilde{\mathscr{T}} = \sum_{r=1}^{R} \mathbf{a_r} \circ \mathbf{b_r} \circ \mathbf{c_r} \ , \tag{3.1}$$

subjected to non-negativity constraints on $\mathbf{a_r}$, $\mathbf{b_r}$ and $\mathbf{c_r}$. The number R of terms in the decomposition controls the complexity of the model: for small values of R, $\tilde{\mathscr{T}}$ is a crude approximation of $\mathscr{T}$, whereas for high values of R the decomposition yields a good approximation but eventually overfits $\mathscr{T}$. The choice of R is usually set by means of heuristics or quality metrics for the decomposition [53]. The vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_R$, $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_R$ and $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_R$ can be arranged into matrices $\mathbf{A} \in \mathbb{R}^{N \times R}$, $\mathbf{B} \in \mathbb{R}^{N \times R}$ and $\mathbf{C} \in \mathbb{R}^{S \times R}$. Rows correspond to the nodes of the network,

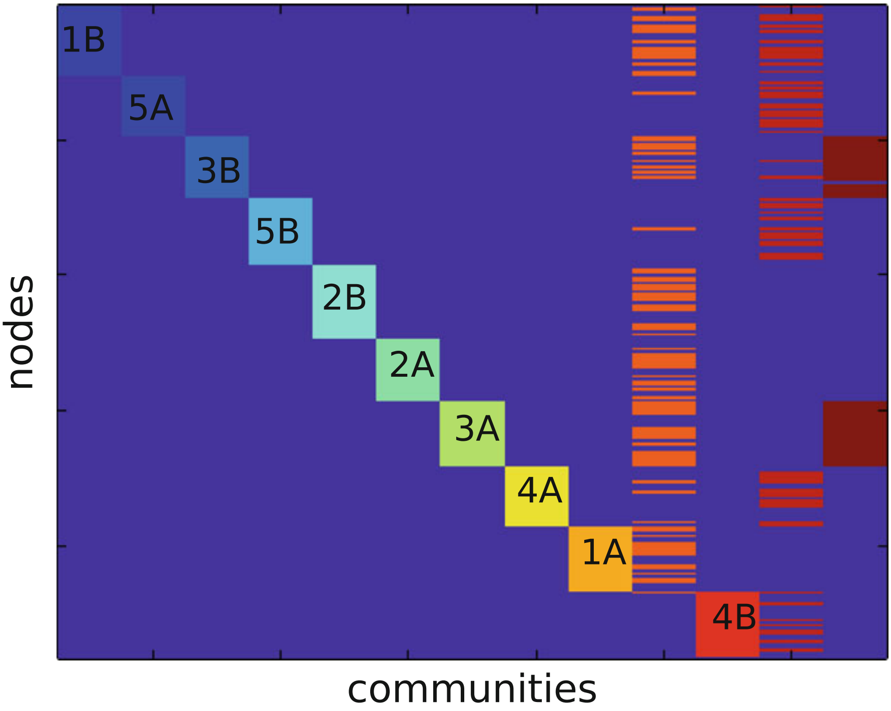


Fig. 3.4 Component-node matrix obtained via non-negative tensor factorization, for $R = 13$. Rows correspond to network nodes and columns to components, here regarded as mesoscopic structural features of the network. The matrix is obtained from the factor **A** by classifying each node as belonging (*light colours*) or not belonging (*dark blue*) to a given component. Node order has been rearranged to expose the block structure of the matrix. Colours identify components, and the structures that correspond to school classes are annotated with the corresponding class name. From [53]

while columns correspond to terms of the decomposition: specifically, the elements a_{ir} and b_{ir} relate individual nodes to components, while the elements c_{kr} associate each component with the times k it spans and can be regarded as an activity level of that component over time. As an illustration, in Fig. 3.4 we display matrix **A** for a sample decomposition of the high-resolution school social network of [33], obtained via non-negative tensor factorization. The method detects both cohesive structures corresponding to school classes and components that describe mixing patterns of the classes induced by scheduled social events such as lunch breaks [53].

Overall, this decomposition model can accommodate the description of mesoscale network structures that mix topological and temporal features in complex fashions: cohesive communities, overlapping communities, groups of links that are only active at specific times, abrupt transitions of the community structure, similar connectivity patterns at distant times, and more. The non-negativity constraints make the decomposition purely additive, and hence yield terms that are more interpretable [54] in relation to contextual information or other background knowledge about the network at hand. We notice that non-negative factorization, because of the properties summarized above, has been already proposed for community detection in static networks [55, 56] when dealing with densely overlapping communities.

We finally remark that a central challenge in designing techniques for detecting mesoscopic structures is the ability to validate the obtained results either by running the decomposition on synthetic benchmark networks or by using empirical data for which a ground truth is independently available, (e.g., the case of [53]).

3.5 Modelling Face-to-Face Interactions

The modelling activity concerning time-varying networks of contacts between individuals is quite recent, mainly because it has followed the availability of time-resolved datasets.[1] For instance, Scherrer et al. [27] have proposed a model of Markovian graph dynamics, in which each link can appear or disappear with probabilities depending on the graph state at each time: This model was tuned to reproduce detailed features of a specific dataset. Another approach consists in considering a set of agents, defining rules of interactions between these agents, and studying the statistical properties of the contact network that emerge from these “microscopic” rules. In particular, the model developed in [58, 59] considers N agents who can either be isolated or form groups. Each agent is characterized by his/her coordination number indicating the number of agents interacting with him/her, and the time at which this coordination number last changed. At each time step, an isolated agent can create a link with another isolated agent, and an agent who is part of a group can leave the group or invite an isolated agent to join it. Each such creation or deletion of links occurs with probabilities that can depend on the concerned agents’ status. Interestingly, the introduction of memory effects in the definition of these probabilities is able to generate dynamical contact networks with properties similar to the ones of empirical data sets [58, 59]. In particular, a reinforcement principle can be implemented by considering that the probability that an agent changes his/her state decreases with the time elapsed since his/her last change of state: This is equivalent to the assumption that the longer an agent is interacting in a group, the smaller is the probability that s/he will leave the group, and that the longer an agent is isolated, the smaller is the probability that s/he will form a new group. As a result, the distributions of contact durations and of time intervals between successive contacts of an individual are power-law distributed, and the aggregated contact networks display features similar to the empirically observed ones [58, 59].

Vestergaard et al. [60] consider a similar model in which, for each pair of agents, the probabilities of creation and deletion of links between agents depend on the

[1] See also [57] for more abstract modelling of adaptive networks.

time elapsed since the last evolution of the involved agents. The model considers four different "memory mechanisms" inspired by empirical evidence showing that long-term memory effects akin to self-reinforcing effects are present in the creation and disappearance of links in contact networks. For instance, more active agents tend to create more new contacts and are more attractive to other agents initiating new contacts; moreover, one of the mechanisms captures the fact that one tends to interact more often with close acquaintances. While all these memory effects are combined in empirical data, the modeling framework of [60] makes it possible to explore their individual roles both analytically and numerically. The model analysis shows how each memory mechanism by itself can lead to the emergence of some heterogeneity in the temporal characteristics of the contact networks, as quantified by broad distributions of, e.g., contact durations or inter-contact times. Interestingly however, the whole empirical phenomenology is retrieved only when all four memory mechanisms are introduced into the model.

Another model of interacting agents is put forward in [61, 62]: agents perform here random walks in two dimensions, and two agents are considered as in contact if they are within a certain distance d of each other. The main ingredient of the model is that each agent i is characterised by an intrinsic "attractiveness" $a_i \in [0, 1]$ that can be interpreted as due, for instance, to social status. When an agent is in contact with other agents, s/he can either perform a random walk step or keep the interaction by staying immobile, and the probability to maintain the contact is proportional to the attractiveness of the most attractive neighbour. Agents can also be active (i.e., can have contacts) or inactive with certain probabilities, to mimic the fact that in empirical datasets, individuals can leave the premises and stop having contacts, or come back and start again interacting. The mechanism is illustrated in Fig. 3.5 and leads to heterogeneous distributions of contact durations, of inter contact times and of aggregated contact durations very similar to empirical data (see Fig. 3.6).

3.6 Conclusions and Open Problems

Face-to-face interactions are a crucial element in the fabric of social connectivity. Their properties and their dynamics entangle many complex aspects that comprise the free agency of individuals, social coordination, human mobility and dynamics under spatial constraints, the interplay of stochasticity and deterministic activity patterns, social network structure, organizational structure, multi-layer and time-varying social networks, and more. On top of this, face-to-face interactions mediate and constrain important dynamical processes, such as information diffusion and epidemic spread of infectious agents that are transmissible during a face-to-face interaction. The research agenda on face-to-face interactions, of course, cannot be fully decoupled from domain-specific aspects, but—as it is usually the case for many complex systems—it is possible to discover and exploit summarized data representations, statistical regularities, stylised facts, and minimal models that reproduce a set of observations across diverse contexts.

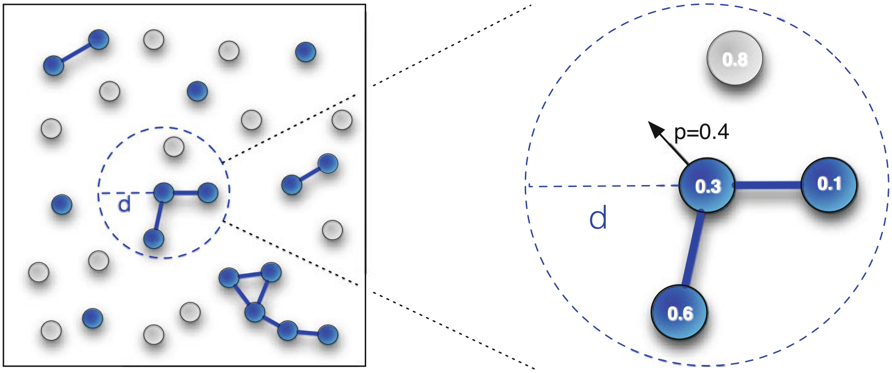


Fig. 3.5 Illustration of the mechanism of interaction of [61]. Each *circle* represents an agent. *Left*: *Dark* agents are active, *grey* (light) agents do not move nor interact. Agents interact if they are within a distance d, and are then connected by a link. *Right*: Each agent is characterized by a number representing attractiveness. The probability for the central individual to move is $p = 1 - \max(0.1, 0.6) = 0.4$ since the attractiveness of the inactive agent is not taken into account. Reprinted figure with permission from Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras, Phys. Rev. Lett. 110, 168701 (2013). Copyright 2013 by the American Physical Society [61]

The research agenda we envision, thus, starts by building an “atlas” of human contacts, which is incrementally assembled by adding map after map of human encounters, obtained by measuring face-to-face interactions in a variety of social contexts, at different points in time, at different scales, and using different proxies to assess individual interactions. The availability of these empirical datasets allows to make progress in the direction of the following goals:

- To learn which proxy is best suited to measure a given type of close-range interaction in a given context, and how different proxies relate to one another when used to quantify the same face-to-face interactions. We illustrated some of these points in Sect. 3.2.
- To uncover statistical regularities, as discussed in Sect. 3.3. The ultimate goal is not to empirically quantify all interactions in any given environment, but rather to learn what should be measured and what we do not need to measure every time.
- To design summarised data representations such as the contact matrices and aggregated networks discussed in Sect. 3.4 that, ideally, retain only the essential information and generalise well to other environments or social contexts.
- To devise minimal dynamical models, like those described in Sect. 3.5, that reproduce a set of important stylised facts and observed statistical properties under minimal assumptions. Models like these are precious to generate synthetic but realistic interaction networks, and to gain insight into the deep mechanisms that are responsible for the observed behaviors.

All of the above points are aimed at achieving parsimonious representations of the empirical data and parsimonious mathematical models for selected observables. However, it is important to remark that whereas simple generative models can

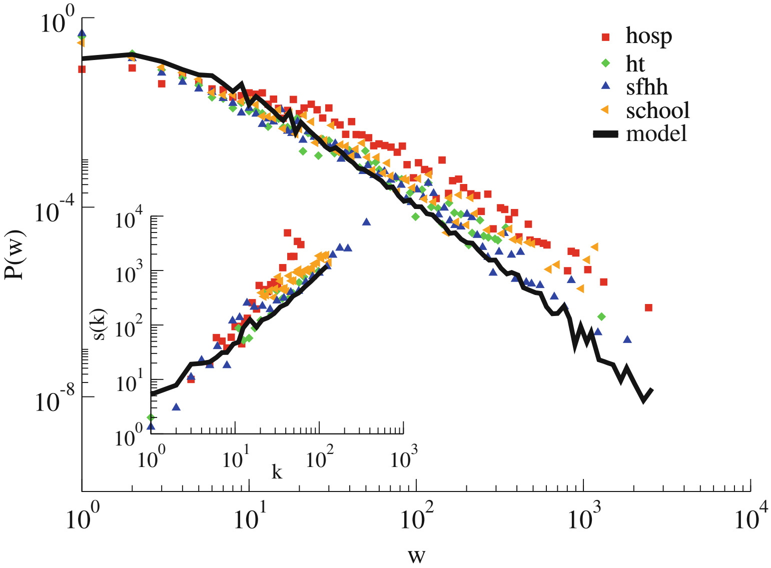


Fig. 3.6 Comparison of the model of [61] with empirical data. *Main figure*: distribution $P(w)$ of links' weights (i.e., aggregated contact durations between pairs of individuals) in the aggregated contact network. *Inset*: average strength s of nodes of degree k in the aggregated network, i.e., average total time in contact (s) of agents having had contacts with k other agents. The datasets correspond to contacts gathered by the SocioPatterns collaboration [21] in a hospital ("hosp"), conferences ("ht" and "sfhh") and in a primary school ("school"). Reprinted figure with permission from Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras, Phys. Rev. Lett. 110, 168701 (2013). Copyright 2013 by the American Physical Society [61]

reproduce some or even many of the observed statistical distributions, the rich multi-level structure that is visible in face-to-face interaction data cannot emerge from such models: when aiming at realistic scenarios, both in a data mining perspective and in a mathematical modeling perspective, there are specificities of the system at hand that we cannot ignore. Because of this, is it important to develop and validate techniques for detecting structures and correlated activity patterns of face-to-face interactions, as discussed in Sect. 3.4.3. Many highly relevant ideas and methods rooted in the domains of data mining and machine learning can be brought to bear on network science. The design of mathematical models that naturally incorporate the observed longitudinal structures, mesoscopic structures, and correlated activity patterns is an outstanding problem.

In this chapter we often discussed, explicitly or implicitly, epidemic processes over face-to-face interaction networks. This disciplinary focus arises from two reasons:

- A need for simplicity. Biological contagion processes unfold over face-to-face interaction in a mechanistic fashion. To describe their dynamics we need not take into account complex cultural attributes of the individuals that may play a crucial role when dealing with, e.g., information spreading over face-to-face encounters. Moreover, when dealing with airborne infectious agents, the network of close-range encounters in space is generally regarded to be the relevant network for the epidemic process. The same does not hold for, e.g., information diffusion, as face-to-face encounters are just one of many information exchange modalities among humans and the relevant network structure for this process is likely to be a multi-layer network.

- Moving from understanding to control. Controlling and mitigating epidemic processes on face-to-face interaction networks is a challenge that combines data-intensive approaches and mathematical models, with a potentially huge real-world impact. Nosocomial infections alone are a huge burden, both financially and in terms of individual health outcomes, and they occur in a context, the hospital, where it is comparatively easy to measure face-to-face interactions and put them in relation with infection surveillance and microbiological data. In general, there is an opportunity to use knowledge on high-resolution social networks to design mitigation strategies and targeted interventions. In [63], for example, we investigated targeted class-closure strategies for mitigating the epidemic of a flu-like disease in schools.

Despite the recent important advances that we have in part described in this chapter, many other open problems and challenges remain [64]. They include further measurements of face-to-face interactions at different scales and in different contexts, with in particular the comparison and integration of different measurement strategies and the development of means to compensate for missing data due to sampling issues and to the finiteness of the population studied. A crucial challenge, in the context of understanding infectious disease dynamics over face-to-face contact networks, regards also the combination of contact data with virological data to better understand the links between contacts and infection events and to better assess the relative importance of different routes of transmission of various infectious diseases. Another important open problem lies in measuring, understanding, and modeling the reactive aspects of social contact in relation to disease status. This entangles the biological contagion dynamics and behavioural/cultural aspects, greatly increasing the complexity of the dynamics [10, 65, 66].

Let us finally note that, in order to make progress in the research agenda described in this section, continued data gathering efforts using various strategies and in contexts as diverse as possible remain essential, as well as the availability of the corresponding datasets for the research community [21].

Acknowledgements This work is partially supported by the Lagrange Project of the ISI Foundation funded by the CRT Foundation to AB, and CC, by the Q-ARACNE project funded by the Fondazione Compagnia di San Paolo to CC, by the HarMS-flu project (ANR-12-MONU-0018) funded by the French ANR to AB, and by the FET Multiplex Project (EU-FET-317532) funded by the European Commission to AB and CC.

References

1. Read, J. M., Edmunds, W. J., Riley, S., Lessler, J., & Cummings, D. A. T. (2012). Close encounters of the infectious kind: Methods to measure social mixing behaviour. *Epidemiology and Infection, 140*, 2117–2130.
2. Bernard, H., Fischer, R., Mikolajczyk, R. T., Kretzschmar, M., & Wildner, M. (2005). Nurses contacts and potential for infectious disease transmission. *Emerging Infectious Diseases, 15*(9), 1438–1444.
3. Edmunds, W. J., O'callaghan, C. J., & Nokes, D. J. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London B, 264*, 949–957.
4. Read, J. M., Eames, K. T. D., & Edmunds, W. J. (2008). Dynamic social networks and the implications for the spread of infectious disease. *Journal of the Royal Society Interface, 5*, 1001–1007.
5. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine, 5*, e74.
6. Mikolajczyk, R. T., Akmatov, M. K., Rastin. S., & Kretzschmar, M. (2008). Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiology and Infection, 136*(6), 813–822.
7. Danon, L., House, T. A., Read, J. M., & Keeling, M. J. (2012). Social encounter networks: Collective properties and disease transmission. *Journal of the Royal Society Interface 9*, 2826–2833.
8. Danon, L., Read, J. M., House, T. A., Vernon, M. C., & Keeling, M. J. (2013). Social encounter networks: Characterizing Great Britain. *Proceedings of the Royal Society B, 280*, 20131037.
9. Conlan, A. J. K., Eames, K. T. D., Gage, J. A., von Kirchbach, J. C., Ross, J. V., Saenz, R. A., et al. (2011). Measuring social networks in British primary schools through scientific engagement. *Proceedings of the Biological Sciences, 278*, 1467–1475.
10. Van Kerckhove, K., Hens, N., Edmunds, W. J., & Eames, K. T. D. (2013). The impact of illness on social networks: implications for transmission and control of influenza. *American Journal of Epidemiology, 178*, 1655–1662.
11. Smieszek, T., Burri, E. U., Scherzinger, R., & Scholz, R. W. (2012). Collecting close-contact social mixing data with contact diaries: Reporting errors and biases. *Epidemiology and Infection, 140*, 744–752.
12. Smieszek, T., Barclay, V. C., Seeni, I., Rainey, J. J., Gao, H., Uzicanin, A., et al. (2014). How should social mixing be measured? Comparing survey- and sensor-based methods. *BMC Infectious Diseases, 14*, 136.
13. Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature, 429*, 180–184.
14. Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F. C., Zagheni, E., Merler, S., et al. (2010). Little Italy: An agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological data. *PLoS Computational Biology, 6*(12), e1001021.
15. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., & Merler, S. (2012). Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS Computational Biology, 8*(9), e1002673.
16. Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J., & Diot, C. (2005). Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-Tolerant Networking* (pp. 244–251), Philadelphia, Pennsylvania, USA.
17. O'Neill, E., Kostakos, V., Kindberg, T., Fatah gen Schieck, A., Penn, A., Fraser, D. S., et al. (2006). Instrumenting the city: Developing methods for observing and understanding the digital cityscape. *Lecture Notes in Computer Science, 4206*, 315–322.
18. Pentland, A. (2008). *Honest signals: How they shape our world*. Cambridge, MA: MIT Press.

19. Salathé, M., Kazandjieva, M., Lee, J. W., Levis, P., Feldman, M. W., & Jones, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences, 107*(51), 22020–22025.
20. Hashemian, M., Stanley, K., & Osgood, N. (2010). Flunet: Automated tracking of contacts during flu season. In *Proceedings of the 6th International Workshop on Wireless Network Measurements* (pp. 557–562), Avignon, 1–3 June 2010.
21. http://www.sociopatterns.org/.
22. Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. -F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS One, 5*, e11596.
23. Hornbeck, T., Naylor, D., Segre, A. M., Thomas, G., Herman, T., & Polgreen, P. M. (2012). Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *The Journal of Infectious Diseases, 206*, 1549.
24. Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., et al. (2014). Measuring large-scale social networks with high resolution. *PLoS One, 9*(4), e95978.
25. Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports, 519*, 97–125.
26. Karagiannis, T., Le Boudec, J. -Y., & Vojnovic, M. (2007). Power law and exponential decay of inter contact times between mobile devices. In *Mobicom 07* (p. 183).
27. Scherrer, A., Borgnat, P., Fleury, E., Guillaume, J. -L., & Robardet, C. (2008). Description and simulation of dynamic mobility networks. *Computer Networks, 52*, 2842.
28. Barrat, A., Cattuto, C., Tozzi, A. E., Vanhems, P., & Voirin, N. (2014). Measuring contact patterns with wearable sensors: Methods, data characteristics and applications to data-driven simulations of infectious diseases. *Clinical Microbiology and Infection, 20*, 10–16.
29. Barabási, A. -L. (2010). *Bursts: The hidden pattern behind everything we do*. London: Dutton Adult.
30. Fournet, J., & Barrat, A. (2014). Contact patterns among high-school students. *PLoS One, 9*(9), e107878.
31. Isella, I., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J. F., & Van den Broeck, W. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology, 271*, 166–180.
32. www.gephi.org.
33. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., et al. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS One, 6*, e23176.
34. Génois, M., Vestergaard, C., Fournet, J., Panisson, A., Bonmarin, I., & Barrat, A. (2015). Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. Network Science. arXiv:1409.7017.
35. Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van den Broeck, W., et al. (2011). Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One, 6*, e17144.
36. Vanhems, P., Barrat, A., Cattuto, C., Pinton, J. -F., Khanafer, N., Régis, C., et al. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One, 8*(9), 73970.
37. Matson, J. (2012). RFID tags track possible outbreak pathways in the hospital. *Scientific American, 307*. http://www.scientificamerican.com/article/graphic-science-rfids-tags-track-possible-outbreak-pathways-in-hospital/
38. Machens, A., Gesualdo, F., Rizzo, C., Tozzi, A. E., Barrat, A., & Cattuto, C. (2013). An infectious disease model on empirical networks of human contact: Bridging the gap between dynamic network data and contact matrices. *BMC Infectious Diseases, 13*, 185.
39. Eames, K. T. D., Tilston, N. D., & Edmunds, W. J. (2011). The impact of school holidays on the social mixing patterns of school children. *Epidemics, 3*, 103.
40. Gauvin, L., Panisson, P., Cattuto, C., & Barrat, A. (2013). Activity clocks: Spreading dynamics on temporal networks of human contact. *Scientific Reports, 3*, 3099.

41. Yang, J., McAuley, J., & Leskovec, J. (2014). Detecting cohesive and 2-mode communities indirected and undirected networks. In *Proceedings of 7th ACM International Conference on Web Search and Data Mining (WSDM2014)* (pp. 323–332).
42. Chen, Y., Kawadia, V., & Urgaonkar, R. (2013). Detecting overlapping temporal community structure in time-evolving networks. arXiv:1303.7226.
43. Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences, 101*(1), 5249–5253.
44. Greene, D., Doyle, D., & Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM2010)* (pp. 176–183).
45. Bassett, D. S., Porter, M. A., Wymbs, N. F., Grafton, S. T., Carlson, J. M., & Mucha, P. J. (2013). Robust detection of dynamic community structure in networks. *Chaos, 23*(1), 013142.
46. Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J. -P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science, 328*, 876–878.
47. Ronhovde, P., Chakrabarty, S., Hu, D., Sahu, M., Sahu, K. K., Kelton, K. F., et al. (2012). Detection of hidden structures for arbitrary scales in complex physical systems. *Scientific Reports, 2*, 329.
48. De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., et al. (2013). Mathematical formulation of multi-layer networks. *Physical Review X, 3*, 041022.
49. Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review, 51*(3), 455–500.
50. Cichocki, A., Phan, A. H., & Zdunek, R. (2009). *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. New York: Wiley.
51. Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of 22nd International Conference on Machine Learning (ICML2005)* (pp. 792–799).
52. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788–791.
53. Gauvin, L., Panisson, A., & Cattuto, C. (2014). Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach. *PLoS One, 9*(1), e86028.
54. Nickel, M., Tresp, V., & Kriegel, H. P. (2011). A three-way model for collective learning on multi-relational data. In *Proceedings of 28th International Conference on Machine Learning (ICML2011)* (pp. 809–816).
55. Yang, J., & Leskovec, J. (2013). Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of 6th ACM International Conference on Web Search and Data Mining (WSDM2013)* (pp. 587–596)
56. Wang, F., Li, T., Wang, X., Zhu, S., & Ding, C. (2011). Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery, 22*(3), 493–521.
57. Gross, T., & Sayama, H. (Eds.). (2008). *Adaptive networks: Theory, models and applications*. Springer/NECSI Studies on Complexity Series. Berlin: Springer.
58. Stehlé, J., Barrat, A., & Bianconi, G. (2010). Dynamical and bursty interactions in social networks. *Physical Review E, 81*, 035101(R).
59. Zhao, K., Stehlé, J., Bianconi, G., & Barrat, A. (2011). Social network dynamics of face-to-face interactions. *Physical Review E, 83*, 056109.
60. Vestergaard, C. L., Génois, M., & Barrat, A. (2014). How memory generates heterogeneous dynamics in temporal networks. *Physical Review E*, *90*, 042805.
61. Starnini, M., Baronchelli, A., & Pastor-Satorras, R. (2013). Modeling human dynamics of face-to-face interaction networks. *Physical Review Letters, 110*, 168701.
62. Starnini, M., Baronchelli, A., & Pastor-Satorras, R. (2014). Model reproduces individual, group and collective dynamics of human contact networks. arXiv:1409.0507.
63. Gemmetto, V., Barrat, A., & Cattuto, C. (2014). Mitigation of infectious disease at school: Targeted class closure vs school closure. *BMC Infect. Dis. 14*, 695

64. Eames, K., Bansal, S., Frost, S., & Riley, S. (2015). Six challenges in measuring contact networks for use in modelling. *Epidemics, 10*, 72–77.
65. Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports,1*, 62.
66. Perra, N., Balcan, D., Gonçalves, B., & Vespignani, A. (2011). Towards a characterization of behavior-disease models. *PLoS One, 6*(8), e23084.

Chapter 4
Modeling and Predicting Human Infectious Diseases

Nicola Perra and Bruno Gonçalves

Abstract The spreading of infectious diseases has dramatically shaped our history and society. The quest to understand and prevent their spreading dates more than two centuries. Over the years, advances in Medicine, Biology, Mathematics, Physics, Network Science, Computer Science, and Technology in general contributed to the development of modern epidemiology. In this chapter, we present a summary of different mathematical and computational approaches aimed at describing, modeling, and forecasting the diffusion of viruses. We start from the basic concepts and models in an unstructured population and gradually increase the realism by adding the effects of realistic contact structures within a population as well as the effects of human mobility coupling different subpopulations. Building on these concepts we present two realistic data-driven epidemiological models able to forecast the spreading of infectious diseases at different geographical granularities. We conclude by introducing some recent developments in diseases modeling rooted in the big-data revolution.

4.1 Introduction

Historically, the first quantitative attempt to understand and prevent infectious diseases dates back to 1760 when Bernoulli studied the effectiveness of inoculation against Smallpox [1]. Since then, and despite some initial lulls [2], an intense research activity has developed a rigorous formulation of pathogens' spreading. In this chapter, we present different approaches to model and predict the spreading of infectious diseases at different geographical resolutions and levels of detail. We focus on airborne illnesses transmitted from human to human. We are the carriers of such diseases. Our contacts and mobility are the crucial ingredients to understand

N. Perra (✉)
Northeastern University, Boston, MA, USA
e-mail: n.perra@neu.edu

B. Gonçalves
Aix Marseille Université, Université de Toulon, CNRS, CPT, UMR 7332,
13288 Marseille, France
e-mail: bgoncalves@gmail.com

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_4

and model their spreading. Interestingly, the access to large-scale data describing these human dynamics is a recent development in epidemiology. Indeed, for many years only the biological roots of transmission were clearly understood, so it is not surprising that classical models in epidemiology neglect realistic human contact structures or mobility in favor of more mathematically tractable and simplified descriptions of unstructured populations. We start our chapter with these modeling approaches that offer us an intuitive way of introducing the basic quantities and concepts in epidemiology.

Advances in technology are resulting in increased data on human dynamics and behavior. Consequently, modeling approaches in epidemiology are gradually becoming more detailed and starting to include realistic contact and mobility patterns. In Sects. 4.3 and 4.4 we describe such developments and analyze the effects of heterogeneities in contact structures between individuals and between cities/subpopulations.

With these ingredients in hand we then introduce state-of-the-art data-driven epidemiological models as examples of the modern capabilities in disease modeling and predictions. In particular, we consider GLEAM [3, 4], EpiSims [5], and FLuTE [6]. The first model is based on the metapopulation framework, a paradigm where the inter-population dynamics is modeled using detailed mobility patterns, while the intra-population dynamics is described by coarse-grained techniques. The other tools are, instead, agent-based model (ABM). This class of tools guarantees a very precise description of the unfolding of diseases, but need to be fed with extremely detailed data and are not computationally scalable. For these reasons their use so far has been limited to the study of disease spread within a limited numbers of countries. In comparison, metapopulation models include a reduced amount of data, while the approximated description of internal dynamics allows scaling the simulations to global scenarios.

Interestingly, the access to large-scale data on human activities has also started a new era in epidemiology. Indeed, the big-data revolution naturally results in real time data on the health related behavior of individuals across the globe. Such information can be obtained with tools that either require the active participation of individuals willing to share their health status or that is mined silently from individuals' health related data. Epidemiology is becoming digital [7, 8]. In Sect. 4.6 we introduce the basic concepts, approaches, and results in this new field of epidemiology. In particular, we describe tools that, using search queries, microblogging, or other web-based data, are able to predict the incidence of a wide range of diseases two weeks ahead respect to traditional surveillance.

4.2 Basic Concepts in Mathematical Epidemiology

Epidemic models divide the progression of the disease into several states or compartments, with individuals transitioning compartments depending on their health status. The natural history of the disease is represented by the type of compartments and the transitions from one to another, and naturally varies from disease to disease. In some illnesses, Susceptible individuals (S) become infected and Infectious when coming in contact with one or more Infectious (I) persons and remain so until their death. In this case the disease is described by the so-called *SI* (susceptible-infected) model. In other diseases, as is the case for some sexual transmitted diseases, infected individuals recover becoming again Susceptible to the disease. These diseases are described by the *SIS* (susceptible-infected-susceptible) model. In the case of influenza like illnesses (ILI), on the other hand, infected individuals Recover becoming immune to future infections from the same pathogen. ILIs are described by the *SIR* (susceptible-infected-recovered) model. These basic compartments provide us with the fundamental description of the progression of an idealized infection in several general circumstances. Further compartments can be added to accurately describe more realistic illnesses such as Smallpox, Chlamydia, Meningitis, and Ebola [2, 9, 10]. Keeping this important observation in mind, here we focus on the *SIR* model.

4.2.1 Modeling Transitions Between Compartments

Epidemic models are often represented using chart such as the one seen in Fig. 4.1. Such illustrations are able to accurately represent the number of compartments and the disease's behavior in a concise and easily interpretable form. Mathematically, models can also be accurately represented as reaction equations as we will see below.

In general, epidemic models include two type of transitions, "interactive" and "spontaneous." Interactive transitions require the contact between individuals in two different compartments, while spontaneous transitions occur naturally at a fixed rate per unit time. For example, in the transition between S to I, Susceptible individuals become Infected due to the interaction with Infected individuals, i.e. $S+I \to 2I$. The transition is mediated by individuals in the compartment I, see Fig. 4.1. On the other hand, an Infectious individual can naturally recover from infection after a certain

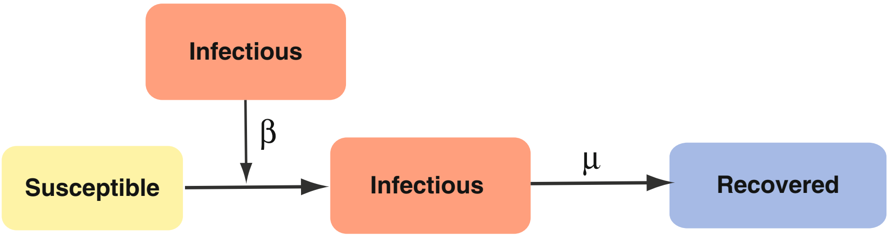


Fig. 4.1 Schematic representation of the SIR model. The transition from S to I is due to the interaction between susceptible and infectious individuals. The transition from I to R is instead spontaneous. The transition rates are β and μ, respectively

amount of time and become Recovered, i.e. $I \rightarrow R$. Individuals are considered to have a fixed recovery rate, μ, defined as the inverse of the average time τ spent in the infected compartment, $\mu = \tau^{-1}$

But how can we model the infection process? Intuitively we expect that the probability of single individual becoming infected must depend on (1) the number of infected individuals in the population, (2) the probability of infection given a contact with an infectious agent and, (3) the number of such contacts. In this section we neglect the details of who is in contact with whom and consider instead individuals to be part of a homogeneously mixed population where everyone is assumed to be in contact with everyone else (we tackle heterogeneous contacts in Sect. 4.3). In this limit, the per capita rate at which susceptible contract the disease, the force of infection λ, can be expressed in two forms depending on the type of population. In the first, often called mass-action law, the number of contacts per individual is independent of the total population size, and determined by the transmission rate β and the probability of randomly contacting an infected individual, i.e. $\lambda = \beta I/N$ (where N is the population size). In the second case, often called pseudo mass-action law, the number of contacts is assumed to scale with the population size, and the transmission rate β, i.e. $\lambda = \beta I$. Without loss of generality, in the following we focus on the first kind of contact.

4.2.2 The SIR Model

The *SIR* framework is the crucial pillar to model ILIs. Think, for example, at the H1N1 pandemic in 2009, or the seasonal flu that every year spread across the globe. The progression of such diseases, from the first encounter to the recovery, happens in matters of days. For this reason, birth and death rates in the populations can be generally neglected, i.e. $d_t N \equiv 0$ for all times t.

Let us define the fraction of individuals in the susceptible, infected, and recovered compartments as s, i, and r. The *SIR* model is then described by the following set of differential equations:

$$\begin{cases} d_t s = -s\lambda \\ d_t i = s\lambda - \mu i \\ d_t r = \mu i \end{cases} \tag{4.1}$$

where $\lambda = \beta i \equiv \beta \frac{I}{N}$ is the force of infection, and $d_t \equiv \frac{d}{dt}$. The first equation describes the infection process in a homogeneous mixed population. Susceptible individuals become infected through random encounters with Infected individuals. The second equation describes the balance between the in-flow (infection process, first term), and the out-flow (recovery process, second term) in compartment i. Finally, the third equation accounts for the increase of the recovered population due to the recovery process. Interestingly, the *SIR* dynamical equations, although apparently very simple, due to their intrinsic non-linearity cannot be solved analytically. The description of the evolution of the disease can be obtained only through numerical integration of the system of differential equations. However, crucial analytic insight on the process can be obtained for early $t \sim t_0$ and late times $t \rightarrow \infty$.

4.2.2.1 Epidemic Threshold

Under which conditions a disease starting from a small number, I_0, of individuals at time t_0 is able to spread in the population? To answer this question let us consider the early stages of the spreading, i.e. $t \sim t_0$. The equation for the infected compartment can be written as $d_t i = i(\beta s - \mu)$, indicating an exponential behavior for early times. It then follows that if the initial fraction of susceptible individuals, $s_0 = S_0/N$, is smaller than μ/β, the exponent becomes negative and the disease dies out. We call this value the epidemic threshold [11] of the *SIR* model. The fraction of susceptibles in the population has to be larger than a certain value, that depends on the disease details, in order to observe an outbreak.

Typically, the initial cluster of infected individuals is small in comparison with the population size, i.e. $s_0 \gg i_0$, or $s_0 \sim 1$. In this case, the threshold condition can be re-written as $\beta/\mu > 1$. The quantity:

$$R_0 \equiv \frac{\beta}{\mu} \tag{4.2}$$

is called the *basic reproductive number*, and is a crucial quantity in epidemiology and provides a very simple interpretation of the epidemic threshold. Indeed, the disease is able to spread if and only if each infected individual is able to infect, on average, more than one person before recovering. The meaning of R_0 is then clear: it is simply the average number of infections generated by an initial infectious seed in a fully susceptible population [10].

4.2.2.2 Disease-Free Equilibrium

For any value of $\mu > 0$, the *SIR* dynamics will eventually reach a stationary, disease-free, state characterized by $i = d_t i = 0$. Indeed, infected individuals will keep recovering until they all reach the R compartment. What is the final number of recovered individuals? Answering this apparently simple question is crucial to quantify the impact of the disease. We can tackle such conundrum dividing the first equation with the third equation in the system 4.1. We obtain $d_r s = -R_0 s$ which in turn implies $s_t = s_0 e^{-R_0 r_t}$. Unfortunately, this transcendent equation cannot be solved analytically. However, we can use it to gain some important insights on the *SIR* dynamics. We note that for any $R_0 > 1$, in the limit $t \rightarrow \infty$, we must have

$s_\infty > 0$. In other words, despite R_0, the disease-free equilibrium of an SIR model is always characterized by some finite fraction of the population in the Susceptible compartment, or, in other words, some individuals will always be able to avoid the infection. In the limit where $R_0 \sim 1$ we can obtain an approximate solution for r_∞ (or equivalently for $s_\infty = 1 - r_\infty$) by expanding $s_\infty = s_0 e^{-R_0 s_\infty}$ at the second order around $r_\infty \sim 0$. After a few basic algebraic manipulations we obtain $r_\infty = \frac{2(R_0-1)}{R_0^2}$ [9].

4.3 Beyond Homogeneous Mixing

In the previous sections we presented the basic concepts and models in epidemiology by considering a simple view of a population where individuals mix homogeneously. Although such approximation allows a simple mathematical formulation, it is far from reality. Individuals do not all have the same number of contacts, and more importantly, encounters are not completely random [12–15]. Some persons are more prone to social interactions than others, and contacts with family members, friends, and co-workers are much more likely than interactions with any other person in the population.

Over the last decade the *network framework* has been particularly effective in capturing the complex features and the heterogeneous nature of our contacts [12–16]. In this approach, individuals are represented by nodes while links represent their interactions. As described in different chapters of the book (see Chaps. 3, 6, and 10), human contacts are not heterogeneous in both number and intensity [12–15, 17] but also change over time [18]. This framework naturally introduces two timescales, the timescale at which the network connections evolve, τ_G and the inherent timescale, τ_P, of the process taking place over the network. Although the dynamical nature of interactions might have crucial consequences on the disease spreading [19–24], the large majority of results in the literature deal with one of two limiting regimens [25, 26]. When $\tau_G \gg \tau_P$, the evolution of the network of contacts is much slower than the spreading of the disease and the network can be considered as static. On the other hand, when $\tau_P \gg \tau_G$, the links are said to be annealed and changes in networks structure are much faster than the spreading of the pathogen. In both cases the two time-scales are well separated allowing for a simpler mathematical description. Here we focus on the annealed approximation ($\tau_P \gg \tau_G$) that provides a simple stage to model and understand the dynamical properties of epidemic processes. We refer the reader to Chap. 3 Face-to-Face Interactions for recent approaches that relax this time-scale separation assumption.

Let us consider a network $G(N, E)$ characterized by N nodes connected by E edges. The number of contacts of each node is described by the degree k. The degree distribution $P(k)$ characterizes the probability of finding a node of degree k. Empirical observations in many different domains show heavy-tailed degree distributions usually approximated as power-laws, i.e. $P(k) \sim k^{-\alpha}$ [12, 13].

Furthermore, human contact networks are characterized by so-called *assortative mixing*, meaning a positive correlation between the degree of connected individuals. Correlations are encoded in the conditional probability $P(k'|k)$ that a node of degree k is connected with a node of degree k' [12, 13]. While including realistic correlations in epidemic models is crucial [27–29] they introduce a wide set of mathematical challenges that are behind the scope of this chapter. In the following, we consider the simple case of uncorrelated networks in which the interdependence among degree classes is removed.

4.3.1 The SIR Model in Networks

How can we extend the *SIR* model to include heterogeneous contact structures? Here we must take a step further than simply treating all individuals the same. We start distinguishing nodes by degree while considering all vertices with the same degree as statistically equivalent. This is known as the degree block approximation and is exact for annealed networks. The quantities under study are now $i_k = \frac{I_k}{N_k}$, $s_k = \frac{S_k}{N_k}$, and $r_k = \frac{R_k}{N_k}$, where the I_k, S_k, and R_k are the number of infected, susceptible, recovered individuals in the degree class k. N_k instead describes the total number of nodes in the degree class k. The global averages are given by $i = \sum_k P(k)\, i_k$, $s = \sum_k P(k)\, s_k$, $r = \sum_k P(k)\, r_k$. Using this notation and heterogeneous mean field (HMF) theory [26], the system of differential equations (4.1) can now be written as:

$$\begin{cases} d_t s_k = -s_k \lambda_k \\ d_t i_k = s_k \lambda_k - \mu i_k \\ d_t r_k = \mu i_k \end{cases} \tag{4.3}$$

The contact structure introduces a force of infection function of the degree. In particular, $\lambda_k = \gamma k \Theta_k$ where γ is the rate of infection per contact, i.e. $\beta = \gamma k$, and Θ_k describes the density of infected neighbors of nodes in the degree class k. Intuitively, this density is a function of the conditional probability that a node k is connected to any node k' and proportional to the number of infected nodes in each class k': $\Theta_k = \sum_{k'} P(k'|k)\, i_{k'}$. In the simple case of uncorrelated networks the probability of finding a node of degree k' in the neighborhood of a node in degree class k is independent of k. In this case $\Theta_k = \Theta = \sum_{k'} (k'-1)\, P(k')\, i_{k'}/\langle k \rangle$ where the term $k' - 1$ is due to the fact that at least one link of each infected node points to another infected vertex [15].

4.3.1.1 Epidemic Threshold

In order to derive the epidemic threshold let us consider the early time limit of the epidemic process. As done in Sect. 4.2.2.1 let us consider that at $t \sim t_0$ the population is formed mostly by susceptible individuals. In the present scenario this implies $s_k \gg i_k$ and $r_k \sim 0 \; \forall k$. The equation for the infected compartment then becomes $d_t i_k = \gamma k \Theta - \mu i_k$. Multiplying both sides for $P(k)$ and summing over all values of k we obtain $d_t i = \gamma \langle k \rangle \Theta - \mu i$. In order to understand the behavior of i around t_0 let us consider an equation built by multiplying both sides of the last equation by $(k-1) P(k) / \langle k \rangle$ and summing over all degree classes. We obtain $d_t \Theta = \gamma (\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}) \Theta - \mu \Theta$. The fraction of infected individuals in each value of k will increase if and only if $d_t \Theta > 0$. This condition is verified when [15]:

$$R_0 \equiv \frac{\beta}{\mu} > \frac{\langle k \rangle^2}{\langle k^2 \rangle - \langle k \rangle} \tag{4.4}$$

giving us the epidemic threshold of an *SIR* process unfolding on an uncorrelated network.

Remarkably, due to their broad-tailed nature, real contact networks display fluctuations in the number of contacts (large $\langle k^2 \rangle$) that are significantly larger than the average degree $\langle k \rangle$ resulting in very small thresholds. Large degree nodes (hubs) facilitate an extremely efficient spreading of the infection by directly connecting many otherwise distant nodes. As soon as the hubs become infected diseases are able to reach a large fraction of the nodes in the network. Real interaction networks are extremely fragile to disease spreading. While this finding is somehow worrisome, it suggests very efficient strategies to control and mitigate the outbreaks. Indeed, hubs are *central* nodes and play a crucial role in the network connectivity [12] and by vaccinating a small fraction of them one is able to quickly stop the spread of the disease and protect the rest of the population. It is important to mention that in realistic settings the knowledge of the networks' structure is often limited. Hubs might not be easy to easily known and other indirect means must be employed. Interestingly, the same feature of hubs that facilitates the spread of the disease also allows for their easy detection. Since high degree nodes are connected to a large number of smaller degree nodes, one may simply randomly select a node, A, from the network and follow one of its links to reach another node, B. With high probability, node B has higher degree than A and is likely a hub. This effect became popularized as the *friend paradox*: on average your friends have more friends than you do [12]. Immunizing node B is then much more effective than immunizing node A. Remarkably, as counter-intuitive as this methodology might seem, it works extremely well even in the case of quickly changing networks [30–32].

4.4 Metapopulation Models

The next step in the progression towards more realistic modeling approaches is to consider the internal structure of the nodes. If each node in the network represents a homogeneously mixed sub-population instead of a single individual and we consider the edges to represent interactions or mobility between the different subpopulations, then we are in the presence of what is known as meta-population. This concept was originally introduced by R. Levins in 1969 [33] for the study of geographically extended ecological populations where each node represents one of the ecological niches where a given population resides.

The metapopulation framework was later extended for use in epidemic modeling by Sattenspiel in 1987. In a landmark paper [34] Sattenspiel considered two different types of interactions between individuals, local ones occurring within a given node, and social ones connecting individuals originating from different locations on the network. This idea was later expanded by Sattenspiel and Dietz to include the effects of mobility [35] and thus laying the foundations for the development of epidemic models at the global scale.

Metapopulation epidemic models are extremely useful to describe particle reaction-diffusion models [36]. In this type of model each node is allowed to have zero or more individuals that are free to diffuse among the nodes constituting the network. In our analysis, as done in the previous section, we follow the HMF approach and consider all nodes of degree k to be statistically equivalent and write all quantities in terms of the degree k. To start, let us define the average number of individuals in a node of degree k to be $W_k = \frac{1}{N_k}\sum_i W_i \delta\left(k_i - k\right)$, where N_k is the number of nodes with degree k and the sum is taken over all nodes i. The mean field dynamical equation describing the variation of the average number of individuals in a node of degree k is then:

$$\frac{\mathrm{d}W_k\left(t\right)}{\mathrm{d}t} = -p_k W_k\left(t\right) + k\sum_{k'} P\left(k'|k\right) p_{k'k} W_{k'}\left(t\right) \tag{4.5}$$

where p_k and $p_{kk'}$ represent, respectively, the rate at which particles diffuse out of a node of degree k and diffuse from a node of degree k to one of degree k'.

With these definitions, the meaning of each term of this equation becomes intuitively clear: the negative term represents individuals leaving the node, while the positive term accounts for individuals originating from other nodes arriving at this particular class of node. The conditional probability $P\left(k'|k\right)$ encodes all the topological correlations of the network. By imposing that the total number of particles in the system remains constant, we obtain:

$$p_k = k\sum_{k'} P\left(k|k'\right) p_{kk'} \tag{4.6}$$

that simply states that the number of particles arriving at nodes of degree k' coming from nodes of degree k must be the same as the number of particles leaving nodes of degree k. The probabilities p_k and $p_{kk'}$ encode the details of the diffusion process [37]. In the simplest case, the rate of movement of individuals is independent of the degree of their origin $p_k = p$ for all values of the degree.

Furthermore, if individuals that are moving simply select homogeneously among all of their connections, then we have $p_{kk'} = p/k$. In this case, the diffusion process will reach a stationary state when:

$$W_k = \frac{k}{\langle k \rangle} \overline{W} \tag{4.7}$$

where $\overline{W} = W/N$, W is the total number of walkers in the system, and N the total number of nodes. The simple linear relation between W_k and k serves as a strong reminder of the importance of network topology. Nodes with higher degree will acquire larger populations of particles while nodes with smaller degrees will have proportionally smaller populations. However, even in the steady state, the diffusion process is ongoing, so individuals are continuously arriving and leaving any given node but are doing so in a way that maintains the total number of particles in each node constant.

In more realistic settings, the traffic of individuals between two nodes is function of their degree [37]:

$$p_{kk'} = w_0 \frac{(kk')^{\theta}}{T_k} \tag{4.8}$$

In this expression θ modulates the strength of the diffusion flow between degree classes (empirical values are in the range $-0.5 \le \theta \le 0.5$ [3]), where w_0 is a constant and $T_k = w_0 \langle k^{1+\theta} \rangle / \langle k \rangle$ is the proper normalization ensured by the condition in Eq. (4.6). In these settings, the diffusion process reaches a stationary state when:

$$W_k = \frac{k^{1+\theta}}{\langle k^{1+\theta} \rangle} \overline{W} \tag{4.9}$$

Note that for $\theta = 0$ this solution coincides with the case of homogeneous diffusion [Eq. (4.7)].

Combining this diffusion process with the (epidemic) reaction processes described above we finally obtain the full reaction-diffusion process. To do so we must simply write Eq. (4.5) for each state of the disease (e.g., Susceptible, Infectious, and Recovered for a simple *SIR* model) and couple the resulting equations using the already familiar epidemic equations. The full significance of Eq. (4.7) now becomes clear: nodes with higher degree have higher populations and are visited by more travelers, making them significantly more likely to also receive an infected individual that can act as the seed of a local epidemic.

In a metapopulation epidemic context we must then consider two separate thresholds, the basic reproductive ratio, R_0, that determines whether or not a disease can spread within one population (node) and a critical diffusion rate, p_c, that determines if individual mobility is sufficiently large to allow the disease to spread from one population to another. It is clear that if $p = 0$ particles are completely unable to move from one population to another so the epidemic cannot spread across subpopulations and that if $p = 1$ all individuals are in constant motion and the disease will inevitably spread to every subpopulation on the network with a transition occurring at some critical value p_c.

In general, the critical value p_c cannot be calculated analytically using our approach as it depends non-trivially on the detailed structure of the network and the fluctuations of the diffusion rate of single individuals. However, in the case of uncorrelated networks a closed solution can be easily found for different mobility patterns. Indeed, in the case where the mobility is regulated by Eq. (4.8) we obtain:

$$p_c = \frac{1}{\overline{W}} \frac{\langle k^{1+\theta} \rangle^2}{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle} \frac{\mu R_0^2}{2(R_0 - 1)^2} \tag{4.10}$$

Interestingly, the critical value of p is inversely proportional to the degree heterogeneity in the network, so that broad tailed networks have very low critical values. This simple fact explains why simply restricting travel between populations is a highly ineffective way to prevent the global spread of an epidemic.

The mobility patterns considered so far are so-called Markovian: individuals move without remembering where they have been nor they have a home where they return to after each trip. Although this is a rough approximation of individuals behavior, Markovian diffusion patterns are allowed to analytically describe the fundamental dynamical properties of many systems. Recently, new analytic results have been proposed for non-Markovian dynamics that include origin-destination matrices and realistic travel routes that follow shortest paths [38]. In particular, the threshold within such mobility schemes reads as:

$$p_c = \frac{1}{\overline{W}} \frac{\langle k^{\eta} \rangle}{\langle k^{1+\eta} \rangle} \frac{\langle k \rangle \mu R_0^2}{2\,(R_0 - 1)^2} \tag{4.11}$$

The exponent η, typically close to 1.5 in heterogeneous networks, emerges from the shortest paths routing patterns [38]. Interestingly, for values of $\theta \leq 0.2$, fixing $\eta = 1.5$, p_c in the case of Markovian mobility patterns is larger than the critical value in a system subject to non-Markovian diffusion. The presence of origin-destination matrices and shortest paths mobility lower the threshold facilitating the global spreading of the disease. Instead, for values of $\theta > 0.2$ the contrary is true.

In these models the internal contacts rate is considered constant across each subpopulation. Interestingly, recent longitudinal studies on phone networks [39] and Twitter mention networks [40] point to the evidence that contacts instead scale super-linearly with the subpopulation sizes. Considering the heterogeneity in population sizes observed in real metapopulation networks, the scaling behavior entails deep consequence in the spreading dynamics. A recent study generalized the metapopulation framework considering such observations. Interestingly, the critical mobility thresholds, in the case of mobility patterns described by Eq. (4.8), changes significantly being lowered by such scaling features of human contacts [40].

Despite their simplicity, metapopulation models are extremely powerful tools in large scale study of epidemics. They easily lend themselves to large scale numerical stochastic simulations where the population and state of each node can be tracked and analyzed in great detail and multiple scenarios as well as interventions can be tested.

The state of the art in the class of metapopulation approaches is currently defined by the global epidemic and mobility model (*GLEAM*) [3, 4]. *GLEAM* integrates worldwide population estimates [41, 42] with complete airline transportation and commuting databases to create a world wide description of mobility around the world that can then be used as the substrate on which the epidemic can spread. *GLEAM* divides the globe into 3362 transportation basins. Each basin is defined empirically around an airport and the area of the basin is determined to be the region within which residents would likely use that airport for long distance travel. Each basin represents a major metropolitan area such as New York, London, or Paris. Information about all civilian flights can be obtained from the International Air Transportation Association (*IATA*) [43] and the Official Airline Guide (*OAG*) [44] that are responsible for compiling up-to-date databases of flight information that airlines use to plan their operations. By connecting the population basins with the direct flight information from these databases we obtain the network that acts as a substrate for the reaction diffusion process.

While most human mobility does not take place in the form of flights, the flight network provides the fundamental structure for long range travel that explains how diseases such as SARS [45], Smallpox [46], or Ebola [47] spread from country to country. To capture the finer details of within country mobility further information must be considered. GLEAM uses census information to create a commuting network at the basin level that connects neighboring metropolitan areas proportionally to the number of people who live in one are but work in the other.

Short-term short-distance mobility such as commuting is fundamentally different from medium-term long-distance airline travel. In one case, the typical timescale is work-day ($8h$) while in the other it is 1 day. This timescale difference is taken into account in *GLEAM* in an effective, mean-field, manner instead of explicitly through a reaction process such as the one described above. This added layer is the final piece of the puzzle that brings the whole together and allows *GLEAM* to describe accurately the spread from one country to the next but also the spread happening within a given country [48].

In Fig. 4.2 we illustrate the progression in terms of detail that we have undergone since our initial description of simple homogeneously mixed epidemic models in a single population. With all these ingredients in place we have a fine grained description of mobility on a world wide scale on top of which we can finally build an epidemic model.

Within each basin, *GLEAM* still uses the homogeneous mixing approximation. This assumption is particularly suited for diseases that spread easily from person to person through airborne means such as ILI. *GLEAM* describes influenza through an *SEIR* model as illustrated in Fig. 4.3. *SEIR* models are a modification of the *SIR* model described above that includes a further compartment, Exposed, to represent

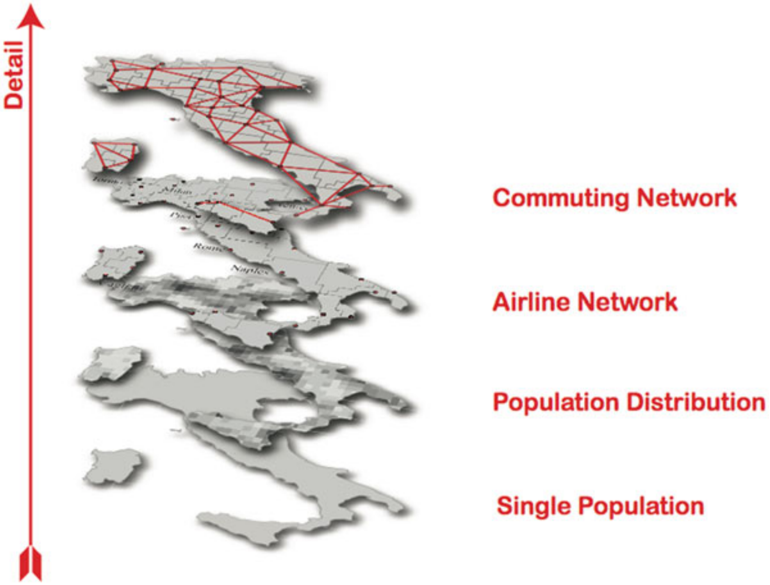


Fig. 4.2 The multilayer structure of *GLEAM*. Each layer increases the level of detail with respect to the previous ones

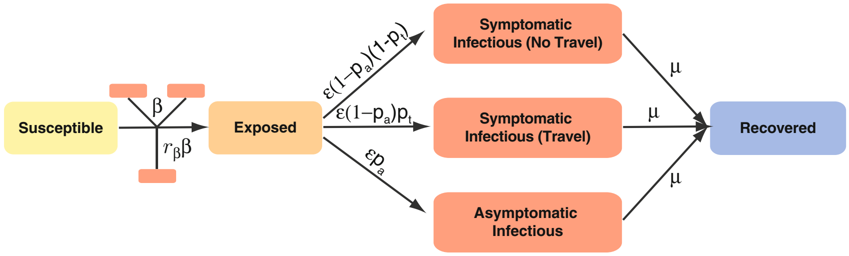


Fig. 4.3 SEIR Epidemic structure used in GLEAM

individuals in the incubation phase of the disease that are already infected but not yet Infectious. *GLEAM* further expands on this model by distinguishing three classes of Infectious individuals based on the severity of the disease. One third of the infectious individuals are asymptotic individuals do not display any symptoms and continue to behave normally while having an infectiousness reduced by a factor $r_\beta = 0.5$. Of the remaining symptomatic individuals, one half is sick enough to decide to not travel or commute while the remaining half continue to travel normally.

Despite their apparent complexity, large scale models such as *GLEAM* are controlled by just a small number of parameters and ultimately, it's the proper setting of these few parameters that is responsible for the proper calibration of the model and validity of the results obtained. Most of the disease and mobility parameters are set directly from the literature or careful testing so that as little as possible remains unknown when it is time to apply it to a new outbreak.

GLEAM was put to the test during the 2009 H1N1 pandemic with great success. During the course of the epidemic, researchers were able to use official data as it was released by health authorities around the world. In the early days of the outbreak there was a great uncertainty about the correct value of the R_0 for the 2009/H1N1 pdm strain in circulation so a methodology to determine it had to be conceived.

One of the main advantages of epidemic metapopulation models is their computational tractability. It was this feature what proved invaluable when it came to determine the proper value of R_0. By plugging in a given set of parameters one is able to generate several hundreds or thousands of *in silico* outbreaks. Each outbreak contains information not only about the number of cases in each city or country as a function of time but also information about the time when the first case occurs within a given country. In general, each outbreak will be different due to stochasticity and by combining all outbreaks generated for a certain parameter set we can calculate the probability distribution of the arrival times. The number of times that an outbreak generated the seeding of a country, say the UK, in the same day as it occurred in reality provides us with a measure of how likely the parameter values used are. By multiplying this probability for all countries with a known arrival time we can determine the overall *Likelihood* of the simulation:

$$\mathscr{L} = \prod_c P_c\left(t_c\right) \tag{4.12}$$

where the product is taken over all countries c with known arrival time t_c and the probability distribution of arrival times, $P_c\left(t\right)$ is determined numerically for each set of input values. The set of parameters that maximizes this quantity is then the one whose values are the most likely to be correct. Using this procedure the team behind *GLEAM* determined that the mostly likely value of the basic reproductive ratio was $R_0 = 1.75$ [49], a value that was later confirmed by independent studies [50, 51].

Armed with an empirical estimate of the basic reproductive ratio for an ongoing pandemic, they then proceeded to use this value to estimate the future progression of the pandemic. Their results predicting that the full peak of the pandemic would hit in October and November 2009 were published in early September 2009 [49]. A comparison between these predictions and the official data published by the health authorities in each country would be published several years later [52] clearly confirming the validity of *GLEAM* for epidemic forecasting in real time. Indeed, the model predicted, months in advance, the correct peak week in 87 % of countries in the north hemisphere for which real data was accessible. In the rest of cases the maximum error reported has been 2 weeks. *GLEAM* can also be further extended to include age-structure [53], interventions and travel reductions.

4.5 Agent-Based Models

The next logical step in the hierarchy of large scale epidemic models is to take the description of the underlying population all the way down to the individual level with what are known as ABM. The fundamental idea behind this class of model is a deceptively simple one: treat each individual in the population separately, assigning it properties such as age, gender, workplace, residence, family structure, etc... These added details give them a clear edge in terms of detail over metapopulation models but do so at the cost of much higher computational cost.

The first step in building a model of this type is to generate a synthetic population that is statistically equivalent to the population we are interested in studying. Typically this is in a hierarchical way, first generating individual households, aggregating households into neighborhoods, neighborhoods into communities, and communities into the census tracts that constitute the country.

Generating synthetic households in a way that reproduces the census data is far from a trivial task. The exact details vary depending on the end goal of the model and the level of details desired but the household size, age, and gender of household members are determined stochastically from the empirically observed distributions and conditional probabilities. One might start by determining the size of the household by extracting from the distribution of household size of the country of interest and selecting the age and gender of the head of the household proportionally to the number of heads of households for that household size that are in each age group. Conditional on this synthetic individual we can then generate the remaining members, if any, of the household. The required conditional probability distributions and correlation tables can be easily generated [54] from high quality census data that can be found for most countries in the world. This process is repeated until enough synthetic households have been generated. Households are then aggregated into neighborhoods by selecting from the households according to the distribution of households in a specific neighborhood. Neighborhoods are similarly aggregated into communities and communities into census tracts.

Each increasing level of aggregation (from household to country) represents a decrease in the level of social contact, with the most intimate contacts occurring at the household level and least intimate ones at the census tract or country level. The next step is to assign to each individual a profession and work place. Workplaces are generated following a procedure similar to the generation of households and each employed individual is assigned a specific household. School age children are assigned a school. Working individuals are assigned to work places in a different community or census tract in a way that reflects empirical commuting patterns.

At this point, we have a fairly accurate description of where the entire population of a city or country lives and works. It is then not entirely surprising that this approach was first used to study in detail the demands imposed on the transportation system of a large metropolitan city. *TRANSIMS*,[1] the TRansportation ANalysis and SIMulation System [55], used an approach similar to the one described above to generate a synthetic population for the city of Portland, in Oregon (OR) and coupled it with a route planner that would determine the actual route taken by each individual on her way to work or school as a way of modeling the daily toll on Portland's transportation infrastructure and the effect that disruptions or modification might have in the daily lives of its population.

EpiSims [5] was the logical extension of *TRANSIMS* to the epidemic world. *EpiSim*s used the *TRANSIMS* infrastructure to generate the contact network between individuals in Portland, OR. Susceptible individuals are able to acquire the infection whenever they are in a location along with one or more infectious individuals. In this way the researchers are capable of observing as the disease spreads through the population and evaluate the effect that measures such as contact tracing and mass vaccination.

More recent approaches have significantly simplified the mobility aspect of this kind of models and simply divide each 24 h period into day time and nighttime. Individuals are considered to be in contact with other members of their workplace during the day and with other household members during the night. In recent years, modelers have successfully expanded the large scale Agent Based approach to the country [6] and even continent level [56].

As the spatial scale of the models increased further modes of long-range transportation such as flights had to be considered. These are important to determine not only the seeding of the country under consideration through importation of cases from another country but also to connect distant regions in a more realistic way. *FluTE* [6] is currently the most realistic large scale Agent-Based epidemic model of the continental United States. It considers that international seeding occurs at random in the locations that host the 15 largest international airports in the US by, each day, randomly infecting in each location a number of individuals that is proportional to the international traffic of those airports.

FluTE is a refinement of a previous model [57] and it further refines the modeling of the infectious process by varying the infectiousness of an individual over time in the *SIR* model that they consider. At the time of infection each individual is assigned one of six experimentally obtained viral load histories. Each history prescribes the individuals viral load for each day of the infectious period and the infectiousness is considered to be proportional to the viral load. Individuals may remain asymptotic for up to 3 days after infection during which their infectiousness is reduced by 50 % with respect to the symptomatic period. The total infectious period is set to 6 days regardless of the length of the symptomatic period.

Given the complexity of the model the calibration of the disease parameters in order to obtain a given value of the basic reproductive ratio, R_0 requires some finesse. Chao et al. [6] uses the definition of R_0 to determine "experimentally" its value from the input parameters. It numerically simulates 1000 instances of

[1] The source code for *TRANSIMS* can be obtained from https://www.code.google.com/p/transims/.

the epidemic caused by a single individual within a 2000 person fully susceptible community for each possible age group of the seeding individual and use it to calculate the R_0^a of each age group a. The final R_0 is defined to the average of the various R_0^a weighted by age dependent attack rate [57]. The final result of this procedure is that the value of R_0 is given by:

$$R_0 = 5.592\lambda + 0.0068 \tag{4.13}$$

where λ is the infection probability per unit contact and is given as input. *FluTE* was a pioneer in the way it completely released its source code,[2] opening the doors of a new level of verifiability in this area. It has successfully used to study the spread of influenza viruses and analyze the effect of various interventions in the Los Angeles County [58] and United States country level [6].

4.6 Digital Epidemiology

The unprecedented amount of data on human dynamics made available by recent advances technology has allowed the development of realistic epidemic models able to capture and predict the unfolding of infectious disease at different geographical scales [59]. In the previous sections, we described briefly some successful examples that have been made possible thanks to high resolution data on where we live, how we live, and how we move. Data availability has started a second golden age in epidemic modeling [60].

All models are judged against surveillance data collected by health departments. Unfortunately, due to excessive costs, and other constraints their quality is far from ideal. For example, the influenza surveillance network in the USA, one of the most efficient systems in the world, is constituted of just 2900 providers that operate voluntarily. Surveillance data is imprecise, incomplete, characterized by large backlogs, delays in reporting times, and the result of very small sample sizes. Furthermore, the geographical coverage is not homogeneous across different regions, even within the same country. For these reasons the calibration and test of epidemic models with surveillance data induce strong limitations in the predictive capabilities of such tools. One of the most limiting issues is the geographical granularity of the data. In general, information are aggregated at the country or regional level. The lack of ground truth data at smaller scales does not allow a more precise selection and training of realistic epidemic models.

How can we lift such limitations? Data, data and more data is again the answer. At the end of 2013 almost 3 billion of people had access to the Internet while almost 7 billion are phone subscribers, around 20 % of which are actively using smartphones. The explosion of mobile usage boosted also the activity of social

[2] http://www.cs.unm.edu/~dlchao/flute/.

media platforms such as Facebook, Twitter, Google+ etc. that now count several hundred million active users that are happy to share not just their thoughts, but also their *GPS* coordinates. The incredible amount of information we create and access contain important epidemiologically relevant indicators. Users complaining about catching a cold before the weekend on Facebook or Twitter, searching for symptoms of particular diseases on search engines, or Wikipedia, canceling their dinner reservations on online platforms like OpenTable are just few examples. An intense research activity, across different disciplines, is clearly showing the potential, as well as the challenges and risks, of such digital traces for epidemiology [61]. We are at the dawn of the digital revolution in epidemiology [7, 8]. The new approach allows for the early detection of disease outbreaks [62], the real time monitoring of the evolution of a disease with an incredible geographical granularity [63–65], the access to health related behaviors, practices and sentiments at large scales [66, 67], inform data-driven epidemic models [68, 69], and development of statistical based models with prediction power [67, 70–78].

The search for epidemiological indicators in digital traces follows two methodologies: active and passive. In active data collection users are asked to share their health status using apps and web-based platforms [79]. Examples are *influenzanet* that is available in different European countries [64], and *Flu near you* in the USA [65] that engage tens of thousands of users that together provide the information necessary for the creation of interactive maps of ILI in almost real time. In passive data collection, instead, information about individuals health status is mined from other available sources that do not require the active participation of users. News articles [63], queries on search engines [74], posts on online social networks [67, 70–73], page view counts on Wikipedia [75, 76] or other online/offline behaviors [77, 78] are typical examples. In the following, we focus on the prototypical, and most famous, method of digital epidemiology, Google Flu Trends (GFT) [80], while considering also other approaches based on Twitter and Wikipedia data.

4.6.1 Social Media Based Epidemic Models

GFT is by far the most famous model in digital epidemiology. Launched in November 2008 together with a Nature paper [80] describing its methodology, it has continuously made predictions on the course of seasonal influenza in 29 countries around the world.[3] The method used by *GFT* is extremely simple. The percentage of *ILI* visits, a typical indicator used by surveillance systems to monitor the unfolding of the seasonal flu, is estimated with a linear model based on search engine queries. This approach is general, and used in many different fields of Science. A quantity of interest, in this case the percentage of ILI visits P, is estimated using a correlated

[3] Data available at http://www.google.org/flutrends.

signal, in this case the *ILI* related queries fraction Q, that acts as surrogate. The fit allows the estimate of P as a function of the value of Q:

$$\text{logit}(P) = \beta_0 + \beta_1 \text{logit}(Q) + \epsilon, \tag{4.14}$$

where $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$, β_0 and β_1 are fitting parameters, and ϵ is an error term. As clear from the expression, the *GFT* is a simple linear fit, where the unknown parameters are determined considering historical data. The innovation of the system lies on the definition of Q that is evaluated using hundreds of billions of searches on Google. Indeed, *GFT* scans all the queries we submit to Google, without using information about users' identity, in search of those that *ILI* related. This is the paradigm of passive data collection in digital epidemiology. In the original model the authors measured the correlation of 50 millions search queries with historic CDC data, finding that 45 of them were enough to ensure the best correlation between the number of searches and the number of *ILI* cases. The identity of such terms has been kept secret in order to avoid changes in users' behavior. However, the authors provided a list of topics associated with each one of them: 11 were associated with influenza complications, 8 to cold/flu remedies, 5 to general terms for influenza, etc. Although the search for the terms has been performed without prior information, none of the most representative terms were unrelated to the disease. In these settings *GFT* showed a mean correlation of 0.97 with real data and was able to predict the surveillance value with 1–2 weeks ahead.

GFT is based on proprietary data that for many different constraints cannot be shared with the research community. Other data sources, different in nature, are instead easily accessible. Twitter and Wikipedia are the two examples. Indeed, both systems are available for download, with some limitations, through their respective *API*s.

The models based on Twitter are built within the same paradigm of *GFT* [67, 71–73, 81]. Tweets are mined in search of *ILI*-related tweets, or other health conditions such as insomnia, obesity, and other chronic diseases [67, 82], that are used to inform regression models. Such tweets are determined either as done in *GFT*, or through more involved methods based on support vector machine (*SVM*) or other machine learning methods that, provided an annotated corpus, find disease related tweets beyond simple keywords matches [67, 71–73, 81]. The presence of *GPS* information or other self-reported geographical data allows the models to probe different granularities ranging from countries [67, 71, 73, 81] to cities [72].

While models based on Twitter analyze users' posts, those based on Wikipedia focus on pages views [75, 76]. The basic intuition is that Wikipedia is used to learn more about a diseases or a medication. Plus, the website is so popular that is most likely one of the first results of search queries on most search engines. The methods proposed so far monitor a set of pages related to the disease under study. Examples are *Influenza*, *Cold*, *Fever*, *Dengue*, etc. Page views at the daily or weekly basis are then used a surrogates in linear fitting models. Interestingly, the correlation with surveillance data ranges from 0.02 in the case of Ebola to 0.99 in for *ILI*s [75, 76], and allows accurate predictions up to 2 weeks ahead. One important limitation of

Wikipedia based methods is the lack of geographical granularity. Indeed, the view counts are reported irrespective of readers' location but the language of the page can be used as a rough proxy for location. Such approximation might be extremely good for localized languages like Italian but it poses strong limitations in the case of global languages like English. Indeed, it is reported that 51 % of pages views for English pages are done in the USA, 11 % in the UK, and the rest in Australia, Canada and other countries [76]. Besides, without making further approximation such methods cannot provide indications at scales smaller than the country level.

Despite these impressive correlations, especially in the case of ILIs, much still remains to be done. *GFT* offers a particular clear example of the possible limitations of such tools. Indeed, despite the initial success, it completely failed to forecast the 2009 H1N1 pandemic [61, 83]. The model was updated in September 2009 to increase the number of terms to 160, including the 40 terms present in the original version. Nevertheless, *GFT* missed high 100 out of 108 weeks in the season 2011–2012. In 2013 *GFT* predicted a peak height more than double the actual value causing the underlying model to be modified again later that year.

What are the reasons underlying the limitations of *GFT* and other similar tools? By construction, *GFT* relies just on simple correlations causing it to detect not only the flu but also things that correlate strongly with the flu such as winter patterns. This is likely one of the reasons why the model was not able to capture the unfolding of an off-season pandemic such as the 2009 H1N1 pandemic. Also, changes in the Google search engine, that can inadvertently modify users' behavior, were not taken into account in *GFT*. This factor alone possibly explains the large overestimation of the peak height in 2013. Plus, simple auto-regressive models using just CDC data can perform as well or better than *GFT* [84]. The parable of *GFT* clearly shows both the potential and the risks of digital tools for epidemic predictions. The limitations of *GFT* can possibly affect all similar approaches based on digital passive data collection. In particular, the use of simple correlations measures does not guarantee the ability of capturing the phenomena across different scales in space and time with respect to those used in the training. Not to mention that correlations might be completely spurious. In a recent study for example, a linear model based on Twitter simply informed with the timeline of the term *zombie* was shown to be a good predictor of the seasonal flu [71].

Despite such observations the potential of these models is invaluable to probe data that cannot be predicted by simple auto-regressive models. For example, flu activity at high geographical granularities, although very important, is measured with great difficulties by the surveillance systems. *GFT* and other spatially resolved tools can effectively access to these local indicators, and provide precious estimates that can be used a complement for the surveillance and as input for generating epidemic models [49, 68].

4.7 Discussion

The field of epidemiology is currently undergoing a digital revolution due to the seemingly endless availability of data and computational power. Data on human behavior is allowing for the development of new tools and models while the commoditization of computer resources once available only for world leading research institutions is making highly detailed large scale numerical approaches feasible at last.

In this chapter, we present a brief review not only of the fundamental mathematical tools and concepts of epidemiology but also of some of the state-of-the-art and computational approaches aimed at describing, modeling, and forecasting the diffusion of viruses. Our focus was on the developments occurring over the past decade that are sure to form the foundation for developments in decades to come.

Acknowledgements BG was partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018).

References

1. Bernulli, D. (1760). Essai dune nouvelle analyse de la mortalité causée par la petite vérole et des advantages de l'inocoulation pur la prévenir. *Mémoires de Mathematique Physique de l'Academie Royale des Sciences, 8*, 1–45.
2. Anderson, R. M., & May, R. M. (1992). *Infectious diseases in humans*. Oxford: Oxford University Press.
3. Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences, 103*(7), 2015–2020.
4. Balcan, D., Gonçalves, B., Hu, H., Ramasco, J. J., Colizza, V., & Vespignani, A. (2010). Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science, 1*(3), 132–145.
5. Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature, 429*(6988), 180–184.
6. Chao, D. L., Halloran, M. E., Obenchain, V. J., & Longini, I. M. (2010). Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology, 6*(1), e1000656.
7. Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection: Harnessing the web for public health surveillance. *New England Journal of Medicine, 360*(21), 2153–2157. PMID:19423867.
8. Salathé, M., Bengtsson, L., Bodnar, J. T., Brewer, D. D., Brownstein, J. S., Buckee, C., et al. (2012). Digital epidemiology. *PLoS Computational Biology, 8*, 7.
9. Bailey, N. T. (1975). *The mathematical theory of infectious diseases*. London: Griffin.
10. Keeling, M. J. & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton: Princeton Univeristy Press.
11. Kermack, W., & McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A, Containing Papers of a Mathematical and Physical Character* (Vol. 115, pp. 700–721)

12. Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
13. Caldarelli, G. (2007). *Scale-free networks*. Oxford: Oxford University Press.
14. Cohen, R., & Havlin, S. (2010). *Complex networks: Structure, robustness and function*. Cambridge: Cambridge University Press.
15. Barrat, A., Barthélemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge: Cambridge University Press.
16. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. -L., Brewer, D., et al. (2009). Computational social science. *Science, 323*, 721.
17. Barabasi, A.-L. (2002). *Linked: How everything is connected to everything else and what it means*. Plume Editors.
18. Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports, 519*, 97.
19. Morris, M. (1993). Telling tails explain the discrepancy in sexual partner reports. *Nature, 365*, 437.
20. Rocha, L. E. C., Liljeros, F., & Holme, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Computational Biology, 7*(3), e1001109.
21. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J. J., & Van den Broeck, W. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology, 271*, 166.
22. Perra, N., Gonçalves, B., Pastor-Satorras, R., & Vespignani, A. (2012). Activity driven modeling of dynamic networks. *Scientific Reports, 2*, 469.
23. Karsai, M., Perra, N., & Vespignani, A. (2014). Time varying networks and the weakness of strong ties. *Scientific Reports, 4*, 4001.
24. Sun, K., Baronchelli, A., & Perra, N. (2014). Epidemic spreading in non-markovian time-varying networks. arxiv:1404.1006.
25. Castellano, C., & Pastor-Satorras, R. (2010). Thresholds for epidemic spreading in networks. *Physical Review Letters, 105*, 218701.
26. Vespignani, A. (2012). Modeling dynamical processes in complex socio-technical systems. *Nature Physics, 8*, 32–30.
27. Boguna, M., & Pastor-Santorras, R. (2002). Epidemic spreading in correlated complex networks. *Physical Review E, 66*, 047104.
28. Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Physical Review E, 66*, 016128.
29. Serrano, M. A., Boguna, M., & Pastor-Satorras, R. (2006). Correlations in weighted networks. *Physical Review E, 74*, 055101(R).
30. Cohen, R., Havlin, S., & ben Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical Review Letters, 91*, 247901.
31. Garcia-Herranz, M., Egido, E. M., Cebrian, M., Christakis, N. A., & Fowler, J. H. (2014). Using friends as sensors to detect global-scale contagious outbreaks. *PLoS One, 9*, 4.
32. Liu, S., Perra, M., Karsai, N., & Vespignani, A. (2014). Controlling contagion processes in activity driven networks. *Physical Review Letters, 112*, 118702.
33. Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the ESA, 15*(3), 237–240.
34. Sattenspiel, L. (1987). Population structure and the spread of disease. *Human Biology, 59*, 411–438.
35. Sattenspiel, L., & Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences, 128*(1), 71–91.
36. Britton, N. F. (1986). *Reaction-diffusion equations and their applications to biology*. New York: Academic.
37. Colizza, V., & Vespignani, A. (2008). Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. *Journal of Theoretical Biology, 251*(3), 450–467.
38. Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports, 1*, 62.

39. Schläpfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., et al. (2014). The scaling of human interactions with city size. *Journal of the Royal Society Interface, 11*, 20130789
40. Tizzoni, M., Sun, K., Benusiglio, D., Karsai, M., & Perra, N. (2014). The scaling of human contacts in reaction-diffusion processes on heterogeneous metapopulation networks. arxiv:1411.7310.
41. Center for International Earth Science Information Network (ciesin), Columbia University, & Centro Internacional de Agricultura Tropical (ciat). (2004). *The gridded population of the world version 3 (gpwv3): Population grids*. Palisades, NY: Socioeconomic Data and Applications Center (sedac), Columbia University (2004).
42. Center for International Earth Science Information Network (ciesin), Columbia University; International Food Policy Research Institute (ifpri); The World Bank; & Centro Internacional de Agricultura Tropical (ciat). (2004). *Global rural-urban mapping project (grump), alpha version: Population grids*. Palisades, NY: Socioeconomic Data and Applications Center (sedac), Columbia University (2004).
43. International Air Transport Association. http://www.iata.org
44. Official Airline Guide. www.oag.com/
45. Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2007). Predictability and epidemic pathways in global outbreaks of infectious diseases: The sars case study. *BMC Medicine, 5*(1), 34.
46. Gonçalves, B., Balcan, D., & Vespignani, A. (2013). Human mobility and the worldwide impact of intentional localized highly pathogenic virus release. *Scientific Reports, 3*, 810.
47. Gomes, M. F. C., Pastore y Piontti, A., Rossi, L., Chao, D., Longini, I., Halloran, M. E., et al. (2014, September 2). Assessing the international spreading risk associated with the 2014 west African Ebola Outbreak. *PLOS Currents Outbreaks* (1st ed.). doi: 10.1371/currents.outbreaks.cd818f63d40e24aef769dda7df9e0da5.
48. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Science, 106*(51), 21484–21489.
49. Balcan, D., Hu, H., Gonçalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., et al. (2009). Seasonal transmission potential and activity peaks of the new influenza a (h1n1): A monte carlo likelihood analysis based on human mobility. *BMC Medicine, 7*(1), 45.
50. Yang, Y., Sugimoto, J. D., Halloran, M. E., Basta, N. E., Chao, D. L., Matrajt, L., et al. (2009). The transmissibility and control of pandemic influenza a (h1n1) virus. *Science, 326*(5953), 729–733.
51. Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., et al. (2009). Pandemic potential of a strain of influenza a (h1n1): Early findings. *Science, 324*(5934), 1557–1561.
52. Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J. J., Balcan, D., Gonçalves, B., et al. (2012). Real-time numerical forecast of global epidemic spreading: Case study of 2009 a/h1n1pdm. *BMC Medicine, 10*(1), 165.
53. Ajelli, M., Gonçalves, B., Balcan, D., Colizza, V., Hu, H., Ramasco, J. J., et al. (2010). Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. *BMC Infectious Diseases, 10*(1), 190.
54. Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415–429.
55. Barrett, C. L., Beckman, R. J., Berkbigler, K. P., Bisset, K. R., Bush, B. W., Eubank, S., Hurford, J. M., Konjevod, G., Kubicek, D. A., Marathe, M. V., et al. (1999). Transims (transportation analysis simulation system). In *Volume 0: Overview. Report LA-UR-99-1658.* Los Alamos, NM: Los Alamos National Laboratory.
56. Merler, S., Ajelli, M., Pugliese, A., & Ferguson, N. M. (2011). Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: Implications for real-time modelling. *PLOS Computational Biology, 7*(9), e1002205.

57. Germann, T. C., Kadau, K., Longini, I. M., & Macken, C. A. (2006). Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Science, 103*(15), 5935–5940.
58. Chao, D. L., Matrajt, L., Basta, N. E., Sugimoto, J. D., Dean, B., Bagwell, D. A., et al. (2011). Planning for the control of pandemic influenza a (h1n1) in los angeles county and the united states. *American Journal of Epidemiology, 173*(10), 1121–1130.
59. Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science, 325*(5939), 425.
60. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2014). Epidemic processes in complex networks. arXiv preprint. arXiv:1408.2701.
61. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science, 343*(6176), 1203–1205.
62. Chan, E. H., Brewer, T. F., Madoff, L. C., Pollack, M. P., Sonricker, A. L., Keller, M., et al. (2010). Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of Sciences, 107*(50), 21701–21706.
63. Health Map. http://www.healthmap.org/.
64. Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., et al. (2014). Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 17–21.
65. Flu Near You. http://www.flunearyou.org.
66. Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology, 7*(10), e1002199.
67. Paul, M. J. & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *ICWSM* (pp. 265–272).
68. Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences, 109*(50), 20425–20430.
69. Zhang, Q., Perra, N., & Vespignani, A. (in preparation). Forecasting seasonal influenza with stochastic microsimulations models assimilating digital surveillance data.
70. Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS One, 6*(5), e19467.
71. Bodnar, T., & Salathé, M. (2013). Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web Companion* (pp. 699–702). International World Wide Web Conferences Steering Committee, 2013.
72. Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012–2013 influenza epidemic. *PloS One, 8*(12), e83672.
73. Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 115–122), ACM.
74. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014
75. Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Priedhorsky, R. (2014). Detecting epidemics using wikipedia article views: A demonstration of feasibility with language as location proxy. arXiv preprint. arXiv:1405.3612.
76. McIver, D. J., & Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Computational Biology, 10*(4), e1003581.
77. Nsoesie, E. O., Buckeridge, D. L., & Brownstein, J. S. (2014). Guess who is not coming to dinner? Evaluating online restaurant reservations for disease surveillance. *Journal of Medical Internet Research, 16*(1), e22.
78. Butler, P., Ramakrishnan, N., Nsoesie, E. O., & Brownstein, J. S. (2014). Satellite imagery analysis: What can hospital parking lots tell us about a disease outbreak? *Computer, 47*(4), 94–97.

79. Wójcik, O. P., Brownstein, J. S., Chunara, R., & Johansson, M. A. (2014). Public health for the people: participatory infectious disease surveillance in the digital age. *Emerging Themes in Epidemiology, 11*(1), 7.
80. Google Flu Trends. http://www.google.org/flutrends/.
81. Signorini, A., Polgreen, P. M., & Segre, A. M. (2010). Using twitter to estimate h1n1 influenza activity. In *9th Annual Conference of the International Society for Disease Surveillance*.
82. De la Torre-Díez, I., Díaz-Pernas, F. J., & Antón-Rodríguez, M. (2012). A content analysis of chronic diseases social groups on facebook and twitter. *Telemedicine and e-Health, 18*(6), 404–408.
83. Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology, 9*(10), e1003256
84. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences, 107*(41), 17486–17490.

Chapter 5
Early Signs of Financial Market Moves Reflected by *Google* Searches

Tobias Preis and Helen Susannah Moat

Abstract The complex behaviour of our society emerges from decisions made by many individuals. In certain combinations, these numerous decisions can lead to sudden catastrophe, as demonstrated during crowd disasters and financial crises. Here, we investigate whether analyses of queries to the search engine *Google* may be able to give us insight into the early information gathering stages of collective decision making in financial markets. Results of the implementation of a profitable hypothetical trading strategy are consistent with the proposal that changes in the volume of online searches for information about a company may be detected before large stock market moves. These results suggest that big data capturing our everyday interactions with the Internet may be able to provide new evidence for the science of collective decision making.

5.1 Introduction

The increasing volumes of "big data" reflecting various aspects of our everyday activities represent a vital new opportunity for scientists to address fundamental questions about the complex social world we inhabit [1–13]. Studies in this area have demonstrated promising links between aggregate online behaviour and collective real world behaviour across a range of data sources such as the search engine Google [14–21], the search engine Yahoo! [22, 23], the online encyclopedia Wikipedia [24–28], the microblogging platform Twitter [29–34] and the photo-sharing website Flickr [35, 36]. For example, a recent study has demonstrated that Internet users from countries with a higher per capita GDP are more likely to search for information about years in the future than years in the past [37, 38].

Such studies have generally interpreted changes in search volume as changes in interest and attention to a subject. This permits compelling explanations of relationships between online and real life behaviour in areas where increases in search activity can be linked to increases in real life activity in a straightforward

T. Preis (✉) • H.S. Moat
Data Science Lab, Behavioural Science, Warwick Business School, The University of Warwick, Coventry, West Midlands CV4 7AL, UK
e-mail: Tobias.Preis@wbs.ac.uk; Suzy.Moat@wbs.ac.uk

© Springer International Publishing Switzerland 2015
B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7_5

manner. For example, an increase in searches for a movie title has been shown to correspond to an increase in attendance figures for that movie [23], and an increase in searches for a company name has been shown to correspond to an increase in trading activity on the stock markets [19].

In some cases however, this simple interpretation does not offer an intuitive prediction for the most interesting aspects of the behaviour in question. Economic decision making in the stock markets is a clear case in point [39–56]. In such situations, we would like to anticipate not only whether more trading may take place [19], but whether traders are generally looking to buy or to sell.

Here, we analyse the performance of a hypothetical trading strategy, where we make decisions to trade stocks in the market depending on changes in search volume for company related terms. We investigate whether increases in searches for company related terms may be interpreted as a sign of increased confidence or concern about the future value of a stock.

5.2 Results

To investigate the relationship between the volume of search queries for company related terms and trading decisions made in the financial markets, we build on the method introduced by Preis, Moat and Stanley, in which they use trading strategies based on search volume data provided by *Google Trends* to identify online precursors for stock market moves [20]. Their analysis of search volume for 98 terms of varying financial relevance suggests that, at least in historic data, increases in search volume for financially relevant search terms tend to precede large losses in financial markets. Similarly, Moat et al. [26, 57] demonstrate that increases in the number of views of *Wikipedia* articles relating either to companies listed in the *Dow Jones Industrial Average* (DJIA) or more general financial topics tend to be followed by stock market falls. Conversely, Moat et al. find no such relationship between stock market behaviour and changes in the number of views of *Wikipedia* articles relating to the less financially relevant topic of actors and filmmakers. The method introduced by Preis, Moat and Stanley was also used by Curme et al. [24] to quantify the semantics of search behaviour before stock market moves.

Here, we consider data on *Google* search queries from the public service *Google Trends*[1] alongside prices $p_i(t)$ for all components i of the DJIA on the first trading day of week t. Specifically, we analyse data generated between the first week in January 2004, the earliest date for which *Google Trends* makes search volume data publicly available, and the last week in September 2012, when we collected the data. We use the *Google Trends* service to calculate how many searches $n_i(t-1)$ have been carried out for the company name or ticker symbol relating to stock i

[1] We accessed the *Google Trends* website (http://www.google.com/trends) on 30th September 2012. Note that *Google* changed the format in which search volume is provided at the end of 2012.

in week $t - 1$, where *Google* defines weeks as ending on a Saturday, relative to the average number of searches carried out on *Google* for this keyword during the whole period considered. Details of the company names and ticker symbols used are provided in the *Appendix*. To quantify changes in information gathering behaviour, we use the relative change in search volume:

$$\Delta n_i(t, \Delta t) = n_i(t) - N_i(t-1, \Delta t) \tag{5.1}$$

with

$$N_i(t-1, \Delta t) = \frac{n_i(t-1) + n_i(t-2) + \cdots + n_i(t-\Delta t)}{\Delta t}, \tag{5.2}$$

where t is measured in units of weeks.

We explain our analysis using the example of *Cisco Systems, Inc.* stocks. Following Preis et al. [20] and Moat et al. [26], we first implement a hypothetical investment strategy using search volume data for the company name (in this case, "**Cisco Systems**") relating to a specific stock i (in this case, *Cisco Systems, Inc.* stocks) where we sell the stock i at the closing price $p_i(t)$ on the first trading day of week t, if $\Delta n_i(t-1, \Delta t) > 0$, and buy the stock i at price $p_i(t+1)$ at the end of the first trading day of the following week to close the position. Note that mechanisms exist which make it possible to sell stocks in a financial market without first owning them. If instead $\Delta n_i(t-1, \Delta t) \leq 0$, then we buy the stock i at the closing price $p_i(t)$ on the first trading day of week t, and close the position by selling the stock i at price $p_i(t+1)$ at the end of the first trading day of the coming week. At the beginning of trading, we set the value of all 30 portfolios to an arbitrary value of 1. Each portfolio trades one DJIA component. If we open and close a short position—selling at the closing price $p_i(t)$ and buying back at price $p_i(t+1)$—then the cumulative return R changes by $\log(p_i(t)) - \log(p_i(t+1))$. If we open and close a long position—buying at the closing price $p_i(t)$ and selling at price $p_i(t+1)$—then the cumulative return R changes by $\log(p_i(t+1)) - \log(p_i(t))$. In this way, buy and sell actions have symmetric impacts on the cumulative return R of a strategy's portfolio. In addition, we neglect transaction fees, since the maximum number of transactions per year when using our strategy is only 104, allowing a closing and an opening transaction per week. The implementation of this strategy for *Cisco Systems, Inc.* stocks is depicted in Fig. 5.1a. We find that this strategy would have increased an investor's portfolio by 131 %.

Augmenting the analyses performed by Preis et al. [20] and Moat et al. [26], we compare this to a reversed strategy too. In the reversed strategy, we buy the stock i at the closing price $p_i(t)$ on the first trading day of week t, if $\Delta n_i(t-1, \Delta t) > 0$. Otherwise, we sell. Positions are then closed at on the first trading day of the following week. Figure 5.1b illustrates an implementation of this strategy for *Cisco Systems, Inc.* stock. We find that this approach would have resulted in an overall loss to an investor's portfolio of 57 %, in contrast to the profit made by the previous strategy.

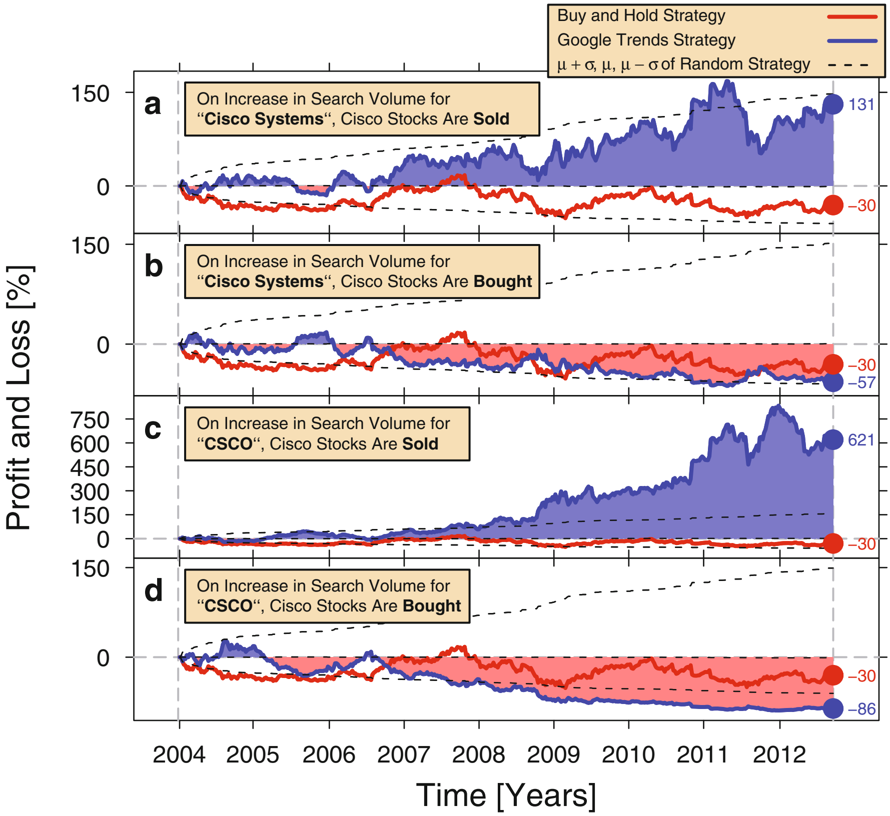


Fig. 5.1 Comparison of cumulative profit and loss for hypothetical trading strategies based on *Google Trends* data. Cumulative profit and loss for strategies based on search query volume data with an aggregation interval of $\Delta t = 2$ weeks, plotted as a function of time. We illustrate the different strategies using the stock price time series of the Cisco Systems, Inc. as an example. We compare all strategies to the distribution of returns generated by a random strategy. The mean profit from random strategies is 0 %. The intervals of −1 and +1 standard deviations around this mean, depicted with *dashed lines*, correspond to the random strategy return distribution. We also depict returns from a "buy and hold" strategy, implemented by buying a stock and selling it at the end of the hold period, to show the underlying movement in stock price, which in this case falls by 30 %. (**a**) On an increase in weekly search volume for the company name "**Cisco Systems**", we sell stocks of the Cisco Systems, Inc. Otherwise, we buy. (**b**) On an increase in the weekly search volume for the company name "**Cisco Systems**", we buy stocks of the Cisco Systems, Inc. Otherwise, we sell. In other words, we reverse the strategy used in (a). This reversed strategy results in a negative cumulative return. (**c**) More experienced financial market participants may also search for specific abbreviations of company names which are frequently used in the industry—namely, ticker symbols. On an increase in the weekly search volume for Cisco's ticker symbol "**CSCO**" we sell, and otherwise we buy. In line with results depicted in (a), purchasing stocks following increases in search volume leads to an overall profit. This strategy vastly outperforms the random strategies. (**d**) On an increase in the weekly search volume for Cisco's ticker symbol "**CSCO**" we buy, and otherwise we sell. Similar to the approach depicted in (b), this strategy results in a negative overall return

To confirm that these results represent profits and losses significantly greater than would be expected by chance, we calculate the returns from a random strategy. In the random strategy, a decision is made each week to buy or sell stock i. The probability that the stock will be bought rather than sold is always 50 %, and the decision is unaffected by decisions in previous weeks. For each stock i, we simulate 10,000 independent realisations of this random strategy. At the beginning of trading, we set the value of all 10,000 random strategy portfolios to the value of 1 too.

Returns of the strategies are calculated as the logarithm of percentage profit, following the usual definition. The distribution of profits of the final portfolio resulting from the random investment strategies is close to log-normal. Cumulative returns from the random investment strategy, derived from the logarithm of these portfolio values, therefore, follow a normal distribution. Importantly, while the mean return from a random strategy is by definition 0, the standard deviation of the returns from a random strategy differs between stocks, due to differences in movement of stock price and the associated differences in opportunities for profit and loss.

We implement our *Google Trends* trading strategy for all 30 components of the DJIA. Preis et al. [20] analyse the behaviour of *Google Trends* strategies based on a set of 98 keywords ranging from words with clear financial connotations (e.g., “debt”) to words with less financial connotations (e.g., “colour”). The approach taken here, as in Moat et al.’s [26] analysis of *Wikipedia* page views, allows us to investigate the nature of the relationship between changes in search volume and stock market moves across a well-defined and finite set of search terms with a clear relationship to the stock market index in question.

In order to compare the returns from the 30 *Google Trends* strategies to the returns generated by purely random strategies, we calculate the returns of the *Google Trends* strategies as the logarithm of the portfolio values produced by these strategies, and then normalise these returns. We implement this normalisation by dividing the return of each *Google Trends* strategy by the standard deviation of the random strategy returns for the relevant stock. We then consider the distribution of returns generated by the 30 *Google Trends* strategies. If the returns of the 30 strategies do not differ from the returns which would be expected by chance, then this distribution should be symmetrically distributed around 0, with a standard deviation of 1.

Figure 5.2 shows the cumulative returns R_i for strategies based on *Google Trends* search volumes for company names related to all 30 stocks of the DJIA. For $\Delta t =$ 2 weeks, we find that returns from a strategy where stocks are sold when search volume for company names increases (Fig. 5.2a) are significantly higher overall than the mean return of 0 expected from random strategies (mean $\mu = 0.62$ standard deviations of random strategy returns, $\sigma = 3.87$, $df = 29$, $p < 0.001$, two-sided one-sample t-test). In direct contrast, returns from a strategy where stocks are bought when search volume for company names increases (Fig. 5.2b) are significantly lower overall than the mean return of 0 expected from random strategies ($\mu = -0.62$, $\sigma = 3.87$, $df = 29$, $p < 0.001$, two-sided one-sample t-test). In summary, following increases in search volume for company names, a strategy where stocks are sold leads to profit, whereas a strategy where stocks are bought leads to loss.

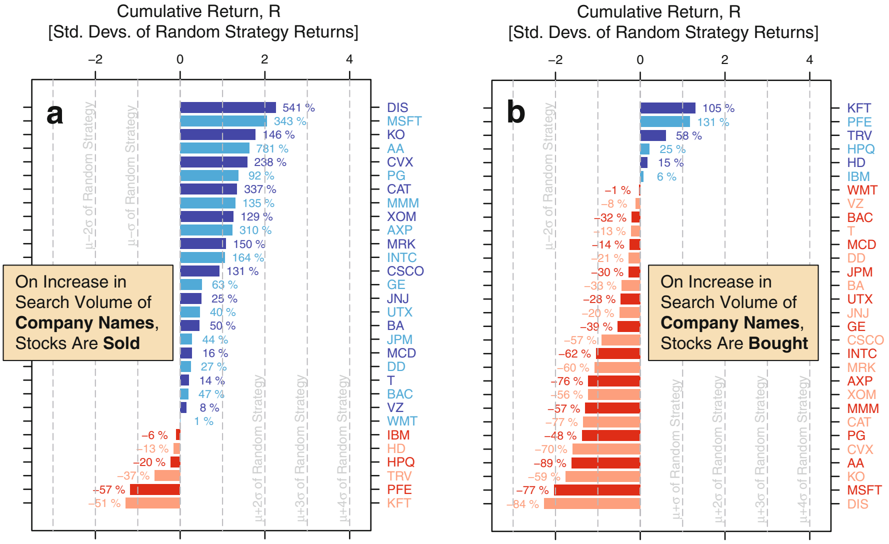


Fig. 5.2 Returns of two hypothetical trading strategies based on *Google Trends* data for company name searches. Distribution of returns gained from strategies based on *Google Trends* search volume for company names for all 30 stocks of the *Dow Jones Industrial Average* (DJIA) using $\Delta t = 2$ weeks. Returns are depicted in terms of standard deviations σ_i of the returns from a random strategy, where a given stock i is bought and sold in an uncorrelated random fashion, and ranked according to this measure. The standard deviation σ_i is determined from results of simulations using 10,000 independent realisations for each stock i. Numbers placed beside the *bars* denote percentage profits and percentage losses. *Dashed lines* correspond to the mean $\mu = 0$ and -3, -2, -1, $+1$, $+2$, $+3$ and $+4$ standard deviations σ of the global distribution of random strategy returns against which all results are normalised. Positive returns are shown in *blue*, and negative returns in *red*, where different shades are used only to improve readability. (**a**) On an increase in search volume for a company name, we sell the corresponding stock; otherwise, we buy. The mean cumulative return across all DJIA components when using this strategy is 0.62 standard deviations above the random strategy mean ($\sigma = 3.87$, $df = 29$, $p < 0.001$, two-sided one-sample t-test). (**b**) On an increase in the weekly search volume for a company name, we buy, and where search volume decreases or does not change, we sell, in direct contrast to the strategy depicted in (a). Consequently, we find a mean cumulative return which is 0.62 standard deviations below the random strategy mean ($\sigma = 3.87$, $df = 29$, $p < 0.001$, two-sided one-sample t-test)

Financial market specialists and more experienced market participants may also search for ticker symbols, specific abbreviations of company names used on stock exchanges. We perform a parallel analysis using search volume for ticker symbols. Figures 5.1c, d illustrate an implementation of both proposed strategies using search volumes for the ticker symbol of the *Cisco Systems, Inc.*, "**CSCO**". Where *Cisco Systems, Inc.* stocks are sold on an increase in search volume for "**CSCO**", the portfolio value is increased by 621 % (Fig. 5.1c). In contrast, buying *Cisco Systems, Inc.* stocks on an increase in search volume for "**CSCO**" would have resulted in an overall loss of 86 % (Fig. 5.1d).

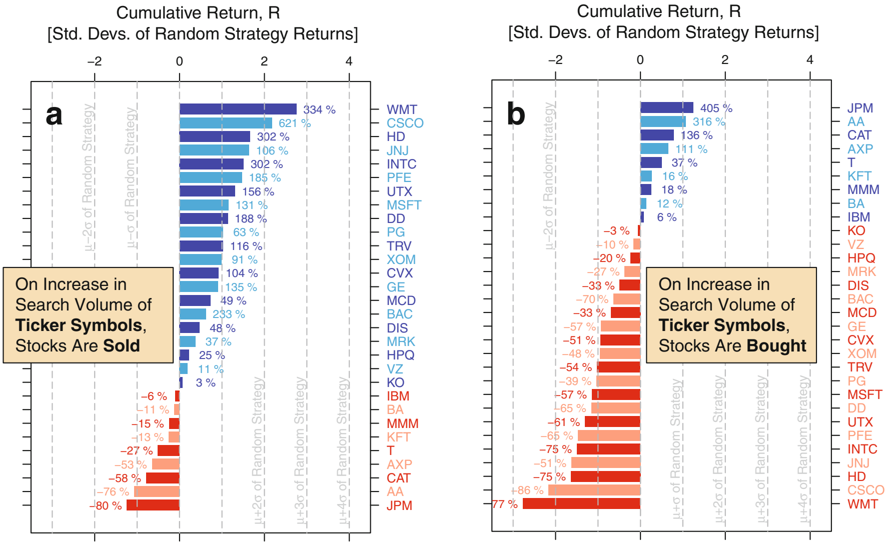


Fig. 5.3 Returns of two hypothetical trading strategies based on *Google Trends* data for ticker symbol searches. Parallel analyses for *Google Trends* data on searches for ticker symbols of DJIA components using $\Delta t = 2$ weeks, depicted in (**a**) and (**b**), provide further evidence that an increase in searches for terms related to companies registered on the stock market is associated with stock price drops in the coming week, rather than stock price rises. (**a**) On an increase in search volume for a ticker symbol, we sell the corresponding stock; otherwise, we buy. The mean cumulative return across all DJIA components when using this strategy is 0.58 standard deviations above the random strategy mean ($\sigma = 3.31$, $df = 29$, $p < 0.01$, two-sided one-sample t-test). (**b**) On an increase in the weekly search volume for a ticker symbol, we buy, and where search volume decreases or does not change, we sell. This strategy leads to a mean cumulative return which is 0.58 standard deviations below the random strategy mean ($\sigma = 3.30$, $df = 29$, $p < 0.01$, two-sided one-sample t-test)

Figure 5.3 shows cumulative returns for strategies based on *Google Trends* search volumes for all DJIA ticker symbols. In line with our results for company names, we find that returns from a strategy where stocks are sold when search volume for ticker symbols increases (Fig. 5.3a) are significantly higher overall than the mean return of 0 expected from random strategies ($\mu = 0.58$, $\sigma = 3.31$, $df = 29$, $p < 0.01$, two-sided one-sample t-test). Returns from a strategy where stocks are bought when search volume for ticker symbols increases (Fig. 5.3b) are significantly lower overall than the mean return of 0 expected from random strategies ($\mu = -0.58$, $\sigma = 3.30$, $df = 29$, $p < 0.01$, two-sided one-sample t-test), again in line with our company name results.

We compare the success of strategies based on company names and strategies based on ticker symbols. We find no difference between returns from a strategy where stocks are sold when search volume for company names increases, and returns from a strategy where stocks are sold when search volume for ticker symbols increases ($\sigma = 0.17$, $df = 29$, $p = 0.87$, two-sided paired t-test). Equally, we find

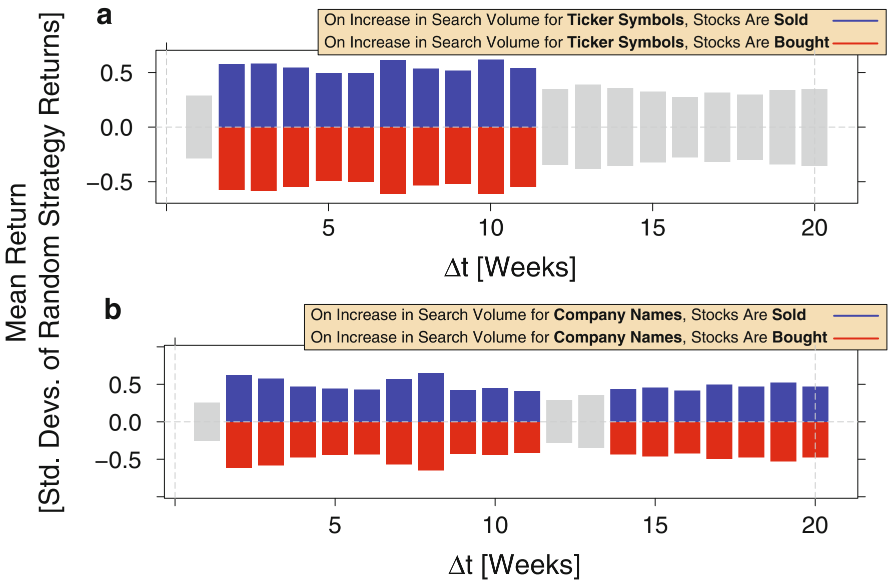


Fig. 5.4 Mean returns of hypothetical trading strategies based on *Google Trends* data for various Δt. We compare the mean cumulative returns for search volume based strategies for various values of Δt to the mean return of 0 expected from random strategies, using two-sided one-sample t-tests with FDR correction for multiple comparisons. Mean values that are significantly different from 0 (t-test, $\alpha = 0.05$, FDR corrected) are depicted in *blue* or *red*. (**a**) Mean returns of trading strategies based on *Google Trends* data for ticker symbols. (**b**) Mean returns of trading strategies based on *Google Trends* data of company names. Hypothetical trading strategies for $\Delta t = 2$ to $\Delta t = 11$ weeks give significantly higher returns than a random strategy using both search volume for ticker symbols and for company names

no difference between returns from a strategy where stocks are bought when search volume for company names increases, and returns from a strategy where stocks are bought when search volume for ticker symbols increases ($\sigma = 0.17$, $df = 29$, $p = 0.87$, two-sided paired t-test). We therefore find no empirical basis to draw any conclusion about preferences to use company names or ticker symbols when gathering information to enhance trading decisions.

For all of our analyses, we report results for $\Delta t = 2$ weeks. Parallel analyses for $\Delta t = 2$ to $\Delta t = 11$ weeks give results with no qualitative differences. Figure 5.4 depicts the mean cumulative returns for search volume based strategies for various values of Δt. We compare the mean return of the strategies to the mean return of 0 expected from random strategies, using two-sided one-sample t-tests with FDR correction for multiple comparisons. Our analyses show that hypothetical trading strategies using values of Δt between 2 and 11 weeks lead to significantly higher returns than a random strategy. It is possible that the weekly change in search volume ($\Delta t = 1$) is too noisy, whereas much higher values of Δt do not allow the algorithm to be sufficiently influenced by the most relevant recent search behaviour.

5.3 Discussion

Large trading datasets from financial markets provide detailed records of decisions made within a complex social system [58]. However, such trading records capture only the final actions which result from the decision making processes. Our investigations provide evidence in line with the intriguing hypothesis that by combining data collected from search engine queries with financial trading data, we may be able to gain insight into an earlier information gathering stage of collective economic decision making. Specifically, a historic analysis of the performance of a hypothetical trading strategy using *Google Trends* data during the period January 2004 to September 2012 suggests that increases in the volume of online searches for information about a company might have been found before drops in the company's stock price. We offer one possible explanation for our results, following Moat et al. [26]. Studies in experimental behavioural science have repeatedly demonstrated that people are loss averse [59]; that is, they consider the impact of losing something that they possess to be much greater than the impact of gaining something of the same value. We suggest that an increase in information gathering activity may therefore indicate an increase in upcoming attempts to sell, as people may be prepared to invest more resources into gathering information before making a potentially higher impact decision. In summary, our findings are consistent with the possibility that online search data may contain early signs of large scale attempts to avert loss, as individuals increase efforts to gather information before trading stocks at a lower price. Our results help demonstrate the notable value of data generated through interactions with the Internet for our understanding of collective decision making.

Acknowledgements We thank Steven R. Bishop (University College London) and H. Eugene Stanley (Boston University) as well as Hankyu Moon, Tsai-Ching Lu, and David L. Allen from HRL Laboratories for comments. This work was partially supported by the Research Councils UK Grant EP/K039830/1. This work was also partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI / NBC) contract number D12PC00285. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

Appendix

The ticker symbols and company names of all components of the *Dow Jones Industrial Average* in September 2012, as used in our search volume analysis, are given in Table 5.1. Ticker symbols are unique identifiers defined at the exchange when stocks are registered for trading. Company names were retrieved from the *Dow Jones Industrial Average* listing at the *Yahoo! Finance* website (http://finance.yahoo.com/q/cp?s=%5EDJI+Components) on 30th September 2012. We removed the following

Table 5.1 Ticker symbols (left) and company names (right) of all components of the *Dow Jones Industrial Average* in September 2012

Ticker symbols	Company name
"AA"	"Alcoa"
"AXP"	"American Express"
"BA"	"Boeing"
"BAC"	"Bank of America"
"CAT"	"Caterpillar"
"CSCO"	"Cisco Systems"
"CVX"	"Chevron"
"DD"	"du Pont"
"DIS"	"Walt Disney"
"GE"	"General Electric"
"HD"	"Home Depot"
"HPQ"	"Hewlett Packard"
"IBM"	"IBM"
"INTC"	"Intel"
"JNJ"	"Johnson & Johnson"
"JPM"	"JP Morgan Chase"
"KFT"	"Kraft Foods"
"KO"	"Coca Cola"
"MCD"	"McDonalds"
"MMM"	"3M"
"MRK"	"Merck"
"MSFT"	"Microsoft"
"PFE"	"Pfizer"
"PG"	"Procter & Gamble"
"T"	"AT&T"
"TRV"	"The Travelers"
"UTX"	"United Technologies"
"VZ"	"Verizon"
"WMT"	"Wal Mart"
"XOM"	"Exxon Mobil"

generic terms from the names retrieved: "Common", "Companies", "Company", "Corporation", "Inc.", "Stock", "(The)", "& Co.", and "& Company". To allow retrieval of *Google Trends* data, we also removed hyphens and apostrophes. Finally, we made the following edits to four company names for which there was otherwise no *Google Trends* data available: "Wal Mart Stores" was replaced with "Wal Mart"; "Verizon Communications" was replaced with "Verizon"; "International Business Machines" was replaced with "IBM"; and "E.I. du Pont de Nemours" was replaced with "du Pont".

References

1. Balcan, D., Goncalves, B., Hu, H., Ramasco, J. J., Colizza, V., & Vespignani, A. (2010). Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computer Science, 1*, 132.
2. Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection—Harnessing the web for public health surveillance. *The New England Journal of Medicine, 360*, 2153.
3. Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., & Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. *Nature, 464*, 1025.
4. Conte, R., Gilbert, N., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., et al. (2012). Manifesto of computational social science. *European Physical Journal-Special Topics, 214*, 25.
5. Johnson, N., Carran, S., Botner, J., Fontaine, K., Laxague, N., Nuetzel, P., et al. (2011). Pattern in escalations in insurgent and terrorist activity. *Science, 333*, 81.
6. King, G. (2011). Ensuring the data-rich future of the social sciences. *Science, 331*, 719.
7. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Computational social science. *Science, 323*, 721.
8. Moat, H. S., Preis, T., Olivola, C. Y., Liu, C., & Chater, N. (2014). Using big data to predict collective behavior in the real world. *The Behavioral and Brain Sciences, 37*, 92.
9. Mondria, J., Wu, T., & Zhang, Y. (2010). The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics, 82*, 85.
10. Perc, M. (2012). Evolution of the most common english words and phrases over the centuries. *Journal of the Royal Society Interface, 9*, 3323.
11. Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports, 2*, 943.
12. Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., & White, D. R. (2009). Economic networks: The new challenges. *Science, 325*, 422.
13. Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science, 325*, 425.
14. Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly, 55*, 107.
15. Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *The Economic Record, 88*, 2.
16. Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance, 66*, 1461.
17. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., et al. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012.
18. Kristoufek, L. (2013). Can Google trends search queries contribute to risk diversification? *Scientific Reports, 3*, 2713.
19. Preis, T., Reith, D., & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: Insights from search engine query data. *Philosophical Transactions of the Royal Society A, 368*, 5707.
20. Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google trends. *Scientific Reports, 3*, 1684.
21. Preis, T., & Moat, H. S. (2014). Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science, 1*, 140095.
22. Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes *PLoS One, 7*, e40014.
23. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 17486.
24. Curme, C., Preis, T., Stanley, H. E., & Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the*

United States of America, 111, 11600.

25. Kristoufek, L. (2013). BitCoin meets Google trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports, 3*, 3415.
26. Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports, 3*, 1801.
27. Moat, H. S., Curme, C., Stanley, H. E., & Preis, T. (2014). Anticipating stock market movements with Google and Wikipedia. In D. Matrasulov, & H. E. Stanley (Eds.), *Nonlinear phenomena in complex systems: From nano to macro scale*. Dordrecht: Springer.
28. Yasseri, T., Kornai, A., & Kertesz, J. (2012). A practical approach to language complexity: A Wikipedia case study. *PLoS One, 7*, e48386.
29. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computer Science, 2*, 1.
30. Ciulla, F., Mocanu, D., Baronchelli, A., Goncalves, B., Perra, N., & Vespignani, A. (2012). Beating the news using social media: The case study of American idol. *EPJ Data Science, 1*, 1.
31. Goncalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on twitter networks: Validation of Dunbar's number. *PLoS One, 6*, e22656.
32. Mocanu, D., Baronchelli, A., Perra, N., Goncalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS One, 8*, e61981.
33. Alis, C. M., Lim, M. T., Moat, H. S., Barchiesi, D., Preis, T., & Bishop, S. R. (2015). Quantifying regional differences in the length of Twitter messages. *PLoS One, 10*, e0122278.
34. Botta, F., Moat, H. S., & Preis, T. (2015). Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science, 2*, 150162.
35. Barchiesi, D., Moat, H. S., Alis, C., Bishop, S., & Preis, T. (2015). Quantifying international travel flows using Flickr. *PLoS One, 10*, e0128470.
36. Preis, T., Moat, H. S., Bishop, S. R., Treleaven, P., & Stanley, H. E. (2013). Quantifying the digital traces of Hurricane sandy on flickr. *Scientific Reports, 3*, 3141.
37. Preis, T., Moat, H. S., Stanley, H. E., & Bishop, S. R. (2012). Quantifying the advantage of looking forward. *Scientific Reports, 2*, 350.
38. Noguchi, T., Stewart, N., Olivola, C. Y., Moat, H. S., & Preis, T. (2014). Characterizing the time-perspective of nations with search engine query data. *PLoS One, 9*, e95209.
39. Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports, 3*, 3578.
40. Fehr, E. (2002). Behavioural science: The economics of impatience. *Nature, 415*, 269.
41. Feng, L., Li, B., Podobnik, B., Preis, T., & Stanley, H. E. (2012). Linking agent-based models and stochastic models of financial markets. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 8388.
42. Gabaix, X., Gopikrishnan, P., Plerou, V., & Stanley, H. E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature, 423*, 267.
43. Haldane, A. G., & May, R. M. (2011). Systemic risk in banking ecosystems. *Nature, 469*, 351.
44. Hommes, C. H. (2002). Modeling the stylized facts in finance through simple nonlinear adaptive systems. *Proceedings of the National Academy of Sciences of the United States of America, 99*, 7221.
45. Johnson, N., Zhao, G., Hunsader, E., Qi, H., Johnson, N., Meng, J., et al. (2013). Abrupt rise of new machine ecology beyond human response time. *Scientific Reports, 3*, 2627.
46. Lillo, F., Farmer, J. D., & Mantegna, R. N. (2003). Econophysics: Master curve for price-impact function. *Nature, 421*, 129.
47. Lux, T., & Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature, 397*, 498.
48. Preis, T., Golke, S., Paul, W., & Schneider, J. J. (2007). Statistical analysis of financial returns for a multiagent order book model of asset trading. *Physical Review E, 76*, 016108.
49. Preis, T., Golke, S., Paul, W., & Schneider, J. J. (2006). Multi-agent-based order book model of financial markets. *Europhysics Letters, 75*, 510.

50. Preis, T., Paul, W., & Schneider, J. J. (2008). Fluctuation patterns in high-frequency financial asset returns. *Europhysics Letters, 82*, 68005.
51. Preis, T., Schneider, J. J., & Stanley, H. E. (2011). Switching processes in financial markets. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 7674.
52. Preis, T., Virnau, P., Paul, W., & Schneider, J. J. (2009). Accelerated fluctuation analysis by graphic cards and complex pattern formation in financial markets. *New Journal of Physics, 11*, 093024.
53. Preis, T. (2010). Simulating the microstructure of financial markets. *Journal of Physics and Chemistry of Solids, 221*, 012019.
54. Preis, T., Kenett, D. Y., Stanley, H. E., Helbing, D., & Ben-Jacob, E. (2012). Quantifying the behavior of stock correlations under market stress. *Scientific Reports, 2*, 752.
55. Sornette, D., & Von der Becke, S. (2011). Complexity clouds finance-risk models. *Nature, 471*, 166.
56. Stanley, H. E., Buldyrev, S. V., Franzese, G., Havlin, S., Mallamace, F., Kumar, P., et al. (2010). Correlated randomness and switching phenomena. *Physica A, 389*, 2880.
57. Moat, H. S., Curme, C., Stanley, H. E.,& Preis, T. (2014). Anticipating stock market movements with Google and Wikipedia. In D. Matrasulov & H. E. Stanley (Eds.), *Nonlinear phenomena in complex systems: From nano to macro scale* (pp. 47–59). Dordrecht, Netherlands: Springer.
58. Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99.
59. Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics, 106*, 1039.

Chapter 6
Online Interactions

Lilian Weng, Filippo Menczer, and Alessandro Flammini

Abstract The ubiquitous use of the Internet has led to the emergence of countless social media and social networking platforms, which generate large-scale digital data records of human behaviors online. Here we review the literature on online interactions, focusing on two main themes: social link formation and online communication. The former is often studied in the context of network evolution models and link prediction or recommendation tasks; the latter combines classic social science theories on collective human behaviors with analysis of big data enabled by advanced computation techniques. But the structure of the network, and the flow of information through the network influence each other. We present a case study to illustrate the connections between social link formation and online communication. Analysis of longitudinal micro-blogging data reveals that people tend to follow others after seeing many messages by them. We believe that research on online interactions will benefit from a deeper understanding of the mutual interactions between the dynamics *on* the network (communication) and the dynamics *of* the network (evolution).

6.1 Introduction

User activity within online socio-technical systems is exploding. Social and micro-blogging networks such as Facebook, Twitter, and Google Plus host the information sharing activity of billions of users every day. Using these network platforms, people communicate ideas, opinions, videos, and photos among their circles of friends and followers across the world. These interactions generate an unprecedented amount of data that can be used as a social observatory, providing a unique opportunity to study the mechanisms behind human interactions with a quantitative approach [1–4].

Research on human online interactions revolves around two main themes: social link formation and online communication. People can build virtual connections with others to subscribe to their messages (i.e., *following* on Twitter and Google

L. Weng (✉) • F. Menczer • A. Flammini
Center for Complex Networks and Systems Research, School of Informatics and Computing, Indiana University Bloomington, Bloomington, IN, USA
e-mail: weng@umail.iu.edu

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7_6

Plus) or to claim a mutual friendships between them (i.e., *friending* on Facebook). The established social links enable people to easily communicate with friends, sharing and spreading information on top of the social network.

Most work on social link formation frames the problem as network evolution modeling, in which each new link creation is driven by predefined mechanisms that resemble the observations from real-world data. Meanwhile, many algorithms have been presented to predict or recommend missing links in a given network.

Communication dynamics is a long-lasting research topic in the social sciences. Many theories of how people interact, exchange ideas, and influence each other have been proposed decades ago. With the availability of big data and advanced computational power, researchers can apply, verify, and enrich classic hypotheses on human behaviors, leveraging the capacity to collect and analyze data on a large scale to reveal patterns of human interactions [3]. New interdisciplinary research fields, namely *computational social science* and *human dynamics*, have emerged in such a scenario [3, 5].

6.1.1 Social Link Formation

Understanding the formation of online social links is a key ingredient for modeling the evolution of online social networks, as the rules for creating new links determine the network structure in time. Various models were introduced to capture the growth and evolution of network topology, as well as different characteristics of complex networks. Most such models focused on defining basic mechanisms that drive link creation [6–10].

6.1.1.1 Classic Network Evolution Models

The random network model is the oldest attempt at characterizing a non-regular network. Although, strictly speaking, it is not an evolutionary model, it can be regarded as such when links are added sequentially to the network. It is characterized by the fact that each link exists independently with the same probability [11]. The random network inspired many subsequent studies in network science, but it was not thought to reproduce several crucial properties of social networks [12], such as the small-world effect [13–15], high clustering coefficient [14, 16, 17], temporal dynamics [18, 19], information propagation [20], and heterogeneous distributions in connectivity patterns [21–25]. Many such characteristics were indeed yet undiscovered at the time when the random network model was proposed.

The small-world effect, also known as "six degrees of separation," originated from the Milgram experiment [13], in which the average length of communication chains between two random individuals was found to be around six—smaller than what expected in a regular network, such as a lattice. The Watts-Strogatz model was designed to reproduce the small-world phenomenon by rewiring each link in a regular network with a small probability [14].

A scale-free network has a power-law degree distribution, commonly seen in many real-world social networks such as the film actor network, the scientific collaboration network, and the citation network [6, 8, 21]. A highly skewed degree distribution in a social network indicates that, although the majority of nodes is poorly connected, there is a consistent group of them (compared to what happens in a random graph) that is extremely well connected, and whose collective connections account for a relevant portion of the entire set of links in the network. In a large group of people, only a few are extremely popular and most others do not have many contacts. Many models have been proposed to reproduce the heterogeneous distribution in connectivity [12, 22–25]. The Barabási-Albert model generates a scale-free network by continuously adding new nodes into the system ("growth") and connecting them with other nodes with preference to high-degree ones ("preferential attachment") [21]. Motivated by the structure of the Web graph, the copying model adds a new node into the network at a time and links it to a random existing node or its neighbors [24, 25]. Another model proposed by Newman, Watts, and Strogatz aimed to build up a random graph with arbitrary degree settings [12]. The *ranking model* grows the network according to a rank of the nodes by any given prestige measure, reproducing arbitrary power-law degree distributions [23].

6.1.1.2 Models with Social Components

The preferential attachment mechanisms in the Barabási-Albert and ranking models have a clear rationale in the social context: people prefer to form edges with well-connected individuals, such as celebrities. However, this prescription alone is not sufficient to reproduce several other important features of real networks. Other models have been put forth to fill the gap, including ingredients such as homophily [26–29], triadic closure [15–17, 30, 31], hierarchical structure [32], and information diffusion [20, 33].

Homophily can be regarded as people's propensity for linking with similar others [26, 28, 29]. The triadic closure mechanism is based on the intuition that two individuals with mutual friends have a higher probability to establish a new contact [30, 31]. This tendency was observed in both undirected and directed online social networks and incorporated into several network growth models [15–17]. In particular, Leskovec et al. tested triadic closure against many other mechanisms in four different large-scale social networks. By using maximum likelihood estimation (MLE) [34], they identified triadic closure as the best rule, among those considered, to explain link creation and to reproduce the clustering coefficient and the degree distribution of the real networks under study [16].

6.1.1.3 Link Prediction

Missing link prediction algorithms aim at inferring new social connections that may happen in the near future given a current snapshot of the network structure. The prediction has practical applications in the online systems; one of the most popular use cases is to provide recommendations for new contacts (i.e., "People You May Know" on Facebook and LinkedIn). Common approaches consider link prediction as a classification task or ranking problem, using node similarity [35, 36], the hierarchical structure of the network [32], random walks [16], supervised random walks [37], graphical models [38], and user profile features [39].

One class of link prediction algorithms is designed on the premise that only the network topology is known. Liben-Nowell and Kleinberg [36] examined and compared a rich set of metrics for quantifying the similarity between a pair of nodes in the social network, where high similarity implies high likelihood of being connected on the basis of homophily. Tested metrics were built upon different network topological features associated with each node, including overlap between neighbor sets, preferential attachment, shortest path distance, and PageRank hitting time. The analysis identified Adamic-Adar similarity [35] as the metric providing the best performance. Clauset, Moore, and Newman [32] explored the observation that real-world networks often exhibit community structure and hierarchical organization. They proposed a link prediction method that uses knowledge about the network's hierarchical structure.

Other link prediction algorithms depend on additional attributes of existing individuals or connections, as well as the topology of the network. Supervised random walks can incorporate the knowledge of nodes and links so that a random walker is guided to follow preferred paths with higher probability [37]. Attributes may include, for example, the number of co-authored papers in a collaboration network and the frequency of interactions between a pair of friends on Facebook. In online systems like Flickr and Last.fm, users can annotate content revealing their topical interests. Schifanella et al. found that users with similar interests in these networks are more likely to be friends and proposed to use the similarity between user annotation metadata as a predictor of missing social links [39].

6.1.2 Communication Dynamics

Early models concerning communication dynamics were inspired by studies of epidemics, assuming that a piece of information could pass from one individual to another through social contacts [40–42]. Recently, starting from observations and theories in social sciences, a wealth of computational models have been proposed to describe human communication.

6.1.2.1 Threshold Model

One class of models is based on the idea of a threshold: people tend to follow the same trends as most of their friends do [43, 44]. A threshold can be defined as the number or fraction of others who must make a decision before a given actor does the same. Many empirical studies have demonstrated the existence of such a threshold in social and behavioral contagion online [45–48]. Threshold models have been widely applied to understand the diffusion of rumors, norms, strikes, voting, educational attainment, migration, and other human behaviors [43, 49, 50], and extended to study the role of competition for finite attention [51].

6.1.2.2 Homophily

The principle of homophily states that similar people are more likely to have contact than dissimilar ones [27, 52, 53]. The existence of homophily in social groups has been supported by various empirical observations and experiments in online settings [26, 29, 39, 54, 55]. Crandall et al. proposed a homophily-based model to predict a user's future activity and interactions with others according to user similarities [56].

A feedback loop has been claimed to result in increasing similarity among users: people grow to resemble their friends because of social (peer) influence, while being more likely to form links with similar people (homophily) [29, 56]. Such a feedback loop could lead to the so-called *echo-chamber* effect, by which people are exposed to limited diversity of opinions in online social networks [57, 58]. Though it is hard to fully distinguish between peer influence and homophily [59], the latter effect contributes to promoting behavioral contagion [54].

6.1.2.3 Weak Tie Hypothesis

Friendships vary in their intensity and intimacy. The concept of tie strength has been introduced to capture this variation: strong ties are our closest confidants and supporters, while weak ties, to whom we feel less close, comprise the majority of our personal networks. Granovetter defined the strength of social ties proportionally to the size of shared social circles and proposed the weak-tie hypothesis [31, 60], according to which weak ties do not carry as much communication as strong ties, but act as bridges between communities and thus as important channels for novel information.

Following up on Granovetter's work, many empirical studies have tested the weak-tie hypothesis [61–69]. Brown and Reingen found an important bridging function of weak ties in word-of-month referral behavior, allowing information to travel from one distinct subgroup of referral actors to another [62]. Gilbert and Karahalios tested several dimensions of tie strength on social media, revealing that both intensity of communication and intimate language are strong indicators of relationship closeness [69]. Onnela et al. analyzed a mobile call network and showed that individuals in clusters tend to communicate more, while ties between clusters have less traffic [68]. Bakshy et al. compared individual adoption rates on Facebook when an external URL shared by friends is or is not included in the newsfeed and found that although stronger ties are individually more influential in persuading others to adopt and spread information, more abundant weak ties are responsible for the propagation of novel information [61].

In summary, strong ties are believed to provide greater emotional support [69, 70] and to be more influential [61, 62, 71], while weak ties provide novel information and connect us to opportunities outside our immediate circles [31, 68, 72].

6.1.2.4 Limited Attention

People have limited attention during communication. This constraint may be related to a cognitive limit on the number of stable social relationships that one can sustain, as postulated by Dunbar [73] and later supported by analyses of Twitter data [74, 75]. Huberman, Romero, and Wu defined friends of a Twitter user as those who have been mentioned at least twice. They found that most users have a very small number of friends compared to a large number of followers, and the friend network is more influential than the follower network in driving Twitter usage [74]. Wu and Huberman analyzed the dynamics of collective attention on Digg.com and modeled the delay of collective attention with a single novelty factor. Their measurements indicated that novelty within groups decays with a stretched-exponential law, suggesting the existence of a natural time scale over which attention fades [76].

6.1.2.5 Communication Dynamics on Evolving Networks

The large majority of studies on communication dynamics consider a static underlying social network, under the assumption that the network evolves on a slower time scale than that characteristic of the information spread. Recent research has addressed the modeling of cases in which the time scales of communication dynamics and network evolution are comparable. These approaches consider the two processes as either independent [19, 77] or coupled [33, 78, 79]. In particular, the studies focused on the former case considered mainly epidemic processes in which links are deleted or rewired according to the disease status of each node [78, 79].

6.2 Case Study: Traffic-Based Social Link Formation

We probe into the effects of information diffusion in shaping the evolution of the social network structure. As a case study, we present a longitudinal analysis of micro-blogging data to better understand the strategies employed by users when expanding their social circles. While the network structure affects the spread of information, the network is, in turn, shaped by this communication activity. This leads us to hypothesize a mechanism whereby people tend to follow others after seeing many messages from them. Interestingly, the coupling of social link formation and information sharing allows to depict a more accurate and comprehensive view of the network evolution [33].

We analyzed a dataset collected from *Yahoo! Meme*,[1] including the entire history of the system from April 2009 until March 2010. A user j following a user i is represented in the follower network by a directed edge $\ell = (i, j)$, indicating j can receive messages posted by i. We adopt this notation, in which the link creator is the target, to emphasize the direction of information flow. In our notation, the in-degree of a node j is the number of people followed by j. Users can repost received messages, which become visible to their followers. When user j reposts content from i, we infer a flow of information from i to j. Each link is weighted by the numbers of messages from i that are reposted or seen by j. At the end of the observation period, the Yahoo! Meme follower network consisted of 128,199 users with at least one edge, connected by a total of 3,485,361 directed edges.

Social micro-blogging networks, such as Twitter, Google Plus, Sina Weibo, and Yahoo! Meme, are designed for information sharing. As illustrated in Fig. 6.1, the dynamics *on* the network directly affects the dynamics *of* the networks, and vice versa. In this case study, we investigate the individual strategies that lead to the creation of new social links. We characterize link creation processes with a set of parameters associated with different link creation strategies, estimated by a Maximum-Likelihood approach [34]. This analysis will show that triadic closure does have a strong effect on link formation, but shortcuts based on traffic are another indispensable factor in interpreting network evolution.

6.2.1 Link Creation Mechanisms

When users post or repost messages, all their followers can see these posts and might decide to repost them, generating spreading paths that, when taken together, form cascade networks. When receiving a reposted message, a Meme user in such a path can see both the *grandparent* (G, the user two steps ahead in the path) and the *origin* (O, original source). A user may decide to follow a grandparent or origin, receiving their future messages directly. These new links create *shortcuts* connecting users at any distance in the network. A triadic closure occurs when a user follows a *triadic node* (Δ, the user two steps away in the follower network). The definitions of different types of link creation mechanisms are illustrated in Fig. 6.2.

[1] Yahoo! Meme was a social micro-blogging system similar to Twitter, active between 2009 and 2012.

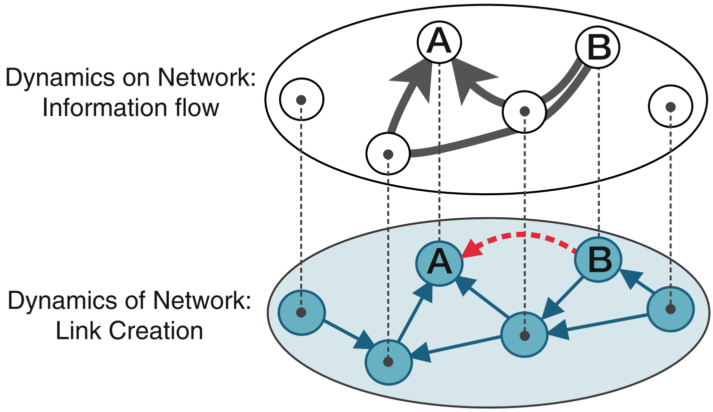


Fig. 6.1 The dynamics *of* and *on* the network are strongly coupled. The *bottom layer* illustrates the social network structure, where the *blue arrows* represent "follow" relationships with the direction of information flow. The *dashed red arrow* marks a newly created link. The *upper layer* depicts the flow of information between people in the same group, leading to the creation of the new link. The social network structure constrains communication patterns, but information propagated through the network also affect how agents behave and ultimately how the network changes and grows

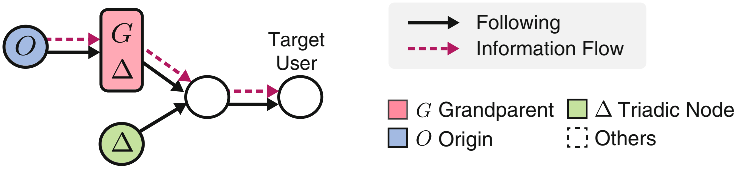


Fig. 6.2 Illustration of link creation mechanisms. A grandparent node is a special case of triadic node, from which or through which information has reached the target user. Therefore traffic-based shortcuts to grandparent nodes are a subset of triadic closures

6.2.1.1 Statistical Analyses of Shortcuts

To quantify the statistical tendency of users to create shortcuts, let us consider every single link creation in the data as an independent event. We test the null hypotheses that links to grandparents, origins, and triadic nodes are generated by choosing targets at random among the users not already followed by the creator.

We label each link ℓ by its creation order, $1 \leq \ell \leq L$, where L is the total number of links. For each link, we can compute the likelihood of following a grandparent by chance:

$$p_G(\ell) = \frac{N_G(\ell)}{N(\ell) - k(\ell) - 1}, \tag{6.1}$$

where $N_G(\ell)$ is the number of distinct grandparents seen by the creator of ℓ at the moment when ℓ is about to be created; $N(\ell)$ is the number of available users in the system when ℓ is to be created; $k(\ell)$ is the in-degree of ℓ's creator at the same moment; and the denominator is the number of potential candidates to be followed.

The indicator function for each link ℓ denotes whether the link connects with a grandparent or not in the real data:

$$\mathbf{1}_G(\ell) = \begin{cases} 1 \text{ if } \ell \text{ links to a grandparent in the data} \\ 0 \text{ otherwise.} \end{cases} \tag{6.2}$$

The expected number of links to grandparents according to the null hypothesis can be then computed as:

$$E_G = \sum_{\ell=1}^{L} p_G(\ell) \tag{6.3}$$

and its variance is given by:

$$\sigma_G^2 = \sum_{\ell=1}^{L} p_G(\ell)\,(1 - p_G(\ell)) \tag{6.4}$$

while the corresponding empirical number is:

$$S_G = \sum_{\ell=1}^{L} \mathbf{1}_G(\ell). \tag{6.5}$$

According to the Lyapunov central limit theorem,[2] the variable $z_G = (S_G - E_G)/\sigma_G$ is distributed according to a standard normal $\mathcal{N}(0, 1)$. For linking to origins (O) or triadic nodes (Δ), we can define z_O and z_Δ similarly. In all three cases, using a z-test, we can reject the null hypotheses with high confidence ($p < 10^{-10}$). We conclude that links established by following grandparents, origins or triadic nodes happen much more frequently than by random connection. These link creation mechanisms have important roles in the evolution of the social network.

[2] Lyapunov's condition, $\frac{1}{\sigma_n^4}\sum_{\ell=1}^{n} E[(X(\ell) - p(\ell))^4] \overset{n\to\infty}{\longrightarrow} 0$ where $X(\ell)$ is a random Bernoulli variable with success probability $p(\ell)$ [80], is consistent with numerical tests. Details are omitted for brevity.

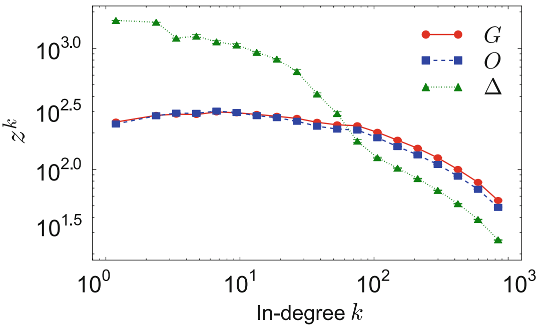


Fig. 6.3 Individual preferences for following grandparents (*red circles*), origins (*blue squares*), and triadic nodes (*green triangles*) change with the in-degree of the link creator

6.2.1.2 User Preference

The variables z_G, z_O, and z_Δ, as defined above, measure how much more likely links of a given type are formed than by chance—in other words, how strong individual preferences are for following grandparents, origins or triadic nodes. To study the dependence of the link formation tendencies on the different stages of an individual's lifetime, let us compute z^k_G, z^k_O, and z^k_Δ for links created by users with in-degree k, that is, those who are following k users at the time when the link is created. Figure 6.3 shows that the principle of triadic closure dominates user behavior when one follows a small number of users ($k < 75$). In the early stages, one does not receive much traffic, so it is natural to follow people based on local social circles, consistently with triadic closure. However, users who have been active for a long time and have followed many people ($k > 75$) have more channels through which they monitor traffic. This creates an opportunity to follow others from whom they have seen messages in the past.

6.2.1.3 Link Efficiency

In information diffusion networks like Twitter and Yahoo! Meme, social links may have a key efficiency function of shortening the distance between information creators and consumers. An efficient link should be able to convey more information to the follower compared to less efficient links. Hence we define the *efficiency* of link ℓ as the average number of posts seen or reposted through ℓ during one time unit after its creation:

$$\eta_{\text{seen}} = \frac{w_{\text{seen}}(\ell)}{T - t(\ell)}, \quad \eta_{\text{repost}} = \frac{w_{\text{repost}}(\ell)}{T - t(\ell)}, \tag{6.6}$$

where $w(\ell)$ is the number of messages seen or reposted through ℓ; $t(\ell)$ is the time when ℓ was created; and T is the time of the last action recorded in our dataset. Both seen and reposted messages are considered, as they represent different types of traffic; the former are what is visible to a user, and the latter are what a user is

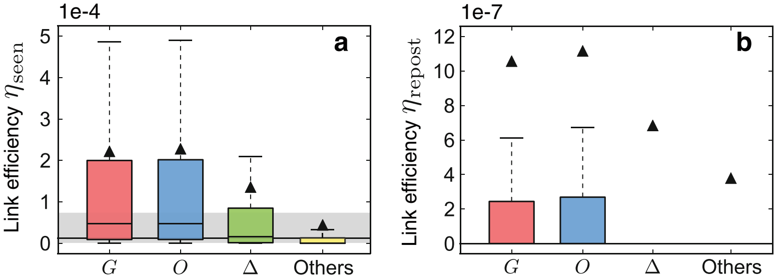


Fig. 6.4 Efficiency of links created according to different mechanisms, or average number of messages (**a**) seen or (**b**) reposted per time unit. Each *box* shows data within lower and upper quartile. Whiskers represent the 99th percentile. The *triangle* and *line* in a box represent the mean and median, respectively. The *black line* and *grey area* across the entire figure mark the median and interquartile range of the measure across all links, respectively

willing to share. We compute the link efficiency of every grandparent, origin, and triadic closure link. As shown in Fig. 6.4, both grandparent and origin links exhibit higher efficiency than triadic closure links, irrespective of the type of traffic. By shortening the paths of information flows, more posts from the content generators reach the consumers.

6.2.2 *Rules of Network Evolution*

To infer the different link creation strategies from the observed data, we characterize users with a set of probabilities associated with different actions, and approximate these parameters by MLE [34]. For each link ℓ, we know the actual creator and the target; we can thus compute the likelihood $f(\ell|\Gamma,\Theta)$ of the target being followed by the creator according to a particular strategy Γ, given the network configuration Θ at the time when ℓ is created. The likelihoods associated with different strategies can be mixed according to the parameters to obtain a model of link creation behavior. Finally, assuming that link creation events are independent, we can derive the likelihood of obtaining the empirical network from the model by the product of likelihoods associated with every link. The higher the value of the likelihood function, the more *accurate* the model.

6.2.2.1 Simple Strategy

We consider five link creation mechanisms and their combinations:

- *Random* (Rand): follow a randomly selected user who is not yet followed
- *Triadic closure* (Δ): follow a randomly selected triadic node
- *Grandparent* (G): follow a randomly selected grandparent

- *Origin* (O): follow a randomly selected origin
- *Traffic shortcut* ($G \cup O$): follow a randomly selected grandparent or origin

Other mechanisms for link creation could be similarly incorporated, such as social balance [81] and preferential attachment [21]. However, preferential attachment is built on the assumption that everyone knows the global connectivity of everyone else, which is not very realistic. The strategies considered here essentially reproduce and extend the copy model [25], approximating preferential attachment with only local knowledge.

To model link creation with a single strategy, we can use a parameter p for the probability of using that strategy, while a random user is followed with probability $1 - p$. The calculation of maximum likelihood, taking the single strategy of grandparents as an example, is as follows:

$$\begin{aligned}\mathscr{L}_G(p) &= \prod_{\ell=1}^{L} (pf(\ell|G,\Theta) + (1-p)f(\ell|\text{Rand},\Theta)) \\ &= \prod_{\ell=1}^{L} \left(p\frac{\mathbf{1}_G(\ell)}{N_G(\ell)} + (1-p)\frac{1}{N(\ell)-k(\ell)-1}\right) \\ &= \prod_{\mathbf{1}_G(\ell)=1} \left(\frac{p}{N_G(\ell)} + \frac{1-p}{N(\ell)-k(\ell)-1}\right) \prod_{\mathbf{1}_G(\ell)=0} \frac{1-p}{N(\ell)-k(\ell)-1}. \end{aligned} \quad (6.7)$$

Note that since a follow action can be ascribed to multiple strategies, it can contribute to multiple terms in the log-likelihood expression. For instance, a link could be counted in both $f(\ell|G,\Theta)$ and $f(\ell|\text{Rand},\Theta)$. For numerically stable computation, we maximize the log-likelihood:

$$\begin{aligned}\log \mathscr{L}_G(p) = &\sum_{\mathbf{1}_G(\ell)=1} \ln\left(\frac{p}{N_G(\ell)} + \frac{1-p}{N(\ell)-k(\ell)-1}\right) \\ &+ \sum_{\mathbf{1}_G(\ell)=0} \ln \frac{1-p}{N(\ell)-k(\ell)-1}. \end{aligned} \quad (6.8)$$

Similar expressions of log-likelihood can be obtained for other strategies (Δ, O, and $G \cup O$).

It is not trivial to obtain the best p analytically, so we explore the values of $p \in (0,1)$ numerically (see Fig. 6.5). Triadic closure dominates as a single strategy, with $p_\Delta = 0.82$. Traffic-based strategies alone account for about 20 % of the links.

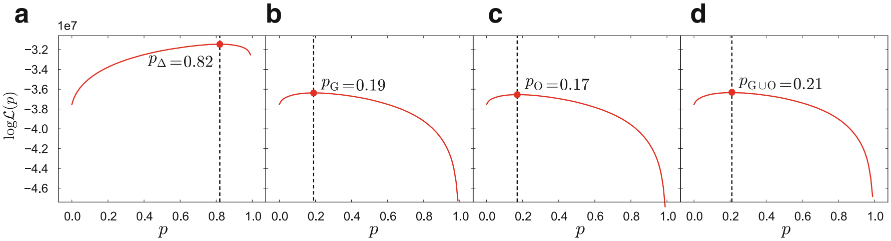


Fig. 6.5 The plot of the log-likelihood $\log \mathscr{L}(p)$ as a function of link creation strategy probabilities for models with a single strategy. The *red circles* mark the maximized $\log \mathscr{L}(p)$. (**a**) Triadic closure, (**b**) Grandfather, (**c**) Origin, (**d**) Grandfather + Origin

6.2.2.2 Combined Strategies

For a more realistic model of the empirical data, let us consider combined strategies with both triadic closure and traffic-based shortcuts. For each link ℓ, the follower with probability p_1 creates a shortcut by linking to a grandparent or an origin ($G \cup O$); with probability p_2 follows a triadic node (Δ); and with probability $1 - p_1 - p_2$ connects to a random node. Taking the combined strategy with grandparent as an example, we compute the log-likelihood as:

$$\log \mathscr{L}_{G+\Delta}(p_1, p_2) = \log \prod_{\ell=1}^{L} \Big[p_1 f(\ell | G \cup O, \Theta) + p_2 f(\ell | \Delta, \Theta) \\ + (1 - p_1 - p_2) f(\ell | \mathrm{Rand}, \Theta) \Big]. \quad (6.9)$$

The detailed derivations of the likelihood functions and the cases of the other combined strategies are omitted for brevity.

Once again, we numerically explore the values of p_1 and p_2 in the unit square to maximize the log-likelihood. The likelihood landscape for the combined strategy considering both grandparents and origins as well as triadic closure is shown in Fig. 6.6. The parameter settings and the maximum likelihood values for all tested models are listed in Table 6.1. We can compare the quality of these models by comparing their maximized $\log \mathscr{L}$'s. The combined models with both traffic shortcuts and triadic closure yield the best accuracy. In these models, triadic closure accounts for 71 % of the links, grandparents and origins for 12 %, and the rest of the links are created at random.

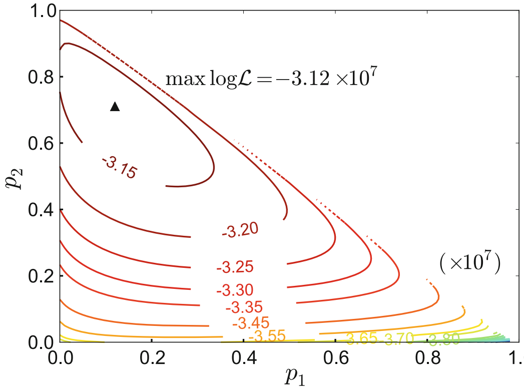


Fig. 6.6 The contour plot of log-likelihood $\log \mathscr{L}(p_1, p_2)$ for the combined strategy of creating traffic shortcuts ($G \cup O$) with probability p_1 and triadic closure links (Δ) with probability p_2. The *black triangle* marks the optimum

Table 6.1 The best parameters in different models and corresponding values of maximized log-likelihood function

Strategy	Model	Parameters		max log $\mathscr{L}$
Single	Δ	$p = 0.82$		-3.15×10^7
	G	$p = 0.19$		-3.64×10^7
	O	$p = 0.17$		-3.65×10^7
	$G \cup$ O	$p = 0.21$		-3.63×10^7
Combined	$G + \Delta$	$p_1 = 0.12$	$p_2 = 0.71$	-3.12×10^7
	O $+ \Delta$	$p_1 = 0.10$	$p_2 = 0.73$	-3.13×10^7
	$G \cup$ O $+ \Delta$	$p_1 = 0.12$	$p_2 = 0.71$	-3.12×10^7

6.3 Discussion

Social link formation and information sharing are two major tracks of research on online interactions. The mechanisms of new link creation determine the topology of linkages among individuals, and the underlying network structure is critical for the dynamics of the diffusion process [6–8, 51]. At the same time, as many social links are driven by the need for more efficient information sharing in social media sites, social link formation is greatly affected by communication activity. Both the evolving structure of the social network and information diffusion have been studied for decades, but the coupling between these dynamical processes has not been well explored. In the present case study, we demonstrate a feedback loop between these two dynamics. While triadic closure is the dominant mechanism for social network evolution, it is mainly relevant in the early stages of a user's lifetime. As time progresses, the traffic generated by communication dynamics on the network becomes an indispensable component for user linking behavior. As users become more active and influential, their links create shortcuts that make the spread of information more efficient in the network.

Studies of online interactions—how social networks evolve and how information spreads—help us gain a better understanding of social influence, user behavior, and network efficiency in the context of online systems. The coupling between dynamics *of* and *on* the network provides us with powerful insights into human interactions in the digital world.

Acknowledgements We thank Jacob Ratkiewicz, Nicola Perra, Bruno Gonçalves, Carlos Castillo, Francesco Bonchi, and Rossano Schifanella for their contributions to the case study presented in this chapter; Yahoo Labs for making the *Meme* data available; and the James S. McDonnell Foundation, National Science Foundation, and DARPA grant W911NF-12-1-0037 for partial support of this research.

References

1. Cho, A. (2009). Ourselves and our interactions: The ultimate physics problem? *Science, 325*, 406.
2. Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. In *Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.
3. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. -L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721–723.
4. Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science, 325*(5939), 425–428.
5. Barabási, A. -L., & Albert, R. (2005). The origin of bursts and heavy tails in human dynamics. *Nature, 435*(7039), 207–211.
6. Albert, R., & Barabási, A. -L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47–97.
7. Barrat, A., Barthélemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge: Cambridge University Press.
8. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167–256.
9. Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
10. Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge: Cambridge University Press.
11. Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5*, 17–61.
12. Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences (PNAS), 99*(Suppl 1), 2566–2572.
13. Milgram, S. (1967). The small world problem. *Psychology Today, 2*(1), 60–67.
14. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.
15. Krackhardt, D., & Handcock, M. S. (2007). Heider vs. simmel: Emergent features in dynamic structure. In E. M. Airoldi, D. M. Blei, S. E. Fienberg, A. Goldenberg, E. P. Xing, A. X. Zheng (Eds.), *Statistical network analysis: models, issues, and new directions* (pp. 14–27). Berlin: Springer.
16. Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 462–470).
17. Romero, D. M., & Kleinberg, J. (2010). The directed closure process in hybrid social-information networks, with an analysis of link formation on Twitter. In *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM).*

18. Perra, N., Gonçalves, B., Pastor-Satorras, R., & Vespignani, A. (2012). Time scales and dynamical processes in activity driven networks. *Nature Scientific Reports, 2*, 469.
19. Rocha, L. E. C., Liljeros, F., & Holme, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Computational Biology, 7*(3), e1001109.
20. Barbieri, N., Bonchi, F., & Manco, G. (2013). Cascade-based community detection. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 33–42).
21. Barabási, A. -L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.
22. Dorogovtsev, S., Mendes, J., & Samukhin, A. (2000). Structure of growing networks with preferential linking. *Physical Review Letters, 85*(21), 4633–4636.
23. Fortunato, S., Flammini, A., & Menczer, F. (2006). Scale-free network growth by ranking. *Physical Review Letters, 96*(21), 218701.
24. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). The web as a graph: measurements, models and methods. *Lecture Notes in Computer Science (LNCS), 1627*, 1–18.
25. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. (2000). Stochastic models for the web graph. In *Proceedings of IEEE Annual Symposium on Foundations of Computer Science* (pp. 57–65).
26. Gallos, L., Rybski, D., Liljeros, F., Havlin, S., & Makse, H. (2012). How people interact in evolving online affiliation networks. *Physical Review X, 2*(3), 031014.
27. McPherson, M., Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*(1), 415–444.
28. Papadopoulos, F., Kitsak, M., Ángeles Serrano, M., Boguña, M., & Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature, 489*(7417), 537–540.
29. Weng, L., & Lento, T. (2014). Topic-based clusters in egocentric networks on Facebook. In *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM).*
30. Simmel, G., & Wolff, K. H. (1950). *The Sociology of Georg Simmel*. New York: The Free Press.
31. Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1.
32. Clauset, A., Moore, C., & Newman, M. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature, 453*(1), 98–101.
33. Weng, L., Ratkiewicz, J., Perra, N., Gonçalves, B., Castillo, C., Bonchi, F., et al. (2013). The role of information diffusion in the evolution of social networks. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 356–364).
34. Cowan, G. (1998). *Statistical data analysis*. Oxford: Oxford Science Publications.
35. Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social Networks, 25*(3), 211–230.
36. Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of American Society for Information Science and Technology (JASIST), 58*(7), 1019–1031.
37. Backstrom, L., & Leskovec, J. (2011). Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 635–644). ACM.
38. Lou, T., Tang, J., Hopcroft, J., Fang, Z., & Ding, X. (2010). Learning to predict reciprocity and triadic closure in social networks. *ACM Transactions on Embedded Computing Systems, 9*(4), 5.
39. Schifanella, R., Barrat, A., Cattuto, C., Markines, B., & Menczer, F. (2010). Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 271–280).
40. Anderson, R. M., May, R. M., & Anderson, B. (1992). *Infectious diseases of humans: Dynamics and control* (Vol. 28). Oxford: Oxford University Press.

41. Daley, D. J., & Kendall, D. G. (1964). Epidemics and rumours. *Nature, 204*(4963), 1118–1119.
42. Goffman, W., & Newill, V. A. (1964). Generalization of epidemic theory: An application to the transmission of ideas. *Nature, 204*(4955), 225–228.
43. Granovetter, M. S. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420–1433.
44. Morris, S. (2000). Contagion. *Review of Economic Studies, 67*(1), 57–78.
45. Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 44–54).
46. Bakshy, E., Karrer, B., & Adamic, L. (2009). Social influence and the diffusion of user-created content. In *Proceedings of ACM Conference on Electronic Commerce* (pp. 325–334).
47. Cosley, D., Huttenlocher, D., Kleinberg, J., Lan, X., & Suri, S. (2010). Sequential influence models in social networks. In *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM).*
48. Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of International Conference on World Wide Web (WWW).*
49. Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–1197.
50. Weng, L., Menczer, F., & Ahn, Y. -Y. (2013). Virality prediction and community structure in social networks. *Nature Scientific Reports, 3*(2522).
51. Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Nature Scientific Reports*, 2(335).
52. Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network1. *American Journal of Sociology, 115*(2), 405–450.
53. McPherson, J. M., & Smith-Lovin, L. (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review, 52*(3), 370–379.
54. Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences (PNAS), 106*(51), 21544–21549.
55. Şimşek, O., & Jensen, D. (2008). Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences (PNAS), 105*(35), 12758–12762.
56. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 160–168).
57. Jamieson, K. H., & Cappella, J. N. (2009). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford: Oxford University Press.
58. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Political polarization on twitter. In *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM).*
59. Shalizi, C., & Thomas, A. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research, 40*(2), 211–239.
60. Granovetter, M. (1995). *Getting a job: A study of contacts and careers*. Chicago: University of Chicago Press.
61. Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of ACM International World Wide Web Conference (WWW)* (pp. 519–528).
62. Brown, J., & Reingen, P. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer Research, 14*(3), 350–362.
63. Friedkin, N. (1980). A test of structural features of granovetter's strength of weak ties theory. *Social Networks, 2*(4), 411–422.

64. Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*(1), 201–233.
65. Levin, D. Z., & Cross, R. (2004). The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. *Management Science, 50*(11), 1477–1490.
66. Lin, N., Ensel, W. M., & Vaughn, J. C. (1981). Social resources and strength of ties: Structural factors in occupational status attainment. *American Sociological Review, 46*, 393–405.
67. Nelson, R. E. (1989). The strength of strong ties: Social networks and intergroup conflict in organizations. *Academy of Management Journal, 32*(2), 377–401.
68. Onnela, J. -P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., et al. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences (PNAS), 104*(18), 7332–7336.
69. Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of ACM International Conference on Human Factors in Computing Systems (CHI)* (pp. 211–220).
70. Wellman, B., & Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American Journal of Sociology, 96*(3), 558–588.
71. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., et al. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295–298.
72. Putnam, R. D. (2001). *Bowling alone: The collapse and revival of American community*. New York: Simon and Schuster.
73. Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology, 9*(10), 178–190.
74. Huberman, B., Romero, D., & Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday, 14*(1), 8.
75. Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PLoS One, 6*(8), e22656.
76. Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences (PNAS), 104*(45), 17599–17601.
77. Perra, N., Baronchelli, A., Mocanu, D., Gonçalves, B., Pastor-Satorras, R., & Vespignani, A. (2012). Random walks and search in time varying networks. *Physical Review Letters, 109*, 238701.
78. Shaw, L. B., & Schwartz, I. B. (2010). Enhanced vaccine control of epidemics in adaptive networks. *Physical Review E, 81*, 046120.
79. Volz, E., & Meyers, L. A. (2009). Epidemic thresholds in dynamic contact networks. *Journal of the Royal Society Interface, 6*, 233241.
80. Billingsley, P. (1995). *Probability and measure* (p. 362). New York: Wiley.
81. Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: reasoning about a highly connected world.* Cambridge: Cambridge University Press.

Chapter 7
The Contagion of Prosocial Behavior and the Emergence of Voluntary-Contribution Communities

Milena Tsvetkova and Michael Macy

Abstract Every day, millions of people write online restaurant reviews, leave product ratings, provide answers to an unknown user's question, or contribute lines of code to open-source software, all without any direct reward or recognition. People help strangers offline as well, as when people anonymously donate blood or stop to help a stranded motorist, but these behaviors are relatively rare compared to the pervasiveness of online communities based on user-generated content. Why are mutual-help communities far more common online than in traditional offline settings that are not mediated by the Internet? We address this puzzle in two steps. We begin with empirical evidence from an online experiment that tests two mechanisms for the contagion of helping behavior: "generalized reciprocity" and "third-party influence". We then use an empirically calibrated agent-based model to show how these mechanisms interact with the rivalness of contributions, that is, the extent to which the benefit from a contribution is limited to just one beneficiary (as when helping a stranded motorist) or benefits many people at once (as when contributing a product review online). The results suggest that the non-rivalness of most user-generated content provides a plausible explanation for the rapid diffusion of helping behavior in online communities.

7.1 Introduction

The health regime we follow [1], the music we listen to [2], the new technologies we adopt [3], the news stories we read [4], and even the likelihood that we vote in an election [5] are all to a large degree influenced by our friends and peers.

M. Tsvetkova (✉)
Department of Sociology, Cornell University, 372 Uris Hall, Ithaca, NY 14853-7601, USA
e-mail: mvt9@cornell.edu

M. Macy
Department of Sociology and Department of Information Science, Social Dynamics Laboratory, Cornell University, 372 Uris Hall, Ithaca, NY 14853-7601, USA
e-mail: m.macy@cornell.edu

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_7

Many human behaviors spread through social contact, including some that are often assumed to be acquired independently, such as obesity and fertility [6].

Prosocial behavior has also been shown to be contagious. Fowler and Christakis [7] found experimental evidence that if you help someone, you not only increase the likelihood that they help others, but that those they help will also help others, and so on, out to three steps. Suri and Watts [8] and Jordan et al. [9] similarly found that generous behavior was contagious at least in direct interaction. These groundbreaking studies have provoked new questions. What are the mechanisms through which prosocial behavior spreads among strangers? How do these mechanisms affect the contagion dynamics? Can they lead to the emergence of cooperation in an initially non-cooperating population?

7.1.1 The Puzzle of Online Generosity

The puzzle of contagious generosity is compounded further by the emergence of online communities with user-generated content, from open source software development to advice forums to Wikipedia [10]. Why are mutual-help communities far more common online than in traditional offline settings that are not mediated by the internet?

We address this puzzle using an empirically calibrated agent based model. The results suggest that the answer may lie in the differences in the rivalness of online and offline public goods involving anonymous contribution. Many offline public goods—like blood donation, charities, and giving up one's seat—are rivalrous, meaning that the contribution transfers resources from the giver to a particular receiver. In contrast, many online public goods, especially in communities based on user-generated content, are non-rival—everyone in the community can benefit from a given contribution. The difference is not limited to the effect of non-rival incentives on the independent probability of contribution by a member of the community. Computer simulation shows that this "within individual" difference is amplified by the "between individual" effects of the contagion dynamics. More precisely, we identify two mechanisms of contagion—"generalized reciprocity" and "third-party influence" (TPI)—and show how these mechanisms interact with differences between rival and non-rival contributions to explain the spread of helping behavior in online communities.

7.1.2 Outline of a Theory of Prosocial Contagion

Previous research has suggested that there are two distinct mechanisms for the contagion of prosocial behavior among strangers: generalized reciprocity and TPI. Generalized reciprocity (GR) refers to cases in which those who *benefit* from a stranger's prosocial behavior behave more prosocially towards another in the future.

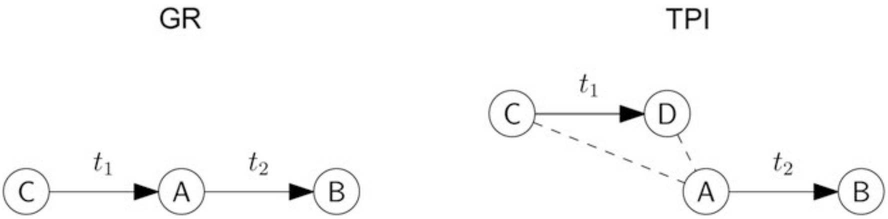


Fig. 7.1 Two mechanisms for the contagion of prosocial behavior. (*GR*) Generalized reciprocity: A helps B because C has helped A. (*TPI*) Third-party influence: A helps B because A has observed C help D

As diagrammed in Fig. 7.1, A helps B because C has helped A [11, 12]. TPI refers to cases in which those who *observe* prosocial behavior by strangers behave more prosocially towards a stranger: A helps B because A has seen C help D. GR characterizes "pay it forward" behavior triggered by a normative or affective response to being helped [13], while TPI characterizes social learning through imitation of others' behavior.

GR and TPI also differ in the pattern of transmission. GR transmits the contagion from person to person through direct contact and hence its contagious effect is constrained to the chain of those who were previously helped. In contrast, TPI has the potential to broadcast the contagion from one person to any number of observers. The interaction of the two mechanisms could generate a powerful self-reinforcing dynamic that dramatically increases the rate of prosocial behavior in an initially uncooperative population.

In this chapter, we summarize an online experiment that distinguished between the behavioral effects of the two contagion mechanisms [14] and use an agent-based model to investigate the contagion dynamics and the population-level outcomes that they entail. The empirical results show that receiving help can increase the willingness to be generous towards others, but observing help can have the opposite effect, particularly among those who have not received help. We use a threshold model with dynamic interaction structure and adaptive behavior to simulate a population of agents with this behavior. The computational experiments indicate that the agents can self-organize in communities based on voluntary contributions in two possible ways. On the one hand, when contributions are rival, a handful of altruists can lead to the emergence of small clusters of contributors as long as agents observe contribution beneficiaries in a relatively large radius (for example, via gossip) and unsatisfied agents are not too mobile. On the other hand, when contributions are non-rival, communities are much more likely to emerge and the level of contributions is higher when agents observe contributors rather than recipients. These two pathways roughly correspond to offline and online interactions. They offer explanation for the fact that cultures of kindness are rare for anonymous face-to-face interactions but common on the Web, for example, in the form of communities based on user generated content.

7.2 Testing Individual Mechanisms

Causal mechanisms are notoriously difficult to observe in natural settings, and controlled diffusion experiments with large groups are highly impractical in traditional laboratory settings. To test the two contagion mechanisms, we therefore designed and conducted a large behavioral experiment online. The experiment used anonymity to isolate the effects of GR and TPI from other cooperation-inducing mechanisms, including direct and indirect reciprocity, as well as peer pressure based on reputation effects. To isolate GR from TPI, we manipulated the extent to which participants received and observed help.

7.2.1 Online Experiment

The study was designed as a sequential two-player investment/gift-exchange game in groups of 150 with random partner selection. In the game, a participant could choose to return part of their payment so that another anonymous participant could benefit.

We first recruited a pool of potential participants by posting a task on the online crowdsourcing platform Amazon Mechanical Turk (AMT). The task invited AMT users to sign up for a study that offered the chance to earn up to $14–21 for doing the same $2–3 10-min task multiple times. The AMT users were informed that they could only participate in the task and earn the promised amount if they were randomly selected from the pool of potential participants. Participants were eligible to be selected multiple times but there was no guarantee that they would be selected even once. If selected, the participant was to receive an e-mail notification with further instructions.

The email invitation informed recipients that they were randomly chosen to participate in the game, which they had to complete within 24 h. Participants were then directed to our website, where they read a description of the game and made a single decision about whether to donate money to benefit a stranger. The game description explained to each participant that they would be paid the amount promised in the original solicitation, which included a "base" payment plus a "bonus" payment. Participants were also told that they were part of a group of 150 AMT users and that only members of this group who received an invitation could actually participate and receive the promised payment. The instructions further informed participants that the study had allocated a limited number of invitations to be distributed to randomly selected participants ("seeds"). The seeds were invited by the experimenters to participate. In addition to these invitations created by the experimenters, each participant who received and accepted an invitation had the option to create a new invitation and allow one more person to participate. However, in order to create a new invitation, the participant had to be willing to donate his or her bonus, even though this would reduce the participant's earnings.

If the participant chose to donate his or her bonus, a recipient of the new invitation (the "invitee") would then be randomly selected from the other 149 AMT users in the group. The instructions explained further that when a participant donated his or her bonus, we supplemented the bonus amount so that the next invited participant received the same base payment and bonus and had the same options: to keep his or her bonus or donate it and create a new invitation for one more participant.

All participants knew that the person who received the donated invitation would not know the identity of the participant who made the donation. Thus, anyone receiving a donated invitation was unable to directly reciprocate or to pass along a favorable reputation. We referred to participants by their AMT worker ID, randomly anonymized in a way that precluded the possibility to identify the same individual and be influenced by reputation.

The experiment involved five manipulations: whether the participant received a donated invitation created by another participant (i.e., being a "link"), the number of times the participant was invited to play the game (ranging from one to six), whether the participant was able to observe donated invitations, the number of donated invitations the participant observed (ranging from zero to 223), and the payment the participant received ($2 base rate and $1 bonus or $1 base rate and $1 bonus).

The observation and payment manipulations were crossed to define four between-individual treatment groups to which participants were randomly assigned. The number of invitations received and observed varied within individuals. Further, some participants were only selected as seeds, others were only selected as invitees, and still others were selected as invitees after having been previously selected as seeds.

7.2.2 Results

After removing data from participants who did not demonstrate an adequate understanding of the instructions, we were left with 518 individuals and 1070 observations. We used random-intercepts logistic regression models of observations nested in individuals to estimate the change in the odds of donating under the different manipulations. The models allow us to adjust for the non-independence of repeated measures and control for the effect of payment level and two other potential confounders, the time elapsed between subsequent interactions and the number of previous interactions. To better isolate the mechanisms, the models pool data only form the relevant treatment conditions: we tested GR in the no-observation condition only, we tested TPI for seeds only, and we tested the interaction of GR and TPI in the observation condition only.

Consistent with GR, participants were more likely to be generous towards a stranger after experiencing generosity (Table 7.1A). However, the effect is limited to the first receipt of generosity as the critical event in triggering GR. The odds of donating do not continue to increase with receiving additional donated invitations.

Table 7.1 Odds ratios for donating across treatments

Manipulation	(A) GR	(B) TPI	(C) GR×TPI
Invitee (receives a donated invitation)	7.006 (0.030)*		0.327 (0.262)
Has previously received donated invitations	0.712 (0.686)		1.021 (0.982)
Seeds			
Observes 0–75		11.414* (0.043)	(Baseline)
Observes 75–150		1.341 (0.787)	0.136 (0.101)
Observes 151+		0.219 (0.280)	0.015* (0.022)
Invitees			
Observes 76–150			19.907* (0.041)
Observes 151+			89.948* (0.026)
High Payment	64.103** (0.007)	2.532 (0.300)	3.235 (0.295)
Time waited (in hours)	0.972* (0.023)	0.992 (0.577)	0.976 (0.075)
Previous participations	0.690 (0.379)	0.784 (0.622)	0.454 (0.171)
Baseline odds	4.305 (0.181)	5.323 (0.100)	268.707*** (0.000)
Number of observations	516	371	554
Number of participants	252	277	266
Wald χ^2	5 df, 11.93* (0.036)	6 df, 6.66 (0.354)	8 df, 11.98 (0.214)

The table reports odds ratios and p values (in brackets) from random-intercept logistic regression models for (A) seeds and invitees in the no-observation treatment by number of donated invitations received; (B) seeds in the observation and no-observation treatments by number of donated invitations observed; and (C) seeds and invitees in the observation treatment by number of donated invitations observed by invitees compared to seeds. Results show that receiving and observing donations initially increases the willingness to help others, and that invitees are less susceptible to a subsequent decline in helping

Two-sided tests: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Consistent with TPI, there was a statistically significant increase in the odds of donating among the seeds who observed between 0 and 75 donated invitations, compared to those who did not observe any (Table 7.1B). However, the level of donation among those who observed more than 75 invitations was not significantly greater than the baseline level. In other words, similarly to GR, the effect of TPI appears to be concave, with most of the effect evident at relatively low levels of observed donation and little subsequent change.

Less intuitively, the effect from observing widespread generosity is significantly different for those who have recently benefited from generosity compared to those who have not. When observing more than 75 donated invitations, the odds of donating decrease for seeds but do not change for invitees (Table 7.1C). This difference between seeds and invitees is statistically significant (χ^2 (1 df) $= 3.88$, $p = 0.049$ for observing 76–150; χ^2 (1 df) $= 5.55$, $p = 0.019$ for observing 151+) and suggests the possibility that seeds succumb to a “free-riding” effect from which invitees are immune due to having been recipients of generosity.

Free riding represents the temptation to refrain from contributions, especially when one becomes aware that others are already contributing. The behavior is common in collective-action situations [15] and is also known as social loafing [16] and as the "bystander effect" or "diffusion of responsibility" [17].

In sum, the experimental results show that receiving and observing generosity can significantly increase the likelihood to be generous towards a stranger. However, the willingness to contribute can be offset by lower perceived need when the level of helping is sufficiently high. This "bystander effect" is especially evident among those who have not themselves benefited from generosity. In other words, the norm to "be generous if that is what others are doing" weakens when the level of helping behavior is high, unless it interacts with the normative obligation to "pay it forward."

The implications of the effects of the two contagion mechanisms for the dynamics of helping cascades are not intuitively obvious. We therefore incorporated the empirical findings in an agent-based model to investigate the macro-level effects of GR and TPI.

7.3 Extrapolating to Population Outcomes

Our model is a threshold model of collective behavior. Such models have been previously used to study the emergence of collective action and the resolution of social dilemmas [18–20]. In this literature, a threshold is the critical number or proportion of contributors at which an individual becomes willing to contribute to a collective action or to join a collective behavior. Depending on the distribution of individual thresholds, cascades are possible in which each additional participant triggers participation by others. It has been established that the emergence of widespread participation critically depends on the composition of the population, and in particular, the existence of a critical mass of altruists, or unconditional contributors.

We model diffusion through the dynamics of selection and influence by relaxing two common assumptions in existing threshold models: fixed interaction structure and fixed individual interests in contributing. Our model assumes that agents both move in space (similarly to [21]) and adapt their behavior (similarly to [19, 22–24]). By combining dynamic interaction structure with adaptive behavior, our model is similar to evolutionary-game models on cooperation [25–30]. In these models, agents choose an action or a strategy in the Prisoner's Dilemma and play it against each of their interaction neighbors. The agents update their behavior by imitating successful neighbors and find more beneficial interaction partners by moving on a spatial grid or rewiring their interaction network. In our model, agents play a gift game with a different number of their neighbors, depending on the rivalness of the exchanged gifts. Influence occurs not because agents imitate others but because they condition their behavior on others' behavior.

7.3.1 Simulation Model

7.3.1.1 Assumptions

The model assumes that agents are heterogeneous with respect to their natural proclivity to condition their contributions on others' behavior and their own outcomes. These proclivities are exogenously predetermined and remain fixed throughout social interactions. In addition to generalized reciprocity, TPI, and free riding, the model assumes two other behavioral mechanisms: unconditional altruism and aspiration. Unconditional altruism captures the extent to which individuals are willing to help strangers regardless of others' behavior or their own outcomes. Aspiration is the expectation about the extent to which one should benefit from others' contributions. Aspiration is the benchmark against which the agent evaluates outcomes as satisfactory [31]. If outcomes are unsatisfactory, the agent can decide to move to a different community (similarly to [21]). We set the aspiration as $\theta_A \sim \text{Uniform}(0, 0.5)$.

Following previous research [24], agents are assigned a level of unconditional altruism that is randomly drawn from a beta distribution: $\theta_{\text{UA}} \sim \text{Beta}(\alpha, \beta)$. The model fixes $\alpha = 5$ and $\beta = 5$. The resulting distribution lacks a critical mass of altruists because the majority of individuals have values close to 0.5. This distribution matches the empirical distribution of behavioral types in the general population, characterized by few unconditional altruists (about 13 %) and a majority of conditional contributors (50–63 %; [32, 33]). Nevertheless, previous analytical work on deterministic threshold models in fixed populations has shown this type of distribution not to favor the emergence of high-levels of contribution [18, 24]. Compared to these earlier models, we start from a lower level of unconditional altruism that is more empirically plausible.

The model assumes that generalized reciprocity $\text{GR} \sim \text{Uniform}(0, 1)$ and TPI $\text{TPI} \sim \text{Uniform}(0, 1)$. The higher the value of GR (TPI) the more the agent's contribution behavior is sensitive to benefits received (observed). For consistency, the free-riding value is always at least as large as the unconditional-altruism value: $\theta_{\text{FR}} = \theta_{\text{UA}} + \text{FR}(1 - \theta_{\text{UA}})$, where $\text{FR} \sim \text{Uniform}(0, 1)$. The higher the value of FR, the lower the observed level of contribution at which the agent refrains from contributing in order to free-ride on others' effort.

The model also assumes that the interaction structure is a square lattice that wraps into a torus. This structure is characterized by a high average clustering, long average path-lengths, and regularity in network positions. The structure is a poor representation for persistent social relations such as friendships and business contacts. However, it is a suitable heuristic for interactions between strangers in geographical space. Further, we assume that an agent's interaction neighborhood does not entirely coincide with the agent's observation neighborhood. In both cases, the neighborhood is a Moore neighborhood (a square with the focal agent in the center) but the radius of the neighborhood can vary. A larger interaction neighborhood corresponds to a larger community size while a larger observation neighborhood corresponds to a higher degree of gossip or centralized broadcasting.

7.3.1.2 Behavioral Rules

The five behavioral mechanisms come together in two separate threshold functions that determine whether agents contribute to a neighbor (or multiple neighbors) from their interaction neighborhood and whether agents move to a new location in their observation neighborhood.

The contribution threshold models the combined effect from receiving and observing others' contributions on one's likelihood to contribute. As in the empirical results, benefiting from others' contributions increases one's likelihood to contribute and decreases one's likelihood to free-ride, while observing others' contributions could increase both one's likelihood to contribute and one's likelihood to free-ride. Following previous models of non-monotonic threshold functions [22–24], the function is characterized by two thresholds: an upward threshold $\theta_{0\to 1}$ and a downward threshold $\theta_{1\to 0}$. The agent contributes as long as the number of received and observed contributions is within these two thresholds. The upward threshold is pre-determined by the agent's unconditional altruism but decreases if the agent experiences TPI. The downward threshold is anchored by the agents' proclivity to free ride but increases if the agent succumbs to generalized reciprocity (Fig. 7.2). More specifically:

$$\begin{aligned}\theta_{0\to 1}(t) &= \theta_{\mathrm{UA}} - \mathrm{TPI} \times M_o(t) \times \theta_{\mathrm{UA}},\\ \theta_{1\to 0}(t) &= \theta_{\mathrm{FR}} - \mathrm{GR} \times M_r(t) \times (1 - \theta_{\mathrm{FR}}),\end{aligned} \tag{7.1}$$

where $M_r(t)$ is the number of contributions the agent remembers receiving and $M_o(t)$ is the proportion of contributions the agent remembers observing in her observation neighborhood. The agent makes a contribution to the benefit of a random neighbor(s) within her interaction neighborhood if the contributions she remembers receiving match or surpass her upward threshold but the contributions she remembers observing do not exceed her downward threshold:

- **Behavior Rule 1:** Contribute if $M_r(t) \geq \theta_{0\to 1}$ and $M_o(t) < \theta_{1\to 0}(t)$.

Similarly, the agent moves with probability μ (mobility) to a random empty site within her observation neighborhood if the contributions she remembers receiving do not match her aspiration:

- **Behavior Rule 2:** Move with probability μ if $M_r < \theta_A$.

Thus, agents who are satisfied with their outcomes tend to stick to the community they have found but unhappy agents tend to move to communities with higher levels of contribution. $M_r(t)$ and $M_o(t)$ are simply the number of contributions the agent received and the proportion of local contributions the agent observed in the previous

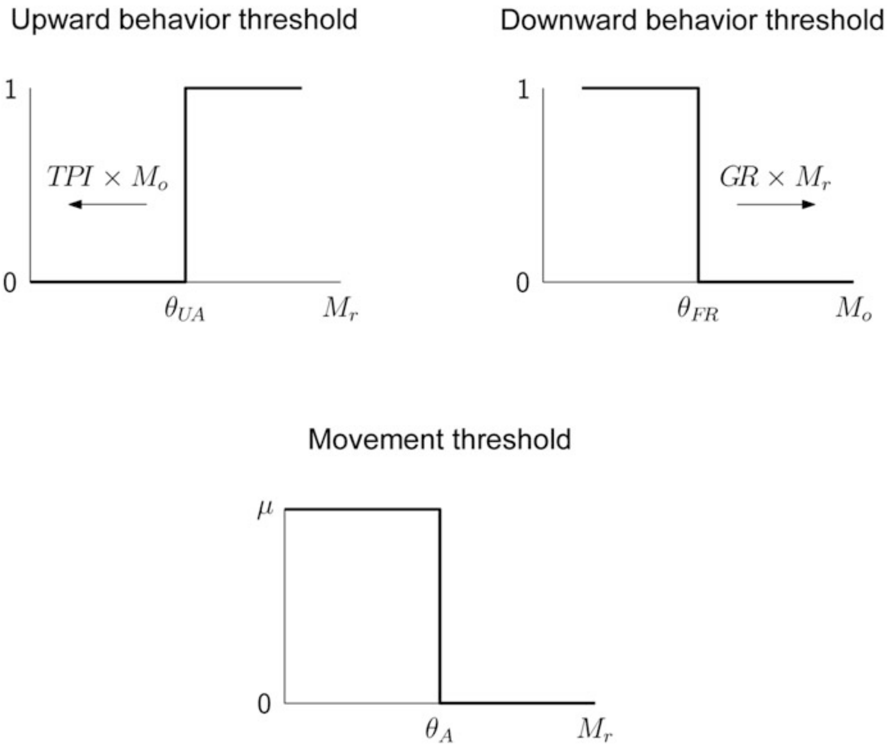


Fig. 7.2 Three thresholds in the simulation model. The upward behavior threshold depends on unconditional altruism (θ_{UA}) but can decrease due to third-party influence (TPI × M_o). The downward behavior threshold depends on the proclivity to free ride (θ_{FR}) but can increase due to generalized reciprocity (GR × M_r). The movement threshold depends on the aspiration (θ_A). The agent makes a contribution to the benefit of a random neighbor(s) within her interaction neighborhood if the contributions she remembers receiving (M_r) match or surpass her upward threshold but the contributions she remembers observing (M_o) do not exceed her downward threshold. The agent moves to a new empty site within her observation neighborhood if the contributions she remembers receiving (M_r) fall below her aspiration

m time periods, where m is the length of memory. More formally, $M_r(t) = \frac{\sum_{t-m}^{t-1} r_t}{m}$ and $M_o(t) = \frac{\sum_{t-m}^{t-1} o_t n_t^{-1}}{m}$, where r_t is the number of times the agent benefited from a contribution at time t, o_t is the number of contributions the agent observed at time t, and n_t is the size of the agent's neighborhood at time t. For the model, $m = 5$ was chosen because this value produced high variability in the results. Increasing constricts the conditions for emergence of contributions since more random events become necessary in an agent's neighborhood in order to convert that agent into a contributor.

Updating is synchronous for both the decision to contribute and to move. At each time period, agents are drawn in random order to decide whether to contribute, given the contributions they observed and the amount of contributions they received up until the last period. Once all agents have had the chance to update their behavior, the agents decide whether to move, given the amount of contributions they have received until the end of the current period. Thus, the model assumes that agents observe and receive contributions within each time period and then decide whether to contribute (Behavior Rule 1) and whether to leave a community (Behavior Rule 2). Since threshold models have been shown not to be robust to noise [34], the model assumes that there is a small probability $\epsilon = 10^{-3}$ that an agent's contribution or movement decision is reversed.

7.3.1.3 Parameter Space

To preclude sensitivity to initial conditions and synchronous updating, the model used behavioral and movement noise, the simulations were run for a sizable agent population, and the results were averaged over multiple repetitions. The fixed parameters in the model (the shape and the range of the distributions and the length of memory) were chosen with the goal to keep them as simple as possible while producing the highest variation in results along the variable parameters.

The computational experiments were run for a population of 1000 agents on a torus (40 % occupied locations). The experiments investigated the average contribution level (i.e., the proportion of contributors) for two different levels of rivalness: we assume that rival contributions benefit one recipient, while non-rival contributions benefit three recipients. The effects of four parameters are explored:

- The mobility $\mu \in [0, 0.05, 0.5]$. This is the probability to move if the agent is unhappy with what she receives from the current community. This parameter represents community turnover. (Turnover could also be adjusted by varying the average aspiration θ_A.)
- The radius of the interaction neighborhood $\in [1, 2, 3, 4, 5, 7, 10, 15]$. Since the model uses Moore neighborhoods, this is equivalent to a maximum of $[8, 24, 48, 80, 120, 224, 440, 960]$ neighbors for each agent. This parameter corresponds to community size.
- The radius of the observation neighborhood $\in [0, 1, 2, 3, 4, 5, 6, 10, 15]$. Since the model uses Moore neighborhoods, this is equivalent to a maximum of $[0, 8, 24, 48, 80, 120, 224, 440, 960]$ neighbors for each agent. This parameter is related to gossip and centralized broadcasting.
- The observation targets $\in$ [recipients, contributors]. Agents observe either the proportion of contributors or the proportion of beneficiaries within their observation neighborhood.

The simulations were run for 5000 periods which was sufficient for convergence to an equilibrium. The equilibrium proportion of contributors was then estimated by averaging the proportion of contributors over the last 1000 periods.The resulting equilibrium proportion of contributors was then averaged over 25 replications for each parameter combination.

7.3.2 Results

Figure 7.3 shows that for non-rival contributions, the equilibrium level of contributing is visibly higher than for rival contributions. Further, for non-rival contributions, the conditions for the emergence of contribution-based communities are significantly less restricted.

When the exchanged contributions are non-rival, the global level of contribution is high over a large range of interaction radii. Widespread contribution fails to emerge only when the interaction radius and/or the observation radius are extremely large. This implies that non-rival exchange allows for relatively large contribution-based communities. For relatively large communities (interaction radius > 1), observed contribution has little effect, and 100 % contribution is possible even when there is no observation (observation radius $= 0$). Overall, observing contributors has a greater effect than observing recipients (right column in Figs. 7.3 and 7.4). Community turnover does not affect outcomes except when the communities are small (interaction radius $= 1$) or when observation is widespread in large communities. In the first case, some mobility is better than no mobility (Fig. 7.4, right) and in the second case, too much mobility is bad (right column in Fig. 7.3).

When the exchanged contributions are rival, only small communities can have high levels of contribution (optimal interaction radius $\sim 2 - 3$; left column in Fig. 7.3). Further, observation is crucial for the emergence of contribution communities: the level of contribution is zero when there is no observation. As the observation radius increases, the level of contribution radically increases initially but eventually starts decreasing slowly (left in Figs. 7.4 and 7.5). The optimal observation radius is between 2 and 5, depending on the target of observation. Compared to observing contributors, observing recipients requires a smaller observation radius to achieve the maximum level of contribution. Finally, the effect of mobility is non-monotonic: low mobility ($\mu = 0.05$) is better than no mobility ($\mu = 0$) or too much mobility ($\mu = 0.5$).

Figure 7.6 identifies the reason for differences between rival and non-rival contributions. Non-rivalness implies that a larger number of individuals can benefit from a single contribution, as when a user is given advice that benefits many others in an online community. This leads to the easy formation of multiple small communities in which contributors benefit and hence continue contributing, despite free-riders who benefit enough to hang around the periphery of the clusters. When contributions are rival and only one individual can benefit from each contribution, contribution-based communities are much less likely to emerge and persist. If they do, this usually happens around a core of unconditional altruists (agents with low θ_{UA} and θ_{FR} high) who form a critical mass. These agents (the blue agents in Fig. 7.6, left column) continue contributing regardless of what others around them do. When outcome-based mobility is relatively low, the agents remain in the neighborhood long enough to have a chance to benefit from a contribution or to observe many others benefiting. (If they were observing contributors instead of recipients, they would have only observed the altruist or the few altruists that started contributing,

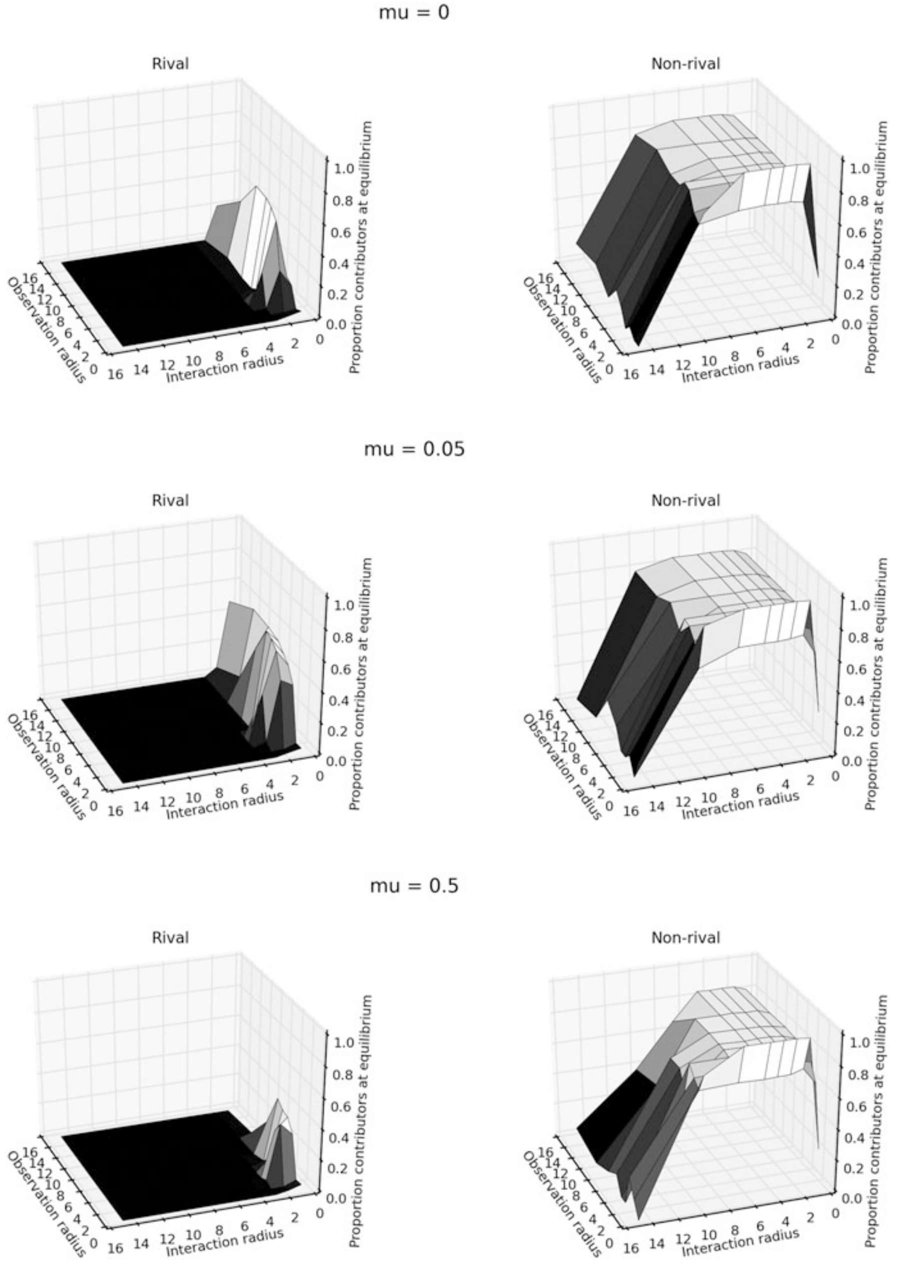


Fig. 7.3 The equilibrium proportion of contributors by observation radius and interaction radius for contributors as observation target and mobility $\mu = 0$ (*top*), $\mu = 0.05$ (*middle*), and $\mu = 0.5$ (*bottom*). Results are shown for rival (*left*) and non-rival (*right*) contributions

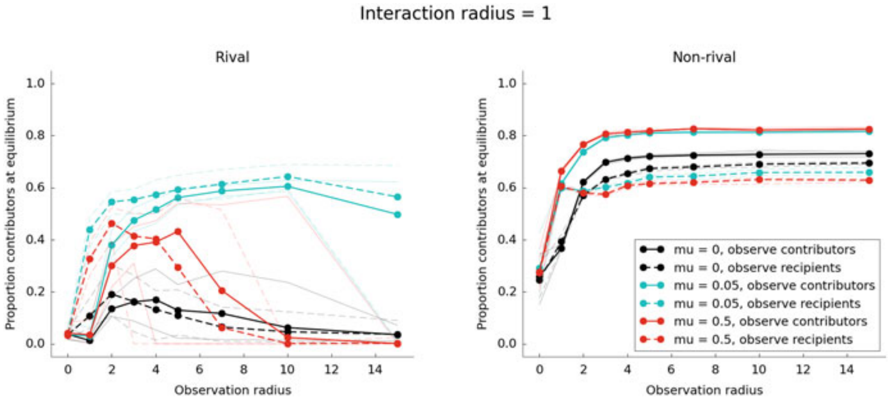


Fig. 7.4 The equilibrium proportion of contributors by observation radius when interaction is constrained to immediate neighbors only (interaction radius = 1). *Line colors* show levels of mobility and line types differentiate the observation target. Results are shown for rival (*left*) and non-rival (*right*) contributions. The *thick lines* show the proportion averaged over 25 replications for that particular parameter combination. The *thin lines* show the minimum and the maximum proportions achieved in the replications

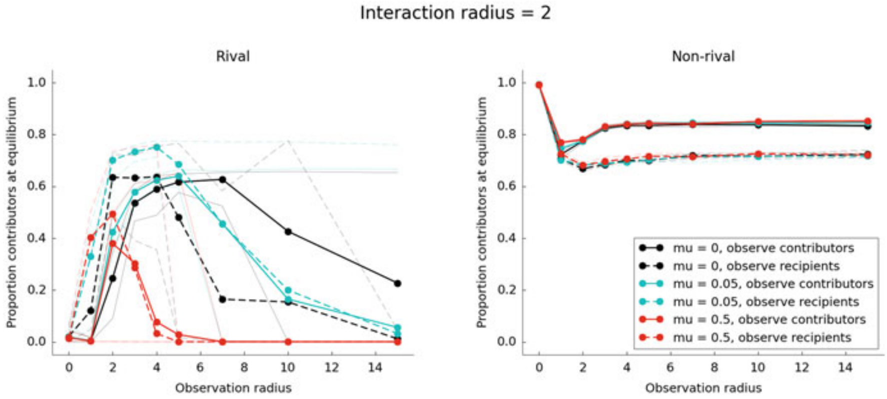


Fig. 7.5 The emergence of contribution by observation radius when interaction is constrained to immediate neighbors and neighbors of neighbors (interaction radius = 2). *Line colors* show levels of mobility and line types differentiate the observation target. Results are shown for rival (*left*) and non-rival (*right*) contributions. The *thick lines* show the proportion averaged over 25 replications for that particular parameter combination. The *thin lines* show the minimum and the maximum proportions achieved in the replications

not the many neighbors who benefit). As a result, a few clusters form around the handful of altruists in the population but the contagion does not spread to agents in other corners of the space.

The differences in the macro-outcomes between rival and non-rival contributions result from the structure of interactions and not from the differences in effect size.

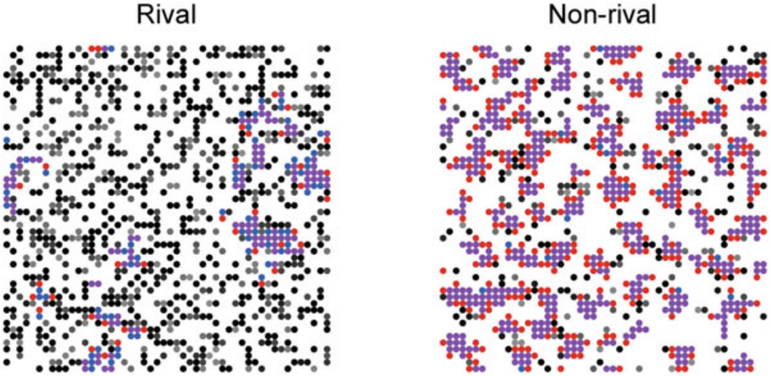


Fig. 7.6 The emergence of contribution communities for rival (*left*) and non-rival (*right*) contributions for interaction radius = 1, observation radius = 5, observing recipients, and mobility. Agents in *blue* contribute but do not benefit, agents in *red* benefit but do not contribute, and agents in *purple* both contribute and benefit

Assuming that the *GR* and *TPI* effects for non-rival contributions are weaker than the *GR* and *TPI* effects for rival contributions does not significantly affect the emergence of non-rival contributions.

7.4 Discussion

Selfless acts of kindness and anonymous voluntary donations can be puzzling, even though they are not uncommon. In daily life, people donate blood, contribute money to charity, hold the door open for the person behind, or vacate a subway seat for an elderly passenger. In the online world, users review services, rank products, or answer strangers' questions on forums. Why do communities vary in the level of member contributions? This study suggests that the answer could lie in the contagion of prosocial behavior. We first presented empirical evidence from an online experiment for the existence and interaction of two distinct mechanisms of contagion—generalized reciprocity and TPI. We then implemented these mechanisms in an agent-based model to investigate the conditions under which they lead to high levels of contributions at the population level.

The empirical results showed that receiving and observing helping behavior can increase the likelihood to help a stranger. However, the willingness to contribute can be offset by lower perceived need when the level of helping is sufficiently high, particularly among those who have not themselves been helped.

We implemented these findings in a threshold model with dynamic interaction structure and adaptive behavior. The computational experiments suggested two alternative pathways to the emergence of contribution-based communities. It is useful to think of these two pathways in the context of rival face-to-face interactions on the one hand and non-rival online contributions on the other hand. In face-to-face interactions, acts of generosity are rival if the benefit is limited to the intended recipient, as happens when holding the door open or vacating one's seat for a stranger. The simulation results show that these contributions can emerge and spread in small and stable communities, that is, communities that are tightly knit and have little turnover. In such communities, hearing about or seeing other people who benefit from the kindness of strangers increases contributions. As a result, a relatively small number of persistent altruists can trigger the spread of helping behavior. In this situation, gossip and newspaper reports about anonymous acts of generosity play an important role. For example, in an office environment, a single active anonymous altruist could trigger a chain of generosity so long as there is sufficient gossip about the level of charitable behavior such that observers come to believe that generosity is normative and conform to this "office culture."

In comparison, non-rival contributions, such as writing a product review on the Web or answering a question in an online forum, are much more likely to emerge and spread across a wider range of conditions, including in much larger groups with high turnover. For example, small esoteric-interest groups and large general-topic online portals could be equally successful user-generated content communities. In such communities, hearing about or seeing other people who contribute sustains high levels of contribution, while awareness of the number of beneficiaries decreases contribution (perhaps due to the belief that there is little need for additional sacrifice).

However, a disclaimer is in order. The chapter provides a plausible explanation for the emergence and persistence of voluntary contribution-based communities but is mainly intended to address the emergence of contribution communities among anonymous individuals. Undoubtedly, once a community forms and anonymity diminishes, cooperation-inducing mechanisms based on social sanctions (for example, reputation systems or long-term-membership privileges) become more prominent and more effective.

Acknowledgements This work was supported by grants SES-1260348 and SES-1303526 from the National Science Foundation. The first author thanks Scott E. Page, John H. Miller, and participants at the Graduate Workshop in Computational Social Science at the Santa Fe Institute for valuable comments and suggestions.

References

1. Centola, D. (2010). The spread of behavior in an online social network experiment. *Science, 329*(5996), 1194–1197.
2. Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science, 311*(5762), 854–856.
3. Rogers, E. M. (2003). *Diffusion of innovations*. New York: Free Press.
4. Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science, 341*(6146), 647–651.

5. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., et al. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295–298.
6. Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. New York: Little, Brown and Company.
7. Fowler, J. H., & Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences, 107*(12), 5334–5338.
8. Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS One, 6*(3), e16836.
9. Jordan, J. J., Rand, D. G., Arbesman, S., Fowler, J. H., & Christakis, N. A. (2013). Contagion of cooperation in static and fluid social networks. *PLoS One, 8*(6), e66199.
10. Kollock, P. (1999). The economies of online cooperation: gifts and public goods in cyberspace. In M. A. Smith, & P. Kollock (Eds.), *Communities in cyberspace* (pp. 220–242). London: Routledge.
11. Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M., & Bonhoeffer, S. (2005). Evolution of cooperation by generalized reciprocity. *Proceedings of the Royal Society B, 272*, 1115–1120.
12. Stanca, L. (2009). Measuring indirect reciprocity: Whose back do we scratch? *Journal of Economic Psychology, 30*(2), 190–202.
13. Bartlett, M. Y., & DeSteno, D. (2006). Gratitude and prosocial behavior. *Psychological Science, 17*(4), 319–325.
14. Tsvetkova, M., & Macy, M. W. (2014). The social contagion of generosity. *PLoS One, 9*(2), e87275.
15. Oliver, P., Marwell, G., & Teixeira, R. (1985). A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology, 91*(3), 522–556.
16. Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65*(4), 681–706.
17. Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*, 377–383.
18. Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology, 83*(6), 1420–1443.
19. Macy, M. W. (1991). Chains of cooperation: Threshold effects in collective action. *American Sociological Review, 56*(6), 730–747.
20. Oliver, P. E. (1993). Formal models of collective action. *Annual Review of Sociology, 19*, 271–300.
21. Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology, 1*(2), 143.
22. Granovetter, M., & Soong, R. (1983). Threshold models of diffusion and collective behavior. *The Journal of Mathematical Sociology, 9*(3), 165–179.
23. Granovetter, M., & Soong, R. (1986). Threshold models of interpersonal effects in consumer demand. *Journal of Economic Behavior & Organization, 7*(1), 83–99.
24. Lopez-Pintado, D., & Watts, D. J. (2008). Social influence, binary decisions and collective dynamics. *Rationality and Society, 20*(4), 399–443.
25. Eguluz, V. M., Zimmermann, M. G., Miguel, M. S., & Cela-Conde, C. J. (2005). Cooperation and the emergence of role differentiation in the dynamics of social networks. *American Journal of Sociology, 110*(4), 977–1008.
26. Biely, C., Dragosits, K., & Thurner, S. (2007). The Prisoner's Dilemma on co-evolving networks under perfect rationality. *Physica D: Nonlinear Phenomena, 228*(1), 40–48.
27. Hanaki, N., Peterhansl, A., Dodds, P. S., & Watts, D. J. (2007). Cooperation in evolving social networks. *Management Science, 53*(7), 1036–1050.
28. Helbing, D., & Yu, W. (2009). The outbreak of cooperation among success-driven individuals under noisy conditions. *Proceedings of the National Academy of Sciences, 106*(10), 3680–3685.

29. Meloni, S., Buscarino, A., Fortuna, L., Frasca, M., Gomez-Gardenes, J., Latora, V., et al. (2009). Effects of mobility in a population of Prisoner's Dilemma players. *Physical Review E, 79*(6), 067101.
30. Fehl, K., van der Post, D. J., & Semmann, D. (2011). Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters, 14*(6), 546–551.
31. Macy, M. W., & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America, 99*(3), 7229–7236.
32. Fischbacher, U., Gachter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters, 71*(3), 397–404.
33. Kurzban, R., & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America, 102*(5), 1803–1807.
34. Macy, M. W., & Tsvetkova, M. (2013). The signal importance of noise. *Sociological Methods & Research*. doi:0049124113508093.

Chapter 8
Understanding the Scientific Enterprise: Citation Analysis, Data and Modeling

Filippo Radicchi and Claudio Castellano

Abstract The large amount of information contained in bibliographic databases has recently boosted the use of citations, and other indicators based on citation numbers, as tools for the quantitative assessment of scientific research. Citations counts are often interpreted as proxies for the scientific influence of papers, journals, scholars, and institutions. Given their importance in practical contexts, the interest in the study of bibliographic datasets is no longer restricted to specialists in bibliometrics but extends to scholars having very different primary fields of research. As a result, the recent past has witnessed a huge production of papers on this topic of research. The present chapter aims at providing a brief overview of the progress recently made in the analysis of bibliographic databases. In the first part of the chapter, we will focus our attention on studies devoted to the statistical description of distributions of citations received by individual publications. The second part is instead devoted at summarizing some recent research efforts towards the modeling of the citation dynamics and the growth of citation networks.

8.1 Introduction

Traditionally, novel scientific knowledge is transmitted personally, as in presentations or private communications, or published in articles and books. Whereas *spoken words fly away*, the *written words* of scientific papers *remain* and thus constitute a concrete way of keeping track of how scientific knowledge is created and disseminated along time. Papers represent tangible products of the scientific enterprise, an immense and complex system whose dynamics depends on complicated processes

F. Radicchi (✉)
Center for Complex Networks and Systems Research, School of Informatics and Computing,
Indiana University, 919 E 10th Street, Bloomington, IN 47408, USA
e-mail: filiradi@indiana.edu

C. Castellano
Istituto dei Sistemi Complessi (ISC-CNR), via dei Taurini 19, 00185 Roma, Italy

Dipartimento di Fisica, "Sapienza" Universitá di Roma,
Piazzale Aldo Moro 2, 00185 Roma, Italy
e-mail: claudio.castellano@roma1.infn.it

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_8

that involve factors at multiple scales, from "macroscopic" economic decisions to "microscopic" self-organized interactions among scientists. In addition to the scientific message, an article contains a wealth of meta-data: title, journal and date of publication, list of authors and their affiliations, bibliographic references, grant funding numbers, just to mention a few. These data are stored in electronic archives, and provide an easily analyzable source of information to determine for example, at the level of the researchers, *when* and *where who* was studying *what*. The meta-data of a paper can be viewed as the most basic pieces of information about the entire scientific enterprise. Taken alone, a single paper cannot tell us so much about the scientific endeavor, but, aggregating the information from the millions of papers published each year, we can reconstruct, piece by piece, how science looked at particular points in space and in time, and infer the missing parts of the big picture.

According to this vision, bibliographic databases represent the starting point for any empirical study of the evolution and dynamics of scientific activity. The analysis of these data has a long tradition in the social sciences. Bibliographic datasets were first analyzed by Lotka in the 1920s [1] and later by Shockley in the 1950s [2] to quantitatively measure the productivity of individual scientists and research laboratories, respectively. Since the pioneering work in the 1960s of Derek de Solla Price [3], who realized that bibliographic data have a natural mathematical representation in terms of directed graphs, the study of co-authorship and citation networks has become the starting point for the formulation of key hypotheses such as the mechanism of cumulative advantage [4] to explain the dynamical patterns of citation accumulation. However, it is only in the last decade that the analysis of bibliographic data has received a boost from advances in information technology and the massive digitalization of documents [5].

In addition to scientific purposes, the use of bibliographic databases is acquiring a practical, and crucial, role in modern science. Citations between scientific publications are in fact commonly used as quantitative indicators for the importance of scientific papers, as proxies for the influence of publications in the scientific community. General criticisms to the use of citation counts have been made [6–8] , and the real meaning of a citation between papers can be very different and context dependent [9]. Nevertheless, a citation can be viewed as a tangible acknowledgment of the citing paper to the cited one. Thus, the more citations a paper has accumulated, the more influential the paper can be considered for its own scientific community of reference. The same unit of measure (i.e., a citation) is commonly used as the basis for the quantitative evaluation of individual scholars [10, 11], journals [12], departments [13], universities and institutions [14], and even entire countries [15]. Especially at the level of individual scientists, numerical indicators based on citation counts are evaluation tools of fundamental importance for decisions about hiring [16] and/or grant awards [17].

The aim of the present chapter is to review recent progresses in the analysis of bibliographic datasets, and in particular in the statistical description of citation distributions. We will consider not only studies about static properties of citation patterns, but we will also review recent analyses aimed at modeling the evolution of citation distributions and predicting the accumulation of citations by individual papers.

8.2 Bibliographic Datasets

There are many bibliographic databases available on the market. Most of these databases are now online and their records can be searched by simple web queries. The *Web of Science* (WoS) database of Thomson Reuters [18] is the oldest and best established commercial source of bibliographic data. WoS indexes papers from every part of the world and from every scientific discipline. Like WoS, other databases store large sets of bibliographic data: *CrossRef* [19], *Scopus* [20], *GoogleScholar* [21], *CiteSeer* [22], *inSpire* [23], and the Eprint archive at www.arxiv.org are just a few examples. These databases do not offer the same coverage of WoS (different journals and conference proceedings are listed depending on the database), but, with the exception of CrossRef and Scopus, they are accessible free of charge.

8.3 Static Models

As previously mentioned, in bibliographic databases, several meta- and relational-information (e.g., journal and year of publication, name and affiliations of the authors, citations to the other papers in the dataset) are associated with each record, so that, from the raw data, various kinds of citation graphs can be generated. The simplest ones are citation networks between papers. Taking the list of references appearing at the end of each article, one can draw directed connections from citing articles to cited ones. The same information can be used to construct citation networks between scientists, journals, and institutions.

8.3.1 Citation Distributions

The primary goal of a large number of empirical studies about citation networks is represented by the characterization of the probability distribution function of citations. This is the probability $P(c)$ that a paper has been cited c times. In the language of network science, measuring the number of citations of a paper means counting the number of incoming links (in-degree) c of a node. In the 1960s, de Solla Price [3] started performing empirical measurements on a relatively small subset of papers and was able to observe that the number of articles with a given number of citations had a broad distribution. Price conjectured a power law scaling $P(c) \sim c^{-\gamma}$ with a decaying exponent $\gamma \simeq 3$. This result was confirmed much later in 1998 by Redner [24]. Redner studied much larger datasets (all papers published in Physical Review D up to 1997 and all articles indexed by Thomson Scientific in the period from 1981 to 1997 and found again that the right tail of the distribution (corresponding to highly-cited papers) shows a power law scaling with $\gamma = 3$.

At the same time, Redner realized that the left part of the distribution was more consistent with a stretched exponential. However, different conclusions were drawn by Laherrére and Sornette [25] in the same year. They studied the dataset of the top 1120 most cited physicists during the period from 1981 to 1997, finding that the whole distribution of citations is more compatible with a stretched exponential $P(c) \sim \exp\left[-c^{\beta}\right]$, with $\beta \simeq 0.3$. The puzzle was seemingly solved by Tsallis and de Albuquerque [26]. By analyzing the same datasets as Redner's plus an additional one composed of all the papers published up to 1999 in Physical Review E, the authors found that the Tsallis distribution $P(c) = P(0)/\left[1 + (\beta - 1)\,\lambda\, c\right]^{\beta/(\beta-1)}$, with $\lambda \simeq 0.1$ and $\beta \simeq 1.5$, consistently fits the entire distribution of citations. However, a new functional form was again attributed to Redner a little later. Redner performed an analysis over all papers published in the 110-years-long history of journals in the Physical Review collection [27], finding that the distribution of citations is best fitted by a log-normal distribution

$$P(c) = \frac{1}{c\sqrt{2\pi\sigma^2}}\,\exp\left\{-\left[\ln c - \mu\right]^2 / \left(2\sigma^2\right)\right\}\;. \qquad (8.1)$$

In subsequent studies, depending on the particular dataset taken under consideration, distributions of citations have been fitted with various functional forms: power-laws [28–32], log-normals [33–36], Tsallis distributions [37, 38], modified Bessel functions [39, 40], and more complicated distributions [41].

8.3.1.1 Universality of Citation Distributions

A typical bias present in many empirical results is the fact that citation distributions are computed without taking into consideration any possible discipline- or age-dependence of the statistics. Please note that we use the term "bias" to indicate the systematic error that is introduced when using raw citation numbers to compare papers belonging to different fields or years of publication. With this term we do not indicate any prejudice, nor we make any claim about the causes of the field dependence empirically observed. Older papers may have more citations than recent ones, not necessarily because of their merits, but because they stayed in the literature longer and had more time to be cited. Even more serious is the bias related to discipline dependence: papers in mathematics and biology are part of two almost disconnected citation networks, which follow different citing behaviors. In [33–36], the authors accounted for these distinctions by analyzing a large number of papers and classifying them according to the date and the journal of publication [33, 36] and the scientific discipline to which they belong [34, 35]. By restricting the statistic to these subsets, the probability that a paper has received c citations is a log-normal distribution. Even more surprisingly, the authors of [34] realized that the only significant difference between different disciplines and years of publication is the average value c_0. When the raw number of citations is replaced by the relative

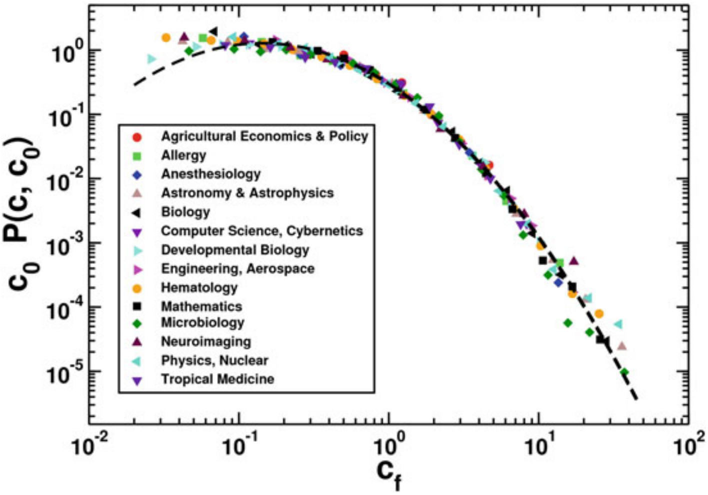


Fig. 8.1 Universality of citation distributions. Each curve refers to papers published in a given year in journals belonging to the same discipline. The disciplines are those identified by ISI Web of Science [18]. The score on the x-axis is the ratio of the number of cites c of a paper by the average number of cites c_0 collected by all papers in that discipline. From [34]

quantity c/c_0, a universal behavior is found and no distinction between curves corresponding to different publication years and scientific disciplines is visible (Fig. 8.1).

Although the role of papers with a null number of citations seems to affect the validity of the universal pattern of citation distributions across disciplines and publication years [42], universal citation distributions have been recently observed in rather different contexts [43], including more refined classification of publications in physics [44] and chemistry [45], impact metrics for scholars [46], as well as journals and institutions [47, 48].

8.3.1.2 Reverse Engineering Approach to the Study of Citation Distributions

In a recent publication, Radicchi and Castellano proposed a novel approach to test the universal behavior of citation distributions across disciplines and years of publication [49]. Their analysis is based on all records appearing in the WoS database and corresponding to six different years of publication ranging from 1980 to 2004 (see Fig. 8.2). Using a quantile–quantile plot, they empirically estimated the function able to map raw citation counts c' of papers within each discipline to the raw citation count c in the set that aggregates publications from all disciplines. They found a power-law relation $c' = ac^{\alpha}$. Such a relation indicates that citation

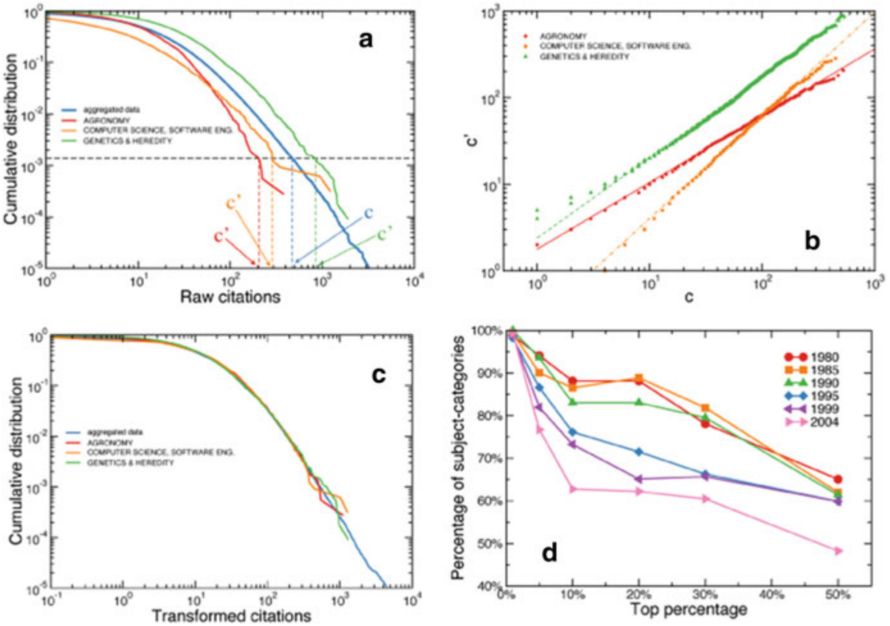


Fig. 8.2 (**a**) Cumulative distribution of raw citation counts for papers published in 1999. The *blue curve* is the reference curve used to estimate the mapping, and is calculated by aggregating all papers from all subject-categories. The *red curve*, the *orange curve*, and the *green curve* are calculated by considering only papers within the subject-categories "Agronomy," "Computer science, software engineering," and "Genetics & heredity," respectively. The figure illustrates the mapping of c' into c. Citation counts c' of single subject-categories are matched with the value of c which corresponds to same value of the cumulative distributions. (**b**) c' are plotted against citation counts of the aggregated data c. The quantities c' and c are related by the power-law relation $c' = ac^{\alpha}$. (**c**) When raw citation numbers are transformed according to $c' \to c = (c'/a)^{1/\alpha}$, the cumulative distributions of different subject-categories become very similar. (**d**) Percentage of subject-categories whose proportion values, after normalization, fall into the 95% confidence interval of values predicted by the statistical null model. Percentage values are plotted as functions of the percentage of top papers considered in the analysis. From [49]

distributions of scientific disciplines form a log-scale-location family of univariate distributions [50]. They also showed that the transformation $c' \to c = (c'/a)^{1/\alpha}$, allows for a collapse between citations distributions of different disciplines that is statistically significant. The empirical transformation obtained by Radicchi and Castellano represents the best way of suppressing biases in citation counts among different domains of science introduced so far. This fact has been proved in later publication of the same authors together with Li and Ruiz-Castillo [51]. In this analysis, the authors compared different renormalization recipes, and proved the superiority of the "reverse engineering" approach using different statistical tests [52].

8.3.2 Citation Networks

Citation networks are directed and, to a great extent, acyclic graphs. The simultaneous presence of directedness and the lack of cycles require the introduction of specific models able to capture the topological properties of citation networks.

These two ingredients are the basis of the theoretical formulation developed by Karrer and Newman [53, 54], where the statistical properties of static acyclic and directed graphs are analyzed in detail. Suppose we have a network composed of N articles (nodes) and that the indices of the nodes are chronologically sorted according to their publication date: $j < i$ means that paper j has been published before paper i. Imagine that both the in- and out-degree sequences of the network are given. This means that the number c_i of papers citing the i-th article and the number r_i of publications cited by paper i are completely specified. The study by Karrer and Newman focuses on the statistical properties of the ensemble of networks that can be constructed by preserving the constraint that all incoming and outgoing stubs are paired, with the restriction that only connections of the type $i \rightarrow j$ with $i > j$ are allowed. This static model is very similar to the one represented by the popular configurational model [55]. A natural variable, fundamental for the analytic treatment of the model by Karrer and Newman, is

$$\lambda_i = \sum_{j=1}^{i-1} c_j - \sum_{j=1}^{i} r_j \ , \tag{8.2}$$

which represents the number of incoming stubs "below" node i still available for connections with outgoing stubs exiting from vertices "above" i. In other words, λ_i counts the number of edges that flow "around" the node i. A necessary and sufficient condition for the construction of the model, assuming that all incoming and outgoing stubs are paired in a way that preserves ordering, is that $\lambda_i \geq 0, \quad \forall \ 1 < i < N$, while $\lambda_1 = \lambda_N = 0$ arise as the natural boundary conditions of the problem. The expected number of connections between nodes i and j can be estimated to be

$$P_{ij} = c_i r_j \, \frac{\prod_{l=i+1}^{j-1} \lambda_l}{\prod_{l=i+1}^{j} (\lambda_l + r_l)} \ , \tag{8.3}$$

for any pair $i < j$, while $P_{ij} = 0$ otherwise. When the network size grows, P_{ij} becomes small and can be considered equal to the probability of observing a citation from j to i.

The model by Karrer and Newman can reproduce some non-trivial properties of real citation networks (Fig. 8.3) and may provide a useful *null model* for testing topological properties of real citation networks including correlations and modular structures. The model by Karrer and Newman is not able to reproduce a very important topological feature of citation networks, represented by a high occurrence of local triangular structures [56]. A simple modification of the rules governing

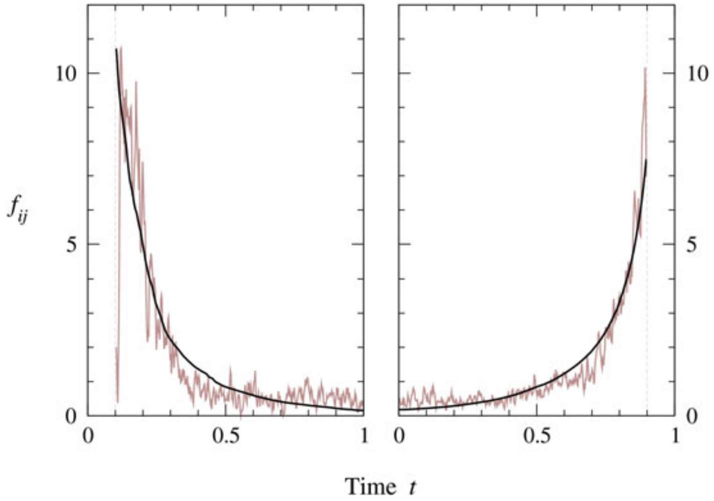


Fig. 8.3 Comparison of the static model by Karrer and Newman with empirical data. One focuses on the function f_{ij}, which is proportional to the connection probability of vertices i and j. The dataset is a citation network of papers on high-energy theory posted on the online eprint archive www.arxiv.org between 1992 and 2003. Papers are ordered from the oldest to the newest. The time of paper i is i/N, and N is the total number of papers. The *left panel* deals with citations from papers at time $t > 0.1$, the *right panel* with citations from papers at time $t < 0.9$. From [53]

the way in which connections are introduced in the network is able to correct this problem. The model by Wu and Holme [57] is very similar in spirit to the one by Karrer and Newman, but adds two new fundamental ingredients. First, the probability that paper i cites paper j is no longer dependent only on topological and time constraints, but is inversely proportional to the age difference between the two papers (aging effect). Second, once the connection between i and j has been established, there is a finite probability that i copies citations from j and therefore creates triangles. The simultaneous presence of these very intuitive and natural ingredients makes the model more representative of real citation networks.

8.4 Dynamical Models

8.4.1 Preferential Attachment

Networks of citations between papers are growing systems with complex topological features: the rate at which new papers are added (published) to the network is almost exponential, while the number of references per paper (out-degree) and the number of citations received (in-degree) are broadly distributed. One of the most surprising features of the growth of citation networks, discovered already by de

Solla Price [4], is related to the mechanism ruling the assignment of citations: the probability that a paper gets cited is proportional to the number of citations it already has received. This mechanism is the so-called cumulative advantage, based on which the "rich get richer," already developed by Yule [58] and Simon [59] in different contexts. The criterion, now widely referred to as "preferential attachment," was recently made popular by Barabási and Albert [60], who proposed it as a general criterion for the emergence of heterogeneous connectivity patterns in networks generated for the description of systems belonging to different scientific domains.

The model by Price [4] anticipated the modern models of network growth. It is very simple: one node (paper) is introduced (published) at each stage of the growth carrying new connections (citations). The average number of citations (mean degree) is m. The rate at which older nodes receive incoming connections is assumed to be linearly proportional to the number of arcs already incident on them and can be simply indicated by $\Pi(c) \sim (1 + c)$. When a sufficiently large number of papers has been published, the probability that an article has received c citations becomes stable and, in the limit of large in-degrees, equals

$$P(c) \sim c^{-2-1/m} , \tag{8.4}$$

which means a power law (or "scale free") distribution with exponent $2 + 1/m$. The exponent of the distribution γ depends on the mean degree m and can therefore be tuned rather arbitrarily.

The Barabási-Albert model [60], in its standard version, considers the total degree, not the in-degree, and yields a power law degree distribution with $\gamma = 3$. Its extension to the directed case is essentially equivalent to the Price model: the attachment rate is $\Pi(c) \sim (A + c)$, where $A > 0$ is a parameter that can be tuned [61, 62]. In this case one has $\gamma = 2 + A/m$, where m indicates the number of new citations introduced by each new paper. The exponent $\gamma = 3$ is recovered by setting $A = m$. The preferential attachment model and its subsequent generalizations not only can predict that the tail of the probability distribution for citations follows a power law, but also that the tail will be predominantly composed of the earliest published papers. This effect, supported by empirical evidence and nicely denominated as "first-mover advantage" [63], reveals that in order to be well cited it is often more convenient to write one of the first papers in a particular topic than the best article in that area.

However, the predominant weakness of the preferential attachment model and its variants is the sensitivity to the assumption that the probability of being cited is simply proportional to the number of citations previously collected. One might consider the general ansatz $\Pi(c) \sim c^{\beta}$ for the attachment probability, with a generic β. The scale-free behavior of $P(c)$ is observed only for $\beta = 1$: for $\beta < 1$, the distribution of citations turns out to be a stretched exponential, and for $\beta > 1$ a condensation of citations is observed and few papers are cited by nearly all other articles [61, 62].

The preferential attachment hypothesis has undergone empirical validation. Jeong et al. [64] considered papers published in Physical Review Letters in 1988 and

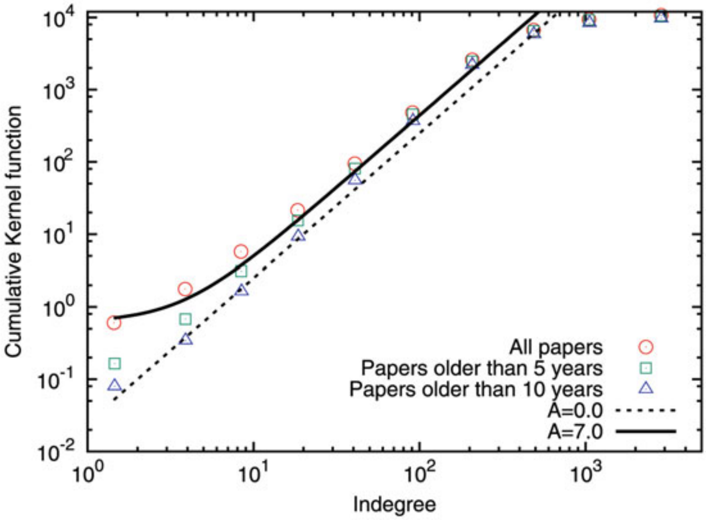


Fig. 8.4 Empirical verification of the validity of the linear preferential attachment mechanism for citation networks. Attachment probability is modeled as $\Pi\,(c) \sim c+7$ (*continuous line*). Standard preferential attachment $\Pi\,(c) \sim c$ is a good descriptor of citation accumulation only for papers older than 10 years. From [32]

all citing articles published later. They divided the time axis into several bins and tested whether the number of citations received up to a certain time was influencing the number of citations received later (Fig. 8.4).

They found that papers are cited with a probability that is nearly a linear function of the number of already-received citations, $\Pi\,(c) \sim c$. A similar result was also observed by Redner [27] by analyzing the whole dataset of publications in journals of the American Physical Society (APS). More recently, Eom and Fortunato reconsider bibliographic data from APS and found signature of growth of citations compatible with linear preferential attachment [32]. A linear attachment probability therefore seems to be a typical characteristic of the evolution of citation networks.

Different conclusions were instead obtained by Golosovsky and Solomon, who tested the linear preferential attachment hypothesis on a citation network composed of more than 40,000 physics papers [65, 66]. They found that citation dynamics of individual papers follows a superlinear preferential attachment $\Pi\,(c) \sim c^{\alpha}$ with exponent α in the range [1.25, 1.30] (see Fig. 8.5).

8.4.2 Aging

An important effect not included in the preferential attachment mechanism is the fact that the probability of receiving citations is time dependent. In the Price model, papers continue to acquire citations independently of their age, while it is reasonable

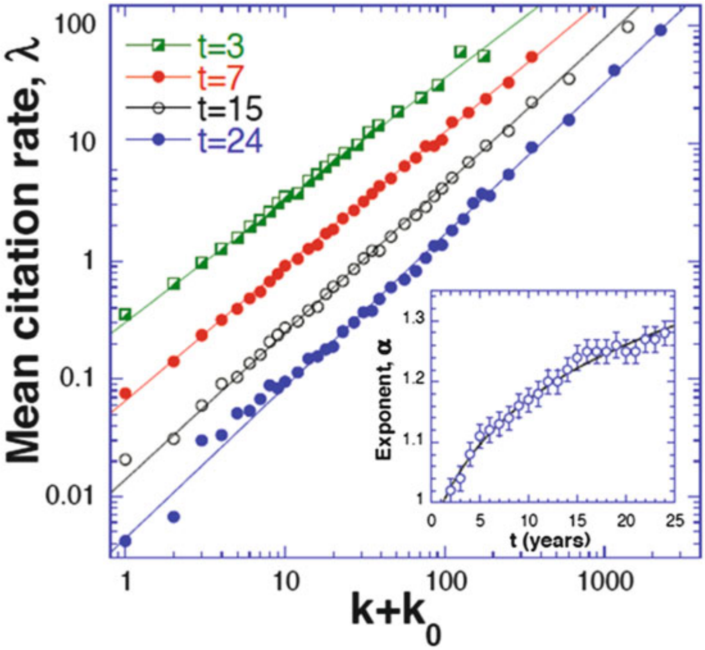


Fig. 8.5 Mean citation rate λ as a function of the number of citations previously accumulated k. t is the number of years after publication. Data are fitted by the function $\lambda \sim (k+1)^{\alpha}$. Best estimates of the exponent α are shown in the *inset*. From [65]

to think and empirically observed [67–70] that the probability for an article to be cited decreases as the age of the same article increases. Some recent papers about growing network models include the aging of nodes as a key feature [67, 70–73]. The probability that a paper receives a citation from a new article can be written as $\Pi(c,t)$, with explicit dependence not only on the number of citations c already received but also on the publication time t. For simplicity, the two effects are generally considered independent of each other and the rate at which papers receive citations becomes separable $\Pi(c,t) \sim K(c) \cdot f(t)$. Various models have been studied by assuming different functional forms for $K(c)$ and $f(t)$. In [71] for example, $K(c) = c$ and $f(t) = t^{\alpha}$. When $\alpha < 0$, the aging effect competes with the preferential attachment mechanism, while for $\alpha > 0$ older nodes are more favored and the age dependence enhances the "rich get richer" effect. The distribution of the number of citations received continues to be a power law for values of $\alpha \geq -1$. In [73], $K(c) = c$ and $f(t) = e^{\alpha' t}$. The model produces power law distributions for the citations only for $\alpha' \leq 0$. A more complicated situation is studied in [72], where $K(c) = c^{\beta}$ and $f(t) = t^{\alpha}$. The limiting distributions for the number of citations are studied in the α-β plane: scale-free distributions arise only along the line $\beta = 1$; for $\beta > 1$, condensation phenomena happen and a few nodes acquire almost all the citations; for $\beta < 1$ and $\alpha \leq -1$, the distribution is a stretched exponential.

8.4.3 Fitness

Wang et al. have recently developed a model aimed at the description of the temporal evolution in the accumulation of citations by individual publications [74]. Their model includes in addition to preferential attachment and aging an intuitive yet fundamental ingredient: a fitness or quality parameter aimed at accounting for the perceived novelty and importance of individual papers. According to the model by Wang et al., the probability that paper p is cited at time t since its publication can be written as

$$\Pi_p(t) \sim \eta_p \, P_p(t) \, c_p(t) \ , \tag{8.5}$$

$\Pi_p(t)$ is given by the product of the three different and independent ingredients of the model. η_p is the fitness associated with paper p; $P_p(t)$ is a log-normal distribution accounting for the aging of the paper p; $c_p(t)$ is the number of citations accumulated by paper p up to time t, and serves to include the linear preferential mechanism in the accumulation of citations. The statistical model summarized in Eq. (8.5) has been tested on the citation histories of papers published in several high-impact journals. providing a very good description of the effective evolution in the accumulation of citations by individual papers.

8.5 Impact Prediction

Apart from the challenging goal of uncovering fundamental mechanisms underlying the production, dissemination and consumption of information and knowledge, the analysis of citation patterns and in particular of their temporal evolution is of crucial importance for the more practical but extremely important questions of (early) impact (or success) prediction. Which papers published today are going to be game changers? How can the success of a research project be assessed only a couple of years after its inception? In a group of young postdocs, which one should be given tenure? Questions like these become more and more relevant and pressing for institutions and funding bodies.

At the level of single publications, the model of Wang et al. [74] offers the possibility to predict the future citations of a paper by simply training the parameter estimation on the early history of the publication, although fitting the model parameters may represent an issue in particular situations [75]. Whereas 5-year training windows are able to provide estimations of the future impact of a publication with a somehow large uncertainty, windows 10-year long generate very accurate predictions (see Fig. 8.6).

An alternative pathway to impact prediction attempts to leverage various types of information available at the time of publication such as words appearing in it or properties of the authors and journal [76, 77]. In the same line of thought, other

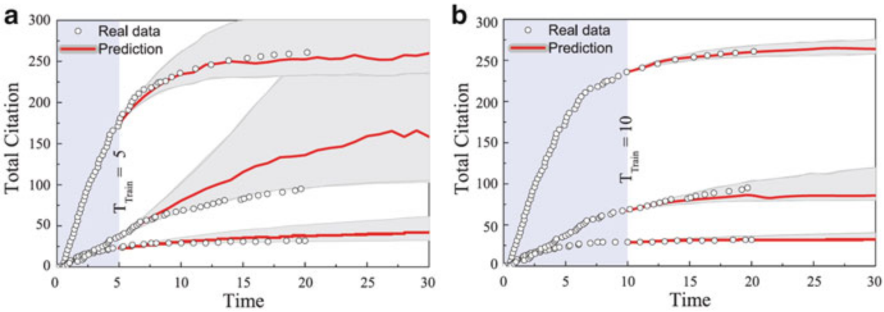


Fig. 8.6 Predictive power of the model by Wang et al. using training windows of 5 (**a**) and 10 (**b**) years (*shaded vertical area*). Increasing the training period decreases the uncertainty of the prediction. From [74]

studies focus instead on the topological properties of the citation networks at the time of publication. Shibata et al. [78] find a correlation between the centrality of a paper and its future citations. Sarigöl et al. [79] reveal instead the existence of correlations between the position of authors in coauthorship networks and the impact of their papers. In particular, the centrality in the coauthorship network of authors of highly cited papers is significantly larger than the centrality of other authors. This probes the social effects underlying scientific success and its predictability. Leveraging this finding, a machine learning approach based only on coauthorship network centrality at the moment of publication is able to predict with high precision whether an article will be highly cited five years after publication. A method exploiting topological features of both citation and coauthorship network is in [80].

In this chapter, we only focused on properties related to citation patterns of individual publications, but it is worth to mention that some research has been done also towards the prediction of the impact at the aggregate level (such as the future impact of a scholar, journal, or institution). This is considerably a more complex problem. For example, a recent attempt to predict the future evolution of the h-index of authors, based on linear regression models applied to their past performance [81], turned out to be largely unsuccessful [82], highlighting that cumulative impact measures (such as the h-index) are not suitable bases for prediction approaches.

8.6 Conclusions

In this chapter, we have briefly presented some recent developments in the rapidly expanding field of citation analysis. The interest in this area of research will surely continue to grow in the next future, for several reasons. From a fundamental point of view, bibliographic databases constitute a very detailed and accurate source of information about a social system (the scientific enterprise) which, at odds with most other data repositories about social systems, spans many decades in time, thus allowing thorough longitudinal investigations. These databases are therefore a veritable treasure trove for studying how individuals and groups interact, compete, cooperate, or clash. From a more practical point of view, it is hard to overestimate the relevance that quantitative indicators about scientific activity will play in the future for decisions about how to allocate resources, at all levels from countries planning how to invest their budget to departments evaluating which researcher to hire. In this respect, it is in the interest of science at large that citation analysis rapidly advances, providing a full spectrum of quantitative indicators whose merits and limits are precisely understood. In this endeavor one of the big challenges is the construction of sensible aggregated measures of scientific performance. This is an extremely difficult task, but it is also sorely needed, as the case of the abysmal quality of world university rankings confirms [83].

Acknowledgements We are indebted to A.Vespignani and S.Fortunato for the core part of the chapter [84]. F. Radicchi acknowledges the support from the NSF grant SMA-1446078.

References

1. Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of Washington Academy of Sciences, 16*(12), 317–324.
2. Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. *Proceedings of the IRE, 45*(3), 279–290.
3. de Solla Price, D. J. (1965). Networks of scientific papers. *Science, 149*(3683), 510–515.
4. de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5), 292–306.
5. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167–256.
6. MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342–349.
7. MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435–444.
8. Adler, R., Ewing, J., Taylor, P. (2009) Citation statistics. *Statistical Science, 24*(1), 1.
9. Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80
10. Hirsch, J. E. (2005). 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), 16569–16572.
11. Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics, 69*(1), 131–152.
12. Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA: The Journal of the American Medical Association, 295*(1), 90–93.
13. Davis, P., & Papanek, G. F. (1984). Faculty ratings of major economics departments by citations. *The American Economic Review, 74*(1), 225–230.
14. Kinney, A. L. (2007). National scientific facilities and their science impact on nonbiomedical research. *Proceedings of the National Academy of Sciences, 104*(46), 17943–17947.
15. King, D. A. (2004). The scientific impact of nations. *Nature, 430*(6997), 311–316.
16. Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review-a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics, 68*(3), 427–440.

17. Bornmann, L., Wallon, G., & Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. *PLoS One, 3*(10), e3480.
18. Web of Science. Available at http://wokinfo.com.
19. CrossRef. Available at http://www.crossref.org.
20. Scopus. Available at http://www.scopus.com.
21. GoogleScholar. Available at http://scholar.google.com.
22. Citeseer. Available at http://citeseerx.ist.psu.edu.
23. inSpire. Available at http://inspirehep.net.
24. Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems, 4*(2), 131–134.
25. Laherrere, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: “fat tails” with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems, 2*(4), 525–539.
26. Tsallis, C., & de Albuquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? *The European Physical Journal B-Condensed Matter and Complex Systems, 13*(4), 777–780.
27. Redner, S. (2005). Citation statistics from more than a century of physical review. *Physics Today, 58*, 49–54.
28. Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628–638.
29. Vázquez, A. (2001). Statistics of citation networks. arXiv preprint cond-mat/0105031.
30. Lehmann, S., Lautrup, B., & Jackson, A. D. (2003). Citation networks in high energy physics. *Physical Review E, 68*(2), 026113.
31. Bommarito, M. J., & Katz, D. M. (2009). Properties of the united states code citation network. Available at SSRN: http://ssrn.com/abstract=1502927 or http://dx.doi.org/10.2139/ssrn.1502927
32. Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS One, 6*(9), e24926.
33. Stringer, M. J., Sales-Pardo, M., & Nunes Amaral, L. A. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS One, 3*(2), e1683.
34. Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268–17272.
35. Castellano, C., & Radicchi, F. (2009). On the fairness of using relative indicators for comparing citation performance in different disciplines. *Archivum immunologiae et therapiae experimentalis, 57*(2), 85–90.
36. Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science and Technology, 61*(7), 1377–1385.
37. Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics, 3*(4), 296–303.
38. Anastasiadis, A. D., de Albuquerque, M. P., de Albuquerque, M. P., & Mussi, D. B. (2010). Tsallis q-exponential describes the distribution of scientific citations – A new characterization of the impact. *Scientometrics, 83*(1), 205–218.
39. van Raan, A. F. J. (2001). Two-step competition process leads to quasi power-law income distributions: Application to scientific publication and citation distributions. *Physica A: Statistical Mechanics and Its Applications, 298*(3), 530–536.
40. Van Raan, A. F. J. (2001). Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions. *Scientometrics, 51*(1), 347–357.
41. Kryssanov, V. V., Kuleshov, E. L., Rinaldo, F. J., & Ogawa, H. (2007). We cite as we communicate: A communication model for the citation process. arXiv preprint cs/0703115.

42. Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology, 63*(1), 72–77.
43. Evans, T. S., Hopkins, N., & Kaube, B. S. (2012). Universality of performance indicators based on citation and reference counts. *Scientometrics, 93*(2), 473–495.
44. Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E, 83*(4), 046116.
45. Bornmann, L., & Daniel, H.-D. (2009). Universality of citation distributions – A validation of Radicchi et al.'s relative indicator cf= c/c0 at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology, 60*(8), 1664–1670.
46. Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics, 7*(4), 924–932.
47. Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & de Nooy, W. (2013). Field-normalized impact factors: A comparison of rescaling versus fractionally counted ifs. *Journal of the American Society for Information Science and Technology, 64*(11), 2299–2309.
48. Chatterjee, A., Ghosh, A., & Chakrabarti, B. K. (2014). Universality of citation distributions for academic institutions and journals. arXiv preprint arXiv:1409.8029.
49. Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS One, 7*(3), e33833.
50. Lawless, J. F. (2011). *Statistical models and methods for lifetime data* (Vol. 362). New York: Wiley.
51. Li, Y., Radicchi, F., Castellano, C., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics, 7*(3), 746–755.
52. Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS One, 8*(3), e58727.
53. Karrer, B., & Newman, M. E. J. (2009). Random acyclic networks. *Physical Review Letters, 102*(12), 128701.
54. Karrer, B., & Newman, M. E. J. (2009). Random graph models for directed acyclic networks. *Physical Review E, 80*(4), 046110.
55. Molloy, M., & Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing, 7*(03), 295–305.
56. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science, 298*(5594), 824–827.
57. Wu, Z.-X., & Holme, P. (2009). Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E, 80*(3), 037101.
58. Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character, 213*, 21–87.
59. Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.
60. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.
61. Krapivsky, P. L., Redner, S., & Leyvraz, F. (2000). Connectivity of growing random networks. *Physical Review Letters, 85*(21), 4629.
62. Dorogovtsev, S. N., Mendes, J. F. F., & Samukhin, A. N. (2000). Structure of growing networks with preferential linking. *Physical Review Letters, 85*(21), 4633.
63. Newman, M. E. J. (2009). The first-mover advantage in scientific publication. *Europhysics Letters, 86*(6), 68001.
64. Jeong, H., Néda, Z., & Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters, 61*(4), 567.
65. Golosovsky, M., & Solomon, S. (2012). Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Physical Review Letters, 109*(9), 098701.
66. Golosovsky, M., & Solomon, S. (2013). The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers. *Journal of Statistical Physics, 151* (1–2), 340–354.

67. Hajra, K. B., & Sen, P. (2004). Phase transitions in an aging network. *Physical Review E, 70*(5), 056103.
68. Hajra, K. B., & Sen, P. (2005). Aging in citation networks. *Physica A: Statistical Mechanics and Its Applications, 346*(1), 44–48.
69. Hajra, K. B., & Sen, P. (2006). Modelling aging characteristics in citation networks. *Physica A: Statistical Mechanics and Its Applications, 368*(2), 575–582.
70. Wang, M., Yu, G., & Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and Its Applications, 387*(18), 4692–4698.
71. Dorogovtsev, S. N., & Mendes, J. F. F. (2000). Evolution of networks with aging of sites. *Physical Review E, 62*(2), 1842.
72. Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E, 63*(5), 056125.
73. Zhu, H., Wang, X., & Zhu, J.-Y. (2003). Effect of aging on network structure. *Physical Review E, 68*(5), 056121.
74. Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.
75. Wang, J., Mei, Y., & Hicks, D. (2014). Comment on "quantifying long-term scientific impact". *Science, 345*(6193), 149–149.
76. Ibáñez, A., Larrañaga, P., & Bielza, C. (2009). Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics, 25*(24), 3303–3309.
77. Livne, A., Adar, E., Teevan, J., & Dumais, S. (2013). Predicting citation counts using text and graph mining. In: *Proceedings of the iConference 2013 Workshop on Computational Scientometrics: Theory and Applications*.
78. Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology, 58*(6), 872–882.
79. Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science, 3*(1), 1.
80. Bertsimas, D., Brynjolfsson, E., Reichman, S., & Silberholz, J. M. (2014). Moneyball for academics: Network analysis for predicting research impact. Available at SSRN: http://ssrn.com/abstract=2374581 or http://dx.doi.org/10.2139/ssrn.2374581
81. Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature, 489*(7415), 201–202.
82. Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., & Fortunato, S. (2013). On the predictability of future impact in science. *Scientific Reports, 3*, 3052.
83. De Nicolao, G. (2014, October). Times higher education world university rankings: Science or quackery?. https://www.aspeninstitute.it/aspenia-online/article/international-university-rankings-science-or-quackery
84. Radicchi, F., Fortunato, S., & Vespignani, A. (2012). Citation networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.) *Models of science dynamics, understanding complex systems* (pp. 233–257). Berlin/Heidelberg: Springer.

Part II
Social Behavior Under Stress

Chapter 9
Behavioral Changes and Adaptation Induced by Epidemics

Piero Poletti, Marco Ajelli, and Stefano Merler

Abstract In this chapter, a modeling framework that explicitly accounts for human adaptations induced by risk perception in the epidemic dynamics is proposed. The diffusion of different behaviors is modeled according to a game theoretical approach and coupled with classic disease transmission models. The developed framework is used to assess the impact of human spontaneous behavioral changes on the natural history of vaccination programs and to investigate how a spontaneous defensive response enacted by susceptible individuals during an epidemic outbreak can affect the course of infection events. The complex interplay between behavioral changes and the epidemic transmission is investigated through the theoretical analysis of the resulting coupled dynamics and highlighted through some illustrative examples based on influenza- and measles-like infections. Our results suggest that human behavioral responses to the risk of infection can either positively or negatively impact the spread of epidemics.

9.1 Risk Perception, Human Behavior and Epidemics

Human infection dynamics is driven by the complex interplay between the transmissibility of a pathogen, the socio-demographic structure of the host population and behavioral patterns of the individuals involved in the chain of infection transmission. Changes in the perceived risk of infection can trigger a spontaneous behavioral response that can dramatically impact the dynamics of an epidemic and the effectiveness of public health policies.

A first interesting phenomenon is represented by the interplay between the perceived risk of infection and parental vaccination choices for childhood diseases

P. Poletti (✉)
Dondena Centre for Research on Social Dynamics, Universitá Commerciale L. Bocconi,
via Rontgen, 1, Milan, Italy
e-mail: piero.poletti@unibocconi.it

M. Ajelli • S. Merler
Center for Information Technology, Bruno Kessler Foundation, via Sommarive,
18, Trento, Italy
e-mail: ajelli@fbk.eu; merler@fbk.eu

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_9

in a system of non-compulsory vaccination. Indeed, decades of immunization at high coverage and the consequent success of vaccination policies in the past have generated the widespread perception that many serious infections do not circulate anymore and that they do not currently represent a concrete risk of illness. As a consequence, the perception of benefits coming from vaccination for preventing childhood disease has dramatically decreased with respect to the perceived risks of suffering vaccine adverse events [1–7]. However, a prolonged period of low vaccine uptake can produce entire birth cohorts of individuals not adequately immunized, making the occurrence of large epidemics in the future just a matter of time. As a matter of fact, large measles outbreaks have occurred recently in several European sites [8, 9] that are seriously threatening the planned deadlines of the WHO measles elimination plan for Europe [10], increasing the cost of the programs in terms of both disease burden and corrective measures such as catch-up campaigns [11].

On the other hand, the awareness of a new epidemic threat has triggered uncoordinated self-imposed measures enacted by the public to reduce the risk of infection, ranging from a larger compliance to better hygienic precautionary behavior to the use of face masks to reduce individual susceptibility, to a change in mobility and contact patterns of the population [12–19]. Examples of such behaviors emerged during the 2003 SARS epidemic, the 2009 influenza pandemic and the 2014 Ebola outbreak. Although it is still unclear to what extent the phenomenon has significantly affected the transmission dynamics in the population, several modeling works have highlighted that spontaneous public response to epidemics can possibly hamper, or at least delay, the spread of infectious diseases [7, 20–25].

Therefore, on the one hand, spontaneous reactive response to risk of infection potentially represents an additional resource to face new epidemic threats but, on the other hand, a misperception of the risk of infection can produce a decrease in public compliance to safe, effective and essential public health policies, such as vaccination.

Although suitable data to calibrate models accounting for human behavioral responses to epidemics are still lacking, a better understanding of how adaptation induced by epidemics can affect the dynamics of transmission and vice versa is crucial to improve model realism and enhance effective public policies and control strategies.

9.1.1 Epidemic Modeling and Game Theory

Spontaneous behavioral responses and adaptations induced by risk perception are explicitly modeled here in the epidemic dynamics by coupling two mutually influencing phenomena: (a) the infection transmission and (b) spontaneous behavioral changes of individuals exposed to the risk of infection. In other words, infection transmission and population behavior are considered as dynamical variables that influence each other.

In this chapter, we look at human behaviors through the lens of game theory, which provides a rich and natural modeling framework. Specifically, we model behavioral changes as driven by imitation mechanisms based on the evaluation of prospective outcomes deriving from alternative decisions (strategies) and cost-benefit considerations (payoffs) [26]. Individuals may choose to switch to a different behavior, depending on cost-benefit assessments based on the perceived risk. Past experience, response to the action of other individuals and changes in exogenous conditions all contribute to the balance. The resulting model consists in the coupling of two dynamical systems, one describing the epidemic transitions and the other one describing the behavioral changes. In principle, there is no reason for the two phenomena to evolve at the same speed; it is therefore crucial to study the model allowing for different time scales, embodied in different time units.

9.1.1.1 Modeling Infection Transmission

We describe in this section a family of models for the spread of a non-fatal infection which includes the possibility of vaccination at birth. The model without vaccination can be derived as a sub-case of the general model and will be used to discuss the case of a new pandemic infection spreading in the population.

The epidemic transitions, whose time unit is t, are modeled according to a $S \to I \to R$ scheme, where S, I, R denote the fraction of susceptible, infective and recovered/immune individuals in the population, respectively. In this simple formulation, susceptible individuals are assumed to develop the infection upon contact with infective individuals and, after recovery, to gain a life-long immunity against reinfection. We assume that individuals' immunizations occur through the use of a "perfect" vaccine administered in a single dose at birth and providing life-long immunity. Recovered individuals are individuals who have developed a resistance to the infection through either a direct experience of the disease or vaccination. For simplicity, we assume that individuals mix with other individuals homogeneously and that the population size is constant over time. Under these simple assumptions, the epidemiological transitions described above can be modeled by the following ordinary differential equations system:

$$\begin{cases} dS/dt = \mu(1-p) - \mu S - \beta SI \\ dI/dt = \beta SI - (\mu + \gamma)I \\ dR/dt = (\gamma - \mu)I + \mu p \end{cases} \tag{9.1}$$

where $1/\mu$ is the average life expectancy of individuals at birth, p denotes the vaccinated proportion among newborn children, β is the transmission rate and $1/\gamma$ defines the average duration of infective period of individuals who acquired the infection.

When considering a rapid epidemic outbreak for which no vaccine is available, as in the case of a new pandemic strain, the fraction of vaccinated individuals is zero and the demographic component of the system can be neglected. In this case, system (9.1) can be rewritten as follows:

$$\begin{cases} dS/dt = -\beta SI \\ dI/dt \ = \beta SI - \gamma I \\ dR/dt = \gamma I \end{cases} \tag{9.2}$$

9.1.1.2 Modeling Behavioral Changes

We now consider that individuals are able to change their behavior spontaneously, following cost/benefit considerations, switching from a non-responsive (*NR*) to a responsive (*R*) behavior and vice versa. We assume that the responsive behavior is associated with a reduced risk of infection. This phenomenon can be cast in the language of evolutionary game theory, by modeling behaviors as strategies in a suitable game, with certain expected payoffs. It is clear that whether it is more convenient to either adopt a responsive behavior or not depends on the state of the epidemic. Of course, the two phenomena, i.e., the infection transmission and the change of behaviors' distribution within the population, may not have the same time scales. In fact, while epidemic transmission can occur only through person-to-person contact, it is fairly reasonable to consider that individuals can access the information required to balance the payoffs of alternative behaviors much more frequently by telephone, email, the Internet and, in general, the media.

The dynamics of behaviors is modeled as a selection dynamics based on imitation [26, 27]. Specifically, we assume that a fraction of the individuals playing strategy *NR* can switch to strategy *R* after having compared the payoffs of the two strategies, at a rate proportional to the difference between payoffs, $\Delta P = P_R - P_{NR}$, with proportionality constant χ. The converse is true for the fraction of the individuals playing *R*.

By assuming a constant rate ω, equal for both behaviors, and denoting by x the fraction of individuals adopting *R*, the resulting dynamics regulating the behavioral changes within the population can be modeled through the following equation:

$$dx(\tau)/d\tau = \omega\chi x(\tau)(1 - x(\tau))\Delta P(\tau) \tag{9.3}$$

where $\tau = t/\alpha$ is the time unit of behavioral changes. Therefore, in the time scale of infection (t), the imitation dynamics driving behavioral changes over time becomes

$$dx(t)/dt = \rho x(t)(1 - x(t))\Delta P(t) \tag{9.4}$$

where $\rho = \omega\chi/\alpha$.

9.2 Uncoordinated Human Behavioral Response to an Epidemic

In this section we have applied the approach defined in the previous section to study how a spontaneous defensive response enacted by susceptible individuals could develop during an epidemic outbreak and affect the course of infection events. In particular, we consider that the susceptible population is divided into two classes of individuals. The first class of individuals adopt self-protective behaviors aimed at reducing their risk of infection, either through a reduction of contacts (e.g., by avoiding crowded places or traveling less) or through a reduction of their susceptibility during their contacts (e.g., by using face masks or enacting precautionary behaviors such as washing hands frequently or following cough/respiratory etiquette).

9.2.1 Model Formulation

In the illustrative example presented in this section, we assume that all susceptible individuals can conform to either one or the other of two different behaviors: responsive *R* or non-responsive *NR*. The first gives the individuals an advantage in terms of reduced risk of infection, yet at some extra cost. For example, avoidance of crowded environments reduces the risk of infection, but also entails disadvantages deriving from greater isolation. Payoffs can be therefore modeled as follows. All individuals pay a cost for the risk of infection, which we assume depends linearly on the fraction of infected individuals, $I(t)$, and it is lower for individuals adopting a responsive behavior. Moreover, individuals enacting *R* pay an extra, fixed cost c. It may be convenient to think of c as costs associated to self-imposable prophylactic measures, such as those deriving from less traveling, working, attending school, etc. The payoffs associated to behaviors *R* and *NR* can be therefore modeled as

$$p_R = -m_R I - c \tag{9.5}$$

$$p_{NR} = -m_{NR} I \tag{9.6}$$

where $m_R < m_{NR}$. The resultant payoff difference can be written as $\Delta P = p_R - p_{NR} = -m_R I - k + m_{NR} I = c(mI - 1)$, where $m = (m_{NR} - m_R)/c > 0$.

On the other hand, while susceptible individuals adopting a non-responsive behavior (i.e., a fraction $1 - x$ of susceptibles S) are assumed to become infected at a rate β, we assume that individuals adopting a responsive behavior (i.e., a fraction x of susceptibles S) become infected at a rate $q\beta$, where $0 \leq q \leq 1$.

According to Eqs. (9.4) and (9.2), the combined dynamics of behavioral changes and epidemic spread can be written as

$$\begin{cases} dS/dt = -[(1-x)+qx]\beta IS \\ dI/dt = [(1-x)+qx]\beta IS - \gamma I \\ dR/dt = \gamma I \\ dx/dt = kx(1-x)(mI-1) \end{cases} \tag{9.7}$$

9.2.1.1 Interpretation of Parameters

In this model formulation the timing of the behavioral response is characterized by parameters m and k. The former describes how the prevalence I is weighted in the payoff functions, i.e., in the balance of the cost associated to the risk of infection and the cost of a self-protection strategy. The latter represents the speed of the imitation process with respect to the disease transmission temporal scale. As a matter of fact, $1/m$ defines the *prevalence threshold* above which individuals reducing contacts have a larger payoff; the larger m, the earlier the *responsive* behavior is perceived as the most convenient choice. On the other hand, k entails the delay (embedded in the imitation dynamics) between the time at which a strategy becomes more convenient and the time at which the strategy becomes widely adopted in the population. In sum, the time at which the transition between the two possible behaviors occurs is driven by m, while the duration of this transition is driven by k. Finally, q is the reduction in the force of infection to which susceptible individuals who adopt a responsive behavior are exposed.

9.2.1.2 Equilibria

System (9.7) admits a continuum of equilibria $(S^\star, 0, 1-S^\star, x^\star))$ with $S^\star \in [0,1]$ and $x^\star \in \{0,1\}$. Notice that, when $S^\star = 1$, the equilibrium (1, 0, 0, 0) is unstable when $R_0 = \beta[(1-x)+qx]/\gamma > 1$ and stable otherwise. R_0 defines the basic reproductive number of system (9.7) and represents the number of new infections that a typical infective individual causes during his/her whole period of infectivity [28]. If we consider the case of a novel pathogen, which can be reflected by the initial condition $(1-I_0, I_0, 0, x_0)$ with I_0 close to 0, and we consider that $x_0 > 0$, then $x \to 0$, and the equilibrium is stable as long as $R_0 < 1$.

By defining $R_0^{NR} = \beta/\gamma$ and $R_0^R = qR_0^{NR}$ we can see that these two quantities represent two reproductive numbers themselves. Indeed, R_0^{NR} is the reproductive number when all susceptible individuals are non-responsive, and R_0^R is the reproductive number when all susceptible individuals are adopting the responsive behavior.

Note that, if the population is initially non-responsive (i.e., x_0 is close to zero), $R_0^{NR} > 0$ represents the threshold condition for observing an epidemic. On the other hand, m defines a prevalence threshold for observing the increase of responsiveness within the epidemic course. Indeed, given that in a classic SIR model $S(t) + I(t) - (1/R_0)\log(S)$ is an invariant and the fraction of susceptibles at the peak is $S = 1/R_0$, it can be easily checked that $I^\star = 1 - 1/R_o^{NR} + 1/R_o^{NR}\log(1/R_o^{NR})$ is the fraction of infected individuals at the peak for the classic SIR model characterized by $R_0 = R_0^{NR}$. Therefore, the condition $1/m < I^\star$ represents the threshold condition for the responsive behavior to become convenient during the course of the epidemic.

9.2.2 Study of Dynamics

Let us now consider $k \to \infty$, which defines the case in which behavioral changes occur extremely faster than the epidemic spreads. In this case the model (9.7) can be rewritten as follows:

$$\begin{cases} dS/dt &= -[(1-x) + qx]\beta IS \\ dI/dt &= [(1-x) + qx]\beta IS - \gamma I \\ dR/dt &= \gamma I \\ \epsilon dx/dt &= x(1-x)\Delta P \end{cases} \tag{9.8}$$

According to this formulation, we can study the case of $\epsilon \to 0$ by approximating the solution of the singularly perturbed initial value problem by the degenerate system defined as follows:

$$\begin{cases} dS/dt &= -[x + q(1-x)]\beta IS \\ dI/dt &= [x + q(1-x)]\beta IS - \gamma I \\ dR/dt &= \gamma I \\ 0 &= x(1-x)\Delta P \end{cases} \tag{9.9}$$

obtained by setting $\epsilon = 0$ and provided that in the last equation of system we use the asymptotically stable equilibrium of the boundary-layer system,

$$dx(s)/ds = x(1-x)\Delta P \tag{9.10}$$

which is obtained by making the transformation of independent variable $s = t/\epsilon$ and then setting $\epsilon = 0$, which implies that $S(s)$, $I(s)$, $R(s)$ are constant [29, 30].

By analyzing the solutions of the last equations where I is a constant, we can easily see that

$x \to 1$ when $I < 1/m$
$x \to 0$ when $I > 1/m$

which means that in the limit case of $\epsilon \to 0$, x becomes discontinuous in I, and the solution of system (9.7) can be locally approximated with the solution of the degenerated system (9.10) where x is constantly either 1 or 0 depending on I. More specifically, this means that as long as $I < 1/m$ the solutions of system (9.7) can be approximated by a classical SIR model driven by R_0^{NR}, while when $I > 1/m$ the solution of system (9.7) can be approximated by a classical SIR model driven by R_0^R.

A full detailed characterization of the dynamics of system (9.7) in the case of $k \to \infty$ can be found in [20].

9.2.2.1 Potential Benefits of Uncoordinated Spontaneous Social Distancing

An interesting result on the dynamics of system (9.7) is that the fraction of susceptible individuals at the end of an epidemic, i.e., S_∞, is a increasing function of m, and $S_\infty \to 1/R_0^n$ when $1/m \to 0$. As a consequence the fraction of susceptible individuals at the end of an epidemic described by system (9.7) is always larger than the one obtained by considering a classical SIR with transmission rate β and thus driven by R_0^{NR}. This means that uncoordinated behavioral response to the risk of infection has the potential to reduce the final attack rate of an epidemic. This result is formally proven in [20].

9.2.3 The Effectiveness of Spontaneous Behavioral Changes in Reducing the Risk of Infection

In order to better understand the interplay between epidemic transmission and the behavioral responses to the risk of infection, we now investigate the effect of different reactions on the epidemic spread by varying, one by one, the parameters regulating the interplay between behavioral changes and the epidemic transmission, starting from an illustrative parameter configuration. In particular, the impact of different parameters on the epidemic dynamics is analyzed in terms of final epidemic size (defined as the total number of infections at the end of the epidemic), daily peak prevalence and peak day.

The parameter set used as a baseline relies on the following assumptions: (a) the adoption of the *responsive* behavior reduces by 15 % the number of potentially infectious contacts, i.e., $q = 0.85$; (b) the *responsive* behavior becomes more convenient when the prevalence becomes larger than 1 % of the population, i.e., $1/m = 0.01$; (c) the delay between the time at which the *responsive* behavior becomes convenient and the time at which more than 50 % of the population becomes responsive is about 5 days, which corresponds to $k = 10$. Finally, parameters merely characterizing the disease transmission process are taken from reliable estimates available for the 2009 H1N1 pandemic influenza. Specifically, R_0 is assumed to be 1.4 and the generation time $(1/\gamma)$ is assumed equal to 2.8 days [21, 31–34]. The initial conditions considered are: $S(0) = 1 - 10^{-3}$, $I(0) = 10^{-3}$, $x(0) = 1 - 10^{-6}$, $R(0) = 0$.

9.2.3.1 Baseline Scenario

The resulting dynamics of system (9.7) as simulated according to the baseline parameter set described above is shown in Fig. 9.1. After an initial growth of the epidemic, the prevalence reaches the threshold $1/m$ and the *responsive* behavior becomes more convenient. As a consequence, when the *responsive* behavior becomes widely adopted in the population, which occurs after a few days, the epidemic growth rate remarkably reduces. As the prevalence decreases below the threshold, the *non-responsive* behavior becomes more convenient and its diffusion produces a heavy tail in the infection dynamics (Fig. 9.1).

9.2.3.2 Quick Reactions Produce Smaller Epidemics

By analyzing the effect of different m on the infection dynamics we can see (Fig. 9.2) that a larger reduction in the final epidemic size and in the daily peak prevalence is observed for larger values of m (which correspond to a smaller prevalence threshold $1/m$). The same effect can be observed by increasing k (which corresponds to considering faster behavioral changes with respect to the transmission dynamics).

However, if the imitation process is too slow (i.e., for small values of k) or the prevalence threshold is too large (i.e., for small values of m), the human response never takes place and the epidemic spreads following the dynamics of the SIR model driven by R_0^n. In particular, the latter corresponds to the case in which an epidemic is not perceived as sufficiently severe to trigger a behavioral response of the population. From a mathematical point of view, this happens when $1/m$ is larger than $I^\star$, i.e., the largest possible daily peak prevalence obtained when all individuals are *not responsive* for the whole course of the epidemic (Fig. 9.2).

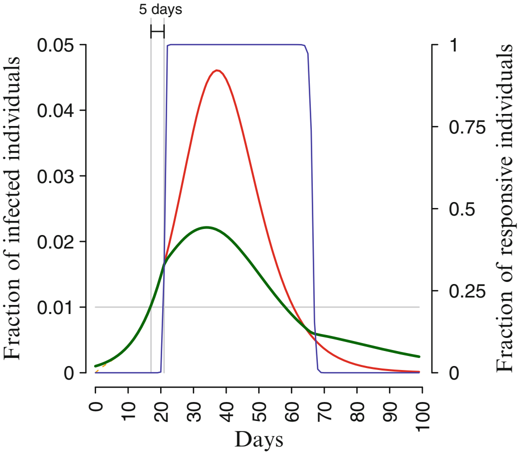


Fig. 9.1 Daily prevalence of infection in the case of no responsiveness ($q = 1$, *red line*), and in the baseline scenario ($q = 0.85$, *green line*) along with the dynamics of x (*blue line*, scale on the *right*). The *horizontal gray line* represents the prevalence threshold $1/m$. The behavioral response appears about 5 days after the prevalence $I(t)$ crosses the threshold $1/m$ producing a lower increase in the prevalence of infection

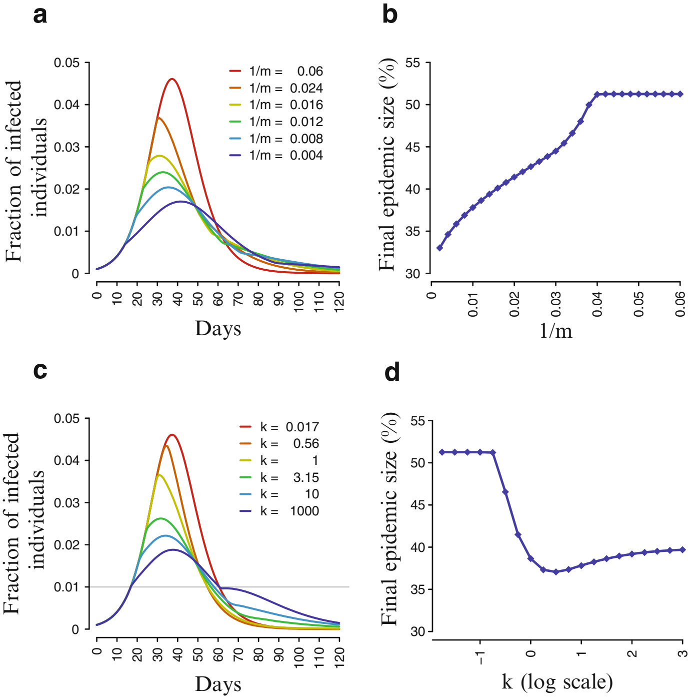


Fig. 9.2 (**a**) Daily prevalence of infections and (**b**) final epidemic size as obtained for different values of the prevalence threshold $1/m$. Other parameters are as in the baseline scenario. (**c**) Same as (**a**) but for different values of k. (**d**) Same as (**b**) but for different values of k

9.2.3.3 A Little Reduction in Contacts Could Make a Big Difference

Different values of q correspond to a different reduction in the risk of infection enacted by individuals when adopting a *responsive* behavior. Thus, it is not surprising that smaller values for q correspond to lower final epidemic sizes and daily peak prevalences (see Fig. 9.3). Moreover, for small values of q, multiple epidemic waves can occur and the possible dynamics of the infections over time becomes quite rich. A proper discussion on the conditions for observing a similar pattern as a consequence of behavioral changes can be found in [20]. Our analysis also shows that a reduction of 100 % in the number of potentially infectious contacts (corresponding to $q = 0$, i.e., total isolation) produces the same effects obtained by considering a reduction of 30 % ($q = 0.7$). This suggests that there exists a threshold for $q^\star$ such that smaller values than $q^\star$ do not determine a larger reduction in the final epidemic size. This is related to the fact that for each $q < 0.7$ we have that $R_0^R = qR_0^{NR} < 1$ since in this example $R_0^{NR} = 1.4$. The threshold for $q^\star$, as $I^\star$, depends on the value of R_0^{NR}.

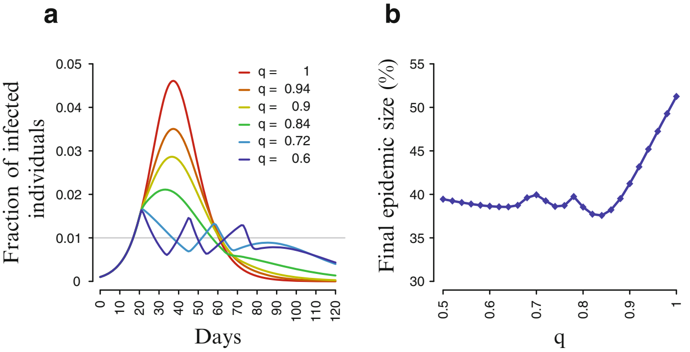


Fig. 9.3 (**a**) Daily prevalence of infections and (**b**) final epidemic size as obtained for different values of the reduction factor q. Other parameters are as in the baseline scenario

9.3 Dynamic Vaccine Demand Under Voluntary Vaccination Programs

In this section we have applied the approach described in Sects. 9.1 and 9.2 to study how the vaccine uptake for childhood infections and the consequent fraction of immunized children can change over time in relation to changes in the perceived risk of infection. Indeed, under voluntary vaccination, high levels of herd immunity might be an incentive for parents to decide not to vaccinate children, ranking the perceived risk of suffering a vaccine side effect (VSE) as much higher than the corresponding risk of infection. In industrialized countries, the incidence of many childhood infections has reduced to negligible levels as a consequence of past immunization efforts. On the other hand, the large number of vaccines routinely administered every year yields steady flows of vaccine-associated side effects (VSEs). Here we model the situation in which individuals can switch between the decisions to vaccinate or not to vaccinate after comparing the perceived risk of disease and the perceived risk of VSEs. Specifically, we model that changes in vaccine coverage for a childhood disease are driven by an underlying imitation process between the parents of the children to be vaccinated, who are divided into the categories *responsive* and *non-responsive*. Responsive individuals represent parents who decide to vaccinate their children at birth. For clarity, we denote hereafter individuals who adopt a responsive behavior as vaccinators (V), and non-responsive individuals as non-vaccinators (NV).

9.3.1 Model Formulation

We assume that parents of newborn children can conform to either one or the other of two different behaviors: to vaccinate (V) or not vaccinate (NV) their children.

The first gives the individuals an advantage in terms of reducing the risk of infection to zero, yet at some cost deriving from the exposure to VSEs.

Specifically, we assume that the cost associated to parents' choice of not vaccinating their children is an increasing function on the fraction of infected individuals, $I(t)$. On the other hand, we assume that the cost associated to the perceived risk of developing VSEs after vaccination depends linearly on the fraction of vaccinated children, $p(t)$. The latter assumption relies on the idea that the public evaluates the risk of VSEs by using the information on the total number of adverse cases in the population. This assumption has the straightforward implication that periods of large vaccine uptake negatively feed back, through an increase in the incidence of VSEs, to parents favorable to vaccination. The payoffs associated to behaviors V and NV can thus be modeled as

$$p_V = -m_V p \tag{9.11}$$

$$p_{NV} = -\tilde{h}(I) \tag{9.12}$$

where $\tilde{h}(I) \geq 0$ and $m_V > 0$. The resultant payoff difference can be written as $\Delta P = p_V - p_{NV} = -m_V p + \tilde{h}(I) = m_V[h(I) - p]$, where $h(I) = \tilde{h}(I)/m_V > 0$.

A noteworthy sub-case of this formulation is when $\tilde{h}(I)$ is assumed to be linear in I, i.e., $p_{NV} = m_{NV} I$. However, some results from the equilibria analysis of the dynamics of vaccine demand in relation to the epidemic spread presented in this section hold under the more general assumption that $\tilde{h}$ is an increasing function of I. Notice that the latter can also take into account the presence of a baseline non-zero risk for any value of I, e.g., $p_{NV} = m_{NV} I + m^0_{NV}$ with $m^0_{NV} > 0$. For instance, it can represent the case when a positive risk is perceived even in the absence of infection. The latter is likely the case when the infection has been locally eliminated by past immunization or reduced to a negligible number of cases, but the risk of reintroduction from abroad is perceived as non-zero.

As for the effect of vaccination on the epidemic spread, we assume that vaccinated newborns (i.e., a fraction p of newborns) gain life-long immunity against the infection while unvaccinated newborns (i.e., a fraction $1 - p$ of newborns) are assumed to result susceptible to the infection. Therefore, according to Eqs. (9.4) and (9.1) the combined dynamics of behavioral changes and epidemic transmission can be written as

$$\begin{cases} dS/dt = \mu(1-p) - \mu S - \beta IS \\ dI/dt = \beta IS - \gamma I - \mu I \\ dR/dt = \gamma I \\ dp/dt = kp(1-p)[h(I) - p] \end{cases} \tag{9.13}$$

Note that for system (9.13) the basic reproductive number becomes pR_0 with $R_0 = \beta/(\gamma + \mu)$ and it defines a critical threshold $p_c = 1/R_0$ for the fraction of vaccinated newborns p over which the infection will be eliminated, i.e., $I \to 0$ for any value of $p > 1/R_0$.

9.3.1.1 Equilibria

The equilibria for system (9.13) are the following:

E1 $(S = 1, I = 0, p = 0)$ is a disease free equilibrium with no vaccinators;

E2 $(S = 0, I = 0, p = 1)$ is a disease free equilibrium where all individuals are vaccinators;

E3 $(S = \frac{1}{R_0}, I = \mu(1 - \frac{1}{R_0})/(\mu + \gamma), p = 0)$ is an endemic equilibrium with no vaccinators;

E4 $(S = 1 - h(0), I = 0, p = h(0))$ is a disease free equilibrium where we recall that $h(0)$ is the perceived risk of being infected in the absence of infection. E4 = E1 when $h(0) = 0$.

E5 $(S = \frac{1}{R_0}, I = I_e, p = h(I_e))$ where I_e is the unique solution of the equation $h(I) = 1 - \frac{1}{R_0} - \frac{\mu+\gamma}{\mu} I$.

By employing standard mathematical techniques, it can be shown that the first three equilibria are always unstable, while E4 is globally asymptotically stable when $h_0 > p_c$ and unstable otherwise.

On the other hand, the stability of E5 depends on the parameter k and the function h. Specifically, E5 is asymptotically stable irrespective of the value of k when $\beta I_e \mu h'(I_e) < (\mu + \beta I_e)(\mu + \beta I_e + 2\sqrt{\beta I_e(\mu + \gamma)})$. Otherwise there are two positive values, k_1 and k_2 with $k_2 > k_1$ such that at $k = k_1, k_2$ there are Hopf bifurcations and the stability of E5 depends on the value of k as follows:

(a) if $0 < k < k_1$ or $k > k_2$ E5 is locally asymptotically stable

(b) if $k_1 < k < k_2$ E5 is unstable and the orbits of $(S, I, p)(t)$ are oscillatory in the sense of Yakubovich [35], which intuitively means that for sufficiently large t all state variables are permanently oscillating with regular or irregular oscillations.

A full detailed characterization of equilibria of system (9.13) can be found in [3].

9.3.1.2 Implications and Parameter Interpretations

The analysis of equilibria of system (9.13) suggests that, if any positive risk of VSEs is perceived by the public and if the infection risk is perceived as close to zero in the presence of negligible levels of prevalence of infection in the population, the elimination of the infection is not possible under voluntary vaccination programs.

Depending on the speed at which individuals change their behavior, i.e., their attitude in vaccinating or not in response to the change in risk perception, the dynamics can result in steady oscillations around a positive sub-optimal coverage and in repeated infection outbreaks.

However, the coverage and infection level of post-vaccination endemic equilibrium E5, namely p_e and I_e, depend on the function h, R_0, the generation time $1/\gamma$ and the life expectancy $1/\mu$, but do not depend on k.

9.3.2 The Impact of Vaccine Side Effects on the Natural History of Immunization Programs

In order to better understand the interplay between epidemic dynamics and vaccination choices, we focus on an illustrative example to explore the impact of human spontaneous behavioral changes on the natural history of vaccination programs.

In particular, hereafter we assume that $\tilde{h}(I)$ is linear in I, i.e., $\tilde{h}(I) = m_{NV}I$ and analyze the dynamics of system (9.13) by setting parameters merely characterizing the disease transmission process according to reliable estimates available in the literature for measles. In particular, R_0 is assumed to be 10, the generation time $(1/\gamma)$ is taken equal to 7 days and the average life expectancy $(1/\mu)$ is assumed to be 75 years [3].

According to the above assumptions, system (9.13) can be rewritten as

$$\begin{cases} dS/dt = \mu(1-p) - \mu S - \beta IS \\ dI/dt \ = \beta IS - \gamma I - \mu I \\ dR/dt = \gamma I \\ dp/dt = kp(1-p)[mI - p] \end{cases} \tag{9.14}$$

where $m = m_{NV}/m_V > 0$.

According to this formulation, we can see that the parameter m drives the balance between the perceived risk associated to each reported case of VSE and the perceived risk associated to each reported case of infection. In particular, it can be easily checked that when mI is larger than p, vaccination is the most convenient strategy to adopt while, when mI is smaller than p, the strategy of not vaccinating will emerge.

Similar to the model presented in Sect. 9.2.1, k represents the relative speed of the imitation dynamics with respect to the dynamics of the infection transmission.

Note that close to the equilibrium E5, the last equation of (9.13) can be approximated as $p' = \Phi(mI - p)$, with $\Phi = kmI_e(1 - mI_e)$. In this case the equation for p can be read as an exponentially fading memory mechanism with average delay $1/\Phi$. This observation provides an idea of how it is possible to estimate—by using reliable data on variations of coverage over time—relevant parameters and functions driving behavioral changes, i.e., k and m.

9.3.2.1 The Balance of Risks from Vaccination and from Infection

In the absence of vaccination, the measles dynamics in the long term leads to an endemic positive prevalence that, according to the parametric set we are considering, results in approximately 0.2 per 1000 individuals. Therefore, as vaccination will reduce the prevalence in the population and given that the unique equilibrium with positive vaccination is E5, i.e., $p_e = mI_e$, it follows that to achieve large equilibrium uptakes for measles, the perceived cost of serious disease has to be at least three orders of magnitude higher than the perceived cost associated to VSEs.

This result is consistent with the epidemiological situation of measles at the beginning of the twentieth century, when the risk of serious sequelae following measles infections was extremely large (100–250 deaths per 100,000 cases of disease), and the absence of a vaccine kept the risk of infection very high. In such circumstances even a large probability of suffering side effects from vaccination could have been tolerated by the community. This however suggests that industrialized countries could face serious difficulties in maintaining high vaccine uptake in the future. Indeed, the remarkable decrease in the risk of serious morbidity and mortality from infectious diseases, together with the current high degree of herd immunity led by decades of sustained vaccination, has dramatically reduced the perceived risk of serious disease from most infections to negligible levels. Although such results are based on the equilibrium analysis of a simple deterministic model with homogeneous mixing and stable demography, more realistic models are not expected to substantially change these important implications.

On the other hand, equilibrium E5 for system (9.14) becomes

$$E5 = (S_e = 1/R_0, \quad I_e = (1 - 1/R_0)/(1 + \gamma/\mu + m), \quad p_e = mI_e) \quad (9.15)$$

Thus, although according to the equilibria analysis the elimination of the disease is impossible (i.e., $I_e > 0$), the infection prevalence at equilibrium can be close to zero. Indeed, it can be observed that I_e decreases with m and $I_e \to 0$ when $m \to \infty$.

The prevalence of infection at equilibrium I_e as obtained for different values of m along with the corresponding coverage level at equilibrium p_e are shown in Fig. 9.4.

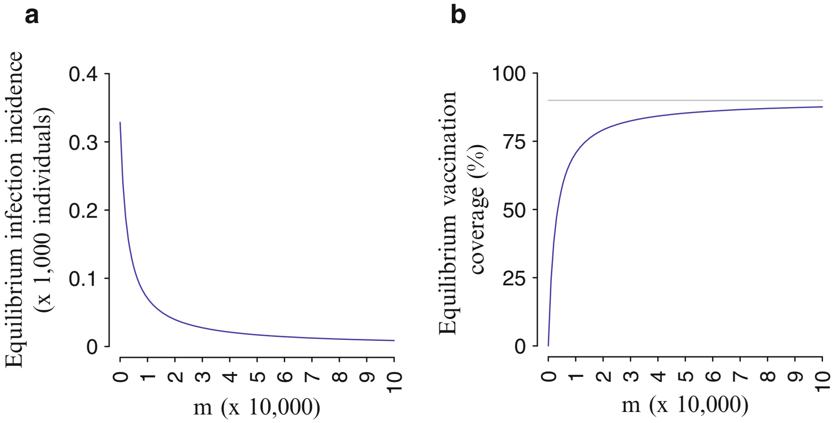


Fig. 9.4 (**a**) Infection prevalence and (**b**) vaccine coverage at equilibrium as obtained for different values of *m* in the illustrative case of a measles-like infection

9.3.2.2 The Introduction of a New Vaccine Under a Voluntary Vaccination Program

We focus now on possible outcomes of introducing a new immunization program based on a voluntary vaccination program. In particular, we consider a childhood infectious disease for which the risk associated with the infection is considered relatively high in the absence of vaccination. Moreover, we assume that a vaccine, for which no side effects are initially known, is made available. However, we assume that the increase of vaccination coverage leads to an increase of reported vaccine adverse events, and we investigate how this could impact, in the long term, the vaccine uptake and the epidemic dynamics.

In particular, we show how different behavioral responses to the risk of infection can affect the epidemic spread, by analyzing the dynamics of system (9.14) for some illustrative values of parameters m and k. Simulations are initialized with a population at the endemic pre-vaccination equilibrium and with few vaccinators in the population. In order to better understand the implications of the analysis of equilibria stability, we assume that a small fraction of new infected cases is steadily imported from outside the study area (namely 0.001 per 1000 individuals per week), which can take into account the risk of infection resurgence after local elimination, for instance as a consequence of international traveling from and to endemic areas. Results obtained by simulating the model with different values of k and of m are shown in Fig. 9.5.

In all simulated scenarios, the vaccine uptake starts increasing, as VSEs are initially negligible with respect to infection disease occurrence. However, as the vaccination coverage increases, the increase of reported VSEs and the decrease of the risk of infection progressively counterbalance the initial convenience of vaccination.

In good agreement with the theoretical results described above, for sufficiently small values of k, after a transient period (the duration of which is inversely proportional to k), the dynamics of infection and vaccine coverage converge to a stable equilibrium. Moreover, for sufficiently large values of m, the epidemic infection prevalence reduces, at least in the very long term, to negligible levels.

On the other hand, for larger values of k the system converges, in epidemiologically reasonable time scales, to a stable limit cycle characterized by oscillations in vaccine uptake around the sub-optimal coverage equilibrium and repeated infection outbreaks. Specifically, in this case, prolonged periods with high coverage and low infection rates, during which the circulation of the infection is essentially sustained by immigration, are followed by periods with low vaccine coverage that periodically induce new epidemic outbreaks, producing in turn an increase of the perceived risk of infection.

We note that, for sufficiently large values of m and k, oscillations can periodically lead the coverage p to values larger than the critical threshold p_c for rather long periods of time. During such periods, despite the endemic persistence of infection, routine vaccination surveys would reveal a satisfactorily high coverage.

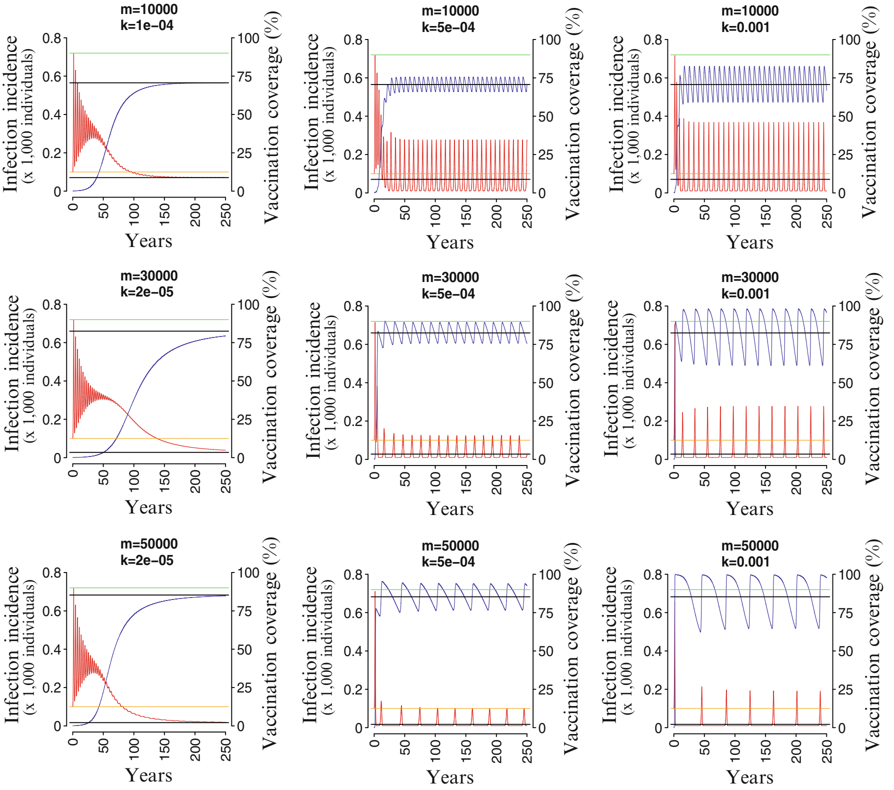


Fig. 9.5 The dynamics of infection prevalence and vaccine coverage as predicted by model simulations for different values of m and k in the illustrative case of a measles-like infection. *Horizontal orange line* represents the infection prevalence at pre-vaccination equilibrium. *Horizontal green line* represents the critical threshold for vaccination p_c. *Black lines* represent values of coverage and infection incidence at equilibrium E5

Our results show that the average uptake is not significantly affected by k and remains close to p_e. However, both the amplitude of oscillations of p and the fraction of total time where $p > p_c$ increase in k. Low values of k yield oscillations that are of small amplitude and make more infrequent the occurrence of outbreaks triggered by the importation of cases and caused by the decline of p. The frequency of oscillations increases with k and decreases with m.

9.3.2.3 The Migration from Compulsory to Voluntary Vaccination

It is important to highlight what can happen if voluntary vaccination is introduced in a population where vaccination is compulsory. In this case, we can assume that compulsory vaccination has led to high vaccination coverage levels, reducing the

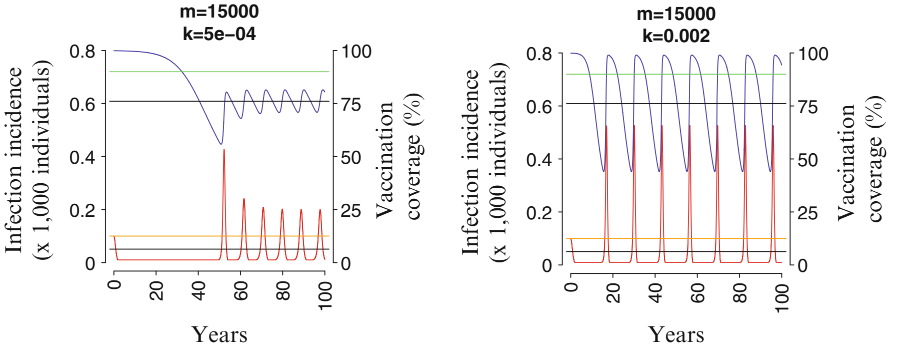


Fig. 9.6 The effect of migration from compulsory to voluntary vaccination on the dynamics of infection prevalence and vaccine coverage, as predicted by model simulations by assuming $m = 15{,}000$ and different values of k in the illustrative case of a measles-like infection

infection to negligible levels. This situation can be modeled by considering that $p(0) = 0.95 > p_c$ and by assuming that the circulation of the infection is initially sustained by immigration only.

Our results (see Fig. 9.6) show that, in this case, vaccination coverage starts declining soon after the immunization program has been converted to voluntary vaccination, and that the spontaneous vaccination choices lead the vaccine uptake to sub-optimal coverage levels, yielding unavoidably to new epidemic outbreaks.

In other words, if local elimination of the infection is achieved due to a period of compulsory vaccination, after which vaccination becomes voluntary, VSEs will progressively induce individuals to switch to non-vaccination, thereby decreasing the vaccination coverage and progressively increasing the probability of infection re-emergence from imported cases.

9.4 Conclusions

The analysis carried out in this chapter has highlighted that human adaptations to the risk of infection can have either a positive or negative impact on the spread of epidemics.

On the one hand, uncoordinated behavioral changes triggered by a new epidemic threat can significantly hamper disease transmission and reduce final epidemic size, even when self-imposed measures produce a small decrease in the force of infection. On the opposite side, risk perception could decrease the public compliance with voluntary immunization programs against childhood infections, leading to repeated infection outbreaks caused by sub-optimal vaccine uptake levels and, in the case of local elimination of the disease, to a progressive increase of the probability of infection re-emergence.

The framework introduced here is fairly general, to be applied to different diseases and to more complex models accounting, for instance, for heterogeneous mixing by age and changes in the demographic structure of the population [36–39]. Our results suggest that behavioral changes driven by the perceived risk of infection can dramatically alter the dynamics of an epidemic and, consequently, the effectiveness of public health policies, pointing out the need for incorporating human behavior in prediction models informing public health decisions.

Acknowledgements The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 283955 (DECIDE).

References

1. Beutels, P., Scuffham, P. A., & MacIntyre, C. R. (2008). Funding of drugs: Do vaccines warrant a different approach? *Lancet Infectious Diseases, 8*(11), 727–733.
2. Bauch, C. T., & Bhattacharyya, S. (2012). Evolutionary game theory and social learning can determine how vaccine scares unfold. *PLoS Computational Biology, 8*(4), e1002452.
3. d'Onofrio, A., Manfredi, P., & Poletti, P. (2011). The impact of vaccine side effects on the natural history of immunization programmes: An imitation-game approach. *Journal of Theoretical Biology, 273*(1), 63–71.
4. d'Onofrio, A., Manfredi, P., & Poletti, P. (2012). The interplay of public intervention and private choices in determining the outcome of vaccination programmes. *PloS One, 7*(10), e45653.
5. Manfredi, P., Posta, P. D., d'Onofrio, A., Salinelli, E., Centrone, F., Meo, C., et al. (2009). Optimal vaccination choice, vaccination games, and rational exemption: An appraisal. *Vaccine, 28*(1), 98–109.
6. d'Onofrio, A., Manfredi, P., & Salinelli, E. (2007). Vaccinating behaviour, information, and the dynamics of SIR vaccine preventable diseases. *Theoretical Population Biology, 71*, 301–317.
7. Manfredi, P., & d'Onofrio, A. (2013). *Modeling the interplay between human behavior and the spread of infectious diseases*. Berlin: Springer.
8. Jakab, Z., & Salisbury, D. M. (2013). Back to basics: The miracle and tragedy of measles vaccine. *Lancet, 381*(9876), 1433–1434.
9. Knol, M., Urbanus, A., Swart, E., Mollema, L., Ruijs, W., van Binnendijk, R. S., et al. (2013). Large ongoing measles outbreak in a religious community in the Netherlands since May 2013. *Euro Surveillance, 18*, 36.
10. Cutts, F. T., Lessler, J., & Metcalf, C. J. (2013). Measles elimination: progress, challenges and implications for rubella control. *Expert Review of Vaccines, 12*(8), 917–932.
11. Simone, B., Balasegaram, S., Gobin, M., Anderson, C., Charlett, A., Coole, L., et al. (2014). Evaluation of the measles, mumps and rubella vaccination catch-up campaign in England in 2013. *Vaccine, 32*(36), 4681–4688.
12. Eastwood, K., Durrheim, D. N., Butler, M., & Jones, A. (2010). Responses to pandemic (H1N1) 2009, Australia. *Emerging Infectious Diseases, 16*(8), 1211–1216.
13. Rubin, G. J., Amlot, R., Page, L., & Wessely, S. (2009). Public perceptions, anxiety, and behaviour change in relation to the swine flu outbreak: Cross sectional telephone survey. *British Medical Journal, 339*, b2651.
14. Schwarzinger, M., Flicoteaux, R., Cortarenoda, S., Obadia, Y., & Moatti, J. (2010). Low acceptability of A/H1N1 pandemic vaccination in French adult population: Did public health policy fuel public dissonance? *PLoS One, 5*(4), e10199.

15. Funk, S., Gilad, E., Watkins, C., & Jansen, V. A. A. (2009). The spread of awareness and its impact on epidemic outbreaks. *Proceedings of the National Academy of Sciences of the United States of America, 106*(16), 6872–6877.
16. Seale, H., Heywood, A. E., McLaws, M., Ward, K. F., Lowbridge, C. P., Van, D., et al. (2011). Why do I need it? I am not a risk! Public perceptions towards the pandemic (H1N1) 2009 vaccine. *BMC Infectious Diseases, 10*(1), 99.
17. SteelFisher, G. K., Blendon, R. J., Bekheit, M. M., & Lubell, K. (2010). The Public's Response to the 2009 H1N1 Influenza Pandemic. *The New England Journal of Medicine, 365*, e65.
18. Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature, 442*, 448–452.
19. Sadique, M. Z., Edmunds, W. J., Smith, R. D., Meerding, W. J., de Zwart, O., Brug, J., et al. (2007). Precautionary behavior in response to perceived threat of pandemic influenza. *Emerging Infectious Diseases, 13*(9), 1307–1313.
20. Poletti, P., Caprile, B., Ajelli, M., Pugliese, A., & Merler, S. (2009). Spontaneous behavioural changes in response to epidemics. *Journal of Theoretical Biology, 260*(1), 31–40.
21. Poletti, P., Ajelli, M., & Merler, S. (2011). The effect of risk perception on the 2009 H1N1 pandemic influenza dynamics. *PLoS One, 6*(2), e16460.
22. Poletti, P., Ajelli, M., & Merler, S. (2012). Risk perception and effectiveness of uncoordinated behavioral responses in an emerging epidemic. *Mathematical Biosciences, 238*, 80–89.
23. Liao, C. M., & You, S. H. (2014). Assessing risk perception and behavioral responses to influenza epidemics: Linking information theory to probabilistic risk modeling. *Stochastic Environmental Research and Risk Assessment, 28*(2), 189–200.
24. Liao, C. M., You, S. H., & Cheng, Y. H. (2015). Network information analysis reveals risk perception transmission in a behaviour-influenza dynamics system. *Epidemiology and Infection 143*(1), 23–36.
25. Cao, L. (2014). Infection dynamics in structured populations with disease awareness based on neighborhood contact history. *European Physical Journal B: Condensed Matter and Complex Systems, 87*(10), 1–10.
26. Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge: Cambridge University Press.
27. Bauch, C. T. (2005). Imitation dynamics predict vaccinating behaviour. *Proceedings of the Royal Society B, 272*, 1669–1675.
28. Diekmann, O., Heesterbeek, J. A. P., & Metz, J. A. J. (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology, 28*(4), 365–382.
29. Hoppensteadt, F. C. (1966). Singular perturbations on the infinite interval. *Transactions of the American Mathematical Society, 123*, 521–535.
30. O'Malley, R. E. (1991). *Singular perturbation methods for ordinary differential equations*. New York: Springer.
31. Ajelli, M., Merler, S., Pugliese, A., & Rizzo, C. (2011). Model predictions and evaluation of possible control strategies for the 2009 A/H1N1v influenza pandemic in Italy. *Epidemiology and Infection, 139*(1), 68–79.
32. Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., et al. (2009). Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science, 324*(5934), 1557–1561.
33. Munayco, C. V., Gomez, J., Laguna-Torres, V. A., Arrasco, J., Kochel, T. J., Fiestas, V., et al. (2009). Epidemiological and transmissibility analysis of influenza A(H1N1)v in a southern hemisphere setting: Peru. *Euro Surveillance, 14*(32), 19299.
34. Baguelin, M., Hoek, A. J. V., Jit, M., Flasche, S., White, P. J., & Edmunds, W. J. (2010). Vaccination against pandemic influenza A/H1N1v in England: A real-time economic evaluation. *Vaccine, 28*(12), 2370–2384.
35. Efimov, D. V., & Fradkov, A. L. (2008). Yakubovich's oscillatority of circadian oscillations models. *Mathematical Biosciences, 216*(2), 187–191.

36. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., et al. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine, 5*(3), e74.
37. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., & Merler, S. (2012). Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS Computational Biology, 8*(9), e1002673.
38. Poletti, P., Melegaro, A., Ajelli, M., del Fava, E., Guzzetta, G., Faustini, L., et al. (2013). Perspectives on the impact of varicella immunization on herpes zoster. A model-based evaluation from three European countries. *PloS One, 8*(4), e60732.
39. Merler, S., & Ajelli, M. (2014). Deciphering the relative weights of demographic transition and vaccination in the decrease of measles incidence in Italy. *Proceedings of the Royal Society B, 281*(1777), 20132676.

Chapter 10
Uncovering Criminal Behavior with Computational Tools

Emilio Ferrara, Salvatore Catanese, and Giacomo Fiumara

Abstract In this chapter we explore the opportunities brought in by advanced social network analysis techniques to study criminal behaviors and dynamics in heterogeneous communication media, along multiple dimensions including the temporal and spatial ones. To this aim, we present *LogViewer*, a Web framework we developed to allow network analysts to study combinations of geo-embedded and time-varying data sources like mobile phone networks and social graphs. We present some use-cases inspired by real-world criminal investigations where we used LogViewer to study criminal networks reconstructed from mobile phone and social interactions to identify criminal behaviors and uncover illicit activities.

10.1 Introduction

The pervasive diffusion of technologically mediated communication channels pushed to unprecedented frontiers the ability of individuals to interconnect and exchange information. Mobile phone networks, social networking and media platforms like Facebook and Twitter, and over-IP messaging systems like Skype and WhatsUp, represent some examples of the multitude of communication media broadly adopted in nowadays society. These phenomena generated lot of interest in the research community. Several aspects of socio-technical systems have been studied [1]: from macroscopic characteristics, like network structure [2–5], to

E. Ferrara (✉)
School of Informatics and Computing, Indiana University Bloomington,
919 E. 10th Street, Bloomington, IN 47408, USA
e-mail: ferrarae@indiana.edu

S. Catanese
Department of Mathematics and Computer Science, University of Messina, viale F. Stagno
DŠAlcontres 31, I-98166 Messina, Italy
e-mail: scatanese@unime.it

G. Fiumara
Department of Mathematics and Computer Science, University of Messina,
viale F. Stagno D'Alcontres 31, I-98166 Messina, Italy
e-mail: gfiumara@unime.it

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational
Social Sciences, DOI 10.1007/978-3-319-14011-7_10

network dynamic, like information diffusion [6–9], from microscopic behaviors, like how individual address their attention [10, 11] and what topics they discuss [12, 13], to social issues, like how people organize and mobilize using technology [14–17] and what effects technological media have at the societal level [18, 19].

One aspect that has vast societal impact is the improper usage of such platforms. Technologies have been long exploited for criminal activities: for example, various studies showed how the Internet has been exploited for cybercrime, terror, and militancy purposes [20–22]. In terms of abuse, mobile communication networks and social media have been mostly studied as vectors for the diffusion of computer viruses and malware [23, 24]. On the other hand, the possibility that such communication channels can be exploited by criminals to organize and coordinate their illicit activities in the physical world has been recently found very real [25, 26]. The ability to detect criminal behavior across different communication media is of paramount importance to avoid abuse and fight crime. For this reason, computational tools and models have been recently proposed to study criminal behavior in online platforms [27–29], social media [30], and mobile phone networks [31, 32]. Usually, such models and techniques are limited to one or few specific use-cases. For example, we recently proposed a tool called *LogAnalysis* that allows an investigator to reconstruct and visualize networks from mobile phone call data [33].

Here we present *LogViewer*, a next-generation Web-based criminal network analysis (CNA) framework that yields advanced social network analysis (SNA) functions, *de facto* extending *LogAnalysis* features to different types of networks, for example phone call networks and social graphs. *LogViewer* allows to study each network from three different angles: (1) static analysis, to investigate the role of nodes and edges, their centrality, and the emerging communities representing potential criminal rings; (2) temporal analysis, to span across different temporal events and study the flow of information over time; finally, (3) spatial analysis, embedding the network in a geographic space to determine physical closeness and locality effects on the network structure. *LogViewer* also allows to create multilayer spatio-temporal networks by merging different network types and to perform the above-mentioned different types of analysis on such a more complex network.

Our framework inherits different visualization layouts and algorithms from *LogAnalysis*: some of them are discussed in detail in our previous work [33]. Here we first give an overview of the basic concepts borrowed by SNA and their meaning in CNA; this includes network centrality measure to identify roles in criminal networks, and community detection to unveil criminal gangs hidden within the network. After that, we present the new features provided by our CNA framework, especially *ad hoc* visualization methods that we devised keeping in mind the needs of law enforcement agencies, analysts and investigators. We illustrate these advanced CNA features by presenting examples or use cases inspired by real investigations, carried out by Italian law agencies, that benefited from the adoption of *LogViewer*.

10.2 *LogViewer* Framework

10.2.1 Architecture and Workflow

LogViewer is a Web-based framework that allows advanced network analysis on criminal networks reconstructed from various data sources, including (mobile) phone data and online social network data. It supports spatio-temporal analysis and it extends, *de facto*, the horizon of possibilities provided by *LogAnalysis* [33].

This framework implements various techniques of network generation, statistical measurement, partitioning (or clustering), and visualization that rely on powerful open-sources tools; the list includes GraphML for data storage, Python network libraries for data import, normalization and network representation like NetworkX[1] and iGraph, the Stanford network analysis project (SNAP) library[2] to efficiently compute network statistics, the Louvain method for network clustering [34], and the Javascript D3.js[3] library for interactive network visualization and exploration.

The architecture of LogViewer is represented in Fig. 10.1. In the following we illustrate the typical workflow to bootstrap a criminal investigation using LogViewer. Let us use the example of data representing a mobile phone call network—the analysis of other sources, such as social network data, follows straightforwardly.

During an investigation, the agency in charge of it will obtain, usually through court warrants, raw data from a Telecommunication Service Provider related to the phone call interactions of a (possibly large) set of suspects involved in a certain criminal activity. Such data are generally provided in different formats: LogViewer

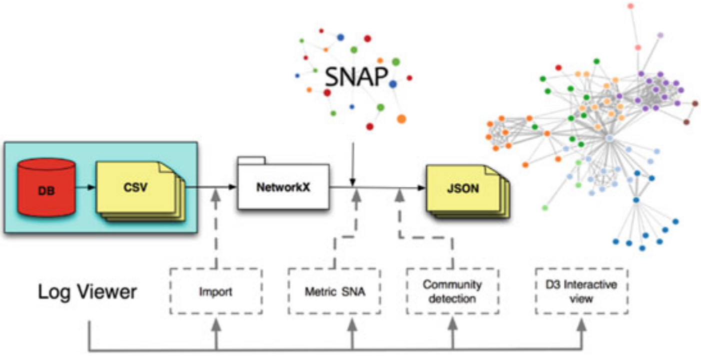


Fig. 10.1 Architecture of LogViewer

[1] http://www.networkx.github.com/.

[2] http://www.snap.stanford.edu/.

[3] http://www.d3js.org/.

allows some degree of standardization, supporting different formats adopted by various European service providers, e.g., Vodafone, Orange, and others.

The analyst can import one (or more) datasets into LogViewer, which will take care of appropriately reconstruct a network representation of such data, where each node corresponds to a given entity (generally speaking, in the mobile phone cases, the framework assumes a one-to-one mapping from phone to person, but it also supports the assignment of multiple phone numbers to the same entity, whereas such information is provided). Interconnections among entities, representing phone calls, are imported as links of this network. Duration and frequency of the calls are encoded in the network representation by means of different weighting systems that can be adopted by the analysts. For example, the raw number of calls between a pair of entities, or the average or total duration, among others, are available metrics that can be used for this purpose. This yields the possibility of performing dynamic network representation and temporal analysis.

In addition, each phone interaction reports geo-referenced data about the location of the caller and the called nodes (e.g., extrapolated from the GPS sensors on the mobile device, or approximated by the telephone cell corresponding to the physical location of the individuals at the time of the call); such information is attached to each event, to allow for spatial analysis. Once the data import procedure is completed, static representation (and spatio-temporal representation when meta-data are available) becomes available through LogViewer's visualization interface.

In the following, let us provide a bit more details about the type of data commonly processed by LogViewer for CNA purposes.

10.2.2 Data and Network Representation

10.2.2.1 Mobile Phone Data

In the context of real-world investigations, mobile phone service providers, upon request by judiciary authorities, release data logs, normally in textual file format, with space or tab separation (CSV format). A typical log file contains, at least, the values shown in Table 10.1.

Similarly, information about owners of SIM cards, dealers of SIM cards and operations like activation, deactivation, number portability are provided by the service providers as additional material to ease and support the investigation activities. Log file formats produced by different companies are heterogeneous. *LogViewer*, first of all, parses these files and converts data into GraphML format. It is an XML valid and well-formed format, containing all nodes and weighted edges, each weight representing the various weighting strategies (e.g., the frequency of phone calls) used to represent the interactions between two connected nodes. GraphML has been adopted both because of its extensibility and ease of import from different SNA toolkits and graph drawing utilities.

Table 10.1 An example of the structure of a phone log file

Field	Description
`IMEI`	IMEI code MS
`Called`	Called user
`Calling`	Calling user
`Date/time start`	Date/time start calling (GMT)
`Date/time end`	Date/time end calling (GMT)
`Type`	SMS, MMS, voice, data etc
`IMSI`	Calling or called SIM card
`CGI`	Lat. long. BTS company

10.2.2.2 Social Graph Data

Another rich source of information that is increasingly becoming adopted during criminal investigation is represented by Online Social Network data. Such types of datasets are provided by the Service Providers (like Facebook or Google) through court warrants to the law enforcement agency, similarly to mobile phone records.

Generally speaking, the datasets obtained by OSN service providers provide user meta-data related to the set of accounts of interest for the criminal investigation, including registration details (e.g., personal information, dates of account creation/deletion, etc.) along with the IP addresses corresponding to the devices used for connection (and/or the GPS coordinates of the mobile device, in case any connection is performed in mobility). Logs include, among other data, the entire history of wall posts and comments, pictures and photographs, check-in events in specific physical locations, the chronology of incoming and outgoing friendship requests, the list of friends (on Facebook) or contacts (followers and followees on Twitter and similar platforms). Some platforms, like Facebook and Twitter, can provide detailed logs of personal interactions, such as chat or personal messages. Possibly, the same set of information is provided about any number of friends/contacts of the given individual target of the criminal investigation, if deemed relevant for the investigation by the judiciary authorities. Such data about the target's neighbors help enriching the amount of information available to LogViewer to perform its analysis.

LogViewer processes these datasets and extracts the information that can be put in form of network representation. For example, when reconstructing a social network, link weighting schemes represent the interactions (e.g., number of wall comments, frequency of chatting, etc.) between a pair of individuals. Although our framework does not yet provide advanced content analysis, such additional information is often adopted by the analysts by using external tools for traditional corpora analysis.

It's worth noting that, in the context of a criminal investigation, the analysts will study social network information with different lens, say with respect to the perspective of phone interactions. This is clearly due to the fact that online friendship, say on Facebook, has a very different meaning if compared to phone interactions. On the other hand, the possibility of performing further analysis on textual content produced by personal interactions (e.g., chat) eases the analysis, say with respect to phone calls monitoring and analysis (which might not be possible whereas recordings are not readily available for investigation purposes or need additional warrants to be accessed).

10.2.3 Data Normalization and Cleaning

Data clean-up usually means the deletion of redundant edges and nodes. This step is very important since datasets often contain redundant information, that crowds graph visualization and biases statistical measures. In these circumstances, redundant edges between the same two nodes are collapsed and a coefficient—i.e., a edge weight—is attached, which expresses the number of calls. Our tool normalizes data after reading and parsing log files whichever format they have been provided among the standard formats (i.e., *fixed width text*, *delimited*, CSV, and more) used by mobile service providers.

10.3 Static Analysis of Criminal Networks

10.3.1 Centrality Measures

LogViewer takes into account the concept of *centrality measure* to highlight actors that cover relevant roles inside the analyzed network [35]. Several notions of centrality have been proposed during the latest years in the context of SNA.

There are two fundamentally different class of centrality measures in communication networks. The first class of measures evaluates the centrality of each node/edge in a network and is called point centrality measure. The second type is called graph centrality measure because it assigns a centrality value to the whole network. These techniques are particularly suited to study phone traffic and criminal networks.

In detail, in *LogViewer* we adopted four point centrality measures (i.e., *degree*, *betweenness*, *closeness*, and *eigenvector* centrality), to inspect the importance of each node of the network.

The set of measures provided in our tool is a selection of those provided by SNA [36]. It could be not sufficient to solve any possible task in phone call network analysis. In fact, for particular assignments it could yet be necessary to use additional tools in support to *LogViewer* and in further evolutions we plan to incorporate new centrality measures.

For each centrality measure, the tool gives the possibility, to rank the nodes/edges of the network according to the chosen criterion. Moreover, *LogViewer* allows to select those nodes that are central, according to the specified ranking, highlighting them and putting into evidence their relationships, by exploiting the node-link layout techniques (discussed in the following). This approach makes it possible to focus the attention of the analysts on specific nodes of interest, putting into evidence their position and their role inside the network, with respect to the others.

In the following we formally describe the centrality measures used in *LogViewer*.

They represent the centrality as an indicator of the activity of the nodes (degree centrality), of the control on other nodes (betweenness centrality), of the proximity to other nodes (closeness centrality), and of the influence of a node (eigenvector centrality).

10.3.1.1 Degree Centrality

The degree centrality of a node is defined as the number of edges adjacent to this node. For a directed graph $G = (N, E)$ with N nodes, and E edges, we can define the in-degree and out-degree centrality measures as

$$C_D(v)_{\text{in}} = \frac{d_{\text{in}}(v)}{N-1}, \quad C_D(v)_{\text{out}} = \frac{d_{\text{out}}(v)}{N-1} \tag{10.1}$$

where $d_{\text{in}}(v)$ is the number of incoming edges adjacent to the node v, and $d_{\text{out}}(v)$ is the number of the outgoing ones.

Since a node can at most be adjacent to $N - 1$ other nodes, $N - 1$ is the normalization factor introduced to make the definition independent on the size of the network and to have $0 \leq C_D(v) \leq 1$.

In and out-degree centrality indicates how much activity is going on and the most active members. A node with a high degree can be seen as a hub, an active node, and an important communication channel.

We chose to include the degree centrality for a number of reasons. First of all, its calculation is computationally feasible even on large networks. Furthermore, in the context of phone call networks it could be interpreted as the chance of a node for catching any information traveling through the network.

Most importantly, in this type of directed networks, high values of in-degree are considered a reliable indicator of a form of popularity/importance of the given node in the network; on the contrary, high values of out-degree are interpreted as a form of gregariousness of the given actor with respect to the contacted individuals.

10.3.1.2 Betweenness Centrality

The communication between two non-adjacent nodes might depend on the others, especially on those on the paths connecting the two nodes. These intermediate elements may wield strategic control and influence on many others.

The core issue of this centrality measure is that an actor is central if she lies along the shortest paths connecting other pairs of nodes. The betweenness centrality of a node v can be defined as

$$B_C(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{10.2}$$

where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ is the number of shortest paths from s to t that pass through a node v.

The importance of the betweenness centrality regards its capacity of identifying those nodes that vehiculate information among different groups of individuals.

In fact, since its definition due to Freeman [37] the betweenness centrality has been recognized as a good indicator to quantify the ability of an actor of the network to control the communication between other individuals and, specifically for this reason it has been included in *LogViewer*.

In addition, it has been exploited by Newman [38] to devise an algorithm to identify communities within a network. Its adoption in the phone traffic networks is crucial to identify those actors that allow the communication among different (possibly criminal) groups.

10.3.1.3 Closeness Centrality

Another useful centrality measure that has been adopted in *LogViewer* is called *closeness centrality*. The idea is that an actor is central if she can quickly interact with all the others, not only with her first neighbors [39]. The notion of closeness is based on the concept of shortest paths (geodesic) $d(u, v)$, the minimum number of edges traversed to get from u to v. The closeness centrality of the node v is defined as

$$C_C(v) = \frac{1}{\sum_{u \in V} d(u, v)} \tag{10.3}$$

Such a measure is meaningful for connected graphs only, assuming that $d(u, v)$ may be equal to a finite value.

In the context of criminal networks, this measure highlights entities with the minimum distance from the others, allowing them to pass on and receive communications more quickly than anyone else in the organization. For this reason, the adoption of the closeness centrality is crucial to put into evidence inside the network, those individuals that are closer to others (in terms of phone communications).

In addition, high values of closeness centrality in such type of communication networks are usually regarded as an indicator of the ability of the given actor to quickly spread information to all other actors of the network. For such a reason, the closeness centrality has been selected to be included in the set of centrality measures adopted by *LogViewer*.

10.3.1.4 Eigenvector Centrality

Another way to assign the centrality to an actor of the network in *LogViewer* is based on the idea that if a node has many central neighbors, it should be central as well. This measure is called *eigenvector centrality* and establishes that the importance of a node is determined by the importance of its neighbors.

The eigenvector centrality of a given node v_i is

$$C_E(v_i) \propto \sum_{u \in N_i} A_{ij} C_E(u) \tag{10.4}$$

where N_i is the neighborhood of the given node v_i, and $x \propto Ax$ that implies $Ax = \lambda x$. The centrality corresponds to the top eigenvector of adjacency matrix A.

In the context of telecom networks, eigenvector centrality is usually regarded as the measure of influence of a given node. High values of eigenvector centrality are achieved by actors who are connected with high-scoring neighbors, which in turn, inherited such an influence from their high-scoring neighbors and so on.

This measure well reflects an intuitive important feature of communication networks that is the influence diffusion and for such a reason we decided to include the eigenvector centrality in *LogViewer*.

10.3.1.5 Clustering Coefficient

The clustering (or transitivity) coefficient of a graph measures the degree of interconnectedness of a network or, in other words, the tendency of two nodes that are not adjacent but share an acquaintance, to get themselves in contact. High clustering coefficients mean the presence of a high number of triangles in the network.

The local clustering coefficient C_i for a node v_i is the number of links among the nodes within its neighborhood divided by the number of links that could possibly exist among them

$$C_i = \frac{|\{e_{jk}\}|}{k_i(k_i - 1)} : v_j, v_k \in N_i, e_{jk} \in E \tag{10.5}$$

where the neighborhood N of a node v_i is defined as $N_i = \{v_j : e_{ij} \in E \wedge e_{ji} \in E\}$, while $k_i(k_i - 1)$ is the number of links that could exist among the nodes within the neighborhood.

In is well-known in the literature [36] that communication networks show high values of clustering coefficient since they reflect the underlying social structure of contacts among friends/acquaintances. Moreover, high values of local clustering coefficient are considered a reliable indicator of nodes whose neighbors are very well connected and among which a substantial amount of information may flow.

For such a reason, *LogViewer* provides the possibility of computing both the global clustering coefficient for any given phone call network and the local clustering coefficient of any given node.

10.3.2 Community Detection in Criminal Networks

A criminal network can be regarded as a special kind of social network in which attention is devoted to secrecy and efficiency, since its members must communicate without being detected [40]. On the other hand, the crucial task of uncovering the functionalities of a criminal organization can be accomplished only by acquiring knowledge about the structure of the underlying criminal network. Criminal networks usually exhibit diversified compositions: hierarchical [41], cellular [42], and flat structures [43] are the most common. One of the most relevant features of real networks is the presence of clustering phenomena, or communities. The detection of communities in criminal networks brings, as a main consequence, the identification of groups and their structures via the information coded in the topology of the corresponding graph.

The problem of finding communities in a network is often expressed as a clustering problem. A widely adopted approach to solve this problem is based on the concept of *network modularity* which can be expressed as follows: given a network, represented by means of a graph $G = (N, E)$, which has been partitioned into m communities, its corresponding value of network modularity is

$$Q = \sum_{s=1}^{m} \left[\frac{l_s}{\mid E \mid} - \left(\frac{d_s}{2 \mid E \mid} \right)^2 \right] \tag{10.6}$$

assuming l_s the number of edges between vertices belonging to the s-th community and d_s is the sum of the degrees of the vertices in the s-th community. High values of Q imply high values of l_s for each discovered community, yielding to communities internally densely connected and weakly coupled among each other.

The network modularity is therefore used as fitness function to solve an optimization problem: among the several methods we mention here the Girvan and Newman (GN) algorithm [44], and an optimized variant known as Newman's algorithm [45], which is fast enough to support interactive real-time adjustments. *LogViewer* provides two strategies for detecting communities, namely the already cited Newman's algorithm and the Louvain method [34], another modularity maximization algorithm that performs very well with larger networks.

We recently discussed in great detail the problem of detecting communities and gangs inside criminal networks [31], and we point the reader's attention toward that work for an in-depth treatment of this topic.

10.3.3 Criminal Network Visualization

Typical network visualization tools rely on the popular force-directed layout [46]. The force-directed model represents the structure of the graph on the same foot as a physical system, in which nodes are physical points subject to various forces; nodes' coordinates (and therefore the layout itself) derive from the search of an equilibrium configuration of the physical system modeled by the algorithm [47]. This particular layout arrangement has the advantage of grouping users in clusters which can be identified according to the heightened connectivity. The Barnes–Hut algorithm [48] associated with this layout simulates a repulsive *N*-body system to continuously update the position of the elements.

To optimize the visualization, it is possible to interactively modify the parameters relative to the tension of the springs (edges). Nodes with low degree are associated a small tension and the elements are located in peripheral positions with respect to high degree nodes. Other parameters can be tuned, such as spring tension, gravitational force, and viscosity. Our goal, in the following, is to suggest two methods to improve force-directed based layouts. As we will show, these techniques are especially well suited for CNA; however, they could potentially be generalized for broader usage in other domains of network analysis—for example, for applications in social and political sciences.

For the usage of traditional network visualization methods in CNA, the reader should consult our recent paper on *LogAnalysis* [33].

10.3.3.1 Focus and Context Based Visualization

The number of edges within a network usually grows faster than the number of nodes. As a consequence, the network layout would necessarily contain groups of nodes in which some local details would easily become unreadable because of density and overlap of the edges. As the size and complexity of the network grow, eventually nodes and edges become indistinguishable. This problem is known as visual overload [49]. A commonly used technique to work around visual overload consists of employing a zoom-in function able to enlarge the part of the graph of interest. The drawback of this operation is the detriment of the visualization of the global structure which, during the zooming, would not be displayed. However, such a compromise is reasonable in a number of situations including, in some cases, the domain of CNA.

During an investigation, it is crucial to narrow down the analysis to the relevant suspects, to efficiently employ human and computational resources. Police officers typically draw some hypotheses about an individual suspect of being part of a criminal organization, or of being involved (or about to) in some crime; they concentrate the initial investigation on this individual, and on that person's social circles, as a ground to build the social network object of analysis. The main role of visual analysis lies in allowing the detection of unknown relations, on the base of the available limited information. A typical procedure starts from known entities, to analyze the relations with other subjects and continue to expand the network inspecting first the edges appearing the most between individuals apparently unrelated. During this procedure, only some nodes are relevant and it is important to focus on them rather than on the network as a whole.

Nevertheless, a spring embedded layout (including force-directed ones) does not provide any support to this kind of focus and analysis. In these situations, *focus and context* visualization techniques are needed to help a user to explore a specific part of a complex network. To this purpose, we here introduce the fisheye and the foci layouts.

10.3.3.2 Fisheye Visualization

Focus and context is an interactive visualization technique [50]. It allows the user to focus on one or more areas of a social network, to dynamically tune the layout as a function of the focus, and to improve the visualization of the neighboring context. The *fisheye view* is a particular focus and context visualization technique which has been applied to visualize self-organizing maps in the Web surfing [51]. It was first proposed by Furnas [52] and successively enriched by Brown et al. [53]. It is known as a visualization technique that introduces distortion in the displayed information.

The fisheye layout is a local linear enlargement technique that, without modifying the size of the visualization canvas, allows to enhance the region surrounding the focus, while compressing the remote neighboring regions. The overall structure of the network is nevertheless maintained. An example of application of this technique is shown in Fig. 10.2. The picture shows a moderately small criminal network reconstructed from phone call interactions of about 75 individuals. The layout on the left panel is obtained by using a force-directed method implemented in our framework, *LogViewer*. The analyst can inspect the nodes of the network, which contains known criminals, suspects, and their social circles. When the focus is applied on a given node, the visualization transitions to the fisheye layout (see the right panel). A tool-tip with additional information about the node appears when the node is selected—it shows the phone number, personal details, address, photo, etc. The layout causes edges among remote nodes to experience stronger distortions than local nodes. The upside of the presented method is the possibility to achieve the three recommendations of Network Nirvana [54] when focusing on a given node: all the nodes' neighbors are clearly visible, the node degree is easily countable, and the edges incident on that node can be identified and followed.

Note that fisheye and force-directed layouts can be used in conjunction. By combining the two methods, our framework efficiently yields focus and context views.

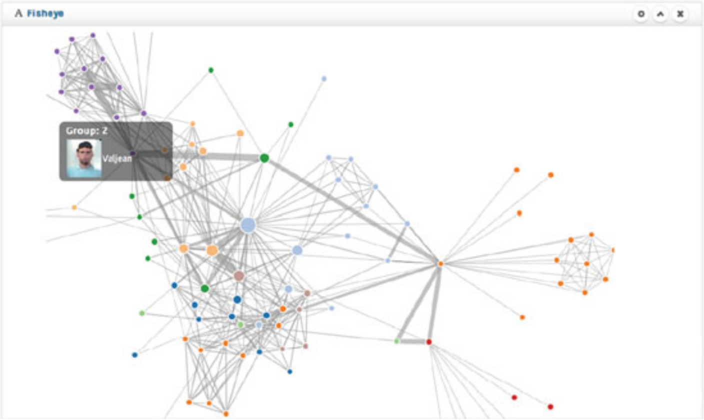


Fig. 10.2 Fisheye visualization

10.3.3.3 Matrix Layout

A network can be represented by using an adjacency matrix in which each cell *ij* represents the edge existing between the vertex *i* and the vertex *j*. In our case, the vertices represent the phone numbers of the users (the caller and the called), and the edges represent their contacts. The natural visualization technique associated with this two-dimensional representation of the graph is the matrix layout. Nevertheless, the efficiency of a matrix diagram strongly depends upon the order of rows and columns: if the nodes that are connected are placed in order, then clusters and connections among communities can be easily identified. As shown in Fig. 10.3, matrix cells can be coded to show additional information: in this case, different colors represent different clusters.

On the contrary of node-link diagrams, matrix layout makes not easily identifiable the paths connecting the vertices. On the other hand, when dealing with highly connected networks, the node-link layout rapidly becomes unreadable as a consequence of the large superposition of nodes and edges.

10.3.3.4 Foci Layout

The *foci layout* implements three network visualization models: force-directed, semantic, and clustered layouts. The latter is based on the Louvain community detection algorithm [34]. Future implementations will explore other methods [55, 56]. Our model supports multilayer analysis of the network through interactive transitions from the force-directed layout, with a single gravitational center, to the clustered one with more force centers placed in predetermined distinct areas. This layout allows to analyze the network on various layering levels depending on

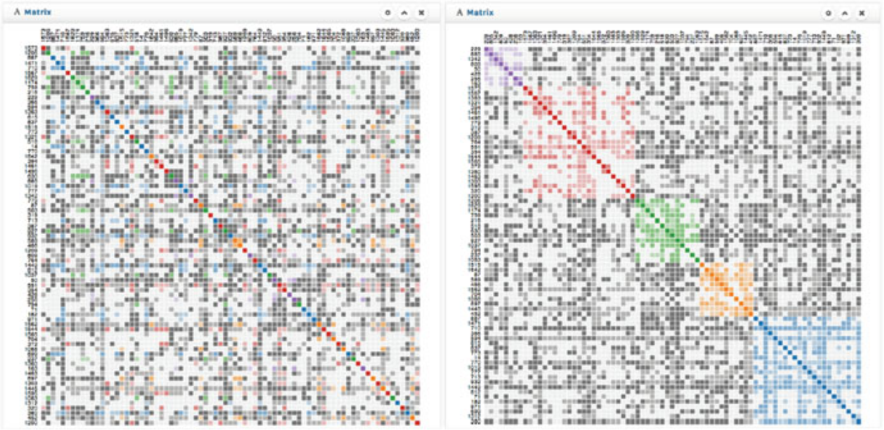
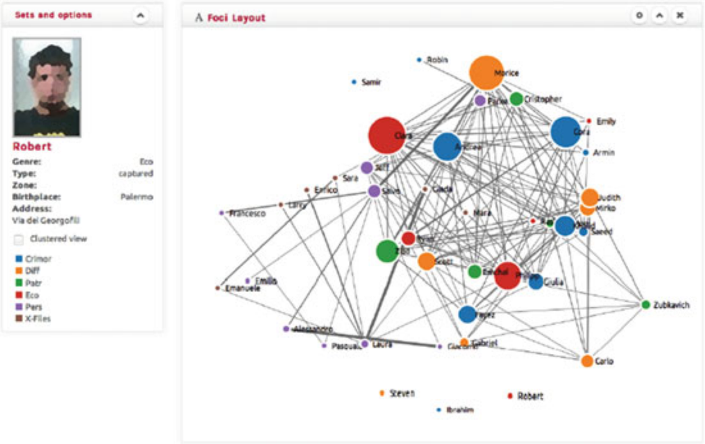


Fig. 10.3 Matrix layout and clustering



Robert
Genre: Eco
Type: captured
Zone:
Birthplace: Palermo
Address:
Via dei Georgofili
□ Clustered view
🟦 Crimor
🟧 Diff
🟩 Patr
🟥 Eco
🟪 Pers
🟫 X-Files

Fig. 10.4 Foci layout

specified node attributes. Figure 10.4 shows the phone traffic network of some clans the previous criminal network, in which the color of the nodes denotes the type of crime committed by the members.

In this example, the clustering truthfully reflects the known territorial division among the groups belonging to the organization. In Fig. 10.4 the focus is on a specific node. Using this layout it is possible to contextually analyze the community structure, the type of committed crime with respect to the members of the clan, and the direct relations of each single individual. This layout integrates also the forth Network Nirvana recommendation, namely the possibility to identify clusters and to highlight the community structure (Figs. 10.5 and 10.6).

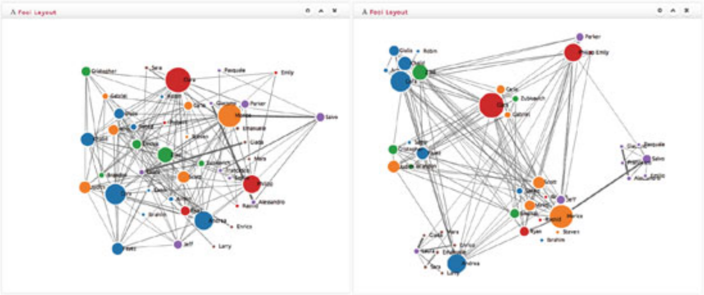
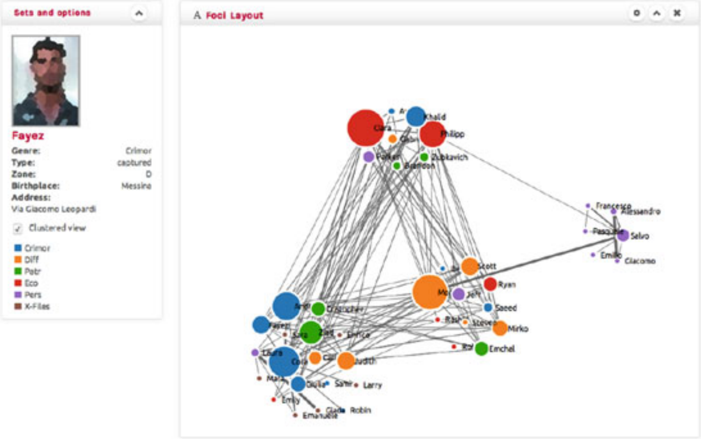


Fig. 10.5 Multi-foci layout



Sets and options

Fayez

Genre: Crimor
Type: captured
Zone: D
Birthplace: Messina
Address:
Via Giacomo Leopardi

☑ Clustered view

- 🟦 Crimor
- 🟧 Diff
- 🟩 Patr
- 🟥 Eco
- 🟪 Pers
- 🟫 X-Files

Fig. 10.6 Filtered and clustered multi-foci layout

10.4 Spatio-Temporal Criminal Networks Analysis

10.4.1 Temporal Network Analysis

Phone call records and online social network data comes with temporal information attached to many events. For example, the time and duration of a call or a chat session, or the timestamp associated with the creation of a given phone contract or account on a social platform, are common meta-data available for investigation.

LogViewer provides extensive support to encode and exploit temporal information, when available, to perform network dynamic and temporal pattern analysis. One example is provided in Fig. 10.7, where we display LogViewer's interface reporting aggregate temporal statistics related to the activity ongoing on a mobile phone network under investigation.

In this example three types of information are displayed: on top, a time series reports the volume of calls per day during the investigation period. It's possible to

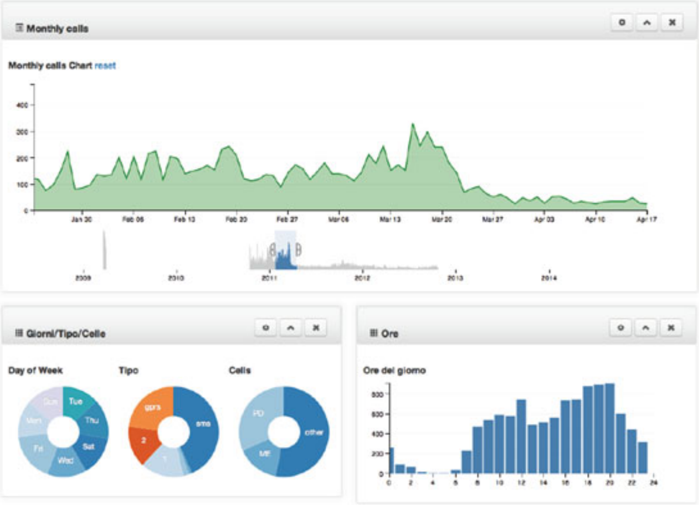


Fig. 10.7 Temporal analysis of a criminal network

see how heterogeneous is this traffic, with a strong attenuation toward the end of the observation period, after a spike coinciding with an actual criminal event in the real world. The analyst has the possibility of zooming in the time series, to select different sub-intervals, to display different types of statistics over time (e.g., total volume of calls, or total duration, etc.) and to filter according to different types of constraints (e.g., showing only the information related to a subset of users, for example a particular cluster). The applied filters are also reported underneath, for example as pie charts that show specific statistics per day of the week, per type of event (e.g., phone calls, texts, video calls, etc.), and per geographic area. Better resolution is provided by histograms that bin the given statistics, say number of calls, per hour of the day.

Another example is provided in Fig. 10.8 that shows a *stream graph* adopted to visualize a sequence of temporal events on an aggregate basis. Stream graphs show the potential of tools that provide dynamics and interactive data exploration. The x axis of the stream graph represents time, whereas the y axis reports an arbitrary metric, say for the example in Fig. 10.8 the total volume of phone interaction, subdivided by type (e.g., calls, texts, Internet sessions, etc.), each displayed using a different color. The stream is proportional to the number of events of each type per unit of time (one bin here is 1 h). LogViewer also implements stacked graphs. Stream and stacked graphs represent especially helpful tools when the analysts want to visually compare extensive metrics that depend on the volume of events in a predetermined period.

By selecting the various temporal analysis tools and filters available, the analyst can dissect the dataset under analysis to obtain granular temporal information or to highlight and let emerge specific patterns of interactions among particular groups

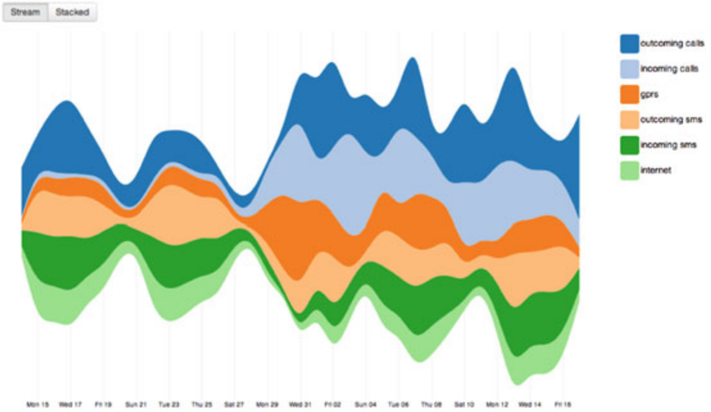


Fig. 10.8 Stream layout of temporal dynamics in a criminal network

of individuals. This, in conjunction with spatial filters that are discussed in the next section, yields the ability to determine when (and where) information flows, and to identify the peeks and lows of interaction activity among the members of a criminal organization, to narrow down investigations towards specific periods of interest (that might concur with events in the real world).

10.4.2 Spatial Network Analysis

Along with temporal information, phone call records and online network datasets report, among others, geographical coordinates of most of the events. Latitude and longitude can be inferred from the BTS (base transceiver station) of a cell, or directly derived from the GPS sensors of enabled devices. In related work [31] we provide some additional detail on the inference mechanism behind the reconstruction of geo-coordinates from BTS cells.

LogViewer encodes, processes, and presents spatial information to derive the mobility patterns of individuals, routine paths and points of interest, reconstructed from the geo-referenced interconnections (both phone calls and online social network sessions and check-ins). Figure 10.9 shows, for example, a case study inspired by a real investigation where nodes, displayed in overlay onto a map, represent areas where, during the observation period, intense contacts among a subset of the population under investigation took place (node sizes encode the volume of calls binned by geographic position). Different filters are provided, along with a slider that allows to "unfold" time and replay the evolution of such network simulating the temporal dimension. The spatial analysis, combined with the temporal filters, allow to observe the dynamic patterns of interconnections among the individuals under observation, and it's especially helpful to locate them in space

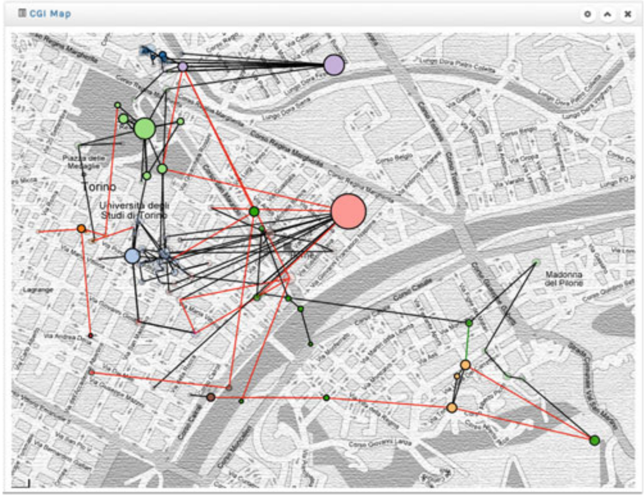


Fig. 10.9 Spatial analysis of a criminal network

and time, that could help in those cases when evidence is needed to prove someone's presence in a determined location during a specific event occurred in the real world (for example, a robbery or a homicide).

10.5 A Use Case Inspired by Real Investigations

LogViewer has been successfully used in real forensic police investigations. Various examples, and the details of the analysis presented here, have been discussed in our latest work [31]: let us summarize few interesting results. Note that, as criminal lawsuits are still in progress, some information has been intentionally obfuscated.

10.5.1 The Initial Configuration

We here discuss a case in which some people allegedly belong to a criminal network. Police determined that phone traffic logs acquired (under court warrants) from the service providers of the suspects might reveal crucial information about their interpersonal relationships and communication dynamics. The logs reflect the phone calls occurred throughout 15 days among these individuals allegedly part of a criminal association responsible of robbery, extortion, and drug illicit trafficking.

From the analysis of the interactions occurred in a given time interval it is also possible to unveil the most important links, in terms of frequencies of relations and flow of information. Links do not necessarily reflect the same type of relations:

different motivations can underlay phone interactions. In lack of advanced methods for conversation analysis (and due to the lack of phone call recordings), content analysis in some cases is impossible. However, the topology of the call networks is precious to reveal possible structural groups and, from there, ascertain the details.

10.5.2 Finding Subgroups

In Fig. 10.10a we show the case study network after the Girvan–Newman (GN) algorithm has been executed and 16 communities have been detected. Different colors of the nodes identify different communities. To improve the clarity of the network visualization, we exploit the clustered view as shown in Fig. 10.10b. This configuration adopts a modified force-directed layout in which nodes of the same community (same colors) form macro-nodes visualized with a circular layout. In such a way, inter-connectivity among communities is captured better. The macro-nodes can be further exposed to reveal intra-community relationships (see Fig. 10.10c).

In this case, we are not interested only in the nodes that occupy prominent positions. Rather, we should focus on those edges whose deletion during the execution of the GN algorithm unveils new structural configurations, which in turn can be investigated using additional information available to police. This analysis will be of central importance for the successful outcome of the investigation.

LogViewer supported this case investigation as follows: first, by automatically parsing interaction data (phone traffic) from heterogeneous sources; then, by abstracting a network representation of such data where nodes represent individuals, being links their interactions—a node-link layout is employed for visualization purpose. Finally, after performing community detection (and visualizing clusters), each member of these groups is analyzed in depth, recursively refining the results.

From clustering, two interesting results follow. First, the more central edges are not always responsible for driving the majority of the information, that is they are not in charge of communications among clusters. They are, however, still important edges from a topological point of view, and *lethal* when regarded inside their group. Secondly, clustering algorithms used to analyze criminal networks help to detect the tightest groups, but the nature of the relations must be carefully evaluated using information which cannot be directly drawn from the mathematical model or its graphical representation. Network metrics applied to our case study reveal that the node with the highest degree (i.e., the highest number of phone calls) has a much lower betweenness centrality than other nodes. In fact, criminal networks heavily employ secrecy to escape investigations and, in particular, a policy of internal communications according to which the most important members issue orders to a very limited number of members which in turn make them known to an increasing number of less important members until the leaves of the network are informed.

In our case study, the nodes having the highest number of communications (i.e., the highest degree) represent the lieutenants of the criminal organization and not the

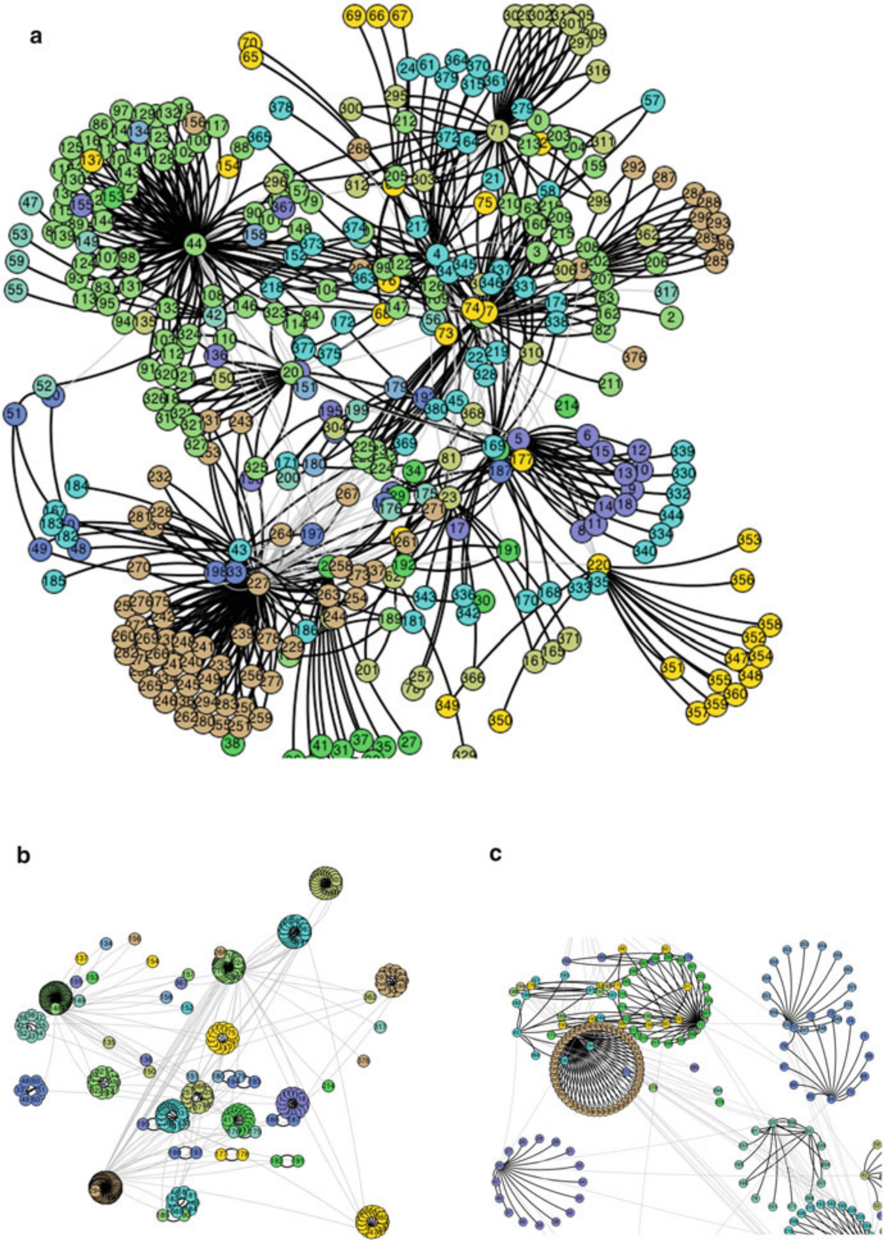


Fig. 10.10 Communities as obtained by using Girvan Newman algorithm. (**a**) Case study network after the GN algorithm. Sixteen communities have been detected. (**b**) Clustered view. Nodes of the same community form macro-nodes visualized with a circular layout. (**c**) Macro-nodes zoom reveals intra-community relationships

boss of the clan, while the edges traversed by the highest number of shortest paths (i.e., having the highest betweenness centrality) represent the most important links among the various groups.

Moreover, the granularity of the clustering allows to identify the optimal members and edges to remove when trying to hinder or disrupt the clan criminal activities.

The next step of analysis is carried out by using the Newman algorithm. Figure 10.11a shows communities embedded in convex hulls. Since the visualization might be cluttered and compromise the interpretation of the results, we here exploited the community compression techniques described above to improve the quality of the representation. For example, by setting the inter-community hop filter to a value of 2, Fig. 10.11b shows the communities, and the respective members, that can be reached from the selected nodes at most in two hops. Figure 10.11c represents the egonet of the selected user, and the summary of communities connected in one hop.

10.5.3 Overlapping Communities

An important aspect in the analysis of communities is represented by the potential overlap of communities. Both the algorithms implemented in *LogViewer* actually perform a partition of the network, thus assigning each of the nodes to exactly one cluster. Often, this is not a correct representation, at least on a semantic basis, of the network. In a specific case such ours, even the algorithmic approaches described in [57, 58] may produce questionable results because of the multiplicity of meanings which can be given to any edge of the network. For this reason, we decided to implement *LogViewer* in such a way to allow the user to choose the level of clustering in order to approximate the results. This feature is illustrated in Fig. 10.12.

In Fig. 10.12a only one cluster has been detected which is composed of the nodes interconnected among the external clusters represented by the nodes "Elio" and "Judy," while in Fig. 10.12b, $Q_{\max}$ has been interactively decreased to a previous lower value. As a consequence, the interconnected nodes are subdivided and new communities emerge.

The in-depth analysis carried out on the members of the clusters interconnected (shown in Fig. 10.12)—and the temporal analysis—allowed the investigators to discover that some clans belonging to the criminal network had worked with a certain degree of autonomy and were responsible for some murders. It turned out that these clans were tasked of committing murders on account of the organization. Figure 10.13 shows the clans at times t_1 and t_2 (all names are fictitious).

Some additional remarks are needed. Applying the GN community detection algorithm without supervision (i.e., only to maximize the modularity) produces a partition according to which the criminal network is composed of 14 clusters. The maximum partition density is 0.014 and the largest community is composed of 84 nodes. This clustering is not coherent with the real structural subdivision

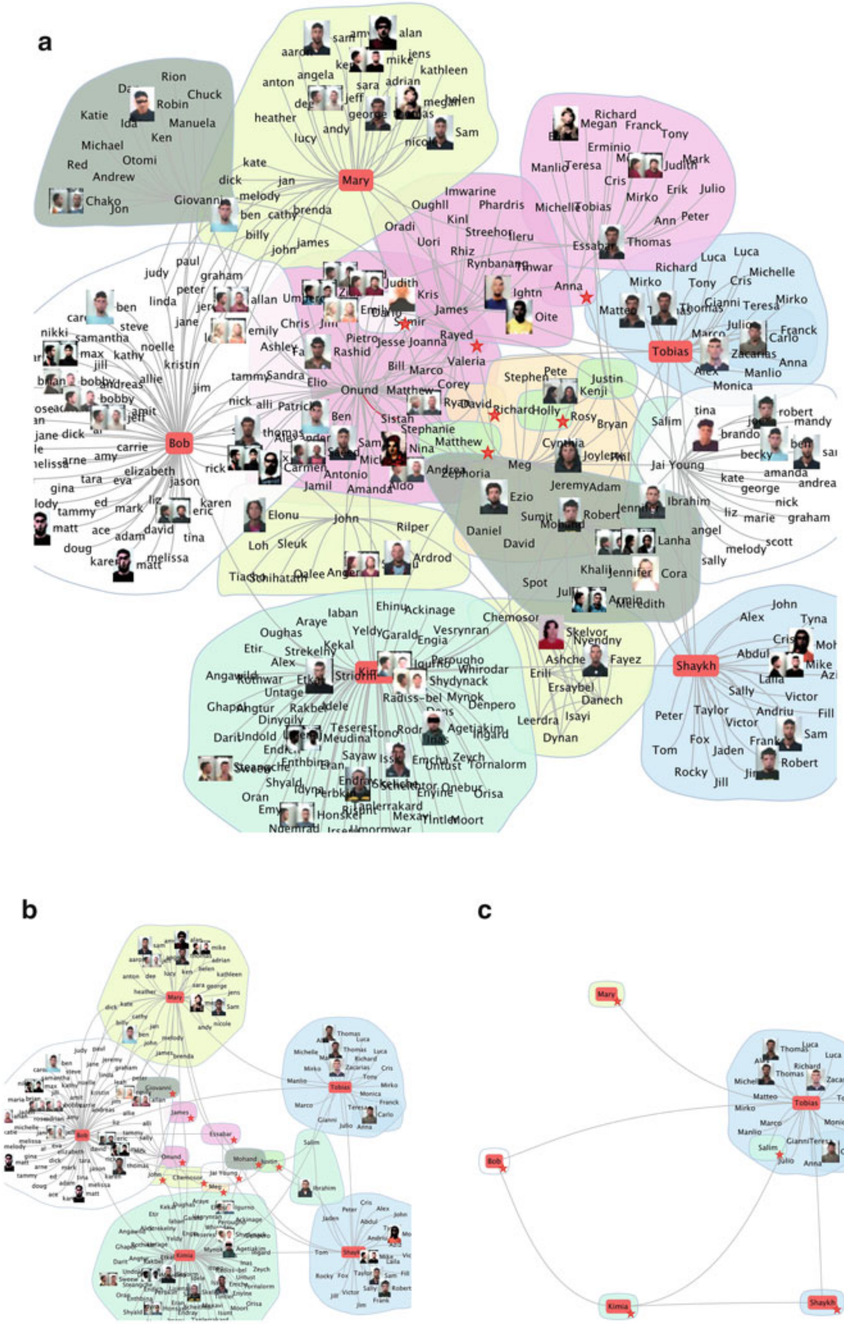


Fig. 10.11 Girvan Newman community detection on case study network. (**a**) Case study network after applying the Newman algorithm. Nineteen communities have been detected, five of which are visualized, centered on the selected node ("Tobias," in *red*). (**b**) Filtered communities with intercommunity hops 2 from Tobias node. (**c**) Egonet of Tobias node (intercommunity hops = 1)

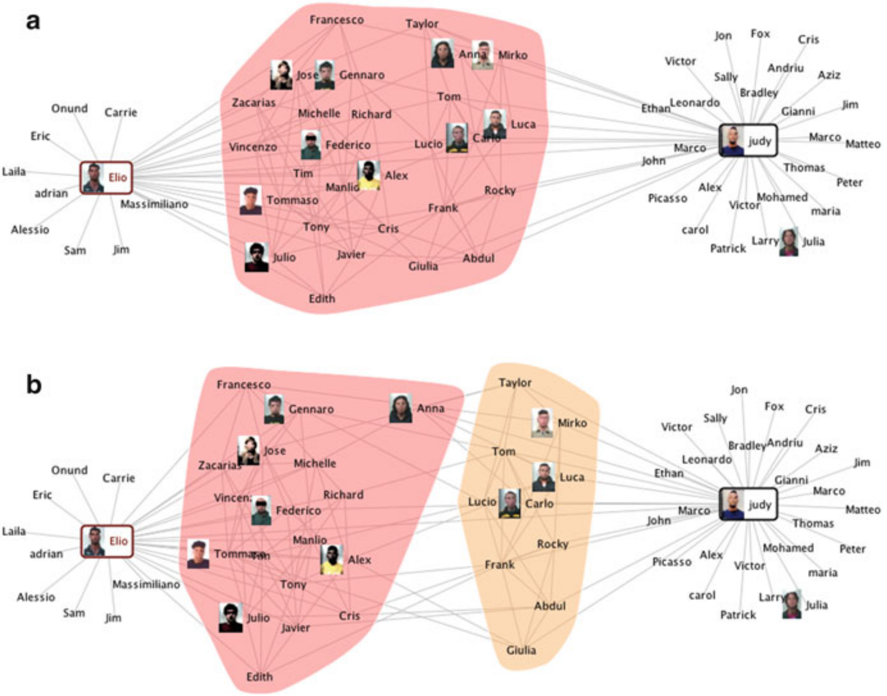


Fig. 10.12 An example of community detection using the Newman algorithm [38]. The convex-hull layout has been adopted for the visualization of the communities. (**a**) $Q_{\max}$ modularity. (**b**) $Q_{\max} - 1$ modularity

of the criminal network, as it emerged from the supervised interactive community detection procedure, combined with additional comparisons and in-depth examinations obtained from other informative sources. Nevertheless, this result was very interesting since important information regarding some members of the criminal network emerged.

In particular, from the analysis of the different levels of clustering selected interactively, and from the observation of the relative variations in the obtained configurations, we identified which elements of the network were affected mostly.

Concluding, the analysis of the distribution of phone calls carried out by each *clan* (see Fig. 10.14) is a method generally very useful to decide if a good level of clustering has been obtained after the execution of the community detection algorithm. The goal of this analysis is twofold: first, it identifies the groups among which the largest number of phone calls, texts, MMS, etc., took place; second, it highlights the peaks of the stream of communications related not to single users but rather to each cluster as a whole, on the occasion of a crime.

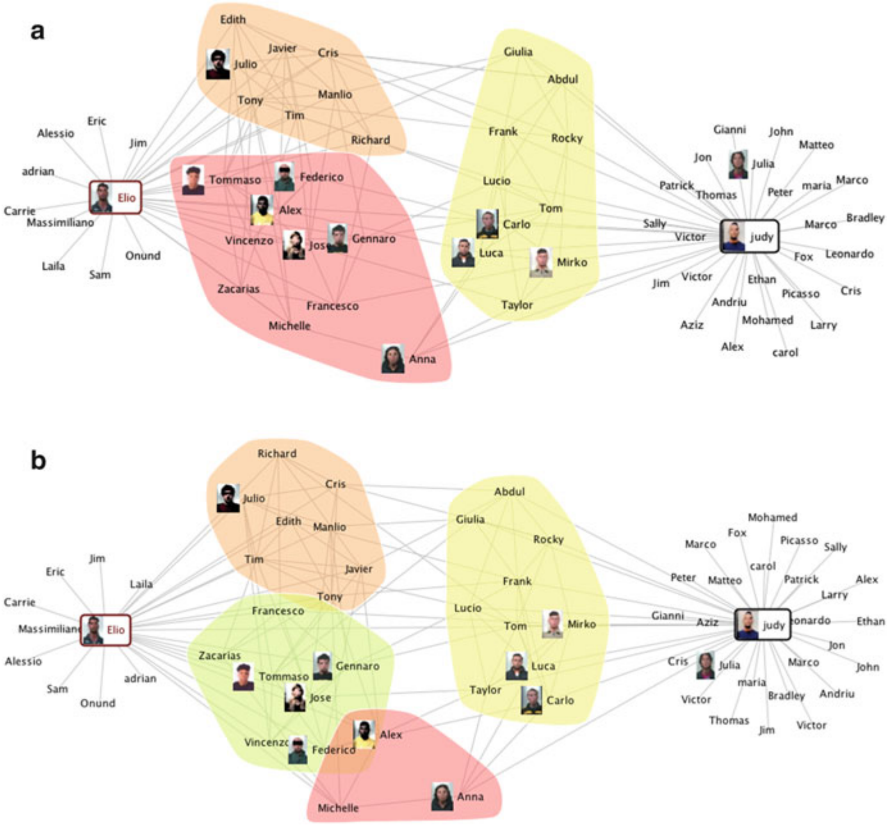


Fig. 10.13 Community detection of a time-varying criminal network. (**a**) The criminal network at time t_1. (**b**) The criminal network at time t_2

10.6 Related Work

In the latest 30 years academic research related to the application of SNA to intelligence and study of criminal organizations has constantly grown. One of the most important studies is due to Malcolm Sparrow [59], related to the application of techniques of network analysis, and the study of network vulnerabilities, for intelligence scopes. He underlined three key aspects of so-called CNA: (1) the importance of SNA for the analysis of criminal data; (2) the potential of added intelligence from network analysis, and (3) the results deriving from the collaboration between the two sectors.

Sparrow defined four features peculiar of criminal networks (CNs): (1) limited dimension—CNs are often composed of at most few thousand nodes; (2) information incompleteness—criminal or terrorist networks are unavoidably incomplete due to fragmentary available information and erroneous information; (3) undefined

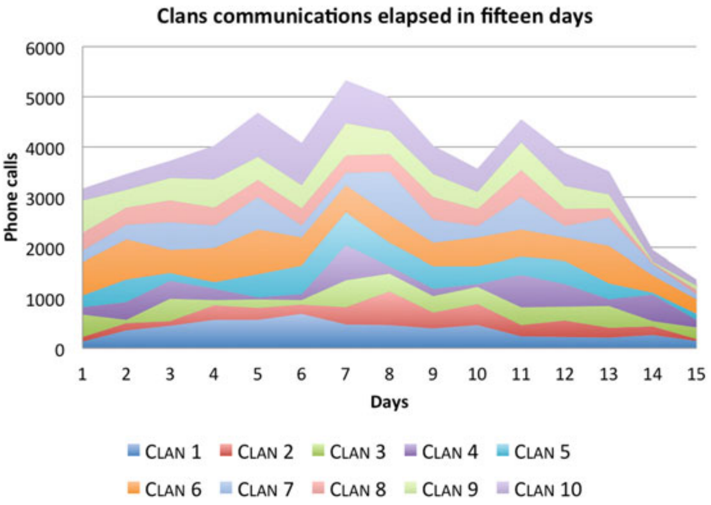


Fig. 10.14 Stacked histogram showing the phone call traffic carried out by each group (or clan) in the time interval of 15 days

borders—it is difficult to determine all the relations of a node; and, (4) dynamics—new connections imply a constant evolution of the structure of the network.

Thanks to Sparrow's work, other authors tried to study criminal networks using the tools of SNA. For example, Baker and Faulkner [60] studied illegal networks in the field of electric plants and Klerks [61] focused on criminal organizations in The Netherlands. In 2001, Silke [62] and Brennan et al. [63] acknowledged a slow growth in the fight against terrorism, and examined the state of the art in the field of CNA.

Arquilla and Ronfeldt [20] summarize prior research by introducing the concept of Netwar and its applicability to terrorism. They illustrate the difference between social networks and CNs, demonstrating the great utility of network models to understand the nature of criminal organizations. Their work shed light on strategies, methods, and systems of information flow for intelligence purpose. The framework proposed by Arquilla and Ronfeldt provided new ground for conceiving network analysis. Nevertheless, they received some criticism due to their theoretical approach. Before 2001-09-11, some criticism can be found in the work of Carley, Reminga and Kamneva [64], devoted to destabilizing initiatives of dynamic terrorist networks.

All these early studies somehow neglected the importance of network visualization, stressing aspects related more to statistical network characterization, or interpretation of individuals' roles rooted in social theory. However, in 2006, a popular work by Valdis Krebs [43] applied graph analysis in conjunction with network visualization theory to analyze the Al Qaeda cell responsible of the 2001-09-11 terrorist attacks in the USA. This work represents a starting point of a series of academic papers in which SNA methods become applied to a real-world cases, differently from previous work where mostly toy models and fictitious networks were used. Krebs' paper is one of the more cited papers in the field of application of SNA to Criminal Networks and it inspired further research in network visualization for the design and development of better SNA tools applications to support intelligence agencies in the fight against terror, and law enforcement agencies in their quest fighting crime.

In criminology and research on terrorism, SNA has been proved a powerful tool to learn the structure of a criminal organization. It allows analysts to understand the structural relevance of single actors and the relations among members, when regarded as individuals or members of (one or more) subgroup(s). SNA defines the key concepts to characterize network structure and roles, such as centrality [37], node and edge betweenness [37, 65], and structural similarity [66]. The understanding of network structure derived from these concepts would not be possible otherwise [36]. The above-mentioned structural properties are heavily employed to visually represent social and criminal networks as a support decision-making processes.

SNA provides key techniques including the possibility to detect clusters, identify the most important actors and their roles and unveil interactions through various graphical representation methodologies [67]. Some of these methods are explicitly designed to identify groups within the network, while others have been developed to show social positions of group members. The most common graphical layouts have historically been the node-link and the matrix representations [68].

Visualization has become increasingly important to gain information about the structure and the dynamics of social networks: since the introduction of sociograms, it appeared clear that a deep understanding of a social network was not achievable only through some statistical network characterization [36]. For all these reasons, a number of different challenges in network visualization have been proposed [54]. The study of network visualization focuses on the solution of the problems related to clarity and scalability of the methods of automatic representation. The development of a visualization system exploits various technologies and faces some fundamental aspects such as: (1) the choice of the layout; (2) the exploration dynamics; and (3) the interactivity modes introduced to reduce the visual complexity.

Recent studies tried to improve the exploration of networks by adding views, user interface techniques and modes of interaction more advanced than the conventional node-link and force-directed [46] layouts. For example, in *SocialAction* [69] users are able to classify and filter the nodes of the network according to the values of their statistical properties. In *MatrixExplorer* [70] the node-link layout is integrated with the matrix layout. Nonetheless, these visualization systems have not been explicitly developed with the aim of the exhaustive comprehension of all properties of the network. Users need to synthesize the results coming from some views and assemble metrics with the overall structure of the network.

Therefore, we believe that an efficient method to enhance the comprehension and the study of social networks, and in particular of criminal networks, is to provide a more explicit and effective node-link layout algorithm. This way, important insights could be obtained from a unique layout rather than from the synthesis derived from some different layouts.

We recently presented a framework, called *LogAnalysis* [31, 33], that incorporates various features of SNA tools, but explicitly designed to handle criminal networks reconstructed from phone call interactions. This framework allows to visualize and analyze the phone traffic of a criminal network by integrating the node-link layout representation together with the navigation techniques of zooming and focusing and contextualizing. The reduction of the visual complexity is obtained by using hierarchical clustering algorithms. In this chapter we discuss three new network layout methods that have been recently introduced in *LogViewer*, namely fisheye, foci, and geo-mapping, and we explain how these methods help investigators and law enforcement agents in their quest to fight crime.

It's worth noting that various tools to support network analysis exist. However, only few of them have been developed specifically for criminal network investigations. We mention, among others, commercial tools like COPLINK [29, 71], Analyst's Notebook,[4] Xanalysis Link Explorer,[5] and Palantir Government.[6] Other prototypes described in academic papers include Sandbox [72] and POLESTAR [73]. Some of these tools show similar features to *LogViewer*, but, to the best of our knowledge, none of them yields the same effective and scalable network visualization with support to criminal networks reconstructed from phone call records.

10.7 Conclusions

In this chapter we presented *LogViewer*, a next-generation Web-based framework that provides advanced features for CNA. We first provided a high-level overview of the workflow that analysts follow to bootstrap a criminal investigation by using a framework like ours, and then we presented some underlying theory behind the network measures, clustering methods, and visualization techniques adopted to uncover criminal behavior in spatio-temporal networks reconstructed from microscopic human interactions (e.g., mobile phone calls or online social network data).

LogViewer paves the way for the creation of a general framework for the identification of criminal activities from digital footprints, however there is a lot to be done yet. In our vision, this framework will extend at least in three fundamental directions in the future: (1) infer roles of individuals in the hierarchical structure of a criminal organization; (2) predict crimes from spatio-temporal patterns of criminal

[4] http://www.ibm.com/software/products/analysts-notebook/.

[5] http://www.xanalys.com/products/link-explorer/.

[6] http://www.palantir.com/solutions/.

activity; (3) predict which individuals within a social network are more exposed to the possibility of turning into criminals in the future, given their social circles and their interactions with existing criminals.

Concluding, from a technical perspective, we are already working to incorporate further sources of network interactions at the microscopic level, such as financial transaction records or face-to-face interactions that might be recorded and tracked through advanced traditional investigation methods.

References

1. Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science, 325*(5939), 425.
2. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (pp. 29–42). ACM.
3. Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. In P. S. Yu, J. Han, & C. Faloutsos (Eds.), *Link mining: Models, algorithms, and applications* (pp. 337–357). Berlin: Springer.
4. Ferrara, E. (2012). A large-scale community structure analysis in Facebook. *EPJ Data Science, 1*(9), 1–30.
5. De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2014). On Facebook, most ties are weak. *Communications of the ACM, 57*(11), 78–84.
6. Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 695–704). ACM.
7. Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 519–528). ACM.
8. Myers, S. A., Zhu, C., & Leskovec, J. (2012). Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 33–41). ACM.
9. Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2013). Traveling trends: Social butterflies or frequent fliers? In *Proceedings of the First ACM Conference on Online Social Networks* (pp. 213–222). ACM.
10. Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 251–260). ACM.
11. Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports, 2.*
12. Conover, M. D., Gonçalves, B., Flammini, A., & Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science, 1*, 6.
13. Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., & Vespignani, A. (2012). Beating the news using social media: the case study of American idol. *EPJ Data Science, 1*(1), 8.
14. González-Bailón, S., Borge-Holthoefer, J., Rivero, A., & Moreno, Y. (2011). The dynamics of protest recruitment through an online network. *Scientific Reports, 1.*
15. Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., & Flammini, A. (2013). The geospatial characteristics of a social movement communication network. *PLoS One, 8*(3), e55957.

16. Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The digital evolution of Occupy Wall Street. *PLoS One, 8*(5), e64679.
17. Varol, O., Ferrara, E., Ogan, C., Menczer, F., & Flammini, A. (2014). Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 ACM Conference on Web Science* (pp. 81–90). ACM.
18. Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on Twitter networks: Validation of dunbar's number. *PLoS One, 6*(8), e22656.
19. Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of babel: Mapping world languages through microblogging platforms. *PLoS One, 8*(4), e61981.
20. Arquilla, J., & Ronfeldt, D. (2001). Networks and netwars: The future of terror, crime, and militancy. *Survival, 44*(2), 175–176.
21. Casey, E. (2011). *Digital evidence and computer crime: Forensic science, computers and the internet*. New York: Academic.
22. Jewkes, Y., & Yar, M. (2013). *Handbook of Internet crime*. London: Routledge.
23. Leavitt, N. (2005). Mobile phones: The next frontier for hackers? *Computer, 38*(4), 20–23.
24. Hypponen, M. (2006). Malware goes mobile. *Scientific American, 295*(5), 70–77.
25. Morselli, C. (2005). *Contacts, opportunities, and criminal enterprise*. Toronto: University of Toronto Press.
26. Morselli, C. (2008). *Inside criminal networks* (Vol. 8). Berlin: Springer.
27. Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. *Communications of the ACM, 48*(6), 100–107.
28. Xu, J., Marshall, B., Kaza, S., & Chen, H. (2004). Analyzing and visualizing criminal network dynamics: A case study. In H. Chen, R. Moore, D. D. Zeng, & J. Leavitt (Eds.), *Intelligence and security informatics* (pp. 359–377). Berlin: Springer.
29. Xu, J. J., & Chen, H. (2005). Crimenet explorer: A framework for criminal network knowledge discovery. *ACM Transactions on Information Systems (TOIS), 23*(2), 201–226.
30. Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from Twitter posts. In S. J. Yang, A. M. Greenberg, & M. Endsley (Eds.), *Social computing, behavioral-cultural modeling and prediction* (pp. 231–238). Berlin: Springer.
31. Ferrara, E., De Meo, P., Catanese, S., & Fiumara, G. (2014). Detecting criminal organizations in mobile phone networks. *Expert Systems with Applications, 41*(13), 5733–5750.
32. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. arXiv preprint. arXiv:1409.2983.
33. Catanese, S., Ferrara, E., & Fiumara, G. (2013). Forensic analysis of phone call networks. *Social Network Analysis and Mining, 3*(1), 15–33.
34. Blondel, V., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.
35. Morselli, C. (2010). Assessing vulnerable and strategic positions in a criminal network. *Journal of Contemporary Criminal Justice, 26*(4), 382–392.
36. Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.
37. Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry, 40*, 35–41.
38. Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6), 066133.
39. Newman, M. (2005). A measure of betweenness centrality based on random walks. *Social Networks, 27*(1), 39–54.
40. Wiil, U. K., Gniadek, J., & Memon, N. (2010). Measuring link importance in terrorist networks. In 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 225–232). Odense: IEEE Computer Society, 9–11 August 2010.

41. Sageman, M. (2004). *Understanding terror networks*. Philadelphia: University of Pennsylvania Press.
42. Todd, M., & Nomani, A. (2011). *The truth left behind: inside the kidnapping and murder of Daniel Pearl*, New York. http://www.publicintegrity.org/2011/01/20/2190/.
43. Krebs, V. (2002). Mapping networks of terrorist cells. *Connections, 24*(3), 43–52.
44. Girvan, M., Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821.
45. Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 26113.
46. Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience, 21*(11), 1129–1164.
47. Brandes, U. (2001). Drawing on physical analogies. In M. Kaufmann, & D. Wagner (Eds.), *Drawing Graphs*. Lecture Notes in Computer Science (Vol. 2025, pp. 71–86). Berlin/Heidelberg: Springer.
48. Barnes, J., & Hut, P. (1986). A hierarchical 0(N log N) force-calculation algorithm. *Nature, 324*, 4.
49. Assa, J., Cohen-Or, D., & Milo, T. (1997). Displaying data in multidimensional relevance space with 2d visualization maps. In *Proceedings of Visualization '97* (pp. 127–134).
50. Leung, Y. K., & Apperley, M. D. (1994). A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction, 1*(2), 126–160.
51. Yang, C., Chen, H., & Hong, K. (2003). Visualization of large category map for internet browsing. *Decision Support Systems, 35*(1), 89–102.
52. Furnas, G. W. (1986). Generalized fisheye views. *SIGCHI Bulletin, 17*(4), 16–23.
53. Sarkar, M., & Brown, M. H. (1994). Graphical fisheye views. *Communications of the ACM, 37*(12), 73–84.
54. Schneider, F., Feldmann, A., Krishnamurthy, B., & Willinger, W. (2009). Understanding online social network usage from a network perspective. In *Proceedings of the 9th SIGCOMM Conference on Internet Measurement Conference* (pp. 35–48). ACM.
55. De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2013). Enhancing community detection using a network weighting strategy. *Information Sciences, 222*, 648–668.
56. De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2014). Mixing local and global information for community detection in large networks. *Journal of Computer and System Sciences, 80*(1), 72–87.
57. Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*, 814–818.
58. Sun, P. G., Gao, L., & Shan Han, S. (2011). Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. *Information Sciences, 181*, 1060–1071.
59. Sparrow, M. K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks, 13*(3), 251–274.
60. Baker, W., & Faulkner, R.. The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American Sociological Review, 58*, 837–860 (1993)
61. Klerks, P., & Smeets, E. (2001). The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the netherlands. *Connections, 24*, 53–65.
62. Slike, A. (2001). The devil you know: Continuing problems with research on terrorism. *Terrorism and Political Violence, 13*, 1–14.
63. Brannan, D. W., Esler, P. F., & Anders Strindberg, N. T. (2001). Talking to terrorists: Towards an independent analytical framework for the study of violent substate activism. *Studies in Conflict and Terrorism, 24*(1), 3–24.
64. Reminga J., Carley, K. M., & Kammneva, N. (1998). *Destabilizing terrorist networks* (Vol. 45). Pittsburgh: Institute for Software Research International .
65. Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology, 25*(2), 163–177.

66. Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology, 1*(1), 49–80.
67. Yang, C. C., Liu, N., & Sageman, M. (2006). Analyzing the terrorist social networks with visualization tools. In *ISI*. Lecture Notes in Computer Science (Vol. 3975, pp. 331–342). Berlin: Springer.
68. Freeman, L. C. (2000). Visualizing social networks. *Journal of Social Structure, 1*, 1–15.
69. Perer, A., & Shneiderman, B. (2006). Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics, 12*, 693–700.
70. Henry, N., & Fekete, J. -D. (2006). Matrixexplorer: A dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics, 12*(5), 677–684.
71. Chen, H., Zeng, D., Atabakhsh, H., Wyzga, W., & Schroeder, J. (2003). Coplink: Managing law enforcement data and knowledge. *Communications of the ACM, 46*(1), 28–34.
72. Wright, W., Schroh, D., Proulx, P., Skaburskis, A., & Cort, B. (2006). The sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '06* (pp. 801–810), New York, NY, USA. ACM.
73. Pioch, N. J., & Everett, J. O. (2006). Polestar: Collaborative knowledge management and sensemaking tools for intelligence analysts. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (pp. 513–521). ACM.

Chapter 11
Modeling Human Conflict and Terrorism Across Geographic Scales

Neil F. Johnson, Elvira Maria Restrepo, and Daniela E. Johnson

Abstract We discuss the nature and origin of patterns emerging in the timing and severity of violent events within human conflicts and global terrorism. The underlying data are drawn from across geographical scales from municipalities up to entire continents, with great diversity in terms of terrain, underlying cause, socioeconomic and political setting, cultural and technological background. The data sources are equally diverse, being drawn from all available sources including non-government organizations, academia, and official government records. Despite these implicit heterogeneities and the seemingly chaotic nature of human violence, the patterns that we report are remarkably robust. We argue that this ubiquity of a particular pattern reflects a common way in which groups of humans fight each other, particularly in the asymmetric setting in which one weaker but ostensibly more adaptable opponent confronts a stronger but potentially more sluggish opponent. We propose a minimal generative model which reproduces these common statistical patterns while offering a physical explanation as to their cause. We also explain why our mechanistic approach, which is inspired by non-equilibrium statistical physics, fits naturally within the framework of recent ideas within the social science literature concerning analytical sociology, as well as setting our results in the wider context of real-world and cyber-based collective violence and illicit activity.

N.F. Johnson (✉)
Department of Physics, University of Miami, Coral Gables, FL 33124, USA
e-mail: njohnson@physics.miami.edu

E.M. Restrepo
Department of Geography, University of Miami, Coral Gables, FL 33124, USA
e-mail: e.restrepo@miami.edu

D.E. Johnson
Lowell House, Harvard University, Cambridge, MA 02138, USA
e-mail: danielajohnsonrestrepo@college.harvard.edu

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7_11

11.1 Introduction

Irrespective of its origin, any given conflict or terrorist campaign will play out as a highly complex dynamical system driven by interconnected actors whose actions are driven by a wide variety of evolving information sources, myriad socioeconomic, cultural, and behavioral cues, and multiple feedback processes. Furthermore, since conflicts and campaigns have a beginning and eventually an end, they will by definition exhibit non-steady state, out-of-equilibrium dynamics. Violent conflict is of course one of humanity's oldest pursuits. However the new technologically enabled mixing of social activity in real and cyber space, together with the fueling of illicit activities by the drug trade and international crime, is blurring the boundaries between terrorism, insurgency, war, so-called organized crime, and common delinquency. In addition to the high-profile current cases of insurgency in Syria and Iraq (e.g., IS Islamic State and its variants), U.S. Secretary of State Hillary Clinton said that the violence by Drug Trafficking Organizations in Mexico may be "morphing into, or making common cause with, what we would call an insurgency" [1]. The United Nations, in its report titled "The Globalization of crime: A transnational organized crime threat assessment" [2], cites a statement by the UN Security Council in which they highlight ".. the serious threat posed in some cases by drug trafficking and transnational organized crime to international security in different regions of the world." Interrelated to the situation in Mexico is that of Colombia, where a thirty-plus year war still awaits a full resolution. Though Marxist in origin, its character has been mixed up by the narcotraffic industry, criminal gangs, mafia cartels, paramilitary groups, the presence of at least two major guerilla organizations, and widespread common delinquency driven by a variety of socioeconomic factors [3]. As such, the struggle faced by state organizations to counteradapt to ever-changing guerilla-narco-crime-cartel innovations is immense. Quoting [3], President Santos outlined new tactics to counteradapt to the guerrillas' adoption of (i) hit-and-run raids using flexible units, (ii) mixing of rebels and criminal gangs and their use of joint activities as mutual needs arise, for example so-called Bacrims which are organized criminal bands, (iii) dressing of insurgents as civilians to merge into the general population, (iv) carrying out small-scale attacks for maximum attention but little risk to themselves. These features (i)–(iv) of an insurgent Red force are not unique to Colombia—they reflect the behaviors likely to be adopted by any present or future armed group on the Red side that is fighting to survive, whether it operates in real space or in the cyberworld, or some future hybrid mix of the two [4–6]. For this reason, these properties (i)–(iv) will play a core mechanistic role in the generic model presented in this paper.

There is an entirely parallel threat which is evolving on the Internet, in terms of transnational attacks in the cyber domain from both sovereign state and non-state actors. This threat is arguably even more urgent than the real-space one, given that cyber 'weapons' (e.g., encounter-network worms or bots) can be assembled very quickly, and transported in principle at the speed of light (i.e., via communications links within fiber-optic networks). The advantage for Red (i.e., an insurgent or illicit organization) is that these cyber-logistics are much easier, quicker, and naturally more clandestine than the physical task of having to transport weapons and/or people from a point of assembly to the place of potential attack. Future predatory threats in real and/or cyberspace are likely to adapt to, and exploit, the rapid, ongoing advances in global connectivity, and hence present clear but evolving dangers to each and every nation state, corporation or legitimate organization.

Irrespective of the precise mix of real-world and cyber terrain, the resulting arms race involving adaptation-counteradaptation by present and future opposing actors (Red vs. Blue) will likely lead to rapid innovation of new predation methods. In addition, the background civilian population, referred to here as Green, cannot a priori be considered as purely passive. It then becomes a three-way struggle between Red, Blue, and Green, with the added feature that there may be many 'shades' of Red with rapidly changing internal allegiances (e.g., current situation in Syria and Iraq). Given this complexity, the possibility for rapid escalation of hybrid real-world attacks, cyber attacks, and cyber-assisted attacks therefore represents an unprecedented future risk which needs to be understood, quantified, mitigated, and controlled—or at least delayed or deflated in terms of its potential impact. But there are many questions that need addressing: How are these national and international threats likely to evolve going forward? Given their finite resources, how can state agencies and countries be best prepared to face this challenge? Are there any likely points of intervention that can be usefully exploited? Without quantitative models of such situations, solutions must be sought purely on the basis of narratives and case-studies, assuming any are available. It is clear that such narratives and case studies could play a crucial role, in particular where very few prior examples are known, or where strong socioeconomic, cultural, or behavioral factors play a key role. But as the amount of available data from such attacks increases, is there anything additional that can be said from a statistical viewpoint? Given that human conflicts and terror campaigns are examples of a highly complex dynamical system driven by interconnected issues and actors, we demonstrate in this chapter that a potentially fruitful approach lies within the framework of the statistical physics of non-equilibrium open systems. We also believe that this data-driven approach to conflict may ultimately shed light back on non-equilibrium statistical physics itself.

Our research draws on multiple disciplines, particularly the quantitative modeling approach of non-equilibrium statistical physics [7–18] and complements recent discussions in the social science literature [19–29]. Our general methodology comprises four steps: (1) Use spatiotemporal datasets with the highest available resolution combined with current narratives from the academic literature, online sources, and the broader national and international media, in order to identify systematic and anomalous behaviors in the ongoing timelines of daily, weekly, and monthly events within a given domain of human predation. (2) Quantify the resulting stylized statistical facts of these multi-component time-series and hence identify statistically significant deviations or anomalies. (3) Carry out a parallel procedure for other predation domains (e.g., provinces or countries) identifying where and when similar stylized facts emerge and, by contrast, where anomalies arise. (4) Develop a generative model of the underlying multi-actor dynamics for the domains of interest (see [30–47] for previous examples of this approach). Our rationale for seeking such patterns across vastly different conflicts is that there are likely to be generic ways in which humans ‘do’ covert group activities—just as in everyday life, both traffic and stock markets exhibit generic statistical features in cities and countries across the globe [27–29].

11.2 Context and Data

Even the simple representation in Fig. 11.1 demonstrates that at any one timestep, the complexity of the actors and their interactions can create a formidably complicated dynamical system. For studies of fatalities, the observable output $\mathbf{x}_i(t)$ can be considered a vector whose elements describe the number of fatalities for each population type (i.e., Red, Blue, Green) at place i at timestep t. More generally, the output $\mathbf{x}_i(t)$ would be a tensor, showing separately the numbers of victims killed and wounded, and the different weapon types used (e.g., improvised explosive device (IED), or suicide bomb, or rocket propelled grenade, or small arms fire). For simplicity, we will tend to refer to the ‘Red’ population as ‘insurgent,’ even though they may be a heterogeneous collection of traditional armed fighters, cyber-gangs, drug cartels, idealistic insurgents, rebels or rioters, and we refer to ‘Blue’ as the ‘coalition military’ or ‘official antiterrorist organization’ even though they

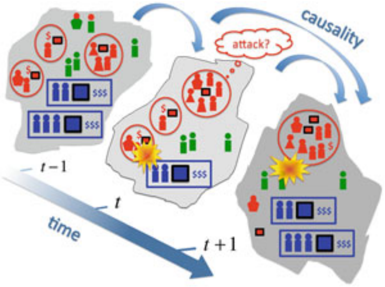


Fig. 11.1 Schematic of the complex spatiotemporal dynamics of modern multi-actor conflict in real and/or cyber space. The result is a complex ecology of interactions and observed events, driven by some dynamically evolving but hidden network of loosely connected *Red* cells featuring non-local interactions aided by electronic communications [31, 46, 47]. At any one time, there may be multiple types of actor, and these may cross different cultural and behavioral boundaries. There is empirical evidence that each population is partitioned into loose temporal cells [3–6]. Occupants of each cell may be geographically separated, but are coordinated through communications channels. Each cell may sporadically coordinate with other cells, or existing coordination within a cell (and hence the cell itself) may fragment in some way—for example, as a result of sensing danger [4–6, 15, 16, 24]. In addition to the traditional *Blue* (e.g., state military, terrorist group, or intelligence organization) and *Red* (e.g., insurgency or hacker group) actors, there is also a background civilian population which is labelled as *Green*, but which may not be passive in the struggle

may be cyber-defense, police, security forces, etc. Setting aside the issue of whether the data recorded has an observational bias or not due to the way it was recorded (e.g., main street bias [44]), there are many other potential complications facing a data-driven research program such as ours. These include, but are not limited to, the following: (1) Heterogeneity of the insurgent force strength (i.e., Red) which is depicted in Fig. 11.1 as various 'types' of fighter, or weapons, or assets including financing. This could also include different cultural, social, and behavioral types within Red. Even the assumption that there is just one Red force can be misleading, as evidenced currently in Colombia (ELN, and FARC) and in the Middle East, particularly Syria with the different rebel factions including ISIS and its variants. In short, it is not just an 'us and them' situation. (2) Heterogeneity of Blue, comprising warfighters, equipment, and money. (3) Heterogeneity of Green, the background civilian population, in terms of tribal or ethnic groups. (4) The non-passive nature of Green due to possible influence, sympathy, or direct recruitment to Red. For example in Fig. 11.1, active support of Red is indicated by two green figures with red heads who then get converted in the next timestep to Red. Or it could simply be that a Green member shows an active failure to support Blue. (5) Changing number of Red members, or Red cells. (6) Finite lifetime of any given Red cell due to endogenous or exogenous factors, such as its implicit fragility in the presence of Blue or when perceiving imminent detection or capture. The grouping dynamics that occur within and between insurgent and terrorist cells, and other illicit group activities, are unlikely to be of the form seen in more open social settings. As stated by Diego Gambetta in his influential book 'Codes of the Underworld', on p. 5. , ". . . . contrary to widespread belief, criminal groups are unstable [4]. In the underworld, people have a higher rate of mobility (and mortality) than most professions." This is also supported in the case of insurgencies by accounts such as by Robb and Kenney [5, 6]. Such fragmentation under danger is also entirely consistent with observed antipredator defenses in birds and mammals [16, 24]. (7) Decisions by Red cells to attack are not made in isolation, nor are they irrespective of the past. Instead there is a complex, possibly unknowable, mix of past events which affect a given cell or its members in particular ways—just as it does in the non-violent world of collective human struggles, e.g. financial market predatory trading [29]. In addition there is the convoluted effect that current and past exogenous and endogenous events and news might have, as is also known from the predatory environment of financial markets [48]. These reactions to past and present events will also likely depend nonlinearly on social, cultural, and behavioral factors. (8) The nature of the observable events themselves: Even if they are accurately recorded, complete information will never be known precisely about who did what and why. For these reasons, the challenges facing anyone such as ourselves who wishes to analyze high-resolution spatiotemporal datasets recording the results of collective human violence, look for common stylized facts, and then finally build minimal mechanistic models, are highly nontrivial. Indeed, the fact that more detailed spatiotemporal data is now becoming available, often down to the daily scale within individual provinces or districts, means that the bar has been raised in terms of what a model needs to achieve in order to be considered consistent with the data.

Our data sources are a mix of real-time media databases, official (government and non-governmental organization) reports, and academic studies [49]. Some of our data was obtained from Uppsala Conflict Data Program. For Afghanistan, the dataset integrates data from icasualties.org with data provided by Marc Herold of the University of New Hampshire and the ITERATE terrorism database. The Iraq data also amalgamates three separate data sets for violent events in Iraq: Iraq Body Count, ITERATE, and icasualities.org. Data for the Peruvian conflict derives from the Truth and Reconciliation Committee. Sierra Leone data comes from Macartan Humphreys of Columbia University. Malcolm Sutton is the source of the data for the Northern Ireland conflict which builds on a large number of sources. For the different Departments within Colombia, the Colombian Conflict Database was kindly provided by the Conflict Analysis Resource Center (CERAC) in Bogota [50]. The American and Spanish civil war data came from the work of Ron Francisco at the University of Kansas. Comparative results for suicides, accidents, homicides, etc. are obtained from analyzing the data of Medicina Legal in Colombia, while those for sexual violence against women come from [51].

In terms of terminology regarding what to call clusters of insurgents, it is common knowledge that a small cluster of people are sometimes called a group, a team or a cell—likewise a larger cluster may also be called a group, a crowd, or even an organization. Similarly, terrorists and insurgencies are sometimes referred to as 'groups' even though this could be the entire entity (e.g., all members of the FARC and their infrastructure) or just a few members who happened to be together on a particular attack. In order to avoid a misunderstanding of what constitutes a group, a cell, and an organization, we adopt the language in which a cell is a cluster of a few Red agents (e.g., insurgents) which carries out a given attack, and organization is the entire Red outfit—even though we stress that we do not want to assign any specific organizational capabilities, or assume that Red is necessarily well organized, or following a hierarchy. Indeed, as we will show, one of the implications of our work is that the cells are loose and transient in terms of their operational activity. This is one of the reasons they are probably so hard to track, in both real and cyber space.

11.3 Theoretical Background

Theoretical attempts to model human conflict mathematically have had a long history. They tend to resemble predator–prey models which themselves are akin to chemical reactions. These models' dynamics are typically evaluated either in the form of continuous differential equations in order to obtain partially analytic results, or through computationally intensive cellular automata or individual-based models on some kind of fixed grid such as a checker-board or static spatial network [18, 52, 53]. Outside the few-particle limit, mean-field mass action equations such as Lotka-Volterra can provide a fair qualitative description of the average behavior, i.e. $d_t N_R(t) = f(N_R(t), N_B(t))$ and $d_t N_B(t) = g(N_R(t), N_B(t))$ where $N_R(t)$ and $N_B(t)$ are the Red and Blue population's strength at time t. However, such population-level descriptions of living systems do not explicitly account for the well-known phenomenon of intra-population group (e.g., cluster) formation [24], leading to intense debate concerning the best choice of functional response terms for $f(N_R(t), N_B(t))$ and $g(N_R(t), N_B(t))$ in order to partially mimic such effects. Analogous mass-action equations have been used to model the interesting non-equilibrium process of attrition (i.e., reduction in population size) as a result of competition or conflict between two predator populations in colonies of ants, chimpanzees, birds, Internet worms, commercial companies, and humans in the absence of replenishment. The term attrition just means that 'beaten' objects become inert (i.e., they stop being involved), not that they are necessarily destroyed.

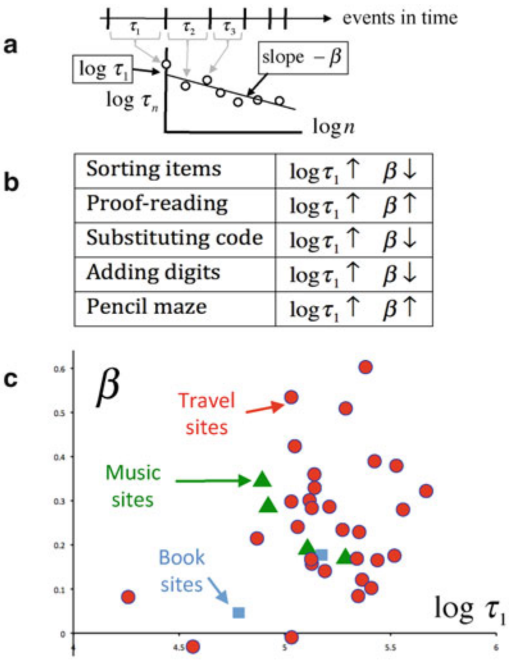
In contrast to the situation a few decades ago, however, there are a number of additional complications in present and future conflicts that challenge such prior models: First, the classic image of a battle being fought between two well-regimented armies lining up at dawn on opposite sides of a field or plain does not describe the fragmented, fluid situation of modern insurgencies [4–6], either in the real or cyber worlds. Second, broadcasting communications now exist in which events and images can be portrayed almost instantly to a broad sector of the global population, thereby possibly influencing the decisions of their elected leaders and respective security forces. Third, personal media resources such as Twitter and Facebook, together with texts and emails, mean that fighters (and potential fighters) who are separated across different streets, or towns, or countries, or continents, can be connected together within a second—and hence they can coordinate their actions such that they begin to behave as one quasi-coherent group (or 'cell'), even though they may never have met each other and may even be geographically located on separate continents. It can also happen that the members of such a cell—who may not be physically connected, but whose actions are somehow coordinated through the use of technology—suddenly lose their collective coherence (e.g., loss of communications, or loss of trust) and hence the cell has effectively fragmented. At the touch of a keystroke or press of a button on a cellphone keyboard, they instantaneously disappear into the background noise generated by everyday human activities. Fourth, the distinction between an insurgent or terrorist (i.e., Red) and the background civilian population (i.e., Green) can be blurred and itself highly fluid. It is no longer the case that a civilian population can be considered some inert background which simply soaks up the violent events as they play out.

Our approach to coping with this complexity considers the underlying ecology as interacting populations of heterogenous agents who operate with covert but dynamically evolving communication networks, and who adapt their strategies in response to external events and news, as well as counteradaptation by the relevant state authorities [30, 31, 34–36, 38]. In so doing, we incorporate the combined effects of intra-population grouping dynamics and inter-population attrition dynamics [7, 24] thereby generating an intriguing non-equilibrium many body problem. Our overall vision of the complex global interaction between gangs, cartels, illicit crime groups, etc. is therefore that of a complex ecology whose dynamics and internal interactions may change and adapt over time, with heterogeneous actors, interactions over space and time, adaptation-counteradaptation, feedback, and movement or communication via some underlying dynamical network. This view is in accordance with the state-of-the-art view of modern violent gangs proposed by Felson [54], and the descriptions of insurgencies by Kilcullen, Robb, and Kenney [4–6]. Our mechanistic methodology is also remarkably consistent with current thinking in the social sciences—in particular, analytical sociology as developed by Hedstrom [55]. In particular, Hedstrom states [55] "The basic idea of a mechanism-based explanation is quite simple: At its core, it implies that proper explanations should detail the cogs and wheels of the causal process through which the outcome to be explained was brought about. Mechanisms consist of entities (with their properties) and the activities that these entities engage in, either by themselves or in concert with other entities. These activities bring about change, and the type of change brought about depends on the properties of the entities and how the entities are organized spatially and temporally." Paraphrasing Hedstrom [55], a basic point of the mechanism perspective is that explanations that simply relate macro-properties to each other are unsatisfactory. He goes on to state that these explanations do not specify the causal mechanisms by which macro-properties are related to each other. It seems that deeper explanatory understanding requires opening up the black box and finding the causal mechanisms that have generated the macro-level observation [55, 56]. He gives the example of a car's engine whose mechanisms and parts are quite visible when the hood is opened [55]. Hedstrom also states that "when one appeals to mechanisms to make sense of statistical associations, one is referring to things that are not visible in the data, but this is different from them being unobservable in principle."

11.4 Timing of Fatal Events and a Dynamical Red Queen

We start by analyzing the timing of events in terms of a generic arms-race struggle of adaptation and counteradaptation between Red and Blue, following [30]. We consider Red (e.g., insurgents) as continually wishing to damage Blue (e.g., kill coalition military). All other things being equal, Red would like to complete successful attacks as quickly as possible so that successive successful attacks become more frequent. We therefore analyze the times for successive fatal days for Blue, finding that they follow an approximate power-law 'progress curve' $\tau_n = \tau_1 n^{-\beta}$ [30]. Here τ_n is the time between the $(n-1)$th and nth fatal day, τ_1 is the time between the first two fatal days, and β describes the subsequent escalation (or de-escalation). A fatal day is one in which Red activity produces at least one death. In particular, we calculated the best-fit power-law progress curve parameters β and τ_1 for each geographical region.

Figure 11.2 shows what one would expect if the relationship between β and τ_1 for Red-Blue events followed that of individuals—more specifically, if the dynamics of events emerging from Red-Blue dyads followed the known patterns of behavior of individuals as studied in the psychology and management literature. In such studies, an individual successfully completes a task that is repeated, just as successive Red



Sorting items	$\log\tau_1 \uparrow$	$\beta \downarrow$
Proof-reading	$\log\tau_1 \uparrow$	$\beta \uparrow$
Substituting code	$\log\tau_1 \uparrow$	$\beta \downarrow$
Adding digits	$\log\tau_1 \uparrow$	$\beta \downarrow$
Pencil maze	$\log\tau_1 \uparrow$	$\beta \uparrow$

Fig. 11.2 In one-sided everyday human activities (i.e., no *Blue* opponent to prevent task completion) there is no clear pattern in the relationship between the progress curve parameters β and τ_1 across individuals. (**a**) *Top*: fitting procedure. Schematic timeline of successive events (i.e. successive completions of task) shown as *vertical bars*. (**b**) *Middle*: existing empirical results in the literature for such tasks. Data from [57]. (**c**) *Bottom*: results for individuals searching Internet sites. Data from [58]. There is no systematic relationship between β and τ_1, in stark contrast to Figs. 11.3 and 11.4 for *Red-Blue* interacting systems. Adapted from [49]

attacks imply that Red has managed to carry out a fatal attack against Blue (i.e. Blue has not managed to stop the attack or prevent fatalities). In the psychology and organizational literature, individuals repeatedly completed tasks such as proof-reading, solving a puzzle, or purchasing something online [57, 58]. Such a task does not change over time, and is hence akin to Blue not counteradapting in any way to resist the next attack. Panel 2(b) summarizes Crossman's classic results showing that for a given type of task (e.g., proof reading), each subject exhibits his/her own β and τ_1. The lack of a generic dependence between β and τ_1 is no surprise given the heterogeneity of individual humans. Figure 11.2c shows that this lack of any linear dependence also arises for humans completing cyber tasks, specifically the navigation of different websites.

By complete contrast, Fig. 11.3 shows that for two-sided conflicts, a remarkable linear relationship emerges between β and $\log\tau_1$ for different geographical regions within each conflict. A specific example for Afghanistan is shown in more detail in Fig. 11.4, showing that the linearity extends to a specific weapon type (i.e., fatalities caused by IEDs)[30].

To explain the suitability of the progress curve $\tau_n = \tau_1 n^{-\beta}$ to describe trends in the timing of fatal events leading to Figs. 11.3 and 11.4, and in particular the observed range of β values, we have developed a dynamical version of the Red Queen evolutionary race [30] as shown schematically in Fig. 11.5. We define *R* to be the lead of the Red Queen (e.g., local insurgency) over the Blue King (e.g., coalition military) opponent, i.e. strategic advantage in an arms race. In general it could be a

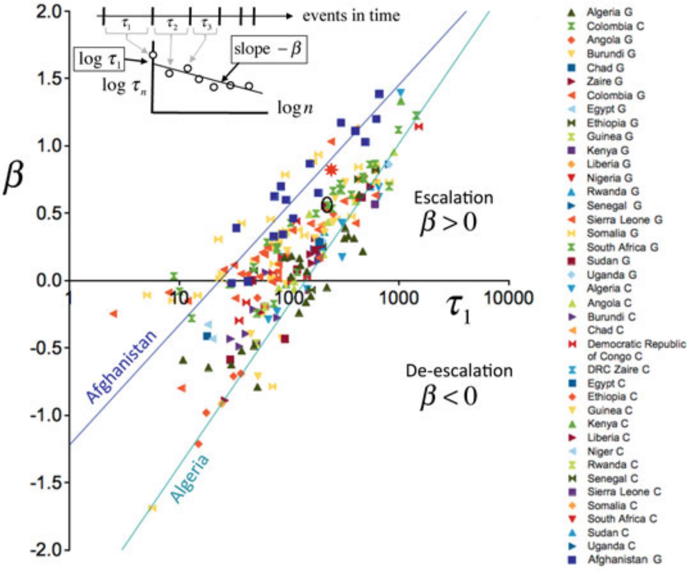


Fig. 11.3 Results from the timing of fatal events across conflicts. For a given symbol (*right panel*), each data-point shows (τ_1, β) on a semi-log plot, where these (τ_1, β) values are obtained from fitting the trend in inter-event times (*upper inset*) within a conflict. Several best-fit *lines* are shown as a guide. Separate *symbols* are used to show that results are insensitive to the precise target of the attacks: while *Blue* represents the overall society that *Red* is attacking, C counts fatal days in terms of *Red* causing civilian casualties while G counts them in terms of *Red* causing state security casualties (e.g., military casualties). *Red star* shows result for global terrorism attacks. Adapted from [49] which contains a detailed discussion of the individual data points

high-dimensional vector since strategic advantage may involve multiple factors, e.g. training, knowledge of local geography, etc. but for simplicity here we represent it as a scalar and hence will deal with a one-dimensional advantage—though we stress that the mathematical nature of random walks in multiple dimensions mean that our analysis and derivation has general validity. The traditional Red Queen story involves her running as fast as she can in order to stay at the same place. This implies that Blue instantaneously and perfectly counter-adapts to any Red advance, such that they are always neck and neck, i.e. $R = 0$ for all time. However, such instantaneous and perfect counter-adaptation is not possible in practice. Indeed, the complex adaptation-counteradaptation dynamics resulting from sporadic changes in the weaponry, skills or numbers of troops and insurgents, or changes in their experience and gathered information, or changes in local sentiment, imply that the temporal evolution of R is likely to be so complex as to appear random. This suggests that we can mimic the complex, jerky ‘walk’ that R undergoes, by a stochastic diffusion process. The key advantage is then that our statistical results do not require knowledge about the precise mechanism causing a given change in R, nor its precise value.

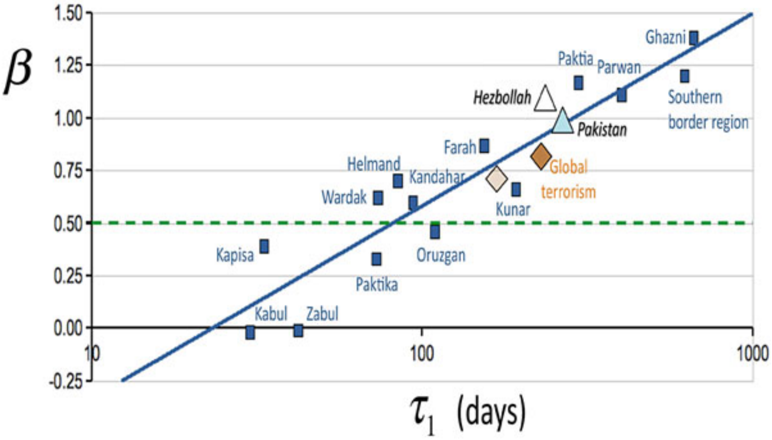


Fig. 11.4 *Solid blue line* shows best linear fit through progress-curve parameter values β and τ_1 on a semi-log plot. Results are shown for individual Afghanistan provinces (*blue squares*) for fatal attacks by insurgents (*Red*) on coalition military (*Blue*). The *green dashed line* shows value $\beta = 0.5$ which is the situation in which there are no correlations in the dynamics of R (see Fig. 11.6). Also shown are the results for global terrorist attacks (*dark diamond* is deduced from the best-fit progress curve for global terrorist group activity when averaged over all attacks while the *light diamond* is an alternative estimate where β and τ_1 are calculated directly by inserting the time intervals between initial attacks into the progress curve formula). *Blue triangle* is suicide bombings from Hezbollah suicide attacks, and the *white triangle* is for suicide attacks within Pakistan (data from *cpost.uchicago.edu/*). These results are based on a slightly smaller dataset than that used in Ref. [30], hence the plot differs slightly in detail

We now discuss the explicit case of a coin-toss stochastic process for R, though our final mathematical expressions are generic. With an outcome of Heads increasing R and Tails reducing it, R will follow a random walk. Given that R is Red's lead, and hence its instantaneous advantage over Blue, it makes sense to use R as a proxy for, and hence set it proportional to, the instantaneous rate of fatal days inflicted by Red. As R tends toward zero, or becomes negative, the time interval between subsequent fatal days diverges. Hence provinces in which R is always positive can have frequent fatal attacks by Red and therefore show up in Fig. 11.3, while provinces in which R is always negative do not. It is reasonable to expect that any significant changes in R (which may be positive or negative, large or small) will occur around days in which Red manages to inflict a fatal attack: Insurgents have by definition become successful at that moment and so this may stimulate a further increase in their strategic advantage R, while Blue's loss may stimulate an effective counter-adaptation effort and hence reduce R. Hence R is predominantly a function of n (i.e., $R(n)$). A well-known mathematical result for large n is that the typical magnitude of R after n steps is given by its root-mean-square value $|R(n)|_{\text{rms}} \sim n^{\beta}$ where $\beta = 0.5$ for any diffusion process in which the changes in $R(n)$ are independent and their distribution has finite variance, even if the changes in $R(n)$ do not have the same size. This follows from the well-known central limit theorem. In the special case that steps in $R(n)$ have the same size, this

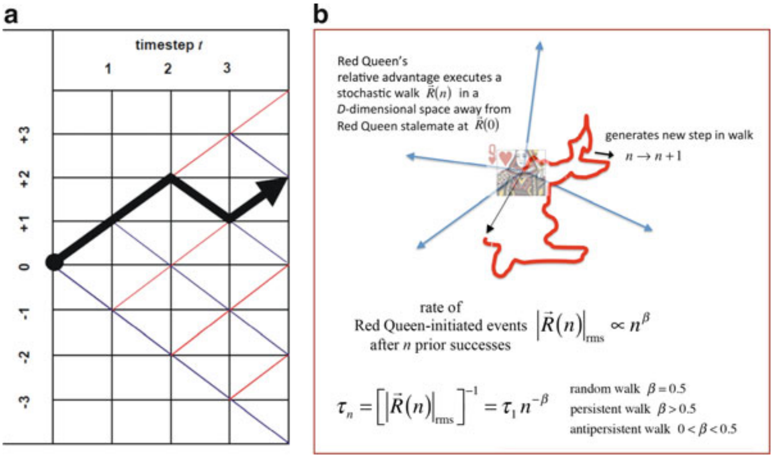

a
timestep t
1
2
3
-3
-2
-1
0
+1
+2
+3
b
Red Queen's relative advantage executes a stochastic walk $\vec{R}(n)$ in a D-dimensional space away from Red Queen stalemate at $\vec{R}(0)$
generates new step in walk
$n \to n+1$
rate of Red Queen-initiated events after n prior successes
$|\vec{R}(n)|_{\text{rms}} \propto n^{\beta}$
$\tau_n = \left[|\vec{R}(n)|_{\text{rms}}\right]^{-1} = \tau_1 n^{-\beta}$
random walk $\beta = 0.5$
persistent walk $\beta > 0.5$
antipersistent walk $0 < \beta < 0.5$

Fig. 11.5 Dynamical *Red Queen* model for the *Red-Blue* struggle. *Red* (e.g., insurgent) advantage R is represented as a vector in a multi-dimensional space whose axes may represent technological, psychological, social, cultural, or behavioral factors. R follows a stochastic walk in this D-dimensional space. Using known results from statistical physics, exact results can be obtained for β under different conditions of correlation, etc. within the walk. For the simplest case of an uncorrelated walk, $\beta = 0.5$. Adapted from [49]

result is equivalent to the statement from elementary statistics that the variance of the sum of uncorrelated variables is equal to the sum of the variances.

It is known from statistical physics that for more general stochastic walks with implicit correlations between changes in $R(n)$, Red's advantage (and hence the rate of Red attacks) will still vary as $|R(n)|_{\text{rms}} \sim n^{\beta}$ but with $\beta \neq 0.5$. Hence the time between attacks will vary as $|R(n)|_{\text{rms}}^{-1} \sim n^{-\beta}$. By definition this is τ_n, hence we have derived theoretically the observed empirical result that $\tau_n \propto n^{-\beta}$ and hence $\tau_n = \tau_1 n^{-\beta}$. Indeed for a wide range of possible correlations within $R(n)$, it is known that $0 < \beta < 1.5$ in agreement with Figs. 11.3 and 11.4. For example, if Blue's counter-adaptation is completely inadequate or absent, R will persistently increase at every step n and hence $|R(n)|_{\text{rms}} \sim n$ which means that $\beta \approx 1$. This is analogous to Red moving steadily forwards at constant velocity while Blue remains stuck at the starting line. If Red gains momentum, R may even start accelerating and hence $\beta > 1$ as observed for a few points in Figs. 11.3 and 11.4. By contrast, effective Blue counter-adaptation to each Red advance means that R stays close to zero, hence $|R(n)|_{\text{rms}}$ is of order 1 (i.e., n^0) and so $\beta \approx 0$. However it is only in the idealized—and highly unrealistic—case where Blue's counter-adaptation is instantaneous and perfect, that R will always be exactly zero. Likewise it is only if Blue proactively produces its own advances that R can become permanently negative and hence that geographical area becomes peaceful.

The unweighted linear least-squares approach that we used to fit the trend in $\log\tau_n$ versus $\log n$ for each geographic area (i.e., for each point in Figs. 11.3 and 11.4), provides an unbiased best estimate in the limit that the residuals approach statistical

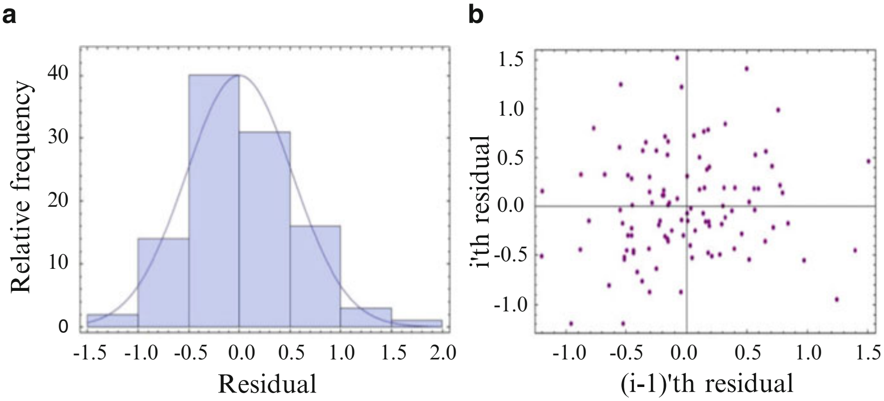


Fig. 11.6 Example of the residuals for the linear fit on the progress curve plot of $\log\tau_n$ vs. $\log n$, typical of the conflicts in Figs. 11.3 and 11.4. Case chosen is Magdalena, Colombia, shown as *black ring* in Fig. 11.3. As can be seen, the residuals are approximately Gaussian distributed (*left*) and show no serial correlation (*right*) which is consistent with the assumption that they are independent and identically distributed variables, and hence the least-square progress curve fit provides unbiased best estimates for $\log\tau_1$ and β. Adapted from [49]

independence with identical distributions (i.i.d.). This does indeed turn out to be a good approximation in our study as demonstrated in Fig. 11.6. The reason it works so well is that the error (i.e., fluctuations) in the underlying τ_n values have a crudely multiplicative form, like a failure process, such that $\tau_n = \mathbf{X}\tau_1 n^{-\beta}$ where $\mathbf{X}$ is a multiplicative noise process of the form $\mathbf{X} = \prod_m^M (1 + \epsilon_m)$ where $\{\epsilon_m\}$ are drawn from a random distribution with finite variance. Taking the logarithm of both sides, and using the well-known result that $\log(1+\epsilon_m) \approx \epsilon_m$ when $\epsilon_m \ll 1$ yields a scatter of points around the line $\log\tau_n$ vs. $\log n$ with residuals that are sums of $\{\epsilon_m\}$. Hence the distribution of the residuals should become Gaussian with no serial correlations, consistent with i.i.d. variables. This in turn suggests that each fatal Red attack can be seen as a failure process in which a set of M processes need to go 'wrong' in order that Red can create its next fatal attack.

Using this theory, we can therefore interpret and compare the entire spectrum of observed β values for different provinces, and also different terrorism domains, in an intuitive and unified way using language concerning the relative advantage between Red and Blue. Most importantly, this broad-brush Red Queen-Blue King theory does not require knowledge of specific adaptation or counter-adaptation mechanisms, and hence bypasses issues such as changes in insurgent membership (i.e., composition, numbers or numbers of cells), technology, learning or skill-set, as well as removing any need to know the hearts and minds of local residents. We also find that a similar picture arises in other situations where an arms-race struggle is underway—for example, for suicide bombings in individual provinces in Pakistan. In all these cases, we stress that a change in Red's lead R might result from a conscious or unconscious adaptation by Red, or by Blue, or both—for example, there may be an increase in Red numbers because of a conscious recruitment campaign or simply due to bad press involving Blue's activity. Likewise *R* may change due to a surge in Blue's numbers or strength, or a change in its tactics or defenses. It does not matter: The precise cause for changes in *R* does not affect the validity of our theory. The fact that the relationships in Figs. 11.3 and 11.4 are linear, suggests an intriguing coupling between the way in which Red and Blue are fighting in each region. If the dynamics were identical within each separate geographic area in a given conflict, all the corresponding (β, τ_1) points would lie on top of each other in Fig. 11.4 (and Fig. 11.3); if they were completely independent, they could in principle lie scattered anywhere in the plane. However the fact that they follow a linear relationship suggests the existence of a weak coupling between them. The origin of such a coupling awaits a future detailed explanation and represents a challenge to existing narratives concerning conflict across different geographic areas.

11.5 Severity of Events and Group Dynamics

Looking across our datasets for different conflicts, we find no evidence for a strong systematic correlation between the timing of fatal events and their severity, which is consistent with reports from other researchers [12]. This lack of correlation provides an important simplification since it enables us to analyze the timing of fatal events separately from their severity. In particular, we find that the event severity distribution is essentially stationary throughout the main portion of each conflict, while the timing of individual events is a non-stationary process with periods of initial escalation or de-escalation as discussed in Section 1.4. We therefore aggregate all events across the main portion of each conflict, checking that the choice of window does not affect our conclusions. Given the ubiquity of power-law forms in other complex systems involving human collective activity, we focus on analyzing the extent to which power-laws provide a good fit to the tail of the severity distribution.

Figure 11.7 summarizes our findings from applying a state-of-the-art maximum likelihood fitting procedure [59] for a power-law $s^{-\alpha}$ to the tail in the distribution of the severity of individual events within a given conflict, across geographic scales ranging from individual departments within a country, to individual countries within a continent, to conflicts across the globe including global terrorism. Figure 11.7b inset illustrates this power-law tail distribution; s is the severity of an individual event which, in the case of violent conflict, is the number killed or injured in an attack; α is the power-law exponent; M is the normalizing factor; p is the goodness-of-fit. It can be seen that most severity distributions approximate to a power law and have a corresponding power-law exponent around 2.5.

Our explanatory model is shown schematically in Fig. 11.8. The most basic version is solved explicitly in the Appendix using a mean-field approach, yielding a steady-state Red cell-size distribution with an approximate power-law tail of the form $n_x \sim x^{-2.5}$ where n_x is the number of cells of strength (size) x. Figure 11.9

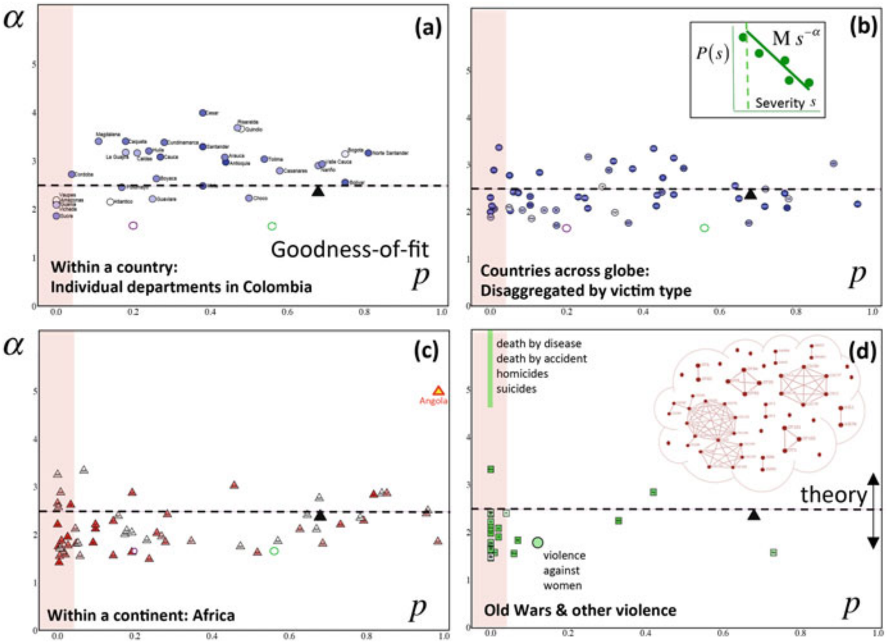


Fig. 11.7 Pattern in distribution of severity per event, shown for multiple conflicts across geographic scales with each datapoint showing the best-fit values for the power-law tail (see inset in (b)). (**a**) shows results for conflict in different spatial regions within a given country (departments in Colombia). (**b**) shows results for high-profile modern conflicts across the globe, including Iraq. (**c**) shows results for countries across a given continent (Africa). (**d**) shows results for conventional wars and other forms of human violence as a comparison. *Inset* shows Red operational network for PIRA in South Armagh, obtained from empirical analysis of available data [49]. Theoretical value of 2.5, shown by *dashed horizontal line*, emerges from a simple version of our theory (see Appendix and Fig. 11.9). *Green ring* is value for entire Africa database. *Black triangle* shows value for global terrorism attacks. *Purple ring* shows value for all interstate wars from 1860–1980. A goodness-of-fit less than 0.05, meaning that it is unlikely that the data have a power-law tail, is shown as a *red shaded area*. Results confirm that one-sided struggles such as suicides and natural deaths do not show the pattern of a power-law tail. The darker the color of each data-point, the larger the total number of victims. Adapted from [49]. Details for each datapoint are given online in the Supplementary Information of [49]

shows that this theoretical result originating from our generative model in Fig. 11.8, is remarkably robust to model variants which relax various assumptions and add additional features to more closely mimic the real world. Also, the network dynamics that it produces are consistent with the most recent and detailed fieldwork available of a Red group: PIRA (the Provisional IRA) who inflicted attacks against the stronger British government forces (Blue) in Northern Ireland from 1969 onwards. A snapshot of the network is shown in Fig. 11.7b. The *coalescence* process in the model mimics the situation in which two cells (or individuals in these cells) initiate a communications link between them of arbitrary range (for example, a

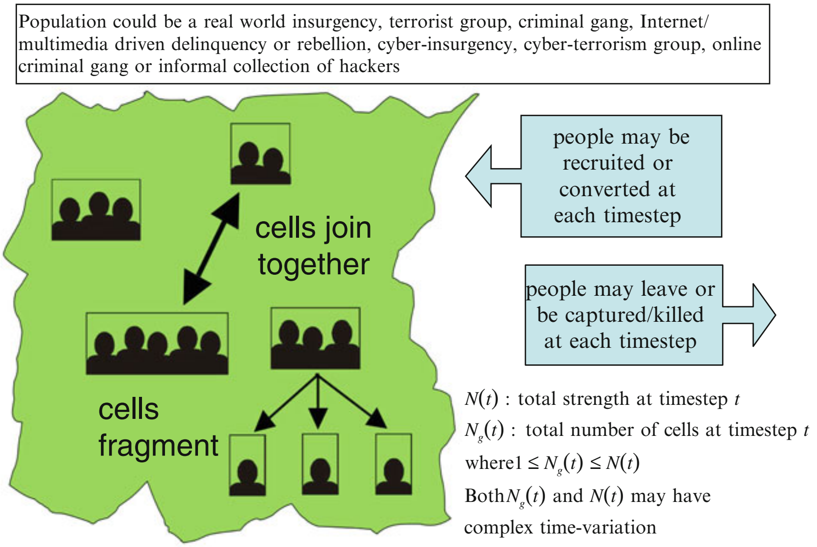
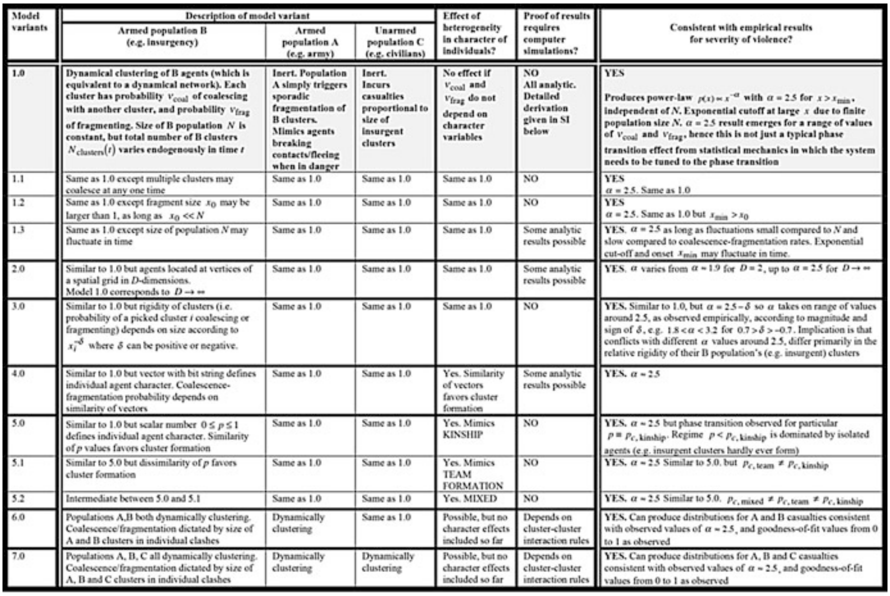

Population could be a real world insurgency, terrorist group, criminal gang, Internet/ multimedia driven delinquency or rebellion, cyber-insurgency, cyber-terrorism group, online criminal gang or informal collection of hackers
cells join together
cells fragment
people may be recruited or converted at each timestep
people may leave or be captured/killed at each timestep
$N(t)$: total strength at timestep t
$N_g(t)$: total number of cells at timestep t
where1 $\leq N_g(t) \leq N(t)$
Both$N_g(t)$ and $N(t)$ may have complex time-variation

Fig. 11.8 Schematic of dynamical grouping within *Red*. See Appendix for mathematical details of our theory and Fig. 11.9 for its generalizations. *Red* has an overall strength $N(t)$ which is distributed into dynamically evolving cells with time-varying size, number and composition. Hence cells can have a wide range of strengths at each time-step t. The total number of cells $N_g(t)$ at time t varies with time, as can the total number of composite objects (i.e., insurgent members, equipment, information) $N(t)$. Since $N_g(t)$ is the number of cells, and $N(t)$ is the total number of objects (e.g., insurgents) these two quantities are fairly independent with the only constraint being that $N_g(t) \geq 1$ (i.e., the smallest number of cells is when every object belongs to this same cell) and $N_g(t) \leq N(t)$ (i.e., the largest number of cells is when every object is isolated). In this example shown, the number of cells of a given size s at this timestep t, prior to fragmentation of the cell of size 3 into 3 cells of size 1, is $n_{s=1}(t) = 0$, $n_{s=2}(t) = 1$, $n_{s=3}(t) = 2$, $n_{s=4}(t) = 0$, $n_{s=5}(t) = 1$, $n_{s\geq 6}(t) = 0$. The total number of insurgents is $N(t) = \sum_s n_s(t) = 1 \times 2 + 2 \times 3 + 1 \times 5 = 13$. The number of cells $N_g(t) = 4$. After fragmentation, $N(t) = 13$ still, but now $N_g(t) = 6$

mobile phone call), and hence the two cells tend to coordinate their actions from then on—albeit maybe loosely. Indeed, the individual agents need not know each other, or be physically present in the same place. The long-range nature of the coupling makes it a reasonable description for physical insurgencies and crime groups using modern communications in real space, as well as cells acting in cyberspace—or any mix of the two [6]. Indeed, the language of what is a cell and what is a group, and what is crime and what is insurgency, becomes somewhat irrelevant since the mechanistic operational details are now very similar. The *fragmentation* process may arise for a number of social or situational reasons, from breakdown in trust within the cell [4] through to detection of imminent danger [6, 24]. It is well documented that groups of objects (e.g., animals, people) may suddenly scatter in all directions (i.e., complete fragmentation) when its members sense danger, simply out of fear [24] or in order to confuse a predator [24]. Or they may fragment following a clash in which the cell perceives that it is losing. As confirmed in



Model variants	Description of model variant			Effect of heterogeneity in character of individuals?	Proof of results requires computer simulations?	Consistent with empirical results for severity of violence?
	Armed population B (e.g. insurgency)	Armed population A (e.g. army)	Unarmed population C (e.g. civilians)			
1.0	**Dynamical clustering of B agents (which is equivalent to a dynamical network). Each cluster has probability v_{coal} of coalescing with another cluster, and probability v_{frag} of fragmenting. Size of B population N is constant, but total number of B clusters $N_{\mathrm{clusters}}(t)$ varies endogenously in time t**	**Inert. Population A simply triggers sporadic fragmentation of B clusters. Mimics agents breaking contacts/fleeing when in danger**	**Inert. Incurs casualties proportional to size of insurgent clusters**	**No effect if v_{coal} and v_{frag} do not depend on character variables**	**NO All analytic. Detailed derivation given in SI below**	**YES Produces power-law $p(x) \propto x^{-\alpha}$ with $\alpha = 2.5$ for $x > x_{\min}$, independent of N. Exponential cutoff at large x due to finite population size N. $\alpha = 2.5$ result emerges for a range of values of v_{coal} and v_{frag}, hence this is not just a typical phase transition effect from statistical mechanics in which the system needs to be tuned to the phase transition**
1.1	Same as 1.0 except multiple clusters may coalesce at any one time	Same as 1.0	Same as 1.0	Same as 1.0	NO	YES $\alpha = 2.5$. Same as 1.0
1.2	Same as 1.0 except fragment size x_0 may be larger than 1, as long as $x_0 << N$	Same as 1.0	Same as 1.0	Same as 1.0	NO	YES $\alpha = 2.5$. Same as 1.0 but $x_{\min} > x_0$
1.3	Same as 1.0 except size of population N may fluctuate in time	Same as 1.0	Same as 1.0	Same as 1.0	Some analytic results possible	YES. $\alpha = 2.5$ as long as fluctuations small compared to N and slow compared to coalescence-fragmentation rates. Exponential cut-off and onset $x_{\min}$ may fluctuate in time.
2.0	Similar to 1.0 but agents located at vertices of a spatial grid in D-dimensions. Model 1.0 corresponds to $D \to \infty$	Same as 1.0	Same as 1.0	Same as 1.0	Some analytic results possible	YES. α varies from $\alpha \approx 1.9$ for $D = 2$, up to $\alpha = 2.5$ for $D \to \infty$
3.0	Similar to 1.0 but rigidity of clusters (i.e. probability of a picked cluster i coalescing or fragmenting) depends on size according to $x_i^{-\delta}$ where δ can be positive or negative.	Same as 1.0	Same as 1.0	Same as 1.0	NO	YES. Similar to 1.0, but $\alpha = 2.5 - \delta$ so α takes on range of values around 2.5, as observed empirically, according to magnitude and sign of δ, e.g. $1.8 < \alpha < 3.2$ for $0.7 > \delta > -0.7$. Implication is that conflicts with different α values around 2.5, differ primarily in the relative rigidity of their B population's (e.g. insurgent) clusters
4.0	Similar to 1.0 but vector with bit string defines individual agent character. Coalescence-fragmentation probability depends on similarity of vectors	Same as 1.0	Same as 1.0	Yes. Similarity of vectors favors cluster formation	Some analytic results possible	YES. $\alpha \approx 2.5$
5.0	Similar to 1.0 but scalar number $0 \leq p \leq 1$ defines individual agent character. Similarity of p values favors cluster formation	Same as 1.0	Same as 1.0	Yes. Mimics KINSHIP	NO	YES. $\alpha \approx 2.5$ but phase transition observed for particular $p = p_{c,\mathrm{kinship}}$. Regime $p < p_{c,\mathrm{kinship}}$ is dominated by isolated agents (e.g. insurgent clusters hardly ever form)
5.1	Similar to 5.0 but dissimilarity of p favors cluster formation	Same as 1.0	Same as 1.0	Yes. Mimics TEAM FORMATION	NO	YES. $\alpha \approx 2.5$ Similar to 5.0, but $p_{c,\mathrm{team}} \neq p_{c,\mathrm{kinship}}$
5.2	Intermediate between 5.0 and 5.1	Same as 1.0	Same as 1.0	Yes. MIXED	NO	YES. $\alpha \approx 2.5$ Similar to 5.0. $p_{c,\mathrm{mixed}} \neq p_{c,\mathrm{team}} \neq p_{c,\mathrm{kinship}}$
6.0	Populations A,B both dynamically clustering. Coalescence/fragmentation dictated by size of A and B clusters in individual clashes	Dynamically clustering	Same as 1.0	Possible, but no character effects included so far	Depends on cluster-cluster interaction rules	YES. Can produce distributions for A and B casualties consistent with observed values of $\alpha \approx 2.5$, and goodness-of-fit values from 0 to 1 as observed
7.0	Populations A, B, C all dynamically clustering. Coalescence/fragmentation dictated by size of A, B and C clusters in individual clashes	Dynamically clustering	Dynamically clustering	Possible, but no character effects included so far	Depends on cluster-cluster interaction rules	YES. Can produce distributions for A, B and C casualties consistent with observed values of $\alpha \approx 2.5$, and goodness-of-fit values from 0 to 1 as observed

Fig. 11.9 Effect of generalizations of our simple one-population coalescence-fragmentation model shown schematically in Fig. 11.8, which describes Red dynamics (see Appendix). Adapted from [49]

Fig. 11.9, the precise details of these mechanisms do not matter since they tend to give similar empirical distributions. Interactions are distance-independent in our model as in [9] since we are interested in systems where messages can be transmitted over arbitrary distances (e.g., modern human communications). Bird calls and chimpanzee interactions in complex tree canope structures can also mimic this setup, as may the increasingly longer-range awareness that arises in larger animal, fish, bird, and insect groups [24]. These mechanisms are consistent with observed animal anti-predator behaviors [16, 24] and also criminal gangs [4, 6, 54]. Indeed such fragile dynamical clustering makes sense within an insurgent population, just as schools of fish or animals will go through cycles of build up and then rapid dispersal when a predator approaches [4–6, 16, 24]. The coalescence-fragmentation process is also consistent with current notions of other modern insurgencies as fragmented, transient, and evolving [3, 5, 19]. We recall the phrase of Gambetta [4] "…. contrary to widespread belief, criminal groups are unstable." Further support is again provided by Kenney [6] in *From Pablo to Osama: Trafficking and Terrorist Networks, Government Bureaucracies, and Competitive Adaptation*: "To protect themselves from the police, trafficking enterprises often compartment their participants into loosely coupled networks and limit communication between nodes"; "Trafficking networks …. are light on their feet. They are smaller and organizationally flatter"; "In progressive-era New York, according to historian Alan Block, cocaine trafficking was organized by different networks of criminal entrepreneurs who formed, reformed, split, and came together again as opportunity arose and when they were able"; "loose collection of 'cells' containing relatively small number of cell workers"; "Abu Sayyaf . . operates as a decentralized network of loosely coupled groups that conduct bombings, kidnappings, assassinations, and other acts of political violence in pursuit of a common goal . . ". Kenney also highlights the close connection of traffickers to terrorists: "Al Qaeda share numerous similarities with drug-trafficking enterprises" [6]. The inset in Fig. 11.7d shows a similarly decentralized, clustered structure which is also consistent with jihadist operational networks and other covert networks, e.g. online gold farmers [49]. In both the empirical PIRA network and our model, a link simply denotes some coordinated activity, but is not necessarily related to spatial proximity or acquaintance.

Following recent empirical findings linking size to lethality [60], our model then takes a cell's strength (size) x as proportional to the severity of an event s in which it participates. We then assume that the probability that a given cell is involved in a given event is set by exogenous factors (i.e., in the right place at the right time). Hence the shape of the distribution of event severities can be mimicked by randomly picking cells and setting the severity equal to the cell strength x. As a result, the tail distributions for the event severities and the cell strength will be approximately the same. Our simple model is therefore able to reproduce the observation in Fig. 11.7 that the distribution of severities has an approximate power-law tail $s^{-\alpha}$ with $\alpha \sim 2.5$. This result is robust to many generalizations (see Fig. 11.9), including the picking of cells for events [31]. One might wonder if our coalescence-fragmentation model falls down on the basis that an approximate power-law severity distribution exists from the outset of their empirical dataset for each terrorist organization [12] and yet the coalescence-fragmentation process may need time to converge to its steady-state power-law distribution. However this is not the case. First, the $N(t)$ initial members can be coalescing and fragmenting before any violent event is undertaken—indeed, there are many examples of underground organizations and US-based militia who spend many years evolving without any noticeable violent activity. No external event may be observed, but there is still a dynamical network of groups evolving in the background. Most importantly, any such organization will undoubtedly already have several existing clusters of contacts, hence it is not the case that the distribution has to build up from all isolated agents. A nascent insurgent, criminal, or cyber group could be created effectively instantly from such an existing structure. Second, numerical simulations show that the fat-tailed distribution develops very quickly, even if we start with isolated agents. Third, it is not the case that starting from day 1 of a given organization, all fatal events are recorded in the database. An alternative candidate model proposed in [11] is simply a combination of phenomenological broad-brush factors which happen to give a power-law, but without any specific justification for yielding the observed exponent value of 2.5. Instead, the parameters of this model [11] need to be picked in order to obtain the observed power-law exponent value of 2.5. In reality, a continuum of values—including values well away from 2.5—are just as likely within that model [11]. Nor is there any quantitative evidence to support this alternate mechanism, e.g. studies of PIRA show that variations in the number of actors can be largely unrelated to variations in the lethality of the organization.

11.6 Outlook

Our modeling approach was characterized by two stages: First, our broad-brush dynamical Red Queen theory describes the timing of fatal events [30]. This theory and analysis does not depend on the precise mechanism which changes Red's lead at any one time. Second, we provide a plausible mechanistic model of Red's internal dynamics comprising dynamically evolving cells in some loose and sporadically changing structure. This model describes the severity of fatal events. The simplicity of our approach allows a range of analytic mathematical analysis to be performed for both the severity [35] and timing of fatal events [29]. Finally we comment on the comparison to cyber-gangs and street gangs. We found that when we analyzed the empirical distributions for Long Beach street gang sizes and online guild sizes for World of Warcraft [37], the empirical distributions were not power-law like. This can be explained by the fact that our data comprised the actual membership of online guilds and gangs, as well as street gangs, as opposed to the number of objects who happen to be coordinated (e.g., online, or on the street) at any one time. The latter is likely to vary rapidly and spontaneously every day as members come online or onto the street, however the underlying membership would be expected to change more slowly over timescales of months. In addition, when individuals leave a street gang or an online guild, it is unlikely that this happens because the entire gang or guild is disbanding—hence the fragmentation process in our model would be less realistic. Indeed, it is known that fission processes involving the partial dismantling of a large cell into just a few randomly chosen splinter-cells tend to generate non-power-law distributions, as observed for street gangs and online guilds [37].

Acknowledgements We are extremely grateful to the many collaborators that have made many of these results possible, including Pablo Medina, Mike Spagat, John Horgan, Paul Gill, Brian Tivnan, Pak Ming Hui, Spencer Carran, Juan Camilo Bohorquez, Roberto Zarama, Guannan Zhao, Pedro Manrique, and Hong Qi and all other co-authors on the cited papers. NFJ gratefully acknowledges a grant from the Office of Naval Research (ONR): N000141110451. The views and conclusions contained in this paper are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any of the above named organizations, to include the U.S. government.

Appendix

Here we consider the basic version of our model, stripped down to a simple form with no decision-making, and only one population—the Red insurgency. Instead of having cells fragment when interacting with Blue, or when sensing imminent danger, we simply assign a probability for them to fragment. The resulting model yields an exponentially cutoff 2.5-exponent power-law for the distribution of cell sizes. We note that generalizations of this model have appeared in the literature—in particular, [35] contains a number of relevant generalizations, including a variable number of agents in time $N(t)$. A later paper [61] reached similar conclusions to our earlier publication [35] concerning the remarkable robustness of the 2.5 exponent to variations in the model mechanisms. Analysis of a simple version of this model was completed earlier by d'Hulst and Rodgers [8], and real-world applications have focused on financial markets—however the derivation below features general values ν_{frag} and ν_{coal}.

At each timestep, the internal coherence of a Red population of N entities (which we refer to as an 'agents' to acknowledge application to human and/or cyber systems) comprises a heterogenous soup of cells. Within each cell, the component entities have a strong intra-cell coherence. Between cells, the inter-cell coherence is weak. An agent i is then picked at random—or equivalently, a cell is randomly selected with probability proportional to size. Let s_i be the size of the cell to which this agent belongs. With probability ν_{frag}, the coherence of a given cell fragments completely into s_i cells of size one. If it doesn't fragment, a second cell is randomly selected with probability again proportional to size—or equivalently, another agent j is picked at random. With probability ν_{coal}, the two cells then coalesce (or develop a common 'coherence' in terms of their thinking or activities). As discussed in the main text, Kenney provides a wealth of case-study support for thinking of an insurgency as a loose soup of fragile cells [6], as do Gambetta [4] and Robb [5].

The Master Equations are as follows: The equation for the number of cells (i.e., clusters) of strength (i.e., size) s for $s \geq 2$ and $s = 1$ are, respectively:

$$\frac{\partial n_s}{\partial t} = \frac{\nu_{\text{coal}}}{N^2} \sum_{k=1}^{s-1} k n_k (s-k) n_{s-k} - \frac{\nu_{\text{frag}} s n_s}{N} - \frac{2\nu_{\text{coal}} s n_s}{N^2} \sum_{k=1}^{\infty} k n_k \ , \tag{11.1}$$

$$\frac{\partial n_1}{\partial t} = \frac{\nu_{\text{frag}}}{N} \sum_{k=2}^{\infty} k^2 n_k - \frac{2\nu_{\text{coal}} n_1}{N^2} \sum_{k=1}^{\infty} k n_k . \tag{11.2}$$

Here ν_{coal} and ν_{frag} are the probabilities per timestep (i.e., rates) of coalescence of two cells, or fragmentation of a cell, respectively. To simplify the limits of the sums, we extend the upper limit to infinity, which is a good approximation for large N. Terms on the right-hand side of Eq. (11.1) represent all the ways in which n_s can change. In the steady state:

$$s n_s = \frac{\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})N} \sum_{k=1}^{s-1} k n_k (s-k) n_{s-k} \ , \quad s \geq 2 \ , \tag{11.3}$$

$$n_1 = \frac{\nu_{\text{frag}}}{2\nu_{\text{coal}}} \sum_{k=2}^{\infty} k^2 n_k \ . \tag{11.4}$$

Consider

$$G[y] = \sum_{k=0}^{\infty} kn_k y^k = n_1 y + \sum_{k=2}^{\infty} kn_k y^k \equiv n_1 y + g[y] \, , \tag{11.5}$$

where y is a parameter and g[y] governs the cell size distribution n_k for $k \geq 2$. Multiplying Eq. (11.3) by y^s and then summing over s from 2 to ∞, yields:

$$g[y] = \frac{\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})N} G[y]^2 \, , \tag{11.6}$$

i.e.

$$g[y]^2 - \left(\frac{\nu_{\text{frag}} - 2\nu_{\text{coal}}}{\nu_{\text{coal}}} N - 2n_1 y \right) g[y] + n_1^2 y^2 = 0 \, . \tag{11.7}$$

From Eq. (11.5), $g[1] = G[1] - n_1$. Substituting this into Eq. (11.7) and setting $y = 1$, we solve for $g[1]$

$$g[1] = \frac{\nu_{\text{coal}}}{\nu_{\text{frag}} + 2\nu_{\text{coal}}} N \, . \tag{11.8}$$

Hence

$$n_1 = N - g[1] = \frac{\nu_{\text{frag}} + \nu_{\text{coal}}}{\nu_{\text{frag}} + 2\nu_{\text{coal}}} N \, . \tag{11.9}$$

Substituting this into Eq. (11.7) yields

$$g[y]^2 - \left(\frac{\nu_{\text{frag}} + 2\nu_{\text{coal}}}{\nu_{\text{coal}}} N - \frac{2N(\nu_{\text{frag}} + \nu_{\text{coal}})}{\nu_{\text{frag}} + 2\nu_{\text{coal}}} y \right) g[y] + \frac{(N(\nu_{\text{frag}} + \nu_{\text{coal}}))^2}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^2} y^2 = 0 \, . \tag{11.10}$$

We then solve this quadratic for $g[y]$

$$g[y] = \frac{(\nu_{\text{frag}} + 2\nu_{\text{coal}})N}{4\nu_{\text{coal}}} \left(2 - \frac{4(\nu_{\text{frag}} + \nu_{\text{coal}})\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^2} y - 2\sqrt{1 - \frac{4(\nu_{\text{frag}} + \nu_{\text{coal}})\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{frag}})^2} y} \right) , \tag{11.11}$$

which can be easily expanded

$$g[y] = \frac{(\nu_{\text{frag}} + 2\nu_{\text{coal}})N}{2\nu_{\text{coal}}} \sum_{k=2}^{\infty} \frac{(2k-3)!!}{(2k)!!} \left(\frac{4(\nu_{\text{frag}} + \nu_{\text{coal}})\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^2} y \right)^k . \tag{11.12}$$

Comparing with the definition of $g[y]$ in Eq. (11.5) shows that

$$n_s = \frac{\nu_{\text{frag}} + 2\nu_{\text{coal}}}{2\nu_{\text{coal}}} \frac{(2s-3)!!}{s(2s)!!} \left(\frac{4(\nu_{\text{frag}} + \nu_{\text{coal}})\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^2} \right)^s N . \tag{11.13}$$

We now employ Stirling's series

$$ln[s!] = \frac{1}{2} ln[2\pi] + \left(s + \frac{1}{2} \right) ln[s] - s + \frac{1}{12s} - \ldots . \tag{11.14}$$

Hence for $s \geq 2$:

$$n_s \approx \left(\frac{(\nu_{\text{frag}} + 2\nu_{\text{coal}})e^2}{2^{3/2}\sqrt{2\pi}\,\nu_{\text{coal}}} \right) \left(\frac{4(\nu_{\text{frag}} + \nu_{\text{coal}})\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^2} \right)^s \frac{(s-1)^{2s-3/2}}{s^{2s+1}} N , \tag{11.15}$$

which implies that

$$n_s \sim \left(\frac{\nu_{\text{coal}}^{s-1}(\nu_{\text{frag}} + \nu_{\text{coal}})^s}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^{2s-1}} \right) s^{-5/2} . \tag{11.16}$$

In the limit $s \gg 1$, this is formally equivalent to saying that

$$n_s \sim \exp(-s/s_0) s^{-5/2} \tag{11.17}$$

where

$$s_0 = - \left[\ln \left(\frac{4(\nu_{\text{frag}} + \nu_{\text{coal}})\nu_{\text{coal}}}{(\nu_{\text{frag}} + 2\nu_{\text{coal}})^2} \right) \right]^{-1} \tag{11.18}$$

characterizes the exponential cut-off which appears at very high s. For large cell sizes (i.e., large s such that $s \sim O(N)$) the power law behavior is masked by the exponential function. The equilibrium state for the distribution of cell sizes can therefore be considered a power-law with exponent $\alpha \sim 5/2 = 2.5$, together with an exponential cut-off. In the human context, the fact that the interactions are effectively distance-independent as far as Eq. (11.1) is concerned, captures the fact that we wish to model systems where messages can be transmitted over arbitrary distances (e.g., modern human communications). A justification for choosing a cell with a probability which is proportional to its size, is as follows: a cell with more members has more chances of initiating an event. It will also be more likely to find members of another cell more frequently, and hence be able to synchronize with them—thereby synchronizing the two cells. It is well documented that cells of living objects (e.g., animals, people) may suddenly scatter in all directions (i.e., complete fragmentation) when its members sense danger, simply out of fear or in order to confuse a predator [62]. This model also offers an explanation for Richardson's finding [17] that the distribution of approximately 10^3 gangs in Chicago, and in Manchoukuo in 1935, separately followed a truncated power-law with $\alpha \approx 2.3$.

References

1. Mexico should call in the Marines. The Washington Post, Friday 26 November 2010.
2. The Globalization of crime: A transnational organized crime threat assessment. United Nations Office on Drugs and Crime. United Nations publication No. E.10.IV.6 (2010). ISBN:978-92-1-130295-0.
3. BBC News, 8 August 2011. www.bbc.co.uk/news/world-latin-america-14441241.
4. Gambetta, D. (2009). *Codes of the Underworld: How criminals communicate*. Princeton: Princeton University Press.
5. Robb, J. (2007). *Brave new war: The next stage of terrorism and the end of globalization*. New York: Wiley.
6. Kenney, M. (2007). *From Pablo to Osama: Trafficking and terrorist networks, government bureaucracies, and competitive adaptation*. Philadelphia: Pennsylvania State University Press.
7. Ispolatov, I., Krapivsky, P. L., & Redner, S. (1996). War: The dynamics of vicious civilizations. *Physical Review E, 54*, 1274.
8. D'Hulst, R., & Rodgers, G. J. (2000). Exact solution of a model for crowding and information transmission in financial markets. *International Journal of Theoretical and Applied Finance, 3*, 609.
9. Eguiluz, V., & Zimmermann, M. (2000). Transmission of information and herd behavior: An application to financial markets. *Physical Review Letters, 85*, 5659.
10. Galam, S. (2002). Minority opinion spreading in random geometry. *European Physical Journal B, 25*, 403.
11. Clauset, A., Young, M., & Gleditsch, K. S. (2007). On the frequency of severe terrorist events. *Journal of Conflict Resolution, 51*, 1.
12. Clauset, A., & Gleditsch, K. S. (2011). The developmental dynamics of terrorist organizations. http://www.arxiv.org/abs/0906.3287.
13. Galam, S., & Mauger, A. (2003). On reducing terrorism power: A hint from physics. *Physica A, 323*, 695.
14. Lim, M., Metzler, R., & Bar-Yam, Y. (2007). Global pattern formation and ethnic/cultural violence. *Science, 317*, 1540.
15. Couzin, I. D., Krause, J., Franks, N. R., & Levin, S. A. (2005). Effective leadership and decision making in animal groups on the move. *Nature, 433*, 513.
16. Couzin, I. D. (2006). Behavioral ecology: Social organization in fission–fusion societies. *Current Biology, 16*, R170.
17. Richardson, L. F. (1960). *Statistics of deadly quarrels*. Pacific Grove: Boxwood reprint edition.
18. Lanchester, F. W. (1956). Mathematics in warfare. In Newman, J. R. (Ed.), *The world of mathematics* (Vol. 4, p. 2138). New York: Simon and Schuster.
19. Kilcullen, D. (2009). *The accidental guerrilla: Fighting small wars in the midst of a big one*. Oxford: Oxford University Press.
20. Kalyvas, S. N. (2006). *The logic of violence in civil war*. Cambridge: Cambridge University Press.

21. Buhaug, H., Cederman, L. E., & Gleditsch, K. S. (2013). *Grievances and inequality in civil wars*. Cambridge: Cambridge University Press.
22. Johnson, D. D. P., & Tierney, D. (2011). The Rubicon theory of war: How the path to conflict reaches the point of no return. *International Security, 36*, 7.
23. Bohannon, J. (2011). Counting the dead in Afghanistan. *Science, 331*, 1256.
24. Caro, T. (2005). *Antipredator defenses in birds and mammals*. Chicago: University of Chicago Press.
25. Horgan, J. (2005). *Psychology of terrorism*. London: Routledge.
26. McCulloh, I. A., Carley, K. M., & Webb, M. (2007). Social Network monitoring of Al-Qaeda. *Network Science, 1*, 25.
27. Bouchaud, J. P., & Potters, M. (2004). *Theory of financial risk and derivative pricing: From statistical physics to risk management*. Cambridge: Cambridge University Press.
28. Mantegna, R. N., & Stanley, H. E. (1995). Scaling behaviour in the dynamics of an economic index. *Nature, 376*, 46.
29. Johnson, N., Jefferies, P., & Hui, P. (2003). *Financial market complexity*. Oxford: Oxford University Press.
30. Johnson, N., Carran, S., Botner, J., Fontaine, K., Laxague, N., Nuetzel, P., et al. (2011). Pattern in escalations in insurgent and terrorist activity. *Science, 333*, 81.
31. Bohorquez, J. C., Gourley, S., Dixon, A., Spagat, M., & Johnson, N. (2009). Common ecology quantifies human insurgency. *Nature, 462*, 911.
32. Johnson, N. F., Manrique, P., & Hui, P. M. (2013). Heterogeneity in conflict dynamics. *Journal of Statistical Physics*. doi:10.1007/s10955-013-0706-z.
33. For full details, see link http://www.mathematicsofwar.com. on our research group website.
34. Zhao, Z., Bohorquez, J. C., Dixon, A., & Johnson, N. F. (2009). Anomalously slow attrition times for asymmetric populations with internal group dynamics. *Physical Review Letters, 103*, 148701.
35. Ruszczycki, B., Zhao, Z., Burnett, B., & Johnson, N. F. (2009). Relating the microscopic rules in coalescence-fragmentation models to the cluster-size distribution. *European Physical Journal, 72*, 289.
36. Johnson, N. F., Ashkenazi, J., Zhao, Z., & Quiroga, L. (2011). Equivalent dynamical complexity in a many-body quantum and collective human system. *AIP Advances, 1*, 012114.
37. Johnson, N. F., Xu, C., Zhao, Z., Ducheneaut, N., Yee, N., Tita, G., et al. (2009). Human group formation in online guilds and offline gangs driven by a common team. *Physical Review E, 79*, 066117.
38. Dixon, A., Zhao, Z., Bohorquez, J. C., Denney, R., & Johnson, N. (2010). Statistical physics and modern human warfare. In L. Naldi, et al. (Eds.), *Mathematical modeling of collective behavior in socio-economic and life sciences*. Boston: Birkhauser.
39. Zhao, Z., Kirou, A., Ruszczycki, B., & Johnson, N. F. (2009). Dynamical clustering as generator of complex system dynamics. *Mathematical Models and Methods in Applied Sciences, 19*, 1539.
40. Zhao, Z., Calderon, J. P., Xu, C., Zhao, G., Fenn, D., Sornette, D., et al. (2010). Effect of social group dynamics on contagion. *Physical Review E, 81*, 056107.
41. Johnson, N. (2008). Mathematics, physics and crime. *Policing, 2*, 160.
42. Johnson, N. F. (2006). The mother (nature) of all wars: Conflict, global terrorism and complexity science. *APS News*, November 2006.
43. Johnson, N. F. (2008). Complexity in human conflict. In D. Helbing (Ed.), *Managing complexity: Insights, concepts, applications* (p. 303). Berlin: Springer.
44. Johnson, N. F., Spagat, M., Gourley, S., Onnela, J., & Reinert, G. (2008). Bias in epidemiological studies of conflict mortality. *Journal of Peace Research, 45*, 653.
45. Johnson, N. F., Smith, D. M. D., & Hui, P. M. (2006). Multi-agent complex systems and many-body physics. *Europhysics Letters, 74*, 923.
46. Johnson, N., Spagat, M., Restrepo, J., Bohorquez, J., Suarez, N., Restrepo, E., et al. (2005). From old wars to new wars and global terrorism. e-print. arXiv:physics/0506213.

47. Johnson, N. F., Spagat, M., Restrepo, J. A., Becerra, O., Bohorquez, J. C., Suarez, N., et al. (2006). Universal patterns underlying ongoing wars and terrorism. e-print. arXiv:physics/0605035.
48. McDonald, M., Suleman, O., Williams, S., Howison, S., & Johnson, N. F. (2008). Impact of unexpected events, shocking news and rumours on foreign exchange market dynamics. *Physical Review E, 77*, 046110.
49. Johnson, N. F., Medina, P., Zhao, G., Messinger, D. S., Horgan, J., Gill, P., et al. (2013). Simple mathematical law benchmarks human confrontations. *Scientific Reports, 3*, 3463.
50. Restrepo, J., Spagat, M., & Vargas, J. (2006). The severity of the Colombian conflict: Cross-country datasets versus new micro data. *Journal of Peace Research, 43*, 99.
51. Kappler, K. E., & Kaltenbrunner, A. (2012). The power laws of violence against women: Rescaling research and policies. *PLoS One, 7*, e40289.
52. Epstein, J. (1997). *Nonlinear dynamics, mathematical biology and social sciences*. Reading: Addison-Wesley.
53. MacKay, N. J. (2006). Lanchester combat models. *Mathematics Today: Bulletin of the Institute for Mathematics and Its Applications, 42*, 170.
54. Felson, M. (2007). *Crime and nature*. Thousand Oaks: Sage Publications.
55. Hedstrom, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review Sociology, 36*, 49.
56. Norkus, Z. (2005). Mechanisms as miracle makers? The rise and inconsistencies of the ‘mechanismic approach’ in social science and history. *History and Theory, 44*, 348.
57. Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics 2*, 153.
58. Johnson, E., Bellman, S., & Lohse, G. L. (2003). Cognitive lock-in and the power law of practice. *Journal of Marketing, 67*, 62.
59. Clauset, A., Shalizi, C., & Newman, M. E. J. (2007). Power-law distributions in empirical data. *SIAM Review, 51*, 661.
60. Asal, V., & Rethemeyer, R. K. (2008). The Nature of the beast: Organizational structures and the lethality of terrorist attacks. *Journal of Politics, 70*, 437.
61. Clauset, A., & Wiegel, F. W. (2010). A generalized aggregation-disintegration model for the frequency of severe terrorist attacks. *Journal of Conflict Resolution, 54*, 179.
62. Humphries, D. A., & Driver, P. M. (1970). Protean defence by prey animals. *Oecologia, 5*, 285.

Chapter 12
Event-Related Crowd Activities on Social Media

Yu-Ru Lin

Abstract Social media like Twitter has been prevalently used for observing the behaviors of a large number of people. The availability of such behavioral trace data leads to an emergent interest in studying "crowds" in many contexts related to real-world events ranging from emergencies to ceremonies. There is, however, a lack of understanding about how different questions regarding crowd activities were asked and approached. This chapter provides a lens into event-related crowd activities within the social media domain by classifying literature into three themes: (1) crowds as event sensors, (2) crowds as event predictors, and (3) crowd characterization around events. Using the classification, it can be revealed that there is a gap between understanding and harnessing crowd activities. Several theoretical and methodological questions and implications are discussed. This chapter concludes with suggested future directions toward better gaining insights from social media crowds.

12.1 Introduction

In recent years, the quantity of human behavior that is being recorded and potentially made available for research has soared—everything we do, from sending emails to spending our money, leaves *digital traces* in some database [1]. The content generated by hundreds of millions of users on social media such as Twitter, Facebook, Foursquare, and other online platforms presents a vast source of continuous data streams of human activities, offering an unprecedented opportunity to observe the behaviors of a large number of people in a variety of situations.

Recent research along this line has used the term *crowd* in many contexts related to real-world events, ranging from emergencies [2], political events [3], sports events [4], ceremonies, and local festivals [5]. In spite of the emergent interest in studying crowds on social media, there is a lack of understanding of how current studies contribute to different aspects of crowd activities. What types of questions

Y.-R. Lin (✉)
University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: yurulin@pitt.edu

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7_12

regarding social media crowd activities were asked? How were these questions approached? Hence, this chapter aims to provide a new lens into event-related crowd activities within the social media domain. With a focus on crowd activities, existing research can be classified into three themes:

(1) Crowds as event sensors: utilizing Twitter crowds' messages to detect the occurrence of an event or summarize the content of an event.
(2) Crowds as event predictors: utilizing Twitter crowds' activities or opinions to predict the outcome of an event that is expected to happen in the near future.
(3) Crowd characterization around events: understanding Twitter crowds' activities and their changes during or after an event.

As will be discussed, based on this classification, it can be revealed that there is a gap between *understanding* and *harnessing* crowd activities on social media. In this chapter, a background on crowd activities is provided (Sect. 12.2), followed by detailed discussions for each research theme (Sects. 12.3–12.5). Several theoretical and methodological questions and implications are also discussed (Sect. 12.6). This chapter concludes with suggested future directions toward better gaining insights from social media crowds.

12.2 Background

Notions of Crowd The term *crowd* has been used in various contexts. In the sociology literature, *crowd* was studied in the context of how *collective behavior* differs from *normal* social behavior [6]. Early sociologists, including Gustave Lebon [7], referred to *collective behavior* as "spontaneous social behavior directed by aroused emotion that distorts people's normal critical abilities". In this view, crowds and collective behavior are characterized by aroused emotions as a response to new and ambiguous conditions [6]. Other scholars, however, suggested that there might be no qualitative difference between collective behavior and other forms of social behavior. For example, Blumer [8] referred to a broad sense of collective behavior as "the behavior of two or more people who are acting together." Milgram and Toch [9] defined the crowd as "people in sufficiently close proximity that the fact of aggregation comes to influence behavior" without mentioning emotion. This broad definition covers the adaptive aspects of people's behavior, which allows considering collective behavior as an emergent social behavior that is "an adaptive response to new or ambiguous condition" [10].

Over the past century, theories to explain crowds and collective behavior have been developed. These theories cover a spectrum of perspectives, ranging from viewing crowds as irrational people resulting from a hypnotic influence [7] to viewing crowds as rational beings who act based on a shared interest [11]. It is important to note that the latter perspective shifts focus from *collective behavior* to *collective action*. In the literature, *collective action* refers to politically oriented social movements where actions are purposeful and directed toward reasonable ends [6].

In this chapter, *crowd* broadly refers to "a large number of people who gather together in the same space at the same time," where the space is an online social media platform. This chapter uses the term *crowd activity* instead of *collective behavior* or *collective action* to avoid confusion with the use of these terms in sociological literature.

Studying Crowds on Social Media The sociological studies of collective behavior and collective action have relied on information gathered from historical material, surveys, and official statistics (e.g., riot or burglary reports from the bail agency), experiments, participant observation, and computer simulation [6]. In the last decade, the increasing use of social media in people's everyday social lives has offered researchers an unprecedented opportunity to examine the behavior of a large number of people. Twitter in particular has emerged as a powerful channel for communication during political and social protests; examples include the 2011 Arab Spring protests [12, 13] and the Occupy movement [14, 15]. The communication data on Twitter are a sequence of short (140-character) text messages, which are easy to process and gather through Twitter APIs [16, 17]. Compared to other social media platforms that have complex privacy settings, the publicly available Twitter has been a predominant choice for researchers to collect research datasets. Recent work [18, 19] classified this ever-growing body of academic work on Twitter in terms of domains, analysis methods, data size, and ethical concerns [18]. Instead of providing a general classification scheme, this chapter focuses on crowd-related studies using Twitter data. The studies discussed below are not meant to be comprehensive, but rather representative of the types of research questions and approaches regarding crowd activities on Twitter.

12.3 Crowds as Event Sensors

In the first research theme, Twitter is primarily utilized as part of early detection systems for detecting real-world events. The occurrence of a real-world event may be planned [20] (e.g., conferences and sports events [4, 21]) or unexpected (e.g., emergencies [22]). Much effort has been devoted to detecting unexpected events such as earthquakes and public health problems. The effectiveness of these systems relies on capturing distinct behavioral signatures at the individual level associated with the emergency event under examination. For example, people tweet with specific words after experiencing an earthquake—most obviously, the use of the word "earthquake," but likely also phrases like "was that an earthquake?" [23]. When people feel sick, they are likely to express their travails to others. A sudden increase in the volume of certain terms or phrases that exceeds their usual fluctuations can be used as an indicator for detecting events. Thus, this body of work considers Twitter users as *diffuse* or *casual* crowds acting like a large set of sensors distributed at different places and their continuously producing words as sensor readings which can be analyzed and aggregated to determine the occurrence of an event. A few studies were selected as examples for different types of events.

Earthquakes Tremendous effort has been made toward utilizing Twitter's geocoded tweets for real-time earthquake detection [24–26]. Sakaki et al. [24] proposed an algorithm to monitor users' tweets for detecting earthquakes (as well as typhoons in Japan). Researchers from the U.S. Geological Survey (USGS) [25, 26] developed an earthquake detection system based on tweet-frequency time series. The detections are faster than seismographic detections, with 75 % occurring within 2 min. These studies indicate that a Twitter-based earthquake detection and characterization system is worth investigation, but limitations exist, such as heterogeneous population density and coverage and various sources of noise [27].

Public Health Twitter has also been extensively used for early detection of emerging public health problems [28–34]. Quincey and Kostkova [32] collected tweets that contained instances of the keyword "flu" in a week during the swine flu pandemic. Their study suggested that the co-presence of other words in tweets can be used by public health authorities to gather information regarding disease activity, early warning, and infectious disease outbreak. Culotta [30] reported a correlation of 0.78 of these messages with the CDC statistics. Gomide et al. [31] were able to predict the number of dengue cases by leveraging tweet content and spatiotemporal information. Signorini et al.[34] tracked time-evolving public sentiments about H1N1 or swine flu, and studied the probability of using the Twitter stream for real-time estimation of weekly influenza-like illness (ILI) statistics generated by the CDC.

Other Events For general event understanding, Becker et al. [35] used an online clustering technique to identify real-world event content on Twitter. Other work attempted to use Twitter messages to build automatic summarization of an event [36, 37], including monitoring people's feedback to an event [3, 38–40]. A special line of research focuses on sports events, aiming to generate a journalistic summary by monitoring people who tweeted about the events while watching the sports games [21, 36, 41]. Another branch of work investigates the use of Twitter during emergency events for better crisis management [42–45].

The capacity of detecting exogenous events often relies on how well the patterns of *normal* activities can be captured. Studies have revealed a regularity of patterns using large-scale Twitter data and geocoded information. For example, Dodds et al. [46–48] have observed temporal and spatial variations in individuals' happiness. Golder and Macy [49] found that the mood of Twitter users has diurnal and seasonal patterns. Grinberg et al. [50] combined Foursquare and Twitter data to extract diurnal patterns, such as eating, shopping, etc., in NYC. Hasan et al. [51] identified different activity categories in NYC using check-in tweets, which can be used to detect local events [5].

Characteristics and Challenges in Sensing Events from Twitter Crowds The signals produced by social media crowds, i.e., their messages, tend to be influenced by other happenings relevant or irrelevant to the event of interest. An early detection system based on distinctive linguistic features alone, however, would likely result in many spurious inferences that an event was occurring. Prominent news may trigger an avalanche of tweets about the event from different places. A key additional element for event detection is that the underlying phenomenon has a distinctive spatiotemporal signature. For instance, earthquakes originate at a point and spread, as a wave, outwards. A spike in the occurrence of the word "earthquake" that did not spread in this fashion would not be indicative of an earthquake. Such spatiotemporal information can be effectively incorporated to improve the accuracy of event sensing [24, 31].

Caution should be exercised when collecting data on Twitter. Twitter provides different types of APIs that allow anyone to collect large amounts of data easily; however, one may not be aware of the systematic bias in their data collection bias. The "Streaming API" [16] allows for retrieving up to 1 % sample of data with given parameters. However, when the data matching the given parameters exceed 1 %, the retrieving result may no longer represent the overall activity on Twitter [52]. For example, Morstatter et al. found that a filtering process in Twitter's Streaming API can cause a misrepresentation of top hashtags in the data [52]. Twitter's "Rest API" [17], on the other hand, allows users to query subsets of data under the rate and data limit conditions. This requires researchers to provide criteria for obtaining a sub-sample. It is thus important to understand how different sampling strategies may affect the analysis results [53, 54]. De Choudhury et al. [53] discussed how sampling based on topology and attributes leads to different analysis results in discovering information diffusion.

Identifying credible information sources to accurately detect or better summarize events poses another challenge [55–58]. Morstatter et al. [57] studied features that help differentiate tweets originating within a crisis region (i.e., the eyewitness tweets) and tweets originating outside the region, aimed at supporting first responders to understand crisis situations. Diakopoulos et al. [58] proposed methods for assessing information sources from the Twitter stream to generate better journalistic summaries.

12.4 Crowds as Event Predictors

The second set of work aims to utilize Twitter to predict the outcome of an event that is expected to happen at a certain time in the future. One such kind of event whose outcome will have a significant impact on a society is a political election, making the electoral prediction from Twitter data the central interest of this research theme. The logic behind the prediction can be analogous to *taking the pulse* of the society where the society is represented by the crowds on Twitter. By sampling the political opinions from the crowds' tweets, researchers have attempted to predict election outcomes. Thus, this body of work considers a relevant set of Twitter users as a diffuse crowd who unknowingly take part in a virtual opinion poll. Studies can also be clustered into work focusing on election outcome prediction and work concerning opinion poll extraction.

Electoral Prediction A recent article by Gayo-Avello [59] provides a review on state-of-the-art electoral prediction work using Twitter data. It covers several important research aspects ranging from data collection and analysis to performance evaluation. One of the earliest works on this subject [60] utilized the tweet related to different parties running for the German 2009 Federal election. While users' sentiments were extracted and analyzed, the authors concluded that the count of tweets mentioning a party alone reflected the election results. Although the results were later rebutted [61], this work demonstrated a way to evaluate the effectiveness of using Twitter data for electoral prediction based on an actual election outcome (in terms of MAE measure). Gayo-Avello [62] examined how different prediction methods failed to predict the 2008 US presidential election and suggested a number of issues in prior work, including the "file-drawer" effect, biased samples, and quality of sentiment analysis.

Opinion Polls The work by O'Connor et al. [63] started to investigate how Twitter data may be used to substitute for traditional polls. The authors calculated daily sentiment scores with respect to consumer confidence and political opinion. They observed high correlation between the sentiment trends and these indicators. Bollen et al. [64] show the relations of the stock market with the sentiment of the tweets. Kunegis et al. [65] show that tweets with positive arousal, i.e., exciting and intense tweets, are more likely to be retweeted, suggesting a potential source of bias for deriving measures from social media data. Relating to event outcome prediction, Ciulla et al. [66] showed that Twitter activity related to a popular TV show can be used to predict the elimination of contestants, and geolocalized data are crucial for the correct prediction.

Characteristics and Challenges in Taking the Pulse from Twitter Crowds As discussed in [59], reducing noise and improving the sentiment analysis [67] are crucial for better predictions and correlations. It is important to note that the actual election outcome depends on voters who will eventually cast votes; hence, it is critical to identify Twitter crowds that can *unbiasedly* represent the voters or relevant populations.

There are different sources of bias. First, the Twitter user base has been known to have *demographic bias*, where certain demographic groups (e.g., young, male) are over-represented. Mislove et al. [68] analyzed a sample of Twitter users in the United States in terms of their geography, gender, and race/ethnicity, and suggested that post hoc corrections based on the over- and under-representation of different groups could be applied to remove demographic bias and improve electoral predictions. The second type of bias is *self-selection bias*—people tweet on a voluntary basis, and hence most of the data are produced by politically active users. This also includes a so-called "spiral of silence" effect, which states that people are

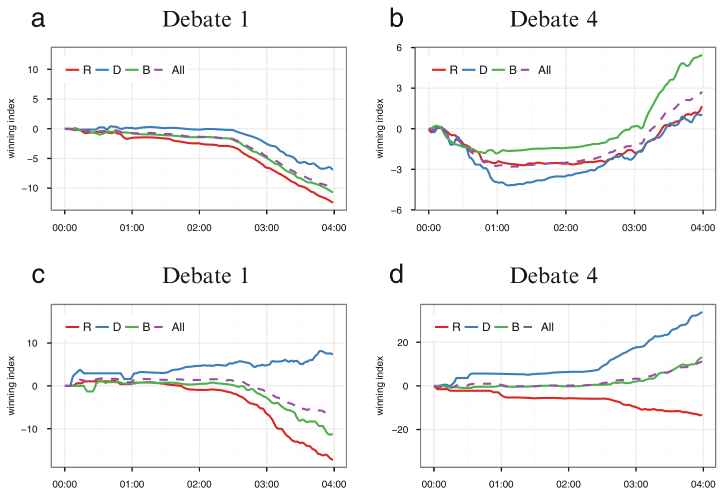


Fig. 12.1 Lin et al. [70] proposed a "computational focus group" framework to track crowds' opinion shifts during events. They developed a real-time system that utilizes prior user behaviors to detect users' biases and then aggregates users' responses with similar biases together. They tracked the behavior streams from these like-minded sub-groups called *focus groups* and presented time-dependent collective measures of their opinions with respect to the 2012 U.S. presidential election debates. These measures control for the response rate and base attitudes of the users, making shifts in opinion both easier to detect and easier to interpret. The figure shows cumulative winning indices in the first and the last presidential debates in 2012. The comparison between focus groups (**a**, **b**) and elite groups (**c**, **d**) revealed that elites tended to favor the ticket of their preferred candidate while focus groups did respond to the debate content. In each panel, the debate started at 01:00 and concluded at 02:30 UTC. Figure reprinted with permission, Copyright 2013 by ACM

more likely to express opinions when they believe others share their views [69]. To overcome the self-selection bias, Lin et al. proposed a novel "computational focus group" framework that utilizes prior user behaviors to detect users' biases and then aggregates users' responses with similar biases together (Fig. 12.1).

An additional challenge of predicting the future event outcome is that people also try to shape expectations via social media. For example, it was observed that most of the Twitter followers of Gingrich in the election of 2012 were in fact manufactured, presumably to create the illusion of enthusiasm for Gingrich's candidacy [71]. Such spoofing is an enormous challenge for making an objective prediction of the election outcomes. If a significant signal regarding the election from Twitter were gleaned, it would create an enormous incentive to manufacture signals, thus negating their signal value in the long run.

12.5 Crowd Characterization Around Events

Unlike the previous themes, where Twitter crowds are utilized either for capturing events or for predicting the event outcomes, the research in the third theme views crowd activities around events as the main subject. This not only concerns what people do and feel before, during, and after an event, but also how people's behavior and feelings change over the event period. In this sense, Twitter users are studied as an *expressive* crowd who gather on social media to cheer, comfort, or show support to each other, and to express other emotions or opinions in response to an event happening in the real world. Some studies focus on online activities corresponding to an offline social or political movement.

Behavioral Response and Change Compared to the abundant results in event detection and prediction, research focusing on characterizing behavioral change in response to external events is relatively sparse. Lin et al. [72] presented the first empirical findings about how individual and collective patterns of Twitter users' behavior changed during the 2012 U.S. presidential debates. The authors introduced a new analytical methodology—rather than relying upon samples from Twitter's APIs, it tracked the behavior across a large but defined population of Twitter users who are known to be extremely engaged in political issues. This method allows meaningful comparisons of collective behavior across distinct political events. As shown in Fig. 12.2, their results exhibited evidence of the "rising stars" dynamic as users' production of information increases without a change in concentration, but users' consumption of information becomes much more concentrated.

A particular branch of work has utilized *hashtags* for tracking content relevant to an external event. Twitter hashtags have been used by users as a label for identifying topically relevant streams of messages or a prompt for commenting and sharing [73]. Analyses of how users adopt a particular hashtag have characterized the differences between "peaky" but ephemeral topics primarily driven by exogenous events versus "persistent conversations" [74, 75]. Lin et al. [76] examined the ecology of multiple hashtags competing for attention following exogenous events (the 2012 U.S. presidential debates). The results showed how the attention of the Twitter crowd triggered by an external shock built up, lasted, and faded away and also discussed factors explaining the variation of dynamics (Fig. 12.3).

Social and Political Movements Owing to Twitter's outreach and immediacy in propagating messages, during the "Arab Spring" and other protests, activists frequently used Twitter along with hashtags to coordinate their actions and garner support [13, 77]. This makes Twitter an ideal place for collecting research data about movements, including the 2008–2009 Iranian protests [78], the 2011 Arab Spring protests [12, 79], and the Occupy movement [14, 15, 77, 80–82]. For example, Conover et al. [14] reported interesting study results based on tweets related to the Occupy Wall Street movement. They found that the geographical patterns exhibited both highly local and inter-state communications where the inter-state communications tended to carry framing languages such as "We are the 99 %."

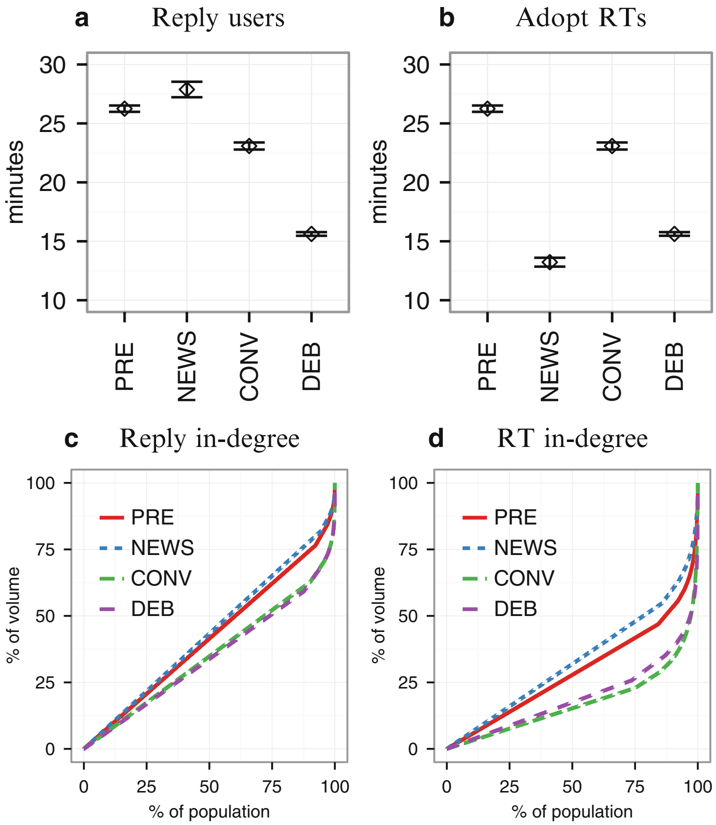


Fig. 12.2 Lin et al. [72] studied crowd activities during "media events" such as political debates that generate conditions of shared attention as many users simultaneously tune in with the dual screens of broadcast and social media to view and participate. (**a**) and (**b**) show that the behavior of users during media events (CONV and DEB) differs significantly from that of news events (NEWS) and baseline (PRE). These differences in the intensity of activity correspond to significant increases in the centralization of activity in the networks as measured by Lorenz curves in (**c**) and (**d**). The findings suggest that while users across the system become more active during media events, this additional activity reflects concentrated attention to a handful of users, hashtags, and tweets

Characteristics and Challenges in Characterizing Crowd Behavior As the core interest is in the evolution or dynamics of crowd activities around a real-world event, it is essential that the analyses capture the behavioral change of a consistent population rather than the shift of populations. The analyses in [72] suggested how to track the same relevant population to achieve a before-after comparison for assessing the change effect. Another issue is that the activity corpus may be unevenly contributed to by individuals within the crowd [72], so it is important that the behavioral statistics reflect any possible skew distribution.

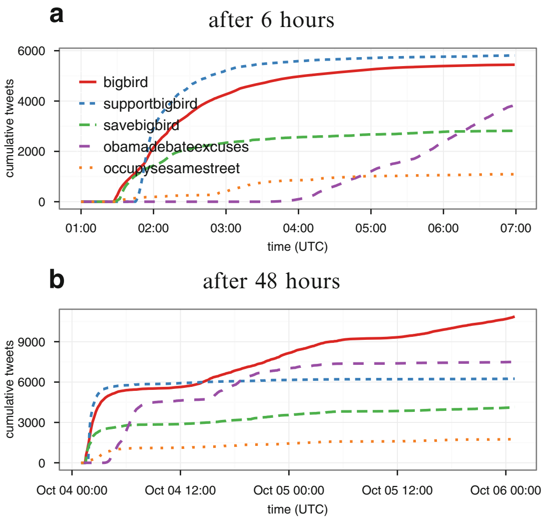


Fig. 12.3 Lin et al. [76] examined crowd response to exogenous events by studying the growth, survival, and context of hundreds of novel hashtags created during political events (the 2012 U.S. presidential debates). (**a**) and (**b**) show cumulative tweet volume of top hashtags in the first debate, over the first 6 and 48 h. Their study revealed an interesting short-term and long-term hashtag dynamics. During the first presidential debate, while the hashtag "#bigbird" was created and rapidly adopted at around 21:30 (1:30 UTC), the hashtag "#supportbigbird" which was created 15 min after took over in about 10 min. As shown in (**a**), after 6 h from the start of the debate, the adoption of "#supportbigbird" was still on top of "#bigbird." However, in a larger time scale as in (**b**), "bigbird" got to the top in the 12th h after the debate. The study further showed that retweets always contribute to faster hashtag adoption; replies extend the life of bursting hashtags while having no effect on other slowly adopted hashtags. This is the first study on the lifecycle of hashtag adoption and use in response to purely exogenous shocks. Figure reprinted with permission, Copyright 2013 by AAAI

12.6 Discussion and Future Directions

Table 12.1 summarizes the three research themes concerning event-related crowd activities on Twitter. The first theme focuses on crowds' messages and the events of interest are the research outcome. The second theme focuses on crowds' activities and opinions; the goal is to capture how an offline population (voters or the public) think or feel. In the third theme, understanding the crowds' activities is the research end. Hence, the first two themes focus on *harnessing* while the third theme focuses on *understanding* the crowd activities on social media. The analytical characteristics reflect how these different research questions lead to different research tasks and technical or analytical issues to be dealt with.

Compared to harnessing the crowds, studies in understanding the social media crowds are relatively few. As harnessing the crowds has led to exciting results, understanding the crowds can provide insights into how collective groups and

Table 12.1 Summary of research on event-related crowd activities

Theme	Research questions	Analytical characteristics
Crowds as event sensors	• How can Twitter users' messages be used to identify events of interest quickly and accurately? • How can Twitter users' messages be used to summarize events?	• Gather relevant event signals through proper tweet collection and sampling, denoising, preselected event lexicon • Detect event occurrences using temporal analysis, spatiotemporal modeling, outlier detection (over trends and regularity) • Generate event summary using sentiment analysis, topic modeling, and assessment of credible sources
Crowds as event predictors	• How can Twitter users' activities or opinions be used to predict event outcome accurately? • How can Twitter users' opinions be used to reflect public or a community's opinion?	• Gather relevant messages using preselected concept lexicon • Sample users to represent offline actors based on (collected or inferred) user attributes • Measure user opinions using sentiment analysis, temporal analysis • Generate prediction by predictive modeling with respect to outcome types (binary, multi-class, or numeric)
Crowd characterization around events	• How do Twitter users act and interact during the events of interest? • How do Twitter users' behavior change during or after the events?	• Identify or sample users participating in the events • Measure user behavior using sentiment analysis, social network analysis • Develop or test theories using temporal analysis, hypothesis testing

society function, especially under stress or emergent situations. This will contribute to better support of crowd management and policy making around planned or unexpected events. Potential directions in the future include:

Connect with Existing Crowd Theories In the sociology literature, many crowd theories, such as contagion, convergence, and emergent norm, have been developed [6, 7, 11]. The behavioral trace data available on Twitter and other social media make it possible to test existing theories empirically. However, it is important that issues in observed data are addressed properly in order to produce valid empirical study results. Experimental methods including natural experimental design, and participant matching can be used to eliminate self-selected samples and create appropriate comparison groups [83].

Explore and Develop New Hypotheses Unlike early sociologists who relied on limited-size datasets, researchers now can gather abundant information from users' behavioral traces left on social media or other digital devices. This provides new opportunities to develop a new and deeper understanding of crowd behavior in various event contexts. One of the challenges lies in how to identify interesting patterns from a huge amount of data. Exploratory and visual analytic tools that leverage data mining techniques to support exploring patterns in user behavioral streams have become a potential solution. For example, a real-time visualization that maps multiple key aspects of information through visual narratives such as "when, where, and how the information diffused during an emergency" can be useful for generating hypotheses [84].

This chapter has reviewed and discussed recent work on event-related crowd activities within the social media domain. A key observation is a gap existing between understanding and harnessing crowd activities in literature. To bridge the gap, more research attention should be directed to theory development and methodological challenges.

Acknowledgements The author would like to thank Dr. David Lazer, the Director of the Lazer Lab at Northeastern University and Dr. Drew Margolin at Cornell University for their valuable insights. Special thanks also go to Dr. Wen-Ting Chung, Ms. Alyx Cole, and the editors for their helpful feedback and comments.

References

1. Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721.
2. Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems, 26*(3), 10–14.
3. Hu, Y., Wang, F., & Kambhampati, S. (2013). Listening to the crowd: Automated analysis of events via aggregated Twitter sentiment. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence* (pp. 2640–2646). Menlo Park: AAAI Press.
4. Tang, A. & Boring, S. (2012). # EpicPlay: Crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (pp. 1569–1572). New York, NY, USA: ACM.
5. Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN '10)* (pp. 1–10). New York, NY, USA: ACM.
6. Miller, D. L. (2013, August). *Introduction to collective behavior and collective action* (3rd ed.). Long Grove: Waveland Press.
7. Le Bon, G. (1897). *The crowd: A study of the popular mind.* New York: Macmillan.
8. Blumer, H. (1951). Collective behavior. In A. M. Lee (Ed.), *New outline of the principles of sociology* (pp. 166–222). New York: Barnes & Noble.
9. Milgram, S., & Toch, H. (1969). Collective behavior: Crowds and social movements. In G. Lindzey, & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 4). Reading: Addison-Wesley.
10. Turner, R. H., & Killian, L. M. (1957). *Collective behavior.* Oxford, England: Prentice-Hall.

11. McPhail, C. (1991). *The myth of the madding crowd*. Piscataway: Transaction Publishers.
12. Hermida, A., Lewis, S. C., & Zamith, R. (2014). Sourcing the Arab Spring: A case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian revolutions. *Journal of Computer-Mediated Communication, 19*(3), 479–499.
13. Starbird, K., & Palen, L. (2012). (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)* (pp. 7–16). New York, NY, USA: ACM.
14. Conover, M. D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., & Flammini, A. (2013). The geospatial characteristics of a social movement communication network. *PLoS One, 8*(3), e55957.
15. Juris, J. S. (2012). Reflections on# Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American Ethnologist, 39*(2), 259–279.
16. Twitter. (2012). The Streaming APIs, September 2012. Accessed 22 Oct 2013.
17. Twitter. (2012). REST API v1.1 Resources, September 2012. Accessed 22 Oct 2013.
18. Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. In *Aslib Proceedings* (Vol. 66, p. 2). Bradford: Emerald Group Publishing Limited.
19. Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation, 69*(3), 384–410.
20. Becker, H., Iter, D., Naaman, M., & Gravano, L. (2012). Identifying content for planned events across social media sites. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)* (pp. 533–542). New York, NY, USA: ACM.
21. Zhao, S., Zhong, L., Wickramasuriya, J., & Vasudevan, V. (2011). Human as real-time sensors of social and physical events: A case study of Twitter and sports games. arXiv preprint. arXiv:1106.4300.
22. Dashun, W., Lin, Y.-R., & Bagrow, J. P. (2014). Learning emergencies from big data. In *Encyclopedia of social networks and mining*. Berlin: Springer.
23. Guy, M., Earle, P., Horvath, S., Turner, J., Bausch, D., & Smoczyk, D. (2014). Social media based earthquake detection and characterization. In *The 2014 KDD Workshop on Learning about Emergencies from Social Information*.
24. Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)* (pp. 851–860). New York, NY, USA: ACM.
25. Guy, M., Earle, P., Ostrum, C., Gruchalla, K., & S. Horvath. (2010). Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. *Advances in Intelligent Data Analysis IX, 6065*, 42–53.
26. Earle, P. S., Bowden, D. C., & Guy, M. (2012). Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of Geophysics, 54*(6), 8 p.
27. Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010). OMG earthquake! Can Twitter improve earthquake response? *Seismological Research Letters, 81*(2), 246–251.
28. Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1568–1576). Stroudsburg: Association for Computational Linguistics.
29. Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One, 5*(11), e14118.
30. Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)* (pp. 115–122). New York, NY, USA: ACM.
31. Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., et al. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the ACM WebSci'11* (pp. 1–8). Koblenz, Germany, 14–17 June 2011.

32. Quincey, E., & Kostkova, P. (2010). Early warning and outbreak detection using social networking websites: The potential of Twitter. In P. Kostkova (Ed.), *Electronic healthcare* (pp. 21–24). Berlin: Springer.
33. Zamite, J., Silva, F., Couto, F., & Silva, M. (2011). Medcollector: Multisource epidemic data collector. In *Transactions on large-scale data- and knowledge-centered systems, IV* (pp. 40–72). Berlin/Heidelberg: Springer.
34. Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS One, 6*(5), e19467.
35. Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. In *ICWSM* (Vol. 11, pp. 438–441).
36. Chakrabarti, D., & Punera, K. (2011). Event summarization using tweets. In *ICWSM*.
37. Tat Chua, F. C. & Asur, S. (2013). Automatic summarization of events from social media. In *ICWSM*.
38. Cheong, M., & Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers, 13*(1), 45–59.
39. Popescu, A.-M., Pennacchiotti, M., & Paranjpe, D. (2011). Extracting events and event descriptions from Twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)* (pp. 105–106). New York, NY, USA: ACM.
40. Ritter, A., Mausam., Etzioni, O., & Clark, S. (2012). Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)* (pp. 1104–1112). New York, NY, USA: ACM.
41. Nichols, J., Mahmud, J., & Drews, C. (2012). Summarizing sporting events using Twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (pp. 189–198). ACM.
42. Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management, 6*(3), 248–260.
43. Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H. W., Mitra, P., et al. (2011). Classifying text messages for the Haiti earthquake. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM2011).*
44. Li, J., & Rao, H. R. (2010). Twitter as a rapid response news service: An exploration in the context of the 2008 China earthquake. *The Electronic Journal of Information Systems in Developing Countries, 42*, 1–22.
45. Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT? In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)* (pp. 71–79). New York, NY, USA: ACM.
46. Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011, December). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One, 6*(12), e26752.
47. Frank, M. R., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2013, September). Happiness and the patterns of life: A study of geolocated tweets. *Scientific Reports, 3*, 2625.
48. Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One, 8*(5), e64417.
49. Golder, S. A., & Macy, M. W. (2011, September). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*(6051), 1878–1881.
50. Grinberg, N., Naaman, M., Shaw, B., & Lotan, G. (2013). Extracting diurnal patterns of real world activity from social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM'13).*
51. Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (UrbComp '13).* New York, NY, USA: ACM.

52. Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.
53. De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? *ICWSM* (Vol. 10, pp. 34–41).
54. Ghosh, S., Zafar, M. B., Bhattacharya, P., Sharma, N., Ganguly, N., & Gummadi, K. (2013). On sampling the wisdom of crowds: Random vs. expert sampling of the Twitter stream. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)* (pp. 1739–1744). New York, NY, USA: ACM.
55. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)* (pp. 675–684). New York, NY, USA: ACM.
56. Gupta, M., Zhao, P., & Han, J. (2012). Evaluating event credibility on Twitter. In *SDM* (pp. 153–164). Anaheim, CA, USA: SIAM.
57. Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., & Liu, H. (2014, March). Finding eyewitness tweets during crises. ACL, *2014*(2014), 23.
58. Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (pp. 2451–2460). New York, NY, USA: ACM.
59. Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review, 31*, 649–679. doi:10.1177/0894439313493979.
60. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM* (Vol. 10, pp. 178–185).
61. Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review, 30*(2), 229–234.
62. Gayo-Avello, D. (2011). Don't turn social media into another 'literary digest' poll. *Communications of the ACM, 54*(10), 121–128.
63. O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (pp. 122–129).
64. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.
65. Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference (WebSci '11)*. New York, NY, USA: ACM.
66. Ciulla, F., Mocanu, D., Baronchelli, A., Gonçalves, B., Perra, N., & Vespignani, A. (2012). Beating the news using social media: The case study of American Idol. *EPJ Data Science, 1*(1), 1–11.
67. Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology, 62*(2), 406–418.
68. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. *Fifth ICWSM* (Vol. 11).
69. Noelle-Neumann, E. (2006). The spiral of silence: A theory of public opinion. *Journal of Communication, 24*(2), 43–51.
70. Lin, Y.-R., Margolin, D., Keegan, B., & Lazer, D. (2013). Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13* (pp. 737–748), Republic & Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
71. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. *Proceedings of ICWSM*.

72. Lin,Y.-R., Keegan, B., Margolin, D., & Lazer, D. (2014). Rising tides or rising stars?: Dynamics of shared attention on Twitter during media events. *PLoS One, 9*(5), e94093.
73. Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what @you #tag: Does the dual role affect hashtag adoption? In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)* (pp. 261–270). New York, NY, USA: ACM.
74. Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)* (pp. 695–704). New York, NY, USA: ACM.
75. Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of collective attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)* (pp. 251–260). New York, NY, USA: ACM.
76. Lin, Y.-R., Margolin, D., Keegan, B., Baronchelli, A., & Lazer, D. (2013, June). #bigbirds never die: Understanding social dynamics of emergent hashtags. In *Seventh International AAAI Conference on Weblogs and Social Media*.
77. Theocharis, Y., Lowe, W., van Deth, J. W., & García-Albacete, G. (2014). Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society, 18*(2), 202–220.
78. Grossman, L. (2009, June 17). Iran protests: Twitter, the medium of the movement. *Time Magazine*.
79. Huang, C. (2011). Facebook and Twitter key to Arab Spring uprisings: Report. *The National. Abu Dhabi Media*, 6.
80. Thorson, K., Driscoll, K., Ekdale, B., Edgerly, S., Thompson, L. G., Schrock, A., Swartz, L., Vraga, E. K., & Wells, C. (2013). YouTube, Twitter and the Occupy movement: Connecting content and circulation practices. *Information, Communication & Society, 16*(3), 421–451.
81. Kenett, D. Y., Morstatter, F., Stanley, H. E., & Liu, H. (2014). Discovering social events through online attention. *PLoS One, 9*(7), e102001.
82. Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The digital evolution of Occupy Wall Street. *PLoS One, 8*(5), e64679.
83. Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge: Cambridge University Press.
84. Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., & Qu, H. (2012). Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2649–2658.

Index

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7