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.

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

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 Alessandro Vespignani
March 2015

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

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.

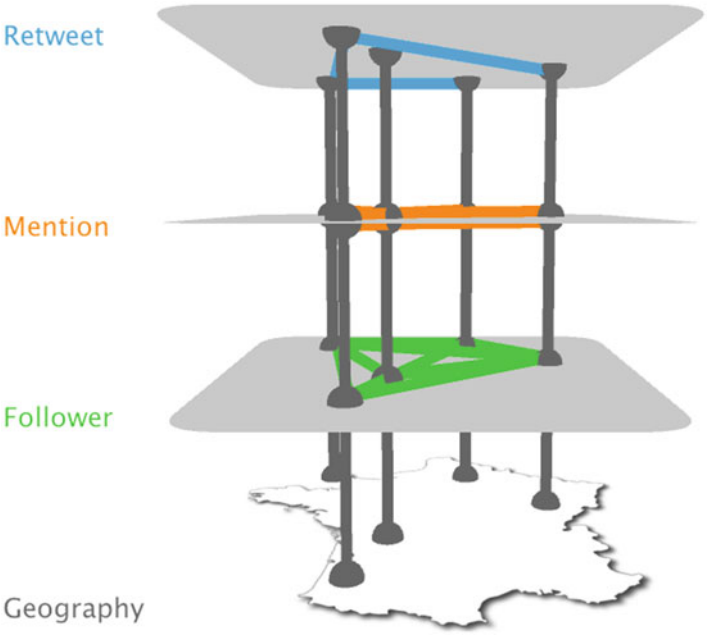

Fig. 1.1 The different layers of Twitter. As example we consider users located in some cities of France

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

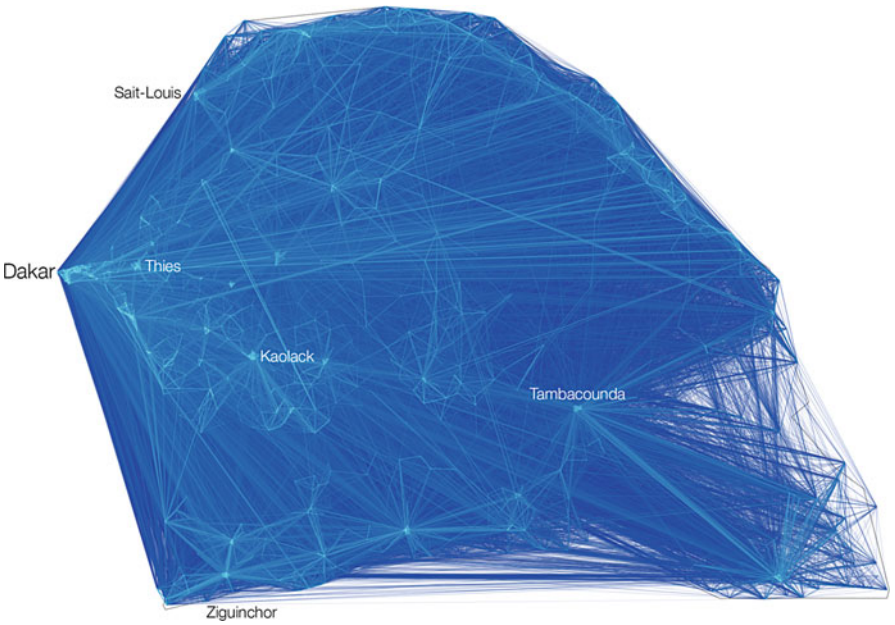

Fig. 1.2 Phone call network during one day in Senegal

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).

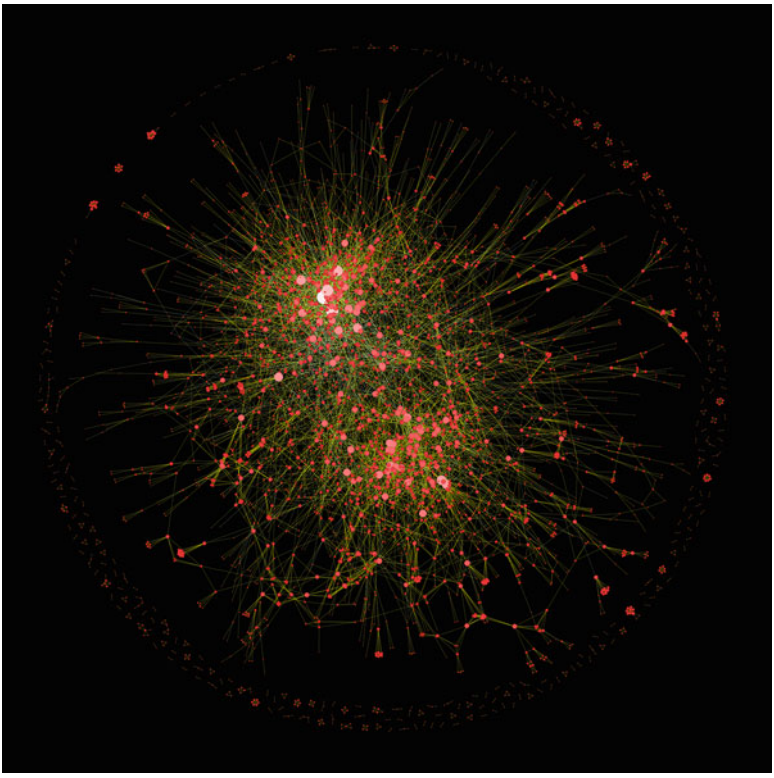

Fig. 1.3 PRL collaboration network for the 1974–2004 period, from [6]

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