BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

APPLICATIONS OF STATISTICAL PHYSICS TO THE SOCIAL AND ECONOMIC SCIENCES

by

ALEXANDER M. PETERSEN

B.Sc., University of Utah, 2003
M.S., Boston University, 2008

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2011

Approved by

First Reader

H. Eugene Stanley, Ph.D.
University Professor and Professor of Physics

Second Reader

William Skocpol Ph.D.
Professor of Physics

Acknowledgments

I would like to acknowledge my parents, Erich and Nancy, for raising me to value education and for supporting me in all my interests. I thank them for attending so many of my baseball games when I was growing up, for making me practice the piano when I wanted to do anything else, for encouraging me and enforcing me in school, and for making family fieldtrips an important part of my childhood. And mostly, I thank you for caring and praying for me, despite my success and/or failure. To Nancy and Erich do I dedicate this dissertation.

I thank my thesis advisor, Professor H. Eugene Stanley, for accepting me into his scientific family, for taking my ideas seriously, for unwavering financial support and perpetual encouragement, and for patiently showing me how to present scientific research in such calculated and efficient language. I will look fondly back on the years of knocking on your office door, and waiting in anticipation to enter and "get to work." Your work ethic and passion for pursuing scientific beauty has infected countless collaborators and former students, and I am both proud and humbled to have worked with you. I am also grateful for your kindness, and the warmth with which you conduct yourself.

I am also thankful to the Boston University Department of Physics faculty for the quality of education they have provided me. In particular, I would like to thank Professor Sid Redner and Professor Bill Klein for their excellent courses on statistical mechanics. It was through their tutelege that I learned and appreciated the fundamental beauty of statistical mechanics and the broad scope of its application to diverse problems. I also thank Sid Redner for financial support and the opportunity I had as a young graduate student to learn from such a creative and careful thinker.

I wish to thank my close collaborators, Prof. Shlomo Havlin, Prof. Boris Podobnik, Dr. Fengzhong Wang, Prof. Woo-Sung Jung, Dr. Orion Penner, Prof. Sauro Succi, Prof. Sergei Buldryev, Prof. Fabio Pammolli, Prof. Plamen Ichanov, and Prof. Sidney Redner for interesting, stimulating discussions, and the pleasure and honor of publishing with them. In particular, I thank you for providing keen and invaluable insight on whatever topic we were analyzing, for teaching me the ins and outs of preparing a manuscript, and all the while having fun while doing science. I would also like to thank Mirtha Cabello, Bob Tomposki, and Jerry Morrow for their crucial administrative help throughout the last seven years.

I must acknowledge all the friends and family who have supported me during my doctoral studies at Boston University: to my Grandparents Nana & Opa and Grandma & Papa who always encouraged me and fed me and made me feel close to home, to my brother Mark for being my little big brother who I'm so grateful for, to Andrew, Benke, Leona, Peter, and Zak for the priceless mementos courtesy of reuniting with old friends, to Orion, Chris, Sebastian and Joel for keeping me wildly entertained, to Sophia for calming me, and to all my classmates during these long years.

I would like to acknowledge my thesis committee, Professor H. Eugene Stanley, Professor William Skocpol, Prof. Plamen Ivanov, Prof. Emanuel Katz and Prof. Anatoli Polkovnikov for their patience and help with the dissertation.

APPLICATIONS OF STATISTICAL PHYSICS TO THE SOCIAL AND ECONOMIC SCIENCES

(Order No.)

ALEXANDER M. PETERSEN

Boston University Graduate School of Arts and Sciences, 2011

Major Professor: H. Eugene Stanley, University Professor and Professor of Physics

ABSTRACT

This thesis applies statistical physics concepts and methods to quantitatively analyze socioeconomic systems. For each system we combine theoretical models and empirical data analysis in order to better understand the real-world system in relation to the complex interactions between the underlying human agents. This thesis is separated into three parts: (i) response dynamics in financial markets, (ii) dynamics of career trajectories, and (iii) a stochastic opinion model with quenched disorder.

In Part I we quantify the response of U.S. markets to financial shocks, which perturb markets and trigger "herding behavior" among traders. We use concepts from earthquake physics to quantify the decay of volatility shocks *after* the "main shock." We also find, surprisingly, that we can make quantitative statements even *before* the main shock. In order to analyze market behavior before as well as after "anticipated news" we use Federal Reserve interest-rate announcements, which are regular events that are also scheduled in advance.

In Part II we analyze the statistical physics of career longevity. We construct a stochastic model for career progress which has two main ingredients: (a) random forward progress in the career and (b) random termination of the career. We incorporate the rich-get-richer (Matthew) effect into ingredient (a), meaning that it is easier to move forward in the career the farther along one is in the career. We verify the model predictions analyzing data on 400,000 scientific careers and 20,000 professional sports careers. Our model highlights the importance of early career development, showing that many careers are stunted by the relative disadvantage associated with inexperience.

In Part III we analyze a stochastic two-state spin model which represents a system of voters embedded on a network. We investigate the role in consensus formation of

"zealots", which are agents with time-independent opinion. Our main result is the unexpected finding that it is the *number* and not the *density* of zealots which determines the steady-state opinion polarization. We compare our findings with results for United States Presidential elections.

Contents

List of Tables

List of Figures

List of Abbreviations

Fed	U. S. Federal Reserve
FOMC	Federal Open Market Committee
MLB	Major League Baseball
NBA	National Basketball Association
PNAS	Proceedings of the National Academy of Science, USA
PRL	Physical Review Letters
NEJM	New England Journal of Medicine
PDF, pdf	Probability Density Function
MC	Monte Carlo
PED	Performance Enhancing Drug

Part I

Introduction

Chapter 1

Interdisciplinary Applications of Statistical Physics: Econophysics and Sociophysics

I often get two questions in succession from people who want to know, sincerely or just for the sake of conversation, what it is that I do. The first question is the commonplace, "So, what do you do?" The second follow-up question, regardless of how I answer the first is, "And that's physics?" My response to the second question is where things typically get interesting, because I'm trying to explain that the physics paradigm is to find a system, make measurements, and develop a theory that explains the empirical data along the lines of the scientific method.

My explanation typically points out that the natural systems were the first systems of real practical use by such a scientific paradigm, because it was actually more feasible to set up reliable and reproducible experiments on closed physical systems, where one can usually (cleverly) eliminate or sufficiently reduce outside influences. In contrast, however, most social or economic systems typically have a lot of background noise because it can be very difficult to effectively close such systems off from their environmental factors. Hence, much more data is needed for open experimental systems (e.g. social and economic systems), so that the fluctuations can be averaged out, and the effective "signal-to-background-noise" increased.

If the person is not from the social or natural sciences, they most likely are vexed by the picture in their mind of college introductory physics, and the torturous memories of drawing free-body diagrams for springs and blocks on inclines. These individuals are typically satisfied with my answer, and are probably wondering why they didn't explore more of the physics courses since econophysics and sociophysics sounds "fun and interesting." The other, more scientifically inclined individuals, can get a bit more defensive, claiming that there are fields that already apply the scientific method to problems in the social and economic sciences. And it is in this scenario that I am in the position of justifying interdisciplinary science, where instead of trying to outperform another field at their own approach, the idea is to combine the expertise of diverse fields in order to analyze problems using hitherto novel methods from other fields.

Interdisciplinary science is a growing field, beginning with mergers like geophysics and biophysics, which applies methods from various fields to address research topics that have traditionally been analyzed by scientists within a specific field. Methods from statistical physics are well-suited for interdisciplinary socioeconomic research due to the similarity between physical systems consisting of interacting particles and socioeconomic systems consisting of interacting agents. In this work, we analyze three socioeconomic systems with emergent complexity arising from the interactions between human agents:

(i) the macroscopic response dynamics in financial markets resulting from speculative traders,

(ii) the career longevity of employees in competitive professions, and

(iii) the Voter model in the presence of "zealots", voters with a fixed opinion, representing quenched disorder in a stochastic opinion model.

We use statistical physics concepts such as scaling analysis, master-equation evolution, and Monte-Carlo simulation to analyze these three systems. The purpose of this research is to demonstrate the utility of these methods in explaining data recorded for various complex social systems and to find statistical laws that quantify the statistical regularities we observe.

Our research is a combination of both theoretical modeling and experimental observation. A wealth of data precisely recording various facets of economic and social phenomena have become available in recent years, allowing scientists to search for statistical laws that emerge from human interactions. We utilize comprehensive data recorded for

(i) U.S. markets over the 2-year period 2001-2002, comprising the top 1,000 companies, and over 1 billion financial transactions,

(ii) Six high-impact journals over the 50-year period 1958-2008 (CELL, the New England Journal of Medicine, Nature, the Proceedings of the National Academy of Science, Physical Review Letters, and Science), comprising 350,000 articles and 600,000 scientists,

(iii) Four professional sports leagues (Major League Baseball, Korean Professional Baseball League, National Basketball Association, English Premier League), comprising over 20,000 athletes and over 200 seasons of league play.

and find several statistical laws. These empirical descriptions aid in the development of appropriate theoretical models, which can provide further insight into the various deterministic and stochastic mechanisms that give rise to real-world phenomena.

Part II

A Statistical Physics Approach to Quantifying Market Dynamics Before and After Financial Shocks

Chapter 2

Motivation

The application of statistical physics methods to economic markets began in the 1990s, mainly due to the availability of large data sets in our burgeoning era of technology and information. The first studies were focussed on determining the stochastic aspects of the time series which described the price of a fundamental asset, such as a stock or stock index, and the associated fractal or Hurst exponents that could describe the fluctuations [1]. Researchers in the 1990s were fascinated by fractal and chaotic systems, and so complex economic markets seemed to be a natural terrain to look for evidence of emergent complexity. However, it became quickly apparent that there was much more room for statistical physics to contribute to the field, which was soon termed "econophysics" by H. Eugene Stanley in the mid 1990s.

It had been known for some time that financial time series were odd creatures as a result of work by mathematicians like Mandelbrot, who had brought these observations to light in the 1960s. However, it was unclear how the odd behavior of financial market time series had anything to do with fundamental economics and/or the theoretical pillars in economics which were the central dogma of economics departments around the world. Some of the highly regarded theories are known as "the efficient market hypothesis", rational agent models, the Black-Scholes equations, optimal portfolio theory, etc. However, most of this work was based on mathematical theory that was not fully stress-tested and that did not compare the ramifications of the theory with empirical data. As a result, the initial econophysics findings were

startling to some, and exciting to others.

Since the mid-1990s, there has been an explosion in the use of physics methods applied to financial markets. Some examples are the application of random-matrix theory from nuclear physics to quantify the cross-correlations of stock fluctuations, agent-based models developed along the lines of statistical physics spin models to model collective herding behavior, path-integral and field-theory methods from quantum mechanics to price options and other financial contracts, and scaling methods from critical phenomena theory to search for cross-market universality.

In this work, we apply methods developed in geophysics to describe the dynamic response of geophysical systems to massive perturbations in which a fault line suddenly releases a large amount of stored energy, commonly known as earthquakes. In financial markets, the corresponding events are financial crashes which can be devastating, with the financial turmoil crippling not only the country of origin, but potentially the international economy. This possibility for worldwide financial crisis epidemic is almost certain with the recent globalization of financial markets. Thus, understanding the dynamic response of financial markets to financial crashes is crucial for the prevention and treatment of financial crises.

However, there is another startling analogy from earthquake physics that makes the problem in finance much more difficult than trying to just prevent "the big one." Just as in earthquake physics there is the Guttenberg-Richter law which says that the frequency of earthquakes follows a power-law, there is an analogous statistical law, the "inverse-cubic law" of financial markets that also quantifies the frequency of financial fluctuations with a power-law. The significance of the power-law formulation lies in the scale-invariance of the underlying systems, such that there is no typical energy or fluctuation scale that defines the system, and so physical shocks (and financial shocks) are predicted to occur across a very large range of energies with a significant probability of occurrence. In the case of financial shocks, not only are financial crashes predicted to occur (contrary to the standard theory of economics) but also, there should also be smaller financial crashes of every size according to the statistical "inverse-cubic" law. This is bad and good. Although it means that there are more than just a few extremely large events to expect, it also means that we can analyze the smaller more frequent events, obtaining a large sample size, and hopefully use our

understanding of the common mechanism of financial shocks to create preventative and therapeutic measures for the "big ones."

So what is a financial crash? Empirically, it corresponds to a time-period, typically ranging from a day to a week, over which there are significant market losses across the entire market. Such a downturn can result from endogenous herding or exogenous events (e.g. the 9/11 World Trade Center attack) which affect the financial outlook. During these events, there is a significant correlated response causing stock prices to plunge across the market. In econophysics, the typical quantity of interest is not necessarily the price $p(t)$, but rather the logarithmic change in the price, also called the returns, $r(t) \equiv \ln p(t) - \ln p(t - \delta t)$ over the time-horizon δt which can range in analysis from a second to a year. This quantity is considered a measure of fluctuation in the system, and in a loose-sense, describes the systems non-equilibrium dynamics.

A group of N coupled stocks comprising an index fluctuate at any given moment, sometimes independently, and sometimes collectively. As a simple null model of response dynamics, a system of harmonic oscillators under a small perturbation are governed by linear response theory, which predicts that the elements will have an exponential decay back to the "equilibrium state" in a fashion that is indistinguishable from the response to noise. This is certainly not the case in financial markets, where stocks individually and markets collectively show evidence of complex systems defined by anomalous signatures often referred to as stylized facts such as long (power-law) memory of autocorrelation and clustering of large events (cascading). These stylized facts are also known to occur in other complex systems in diverse fields, e.g. heart attacks and brain seizures (medicine), earthquakes (geophysical) and solar flares (stellar). Hence, the hope is that there is a common mechanism that can describe complex systems consisting of interacting components, and that there is a common system response to both small and large perturbations. Here we investigate the statistical regularities in the market dynamics at the 1-minute time resolution in order to better understand the cascading dynamics that occur in financial markets.

Chapter 3

Case Study: Quantifying the Market Response to Federal Reserve Announcements

3.1 Summary

We study the behavior of U.S. markets both before and after U.S. Federal Open Market Commission (FOMC) meetings, and show that the announcement of a U.S. Federal Reserve rate change causes a financial shock, where the dynamics after the announcement is described by an analogue of the Omori earthquake law. We quantify the rate $n(t)$ of aftershocks following an interest rate change at time T, and find power-law decay which scales as $n(t-T) \sim (t-T)^{-\Omega}$, with Ω positive. Surprisingly, we find that the same law describes the rate $n'(|t-T|)$ of "pre-shocks" *before* the interest rate change at time T. This is the first study to quantitatively relate the size of the market response to the news which caused the shock and to uncover the presence of quantifiable preshocks. We demonstrate that the news associated with interest rate change is responsible for causing both the anticipation before the announcement and the surprise after the announcement. We estimate the magnitude of financial news using the relative difference between the U. S. Treasury Bill and the Federal Funds

Effective rate. Our results are consistent with the "sign effect," in which "bad news" has a larger impact than "good news." Furthermore, we observe significant volatility aftershocks, confirming a "market underreaction" that lasts at least 1 trading day.

3.2 Introduction

Interest rate changes by the Federal Reserve provide a significant perturbation to financial markets, which we analyze from the perspective of statistical physics [2, 3, 4, 5, 6, 7]. The Federal Reserve board (Fed), in charge of monetary policy as the central bank of the United States, is one of the most influential financial institutions in the world. During Federal Open Market Commission (FOMC) meetings, the Fed determines whether or not to change key interest rates. These interest rates serve as a benchmark and a barometer for both American and international economies. The publicly released statements from the scheduled FOMC meetings provide grounds for widespread speculation in financial markets, often with significant consequences.

In this thesis, we show that markets respond sharply to FOMC news in a complex way reminiscent of physical earthquakes described by the Omori law [8, 9]. For financial markets, the Omori law was first observed in market crashes by Lillo and Mantegna [10], followed by a further study of Weber *et al.* [11], which found the same behavior in medium-sized aftershocks. However, the market crash is only an extreme example of information flow in financial markets. This thesis extends the Omori law observed in large financial crises to the more common Federal Reserve announcements, and suggests that large market news dissipates via power-law relaxation (Omori law) of the volatility. In addition to the standard Omori dynamics following the announcement, we also find novel Omori dynamics before the announcement.

The market dynamics following the release of FOMC news are consistent with previous studies of price-discovery in foreign exchange markets following marcroeconomic news releases [12, 13]. Furthermore, we hypothesize that the uncertainty in Fed actions, coupled with the pre-announced schedule of FOMC meetings, can increase speculation among market traders, which can lead to the observed market underreaction [14]. Market underreaction, meaning that markets take a finite time to readjust prices following news, is not consistent with the efficient market hypothesis;

Several theories have been proposed to account for this phenomenon [15].

We analyze all 66 scheduled FOMC meetings in the eight-year period 2000-2008 using daily data from *http://finance.yahoo.com*. Also, for the two-year period 2001-2002, we analyze the intraday behavior for 19 FOMC meetings using Trades And Quote (TAQ) data on the 1-minute time scale.

The section is organized as follows: In Section 3.3 we describe the FOMC meetings and the Fed interest rate relevant to our analysis. In Section 3.4.1 we analyze the response of the S&P100, the top 100 stocks (ranked by 12-month sales according to a 2002 *BusinessWeek* report) belonging to the 2002 S&P 500 index, over the 2000-2008 period using daily data. Using the relative spread between the 6-Month Treasury Bill and the Federal Funds Effective rate, we relate the speculation *prior* to the FOMC meetings to the daily market volatility, measured here as the logarithmic difference between the intraday high and low price for a given stock on the day of the announcement. In Section 3.4.2 we study high-frequency intraday TAQ data on the 1-min scale for the S&P100, and find an Omori law with positive exponent immediately following the announcement of Fed rate changes. Further, we relate the intraday market response, (quantified by both the Omori exponent and Omori amplitude), to the change in market expectations before and after the announcement.

3.3 FOMC Meetings, Fed Interest Rates and Treasury Bills

There are many economic indicators that determine the health of the U.S. economy. In turn, the health of the U.S. economy sets a global standard due to the ubiquity of both the U.S. dollar and the economic presence maintained through imports, exports, and the *Global* Market [16]. The U.S. Federal Reserve Target rate, along with the Effective "overnight" rate, set the scale for interest rates in the United States and abroad. The Target rate is determined at FOMC meetings, which are scheduled throughout the year, with detailed minutes publicly released from these meetings. The Effective rate is a "weighted average of rates on brokered trades" between the Fed and large banks and financial institutions, and is a market realization of the Target rate [17]. In Fig. 3.1 we plot the Federal interest rates over the 8-year

period 2000-2008.

Our analysis focuses on the FOMC meetings after January 2000. Historically, the methods for releasing the meeting details have varied. In the 1990s, there was a transition from a very secretive policy towards the current transparent policy [18]. Since the year 2000, the Fed has released statements detailing the views and goals of the FOMC. This increase in public information has led to an era of mass speculation in the markets, revolving mainly around key economic indicators such as the unemployment rate, the Consumer Price Index, the money supply, etc. These economic indicators also influence the FOMC in their decision to either change or maintain key interest rates. Speculation has assumed many forms and new heights, evident in the implementation of new types of derivatives based on federal securities. For instance, options and futures are available at the Chicago Board of Trade which are based on Federal Funds, Treasury Bills, and Eurodollar foreign exchange. These contracts can be used to estimate the implied probability of interest rate changes, utilizing sophisticated methods focussed on the price movement of expiring derivative contracts [19, 20, 21, 22, 23].

In the next section, we outline a simple method to measure speculation prior to a scheduled FOMC meeting using the 6-Month Treasury Bill and the Federal Funds Effective ("overnight") rate. These data are readily available and are updated frequently at the website of the Federal Reserve [17]. Because each FOMC meeting is met with speculation (in the weeks before the meeting) and anticipation (in the hours before the announcement), we identify the decision to change or not to change key interest rates as a market perturbation. The market response results from the systematic stress associated with the speculation and anticipation, which are not always in line with the FOMC decision.

3.4 Empirical Results

3.4.1 Response to FOMC Meetings on Daily Time Scale

In this section we analyze the daily activity before and after 66 *scheduled* FOMC meetings over the 8-year period 2000-2008, where scheduled meetings are publicly announced at least a year in advance [17]. We do not consider unscheduled meetings

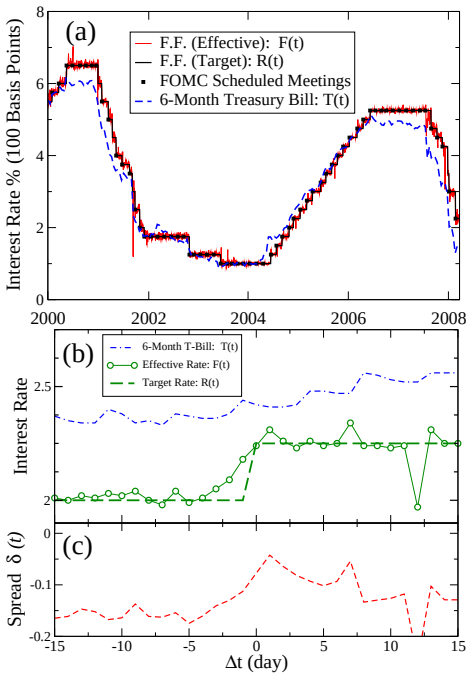

Figure 3.1: An illustration of the close relation between the Treasury Bill and the Federal Funds rate. (a) Time series of the Federal Reserve Target rate, $R(t)$, and the Federal Reserve Effective rate, $F(t)$, for Federal Funds (F.F.) dating from Jan. 2000 to Apr. 2008. The 6-Month Treasury Bill, $T(t)$, tends to anticipate the effective rate, with speculation about future changes causing deviations in the relative values. United States Treasury Bills carry little risk, and are considered to be one of the most secure investments. (b) A typical illustration of the Federal Funds Effective rate and the Treasury Bill, where both gravitate around the Federal Funds Target rate. The change in the relative spread $\delta(t)$, defined in Eq. (3.1), between the Treasury bill and the Federal Funds Effective rate, indicates changes in market speculation. (c) The relative spread, $\delta(t)$, 15 days before and 15 days after the scheduled FOMC meeting on Dec. 14, 2004, which corresponds to $\Delta t = 0$. Note that the average value of the relative spread increases after the announcement, indicating a shift in market consensus and speculation.

resulting in rate change, which contain an intrinsic element of surprise, and are historically infrequent (only 4 unexpected Target rate changes over the same period). Of primary importance, is the FOMC committee's decision to change or not change the Target rate $R(t)$ by some percent $\Delta R(t)$, where the absolute relative change $|\Delta R(t)/R(t-1)|$ has typically filled the range between 0.0 and 0.25. This section serves as an initial motivation for the intraday analysis, and will also serve as a guide in developing a metric that captures market speculation. In this section we use the intraday high-low price range to quantify the magnitude of price fluctuations. In particular, we analyze the companies belonging to the S&P 100, and also the subset of 18 banking and finance companies referred to here as the "Bank" sector.

In Fig. 3.1(a) we plot $T(t)$, the time series for the 6-Month Treasury Bill, along with $F(t)$, the *Federal Funds Effective rate*, and $R(t)$, the *Federal Funds Target rate*, over the 8-year period beginning in January 2000. The relative difference between the 6-Month Treasury Bill and the Federal Funds Effective rate is an indicator of the future expectations of the Federal Funds Target rate [18]. Note that the 6-Month Treasury Bill has anticipatory behavior with respect to the Federal Funds Target (and hence Effective) rates. Other more sophisticated models utilize futures on Federal Funds and Eurodollar exchange, but these markets are rather new, and represent the highly complex nature of contemporary markets and hedging programs [19, 20, 21, 22, 23]. Hence, we use a simple and intuitive method for estimating market speculation and anticipation by analyzing the relative difference between the 6-Month Treasury Bill and the Federal Funds Effective rate.

Fig. 3.1(b) exhibits the typical interplay between the 6-Month T-Bill and the Federal Funds Effective rate before and after a FOMC meeting. The change in the value of the Effective rate results from market speculation, starting approximately one trading week (5 trading days) prior to the announcement. This change follows from the forward-looking Treasury Bill, which in the example in Fig. 3.1(b), is priced above the Federal Funds rate even 15 trading days before the announcement.

In order to quantify speculation and anticipation in the market prior to each scheduled FOMC meeting, we analyze the time series $\delta(t)$ of the relative spread between $F(t)$ and $T(t)$,

$$\delta(t) \equiv \ln\Big(\frac{F(t)}{T(t)}\Big) \ . \tag{3.1}$$

As an example of this relation, in Fig. 3.1(c) we plot $\delta(t)$ for the 15 days before and after a typical FOMC meeting resulting in a rate change. In order to study the speculation preceding the i^{th} scheduled FOMC meeting, we calculate the average relative spread over the $L_1 = 15$ day period. We weight the days in the L_1-day period leading up to the FOMC meeting day exponentially, such that the relative spread on the Δt^{th} day before the announcement has the weight $w(\Delta t) = e^{-\Delta t/\lambda}$. Without loss of generality, we choose the value of $\lambda = 10$ days corresponding to two trading weeks. For the calculation of Θ and Δ, we choose values of L and λ to be on the order of a couple trading weeks prior to the announcement, so that we isolate fresh speculation leading into the meeting. The parameter λ provides an effective cutoff period, after which the weights begin to decrease quickly. Conversely, the weights corresponding to days close to the meeting, $\Delta t = 0$, are effectively constant. The values of Θ and Δ do not change much with varying choice of $L \in [5, 15]$ or $\lambda \in [5, 15]$. Without loss of generality, we choose $L = 15$ days, and $\lambda = 10$ days. We define the speculation metric,

$$\Theta_i = \overline{\delta(t)}_i \equiv \frac{\sum_{\Delta t} \delta(t_i - \Delta t) w(\Delta t)}{\sum_{\Delta t} w(\Delta t)} \ , \tag{3.2}$$

which is a weighted average of $\delta(t)$ before the announcement, where the sums are computed over the range $\Delta t \in [1, L_1]$. The metric Θ_i for the i^{th} FOMC meeting can be positive or negative, depending on the market's forward-looking expectations.

In order to quantify the market response to the speculation Θ_i, we analyze the market volatility around each FOMC meeting. For a particular stock around the i^{th} scheduled FOMC meeting, we take the daily high price $p_{\rm hi}(t_i + \Delta t)$, and the daily low price $p_{\rm low}(t_i + \Delta t)$, for $\Delta t \in [-20, 20]$, where $\Delta t = 0$ corresponds to t_i, the day of the meeting. We then compute the high-low range for each trading day,

$$r(t_i + \Delta t) \equiv \ln\left(\frac{p_{\rm hi}(t_i + \Delta t)}{p_{\rm low}(t_i + \Delta t)}\right) . \tag{3.3}$$

For each stock and each meeting, we scale the range by $\langle r \rangle$, the average range over the 41-day time sequence centered around the meeting day, resulting in the normalized volatility $v(t_i + \Delta t) \equiv r(t_i + \Delta t)/\langle r \rangle$. Similarly, we use $\Phi(t_i + \Delta t)$, the time series for the volume traded over the same period, to compute a weight for each stock corresponding to the normalized volume on the day of the FOMC meeting. We

calculate this weight as $\phi_i \equiv \Phi(t_i)/\langle\Phi\rangle_i$, where $\langle\Phi\rangle_i$ is the average daily volume over the 41-day time sequence centered around the i^{th} meeting day. We use a volume weight in order to emphasize the price-impact resulting from relatively high trading volume, since there are significant cross-correlations between volume change and price change [24]. Finally, we compute the weighted average volatility time series over all stocks and all meetings,

$$\langle v(\Delta t)\rangle \equiv \frac{\sum v(t_i+\Delta t)\phi_i}{\sum \phi_i} \ . \tag{3.4}$$

In Fig. 3.2 we plot the trend of average daily volatility defined in Eq. (3.4) for the 10 days before and after the scheduled announcements.

We observe a peak in $\langle v(\Delta t)\rangle$ on FOMC meeting days, corresponding to $\Delta t = 0$, with a more pronounced peak in the bank sector (Fig. 3.2). Stocks in the bank sector are strongly impacted by changes in Fed rates, which immediately influence both their holding and lending rates. On average there is a 15-20% increase in volatility on days corresponding to FOMC meetings.

In order to quantify the impact of a single FOMC announcement on day t_i, we define the average market volatility

$$V_i = \langle v(t_i)\rangle \equiv \frac{\sum^{(i)} v(t_i)\phi_i}{\sum^{(i)} \phi_i} \ . \tag{3.5}$$

Here, $\langle\cdots\rangle_i$ and $\sum^{(i)}$ refer to the average and sum over records corresponding only to the day t_i. Again, $\phi_i \equiv \Phi(t_i)/\langle\Phi\rangle$ is a normalized weight, where now $\langle\Phi\rangle$ is the average daily volume over the entire 8-year period, since we compare many meetings across a large time span.

In Fig. 3.3 we plot the average volatility V_i of the (a) S&P100 and (b) the subset of 18 banking stocks versus Θ_i. For negative values of Θ_i, for which $T(t) > F(t)$ corresponding to an expected rate increase, we observe a less volatile market response. Conversely, for larger positive values of Θ_i, for which $T(t) < F(t)$ corresponding to a rate cut, there tends to be larger market fluctuations. Hence, the market responds differently to falling and rising rates, where the direction in rate change often reflects the overall health of the economy as viewed by the FOMC. Typically, the FOMC implements rate increases to fight inflation, whereas rate decreases often follow bad economic news or economic emergency. Hence, our findings are consistent with the

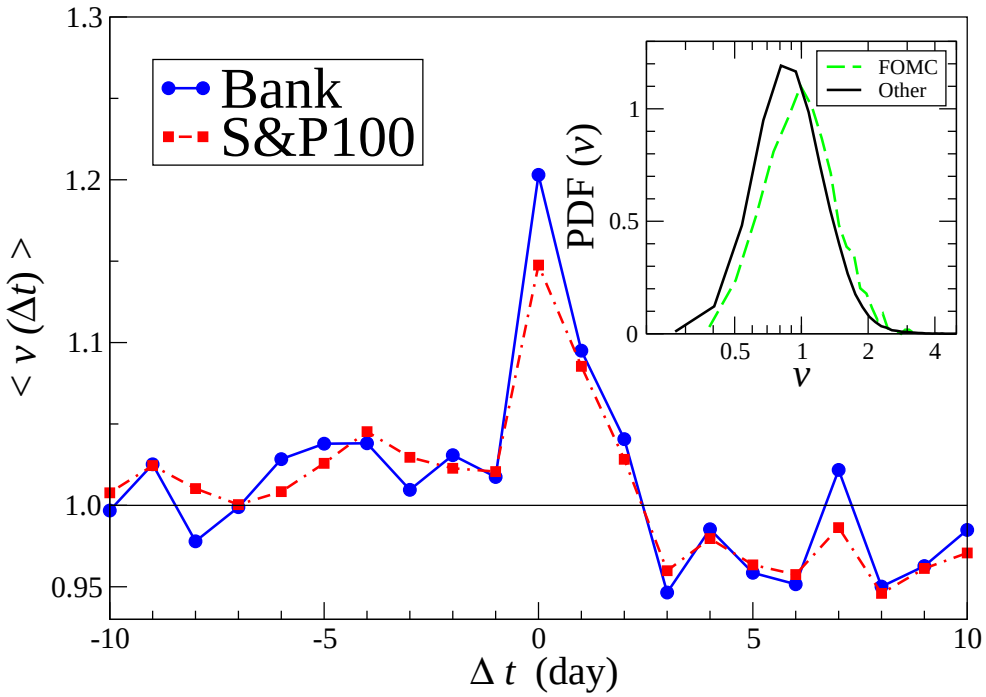

Figure 3.2: Average daily volatility trend defined in Eq. (3.4) exhibits increased market volatility on the day of FOMC meetings, corresponding to $\Delta t = 0$. "Bank" refers to the portfolio of 18 stocks that belong to the S&P100. There is a 15-20% increase in volatility on days corresponding to FOMC meetings. Standard deviation $\sigma(v(\Delta t)) \approx 0.4$ can be assigned to each data point in the time series, and are calculated by randomizing the daily volatility time series of each company. (Inset) Probability density function (pdf) of normalized volatility $v \equiv r(t)/\langle r \rangle$, where the quantity $r(t) \equiv \ln(p_{hi}(t)/p_{low}(t))$ is the range of the price time series of a given stock for a particular day. We plot the pdf of volatility values for the S&P100 on the set of days with FOMC meetings and for the set of all "other" days. The distributions are approximately log-normal, with a shift towards higher average volatility on FOMC days. The average values for the two data sets are $\langle v \rangle_{FOMC} = 1.12$ and $\langle v \rangle_{other} = 1.00$.

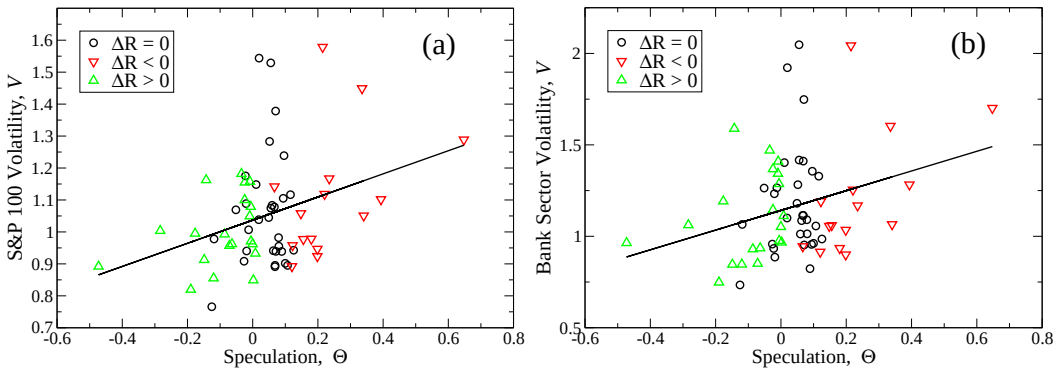

Figure 3.3: Demonstration of the relation between speculation of interest rate change and market volatility in the S&P100 and for the subset of *Banking* stocks. We relate Θ_i, the speculation in the market before a FOMC meeting defined in Eq. (3.2), to the market volatility V_i defined in Eq. (3.5). A large absolute value of Θ_i reflects the high probability that an interest rate change will happen. Interestingly, there are many instances where $\Theta_i \approx 0$, followed by large volatility. These values correspond to FOMC decisions to maintain interest rate levels ($\Delta R = 0$), and suggest a fundamental difference in the dynamics following decisions to change versus decisions not change the Federal Funds Target rate. Also, there is an underlying symmetry in ΔR, since in the case of either a rate increase or a rate decrease, the FOMC also has the option of no increase. Hence, $\Delta R = 0$ can occur as either "good" or "bad" news, whereas typically decisions of $\Delta R > 0$ reflect situations with positive market sentiment whereas decisions of $\Delta R < 0$ reflect situations with negative market sentiment. Hence, the asymmetry in market volatility is consistent with the "sign effect" [13]. Although the correlation between Θ_i and V_i is dominated by residual error, it is nevertheless supporting that the regression captures the crossover at $(\theta, V) = (0, 1)$. Including all data points, the regression correlation coefficient is $r^2 = 0.34$, and the slope of the regression is $m = 0.36 \pm 0.13$ for panel (a) and $r^2 = 0.30$ and $m = 0.54 \pm 0.22$ for panel (b). All linear regressions pass the ANOVA F-test, rejecting the null hypothesis that $m = 0$ at the $\alpha = 0.05$ significance level.

empirical "sign effect", in which "bad" news has a greater impact in markets than does "good" news [13]. Furthermore, there is also a tendency for large average volatility even when Θ_i is small, possibly stemming from the extreme surprise characteristic of some FOMC decisions. In these cases, more sophisticated methods are needed to improve the predictions of market movement.

3.4.2 Intraday response to FOMC decision via an Omori Law

In the previous section we studied the market response on the daily scale. Now we ask the question, "What is the intraday response to FOMC news?" Here we analyze the TAQ data over the 2-year period Jan. 1, 2001 to Dec. 31, 2002. The reported times for the FOMC announcement are listed in Table 1. We find historical information about FOMC meetings using resources at the Federal Reserve websit and using newspaper archives. Intraday time of announcement, T, are often quoted in New York Times finance articles by Richard W. Stevenson the day after FOMC announcements. They are also evident in the intraday Omori plots of $N(t)$ in Fig. 4. Inspired by the non-stationary nature of financial time series, methods have been developed within the framework of non-equilibrium statistical mechanics to describe phenomena ranging from volatility clustering [25, 26, 27] to financial correlation matrices [28, 29, 30].

We use the Omori law, originally proposed in 1894 to describe the relaxation of after-shocks following earthquakes, to describe the response of the market to FOMC announcements. Defined in Ref. [10], the Omori law quantifies the rate $n(t)$ of large volatility events following a singular perturbation at time T. The shock may be *exogenous* (resulting from external news stimuli) or *endogenous* (resulting from internal correlations, e.g. "herding effect") [31, 32, 33, 34, 35]. This rate is defined as,

$$n(|t-T|) \sim |t-T|^{-\Omega} \,, \tag{3.6}$$

where Ω is the Omori power-law exponent.

Here we study the rate of events greater than a volatility threshold q, using the high-frequency intraday price time series $p(t)$. The intraday volatility (absolute returns) is expressed as $v(t) \equiv |\ln(p(t)/p(t-\delta t))|$, where we use $\delta t = 1$ minute. To compare stocks, we scale each raw time series in terms of the standard deviation over

the entire period analyzed, and then remove the average intraday trading pattern as described in Ref. [11]. This establishes a common volatility threshold q, in units of standard deviation, for all stocks analyzed.

In the analysis that follows, we focus on $N(|t-T|)$, the cumulative number of events above threshold q,

$$N(|t-T|) = \int_T^t n(|t'-T|)dt' \equiv \beta(|t-T|)^{1-\Omega} \ , \tag{3.7}$$

which is less noisy compared to $n(|t-T|)$. Using $N(|t-T|)$, we examine the intraday market dynamics for 100 S&P stocks, before ($t < T$) and after ($t > T$) the i^{th} FOMC announcement at T_i, which typically occurs at 2:15 PM ET (285 minutes after the opening bell) for scheduled meetings.

In Figs. 3.4(a,b) we plot the average volatility response $N(t)$ of the $S \equiv 100$ stocks analyzed, where

$$N(t) \equiv \frac{1}{S}\sum_{j=1}^{S} N^j(t) \ . \tag{3.8}$$

This average is obtained by combining the individual Omori responses, $N^j(t)$, of the S stocks. Such averaging does not cancel the Omori law, but allows for better statistical regression. This is especially useful for an Omori law corresponding to large volatility threshold q, where a single stock might not have a sufficient number of events. In Fig. 3.4(c) we plot the trade pattern $N^j(t)$ of Merrill Lynch on Tuesday 08/21/01, and also in Fig. 3.5 for the following three days, demonstrating that the Omori relaxation can persists for several days.

The abrupt change in the curvature of $N(t)$ illustrates the volatility clustering which begins around the time of the announcement T, corresponding to the vertical line at $t = 285$ minutes in Figs. 3.4(a-c). For comparison, we find that the average $\langle N(t)\rangle$ time series calculated from all days without FOMC meetings is approximately linear with time throughout the entire day, indicating that the sudden increase in excess volatility before and after announcement times T results from the FOMC news. Volatility clustering in financial data sampled at the 1-minute scale persists for several months, with a significant crossover in the observed power-law autocorrelations occurring around 600 minutes (≈ 1.5 days) [36, 37, 38].

In order to compare the dynamics before and after the announcement, we first separate the intraday time series $N(t)$ into two time series $N_b(t|t < T)$, and $N_a(t|t >$

FOMC Date	R_{new} (%)	ΔR	$\frac{\Delta R}{R_{old}}$	$\boldsymbol{T}$
01/03/01**	6	-0.5	-0.077	210
01/31/01	5.5	-0.5	-0.083	285
03/20/01	5	-0.5	-0.091	285
04/18/01**	4.5	-0.5	-0.100	90
05/15/01	4	-0.5	-0.111	285
06/27/01	3.75	-0.25	-0.063	285
08/21/01	3.5	-0.25	-0.067	285
09/17/01**	3	-0.5	-0.143	0
10/02/01	2.5	-0.5	-0.167	285
11/06/01	2	-0.5	-0.200	285
12/11/01	1.75	-0.25	-0.125	285
01/30/02	1.75	0	0.00	285
03/19/02	1.75	0	0.00	285
05/07/02	1.75	0	0.00	285
06/26/02	1.75	0	0.00	285
08/13/02	1.75	0	0.00	285
09/24/02	1.75	0	0.00	285
11/06/02	1.25	-0.5	-0.286	285
12/10/02	1.25	0	0.00	285

Table 3.1: Reported times of market perturbations in the form of FOMC news. Dates of 19 FOMC meetings in the 2-year period between Jan. 2001 - Dec. 2002, where the Federal Funds Target rate R_{new} was implemented by the rate change ΔR at T minutes after the opening bell at 9:30 AM ET. The absolute relative change $|\frac{\Delta R}{R_{old}}| \equiv |\Delta R(t)/R(t-1)|$ has typically filled the range between 0.0 and 0.25. Note: Date** refers to *unscheduled* meetings, in which the announcement time did not correspond to 2:15 PM ET ($T = 285$ minutes).

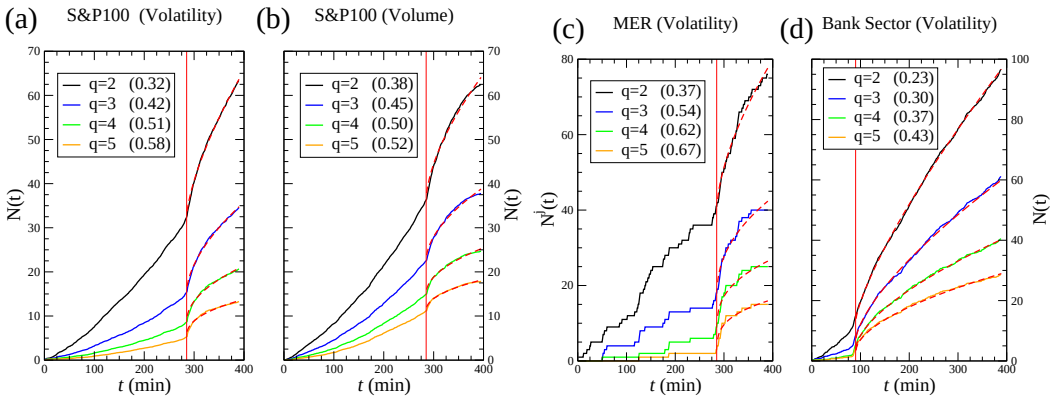

Figure 3.4: The cumulative volatility time series $N(t)$ demonstrates Omori-law response dynamics, here in response to FOMC announcements occurring at the time T indicated by a vertical solid line. The abrupt change in the curvature of $N(t)$ around time $t \approx T$ illustrates the increased volatility caused by the announcement. The significant aftershocks which occur until the end of the trading day are consistent with "market underreaction" [12, 13]. Market underreaction and other market inefficiencies can result from increased levels of uncertainty among traders following market news [14]. Each time series $N(t)$ is calculated for a given volatility threshold q, where larger q values correspond to $N(t)$ curves with smaller amplitude (smaller rate of large volatility events). Panels (a,b,c) illustrate the dynamics around a scheduled announcement made at $T = 285$ minutes (2:15 PM ET). For the S&P 100, we calculate $N(t)$ on 05/15/01 for (a) 1-minute volatility and (b) 1-minute total volume using Eq. (3.8). (c) We calculate $N^j(t)$ for Merrill Lynch (MER) on 08/21/01. (d) The Omori law also occurs for *unscheduled* FOMC announcements, as illustrated for the *Bank* sector $N(t)$ on 04/18/01, when the surprise rate change was announced at $T = 90$ minutes (11:00 AM ET), resulting in raised levels of volatility throughout the entire trading day. For panels (a-d), the dashed lines are power-law fits beginning immediately after the announcement, with the corresponding exponents $\Omega_a(q)$ appearing in parentheses within the legends.

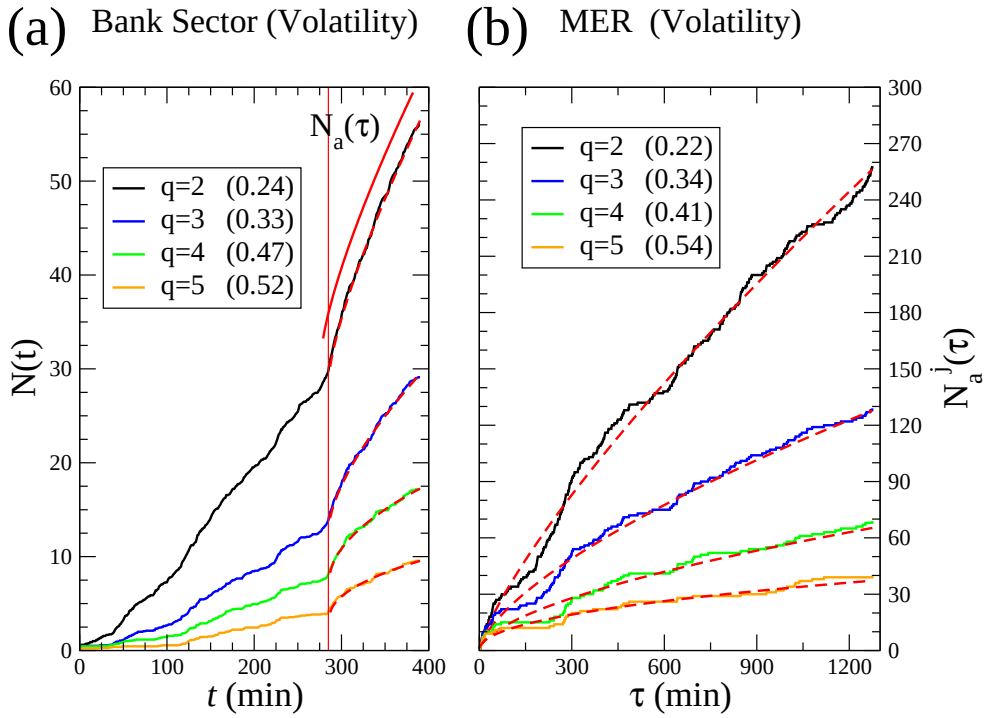

Figure 3.5: The Omori law relaxation can extend for several days. We compare the Omori exponents $\Omega_a(q)$ (indicated in legends) calculated for (a) the time series $N_a(\tau)$ of the *Bank* sector and (b) the time series $N_a^j(\tau)$ of Merrill Lynch for the three days (1275 minutes) after the announcement on Tuesday 08/21/01 at 2:15 pm ($\tau = 0$ minutes corresponding to $T = 285$ minutes). For the remaining 3 days of the trading week, the Omori law relaxation corresponding to an individual stock (MER) is quantitatively similar to the the Omori law relaxation of the *Bank* sector over the final 105 minutes of the initial trading day. We do not use the *Bank* sector $N_a(\tau)$ over the same 1275-minute time period for comparison because "opening effects" occurring during the first 60 minutes of each trading day makes power-law regression of conjoined $N_a(\tau)$ problematic.

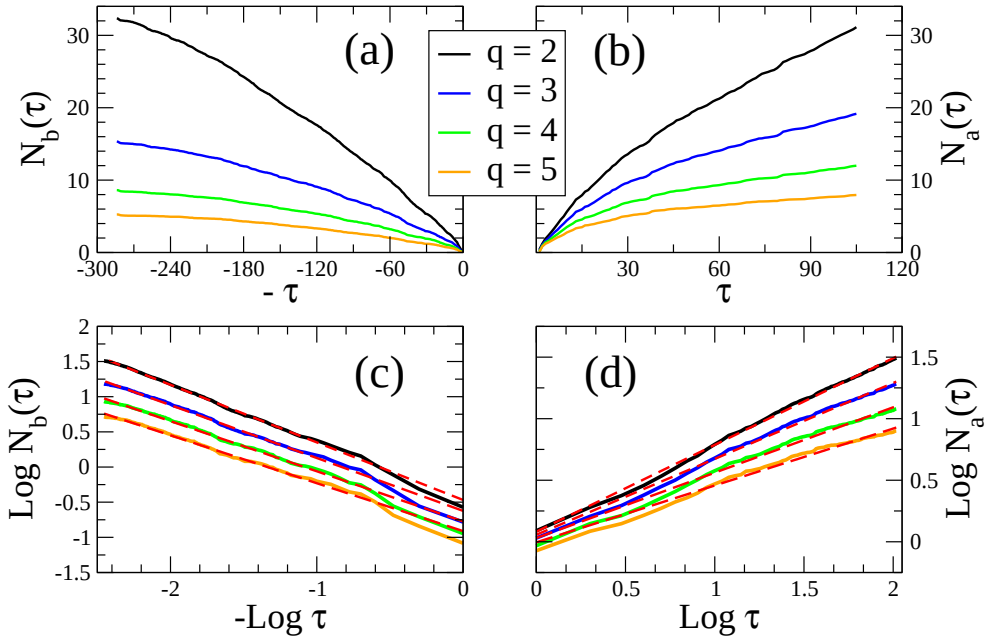

Figure 3.6: An illustration of the method used to calculate (a) $N_{b,i}(\tau) = N_i(T_i) - N_i(|t - T_i|)$ and (b) $N_{a,i}(\tau) = N_i(t - T_i) - N_i(T_i)$ for each intraday time series $N_i(t)$. The displaced time $\tau = |t - T_i|$ is defined symmetrically around the time of the announcement T_i. We plot the same data as in Fig. 3.4 (a), corresponding to the announcement on 05/15/01 which occurred at $T = 285$ minutes. Panels (c) and (d) show that $N_{b,i}(\tau)$ and $N_{a,i}(\tau)$ are approximately linear on logarithmic scale.

T). Then, to treat the dynamics symmetrically around the i^{th} intraday announcement time T_i [35, 39], we define the displaced time $\tau = |t - T_i| \geq 1$ as the temporal distance from the minute T_i. As an illustration, we plot $N(\tau)$ in Fig. 3.6 for the 4 corresponding $N(t)$ curves exhibited in Fig. 3.4(a). We then employ a linear fit to both $N_{b,i}(\tau) = N_i(T_i) - N_i(|t - T_i|)$ and $N_{a,i}(\tau) = N_i(t - T_i) - N_i(T_i)$ on a log-log scale to determine the Omori power-law exponents Ω_b before the news and Ω_a after the news. In analogy, we define the amplitude β before as β_b and after as β_a, as defined in Eq. (4.6).

Typically $\Omega_a > 0$, which reflects the pronounced increase in the rate of events above the volatility threshold q after the time of the announcement. We also observe $\Omega < 0$, which corresponds to a time series in which the pre-shocks or after-shocks farther away from the announcement (for large τ) are dominant over the volatility

cascade around time $\tau \approx 0$. For comparison, $n(\tau)$ is constant for stochastic processes with no memory, corresponding to $\Omega \equiv 0$. Hence, for an empirical value $\Omega \approx 0$, the rate $n(\tau)$ is indistinguishable from an exponential decay for $\tau < \overline{t}$, where $\overline{t}$ is the characteristic exponential time scale. However, for larger values of Ω, the exponential and power-law response curves are distinguishable, especially if several orders of magnitude in τ is available.

For all meetings analyzed, we find that $\Omega \equiv \Omega(q)$ increases with q, meaning that the relatively large aftershocks decay more quickly than the relatively small aftershocks. Hence, the largest volatility values cluster around the announcement time T. For comparison, $\Omega(q)$ values are calculated in [10] using $q = 4, 5, 6, 7$ and in [11] using $q = 3, 4$ for large financial crashes. For our data set, the cumulative probability $P(v > q)$ that a given volatility value is greater than volatility threshold q is $P(v > 3) = 0.18$ and $P(v > 5) = 0.087$. Furthermore, we reject the null hypothesis that $q > 5$ volatilities are distributed evenly across all days, finding that 5% of the volatility values greater than $q = 5$ are found on FOMC meeting days, whereas only 4% are expected under the null hypothesis that large volatilities are distributed uniformly across all trading days. The 25% increase for $q = 5$ indicates that FOMC meetings days are more volatile than other days at the $\alpha \approx 0$ significance level. We also observe that the amplitudes of the Omori law generally obey the inequality $\beta_b < \beta_a$, resulting from the large response immediately following the news.

Although we focus mainly on price volatility $v(t)$ in this thesis, we also observe Omori dynamics in the high-frequency volume time series $\omega(t)$, defined as the cumulative number of shares traded in minute t. In Figs. 3.7 (a-d) for the S&P 100 and Figs. 3.7 (e-h) for the bank sector, we compare the average of Omori exponents Ω_b and Ω_a for both volatility and volume dynamics, and for volatility threshold value $q = 3$. We compute the average Omori exponents using two averaging methods, the "individual" method and the "portfolio" method.

To analyze the time series $N_{a,i}$ after the announcement i, we first average the exponents Ω_a^j obtained for each individual stock j, yielding $\langle \Omega_a \rangle$. This "individual" method provides an error bar corresponding to the sample standard deviation $\sigma(\Omega_a)$. The second "portfolio" method determines a single Ω_a from $N(t)$ in Eq. (3.8). Comparing the open-box (individual method) and closed-box (portfolio method) symbols

in Fig. 3.7, we observe that both methods yield approximately the same average value of Ω_a. Note that for the subset $i = \{1,8\}$ of the unscheduled FOMC meetings, Ω_a is smaller than usual, capturing the intense activity following surprise announcements. Hence, unexpected FOMC announcements can produce an inverse Omori law exhibiting convex relaxation ($\Omega_a < 0$) over a short horizon if the news contains a large amount of inherent surprise. The 8th meeting corresponds to the opening of U.S. markets a week after the terrorist attacks on the World Trade Center on Sept. 11, 2001.

For the time series $N_{b,i}$ before the announcement i, individual stocks often do not have sufficient activity to provide accurate power-law fits. Hence, to estimate the sample standard deviation $\sigma(\Omega_b)$, we produce partial combinations, $\langle N(\tau)\rangle_{b,i} \equiv \frac{1}{M}\sum_{j=1}^{M} N_{b,i}^{j}(\tau)$ using $M \equiv 5$. We then compute a standard deviation $\sigma(\Omega_b)$ from the Ω_b values calculated from $\langle N(\tau)\rangle_{b,i}$. The $\sigma(\Omega_b)$ values correspond to the error bars for $\langle\Omega_b\rangle$ in Fig. 3.7.

We also compute a single Ω_b value from the portfolio average $N_{b,i}(\tau)$, which corresponds to the limit $M = S$. The values of Ω_b using the two methods are consistent. Interestingly, the values of Ω_b calculated from volume data are all close to zero. However, using the Student T-test we reject the null hypothesis that each average value $\langle\Omega_b\rangle$ is equal to zero at the $\alpha = 0.01$ significance level for 15 out of 17 dates.

Fig. 3.7 shows the range of Ω values for each of the 19 FOMC meetings we analyze. There are 8 panels comparing the Ω values (i) between the dynamics before and after T, (ii) between the volatility and volume dynamics, and (iii) between set of all stocks comprising the $S\&P100$ and the set of stocks comprising the banking sector. We hypothesize that the differences in the Omori Ω values, before and after the announcement, are related to the anticipation and perceived surprise of the FOMC news. Furthermore, for the dynamics after the news, we find anomalous negative Ω_a values for two surprise FOMC announcements $i = 1$ and $i = 8$. Also, we find that volume Ω values are more regular across all meeting events, suggesting that volume and price volatility contain distinct market information [40, 41].

In order to find potential variations in the response dynamics for different stock sectors, in Fig. 3.8(a) we compare the Ω_a values after the announcement for 5 approximately equal-sized sectors using volatility threshold $q = 3$. We observe that the

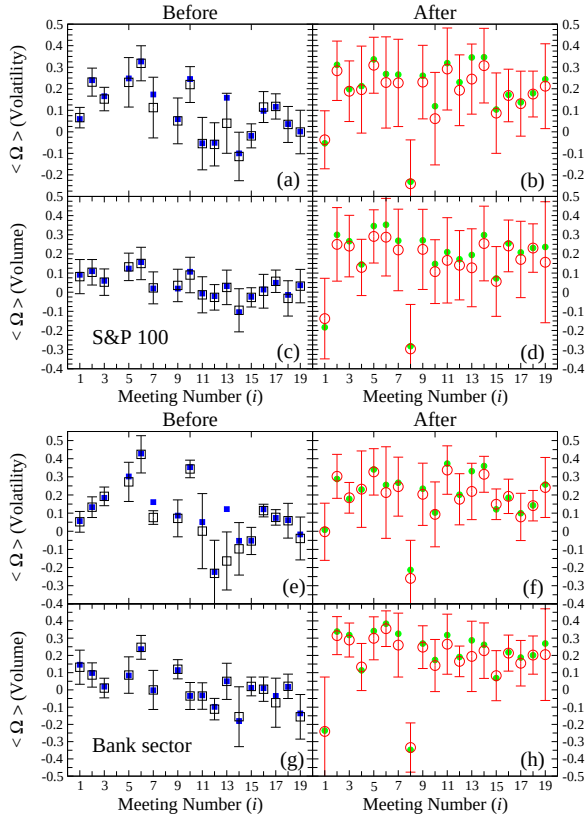

Figure 3.7: Comparison of Omori law exponents for both volatility dynamics and volume dynamics on the day of 19 FOMC meetings during the two-year period Jan. 2001- Dec. 2002. Panels (a-d) correspond to the S&P 100 and panels(e-h) correspond to the bank sector. The similarity in exponents for 1-minute volatility and 1-minute cumulative volume suggest a universal underlying mechanism. Solid symbols (■ and •) refer to Ω computed from $N(t)$. Open symbols (□ and ◯) refer to $\langle\Omega\rangle$ computed from S individual Omori exponents Ω^j, with $S_{bank} = 18$. Note the relatively low value of Ω_a and $\langle\Omega_a\rangle$ for unscheduled FOMC announcements $i = 1$ and 8, which indicates that volatility rate following the announcement increased throughout the day.

differences in the average values of the sectors are fairly small, indicating a broad market response. We also observe that the technology sector (Tech.), composed of hardware, software, and IT companies, often has the largest average Ω_a value. Larger exponents, which correspond to shorter relaxation times, could result from the intense trading in the Tech sector during the Tech/IT bubble, which peaked in the year 2000. In follow-up analysis, we find in [42] that stocks with higher trading activity, quantified as the average number of transactions per minute, have larger Ω^j in response to market shocks, and thus, faster price discovery. In order to compare the variation in the individual values of Ω_a^j, we plot the pdf of exponents for all stocks and meetings in Fig. 3.8(b) using the shifted variable $x^j \equiv \Omega_{a,i}^j - \Omega_{a,i}$. We conclude from a Z-test at the $\alpha = 0.0005$ significance level that Tech sector Omori exponents are larger on average, $\langle x \rangle_{Tech} > \langle x \rangle_{SP100}$.

Motivated by the metric Θ_i defined in Eq. (3.2), which quantifies speculation and anticipation in the market preceding FOMC meetings, we now develop a second metric to describe surprise through the change in market speculation following the announcement. This metric Δ_i compares the anticipation leading up to the announcement with the revised speculation following the FOMC decision. This can be quantified through the relative change in $\delta(t)$, which provides a rough measure of the market stress that is released in the financial shock. Qualitatively, Δ_i relates the average value of the spread before and after the i^{th} scheduled meeting. We define,

$$\Delta_i \equiv \left(\overline{\delta(t)}_{i,a} - \overline{\delta(t)}_{i,b} \right) \times S(\Delta R_i) \tag{3.9}$$

$$\equiv \left(\frac{\sum \delta(t_i - \Delta t)\ w(\Delta t)}{\sum w(\Delta t)} - \frac{\sum \delta(t_i + \Delta t)\ w(\Delta t)}{\sum w(\Delta t)} \right) \times S(\Delta R_i)\ , \tag{3.10}$$

where the sum is computed over the range $\Delta t \in [1, L_2]$ trading days, with $L_2 = 15$ trading days and $\lambda_2 = 10$ trading days. The factor $S(\Delta R_i) = 1$ when the Fed increases or maintains the Target rate $R(t)$, while $S(\Delta R_i) = -1$ when the Fed decreases the Target rate.

In Figs. 3.9 (a-d) we relate the amplitudes $\langle \beta_b \rangle$ and $\langle \beta_a \rangle$, and also the exponents $\langle \Omega_b \rangle$ and $\langle \Omega_b \rangle$ to the speculation metric Θ and the surprise metric Δ. We observe that larger Θ and larger Δ are related to larger amplitude $\langle \beta_b \rangle$ quantifying the preshock dynamics. However, we do not find a statistically significant relation between Θ or Δ and the aftershock parameters, suggesting that the relaxation dynamics following

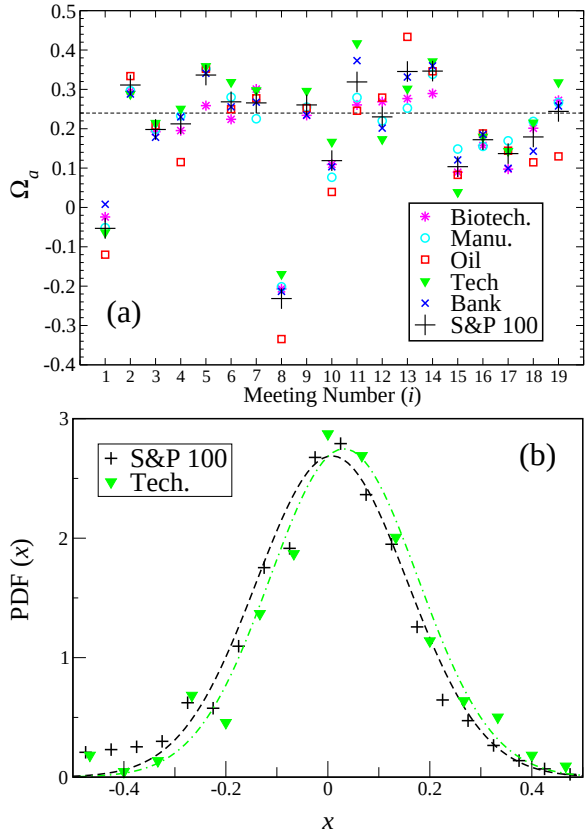

Figure 3.8: (a) A comparison of Ω_a for 5 sectors with volatility threshold $q = 3$ suggests a broad universal market response to FOMC news. The Tech sector tends to have the largest average Ω_a, where large Ω values corresponds to faster relaxation. The horizontal straight line represents the mean $\overline{\Omega}_a = 0.24 \pm 0.08$, averaged over all stocks in the S&P100 and all scheduled meetings (excluding the unscheduled meetings $i = \{1, 4, 8\}$). (b) Probability density function $P(x)$ of the variable $x \equiv x^j_{a,i} = \Omega^j_{a,i} - \langle \Omega_{a,i} \rangle$, which correspond to individual $\Omega^j_{a,i}$ values centered around the average exponent $\langle \Omega_{a,i} \rangle$ of a given meeting i. Tech sector Omori exponents are larger on average, and since larger Ω values correspond to shorter relaxation time, we find that the Tech sector stocks responds more quickly to FOMC news, possibly as a result of relatively intense trading activity among these stocks.

FOMC news are less predictable. Nevertheless, the aftershock dynamics are consistently more pronounced, with $\langle\beta_a\rangle > \langle\beta_b\rangle$. We interpret Figs. 3.9(a) and (c) as follows: when $\Theta < 0$, corresponding to "good" market sentiment and possible rate increase, the dynamics before the announcement have small β_b and small Ω_b reflecting low activity. After the announcement, the values of β_a and Ω_a increase, corresponding to a fast response of medium size. In the case of $\Theta > 0$, corresponding to "bad" market sentiment resulting from speculation of a rate cut, the dynamics before the announcement have large β_b and large Ω_b, corresponding to a strong but quick buildup of volatility. After the announcement, the dynamics have large β_a and small Ω_a, corresponding to a strong and lasting relaxation dynamics. The interpretation of Figs. 3.9(b) and (d) is similar to the interpretation of Figs. 3.9(a) and (c), in that both surprise ($\Delta > 0$) and expected ($\Delta \approx 0$) bad news, correspond to a stronger and longer-lasting relaxation dynamics.

3.5 Discussion

Information flows through various technological avenues, keeping the ever-changing world up-to-date. All news carries some degree of surprise, where the perceived magnitude of the news certainly depends on the recipient. In financial markets, where speculation on investment returns results annually in billions of dollars in transactions, news plays a significant role in perturbing the complex financial system both on large and small scales, reminiscent of critical behavior with divergent correlation lengths [43]. Perturbations to the financial system are easily transmitted throughout the market by the long-range interactions that are found in the networks of market correlations [28, 29, 30]. Afterwards, the effects of the perturbation may persist via the long-term memory observed in volatility time series [36, 37, 38], with fluctuation scaling obeying the empirical Taylor's law [44, 45].

We have shown that the Omori law describes the dissipation of information following the arrival of Federal Open Market Commission (FOMC) news. This type of relaxation is consistent with the substructure of financial crash aftershocks observed on various scales [11]. In particular, we systematically study the dynamical response of the stock market to perturbative information in the form of a Federal

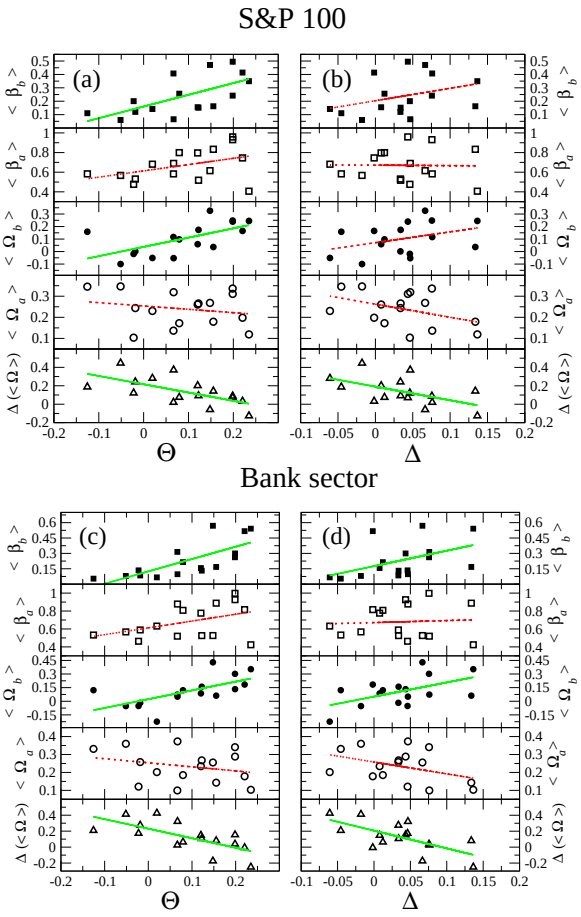

Figure 3.9: The relation between the size of the financial shock, quantified by the S&P 100 volatility Omori law parameters $\langle\beta_b\rangle$, $\langle\beta_a\rangle$, $\langle\Omega_b\rangle$, $\langle\Omega_a\rangle$, and $\Delta(\langle\Omega\rangle) = \langle\Omega_a\rangle - \langle\Omega_b\rangle$, and the size of the FOMC news, quantified through the metrics Θ representing market anticipation and Δ representing market surprise. All trends are consistent with the hypothesis that a strong anticipation of an interest rate change, and the element of surprise inherent in the FOMC decision, result in a market perturbation that is significant in scale, and broad across the market. Linear regressions of S&P100 data (a,b) and bank sector data (c,d) are provided for visual aide. Linear regressions that pass the ANOVA F-test (rejecting null hypothesis that regression slope $m = 0$) at the $\alpha = 0.05$ significance level are solid; regressions that fail to pass the F-test at the $\alpha = 0.05$ significance level are dashed.

Reserve FOMC interest rate announcements, which can be expected (scheduled) or unexpected (as in cases of emergency).

In the case of unexpected news, as in Fig. 3.4(d), a pronounced response may result from reduced market liquidity, since traders do not have ample time to prepare and adjust [13]. Our findings suggest that the dynamics of "rallies" based on other forms of news, such as earning reports, upgrades and downgrades of stocks by major financial firms, unemployment reports, merging announcements etc., might also be governed by the Omori law with parameters that depend on the type of news. The impact of macroeconomic news has been analyzed for foreign exchange markets [13], where it is found that high levels of volatility are present following both scheduled and surprise news.

According to the efficient market hypothesis [15], the time scale over which news is incorporated into prices should be very small. However, consistent with previous studies, we find "market underreaction" [14] evident in the finite time scale (found here to be at least 1 trading day) over which the volatility aftershocks are significant. Moreover, we quantify the dynamics before and after, and show that the Omori parameters are related to investor sentiment [14], here measured by comparing the 6-month the Treasury Bill and the Federal Funds rates.

It is also conceivable that Omori law decay of market aftershocks also exists in the traded volume time series and the bid-ask spread time series [39, 46] . Recently, Joulin *et al.* [33] use a similar method to describe the relaxation of trading following news streaming from feeds such as Dow Jones and Reuters, and compare their findings with the relaxation following anomalous volatility jumps. They also find Omori law relaxation, with exponent $\Omega_a \approx 1$ following a news source, and $\Omega_a \approx 0.5$ following an endogenous jump; interestingly, they find that the amplitude of the Omori law is larger for news sources than for endogenous jumps. For further comparison, Weber *et al.* [11] find $\Omega_a \approx 0.69$ for the 38 market days following the market crash on September 11, 1986. One distinct difference between these studies, is the source of the news: Joulin *et al.* pool together thousands of news sources, some possibly pertaining to only a single stock; we focus on one particular type of news, the FOMC Target rate decision, which has a broad impact on the whole market and economy. It is possible that the difference between *anticipated* news and *idiosyncratic* news is the

important criterion to consider when analyzing market response functions in relation to exogenous events. Here, we find novel dynamics before anticipated announcements.

In the case of FOMC news, speculation can be quantified by measuring the relative difference between the effective Federal Funds rate and the Treasury Bill in the weeks leading up to a scheduled meeting. We develop a speculation metric, Θ, and relate it to V, the volatility on the day of the meetings, finding that the market behaves more erratically when the Treasury Bill predicts a decrease in the Federal Funds Target rate. A rate decrease often occurs in response to economic shocks, whereas a rate increase is often used to fight inflation. Hence, the asymmetric response in Fig. 3.3 to rising and falling rates is consistent with the "sign effect", where it has been found that bad news causes a larger market reaction than good news [13], and that the asymmetry may result from the increased uncertainty in expectations among traders.

We analyze the four Omori-law parameters Ω_b, Ω_a, β_b and β_a calculated for 19 FOMC meetings. We conjecture that the Omori-law parameters are related to the market's speculation, anticipation and surprise on the day of the FOMC meeting. In order to quantify speculation of rate cuts and rate increases, we define the measure Θ, which is the relative spread between the Treasury Bill and the Federal Funds rates, before the meeting. In order to quantify surprise, we develop Δ, which measures the change in the relative spread between the Treasury Bill and the Federal Funds rates, before and after the meeting. We relate both Θ and Δ to the dynamical response of the market on the day of the meeting. We find that relatively small Ω values and relatively large amplitude β values, corresponding to longer relaxation time and large response, follow from "bad" news, as in the case of the market reaction to the World Trade Center attacks in 2001. In all, these results show that markets relax according to the Omori law following large crashes and Federal interest rate changes, suggesting that the perturbative response of markets belongs to a universal class of Omori laws, independent of the magnitude of news.

Chapter 4

Regularities in the Stock Market Response to 219 Financial Shocks

4.1 Summary

We study the cascading dynamics immediately before and immediately after 219 market shocks. We define the time of a market shock T_c to be the time for which the market volatility $V(T_c)$ has a peak that exceeds a predetermined threshold. The cascade of high volatility "aftershocks" triggered by the "main shock" is quantitatively similar to earthquakes and solar flares, which have been described by three empirical laws — the Omori law, the productivity law, and the Bath law. We analyze the most traded 531 stocks in U.S. markets during the two-year period 2001-2002 at the 1-minute time resolution. We find quantitative relations between (i) the "main shock" magnitude $M \equiv \log V(T_c)$ occurring at the time T_c of each of the 219 "volatility quakes" analyzed, and (ii) the parameters quantifying the decay of volatility aftershocks as well as the volatility preshocks. We also find that stocks with larger trading activity react more strongly and more quickly to market shocks than stocks with smaller trading activity. Our findings characterize the typical volatility response conditional on M, both at the market and the individual stock scale. We argue that there is potential utility in these three statistical quantitative relations

with applications in option pricing and volatility trading.

4.2 Introduction

Financial fluctuations have been a topic of study for economists [47, 48], mathematicians [49], and physicists [2, 3, 4, 5, 6, 50]. Here we study financial fluctuations using concepts developed in the field of seismology [8, 9] and analogies from turbulent dynamics in our description of market main shock magnitudes in order to analyze the dynamic response of markets to financial shocks. We identify parallels between energy cascades and information cascades, and also between turbulent bursts and the clustering of volatility [51]. Our results demonstrate three statistical regularities which relate the volatility magnitude $M \equiv \log V(T_c)$ to the market response before and after market shocks.

Common financial "shocks" are relatively smaller in the volatility magnitude, the duration, and the number of stocks affected, than the extremely large and infrequent financial crashes. Devastating financial shocks such as Black Monday (20 October, 1987) have significant aftershocks that can last for several months, and this "dynamic relaxation" is similar to the aftershock cascade following an earthquake [10]. Here we aim to better understand market shocks over a range of M values. While the previous studies have focussed on at most a few large crashes, we use a large data set of 219 financial "main shocks" observed in American markets over the 2-year period 2001-2002. We analyze 531 frequently traded stocks corresponding to approximately 44,000,000 volatility records at a 1-minute time resolution. We find three quantitative relations which enable answering such questions as:

(i) How does the rate of volatility aftershocks decay with time, and how do the decay parameters relate to the main shock magnitude M?

(ii) How many aftershocks above a given threshold can be expected after a main shock of magnitude M?

(iii) What is the relation between the value of the main shock volatility $V(T_c)$ and the second largest aftershock (or preshock)?

These three questions have been studied for geophysical earthquakes, and the corresponding statistical laws are referred to respectively as the Omori law, the productivity law, and the Bath law.

The Omori law was first investigated in the context of financial crashes by Lillo and Mantegna [10], who found a power-law relaxation of fluctuations at a 1-min time resolution for the S&P500 over the 100-day period following the Black Monday crash. Power-law relaxation of aftershocks is also observed for long periods following several other medium-size crashes [11], and also for short periods up to several days following U.S. Federal Reserve interest rate change announcements [52]. One key feature of long-range relaxation dynamics is the scale-free decay of large fluctuations that is typical of a system with memory, and which is complemented by self-similarity in the decay substructure [11].

We find similar perturbation-response dynamics in the intraday volatility (absolute return) time series for many single stocks on numerous days, indicating that markets respond in a common way to perturbations that range in size from everyday market fluctuations to infrequent market crashes. Interestingly, the market is very responsive to Federal Open Market Commission (FOMC) news, either in the form of subtle hints from the Fed or actual rate changes (expected or unexpected), because Fed Target rates serve as a benchmark and barometer for both U.S. and World markets [52]. This connection between macroeconomic factors and financial markets is a tribute to the complexity and connectivity of economic systems. It is a further indicator that news, in addition to complex order-book dynamics, can play a significant role in explaining the large rate of occurence of large fluctuations in markets.

Here we quantify the rate $n(|t - T_c|)$ of aftershocks at time t both before and after a market shock occuring at time T_c. In order to determine T_c, we develop a method for selecting a critical time T_c from a set of candidate times $\{t_c\}$ for which the collective market volatility of S individual stocks is above a given threshold. For 19 particular dates corresponding to days with FOMC announcements, we compare the values of calculated T_c with the reported values of T analyzed in [52], and we find good prediction of T using this method. After this calibration, we study the relaxation dynamics of $S = 531$ stocks, analyzing the Omori law, the productivity law, and the Bath law for the dynamics both before ($t < T_c$) and after ($t > T_c$) the

main market shock.

In Section 4.3 we discuss the data, the quantitative methods used to calculate $n(|t - T_c|)$, and define collective market movement. In Section 4.4.1 we quantify the threshold for selecting candidate cascades and calibrate using known values of T_c corresponding to FOMC meetings. In Section 4.4.2 we describe the method for choosing T_c from each significant cascade we identify. In Section 4.5 we discuss the Omori-law parameters α and Ω, the productivity parameter Π, and the Bath law parameter B. We note that both Π and B are independent of the dynamical model, and hence do not depend on for $n(|t - T_c|)$, the functional form of the relaxation dynamics. For each of the statistical laws, we compare the results we obtain for the market average with the results we obtain for individual stocks.

4.3 Data Analyzed

For the two-year period 2001-2002, we analyze Trades and Quotes (TAQ) data of more than 500 stocks listed on the NASDAQ and NYSE. In order to analyze the most important subset of stocks, we rank each stock by the average number of transactions per minute. We find $S = 531$ stocks with an average of more than 3 transactions per minute, $S = 136$ stocks with an average of more than 10 transactions per minute, and $S = 20$ stocks with an average of more than 50 transactions per minute. Unless otherwise stated, our results correspond to the top $S = 531$ stocks, but all results become more statistically significant for smaller subsets of more heavily traded (larger) stocks.

In this section, we study the volatility $v_j(t)$ of the intraday price time series $p_j(t)$ for stock j. The intraday volatility (absolute returns) is expressed as

$$v_j(t) \equiv |\ln(p_j(t)/p_j(t - \delta t))| \ , \tag{4.1}$$

where here we choose $\delta t = 1$ minute so that we can analyze the dynamics immediately before and immediately after market shocks. To compare stocks, we scale each volatility time series by the standard deviation over the entire period analyzed. We then remove the "U"-shaped intraday trading pattern (averaged over 531 stocks) from each time series. This establishes a normalized volatility in units of standard deviation for all minutes during the day and for all stocks analyzed (see Ref. [11]).

We introduce a volatility threshold q which defines a binary volatility time series $n_j(t)$ for each stock j, which we calculate from the normalized volatility time series $v_j(t)$ as

$$n_j(t) \equiv \begin{cases} 1 & , \quad v_j(t) \geq q \\ 0 & , \quad v_j(t) < q \ . \end{cases} \tag{4.2}$$

We find that a volatility threshold $q \equiv 3\sigma$ is large enough to distinguish between significant fluctuations and normal background activity. We also choose this value $q \equiv 3\sigma$ to provide comparison with the analysis performed in [52]. The rate $n(t)$ measures the fraction of the market exceeding q at time t,

$$n(t) \equiv \frac{1}{S} \sum_{j=1}^{S} n_j(t) \ . \tag{4.3}$$

The rate $n_j(t)$ quantifying the volatility of a single stock j corresponds to the limit $S \to 1$. We define the average market volatility $V(t)$ similarly by

$$V(t) \equiv \frac{1}{S} \sum_{j=1}^{S} v_j(t) \ . \tag{4.4}$$

A market shock at time T_c may result from *exogenous* (external) news or *endogenous* herding [31, 32]. In many cases, the market shocks can be linked to exogenous news using archived news feeds that cover and summarize daily market events [53]. In order to analyze market dynamics symmetrically around a market shock at time T_c, we analyze the per unit time rate $n(|t - T_c|)$ around the time T_c. It has been empirically observed that the response dynamics in financial markets show a power-law decay [10, 11, 52, 39, 33, 46],

$$n(|t - T_c|) \sim \alpha |t - T_c|^{-\Omega} \ , \tag{4.5}$$

where Ω is called the Omori power-law exponent, α is the cascade amplitude, $t < T_c$ corresponds to before the main shock, and $t > T_c$ corresponds to after the main shock. For comparison, $n(|t - T_c|)$ is constant for stochastic processes with no memory, corresponding to $\Omega \equiv 0$. Hence, for an empirical value $\Omega \approx 0$, the rate $n(|t - T_c|)$ is indistinguishable from an exponential decay for $|t - T_c| < \overline{t}$, where $\overline{t}$ is the characteristic exponential time scale. However, for larger values of Ω, the exponential and

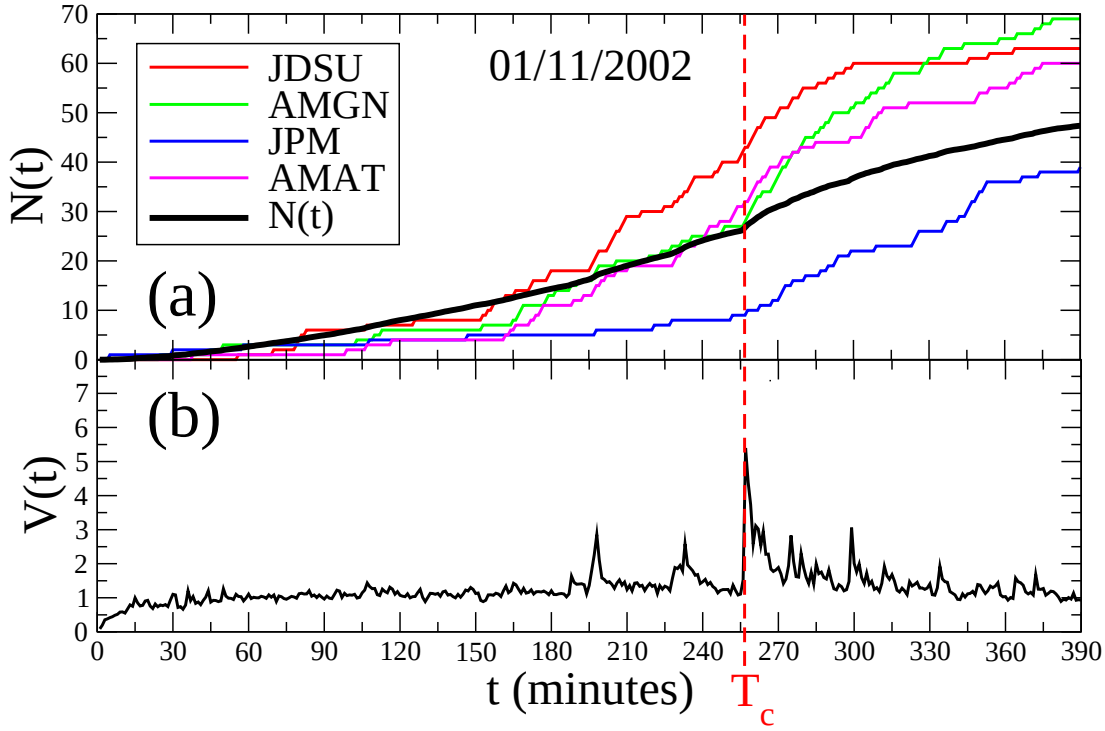

Figure 4.1: Typical volatility curves on 01/11/2002 with market shock at $T_c = 256$ minutes. (a) The cumulative volatility $N^j(t)$ for the stock of several large companies have varying behavior before T_c, but each stock shown begins to cascade soon after T_c. The market average $N(t)$ over all $S = 531$ stocks analyzed demonstrates a distinct change in curvature at $t = T_c$. (b) The average market volatility $V(t)$ demonstrates a sharp peak at T_c, and also two precursor events at $t \approx 190$ and ≈ 230 minutes.

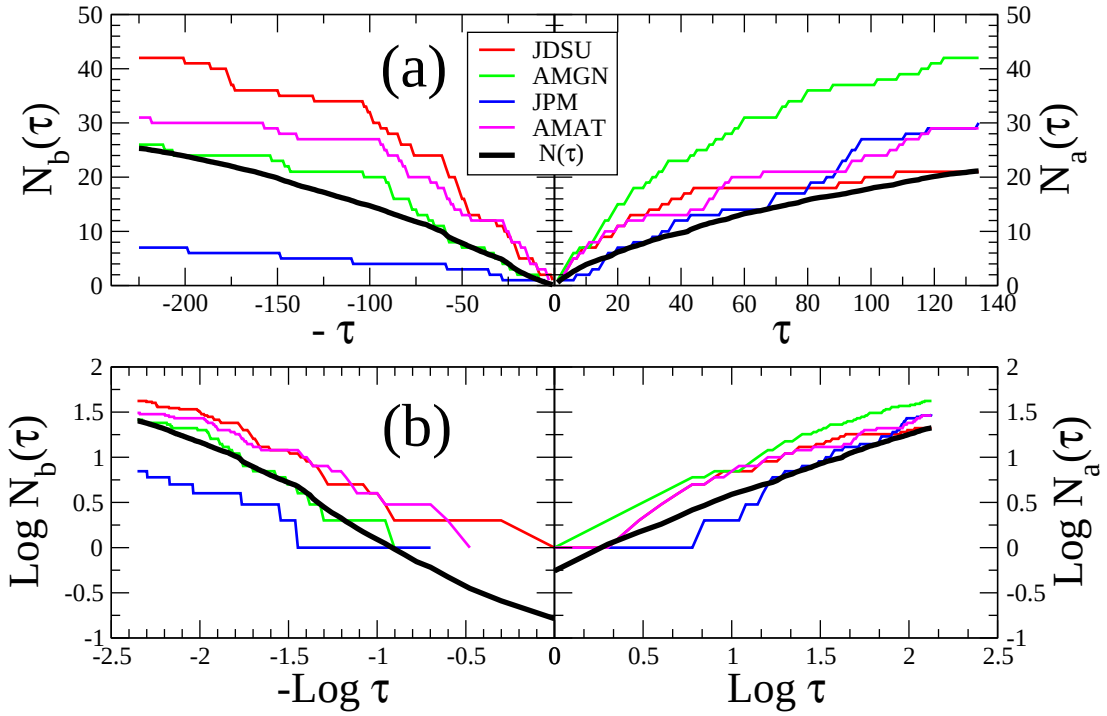

Figure 4.2: (a) An illustration of $N_b(\tau)$ and $N_a(\tau)$ for the same set of curves plotted in Fig. 4.1. The displaced time $\tau = |t - T_c|$ is defined symmetrically around $T_c = 256$ minutes on 01/11/2002. (b) $\log N_b(\tau)$ and $\log N_a(\tau)$ are linear with $\log \tau$ over two orders of magnitude on a logarithmic scale. The Omori parameters in Eq. (4.5) calculated from $N(t)$ are $\Omega_b = 0.09 \pm 0.01$, $\alpha_b = 0.21 \pm 0.01$ and $\Omega_a = 0.32 \pm 0.01$, $\alpha_a = 0.81 \pm 0.01$.

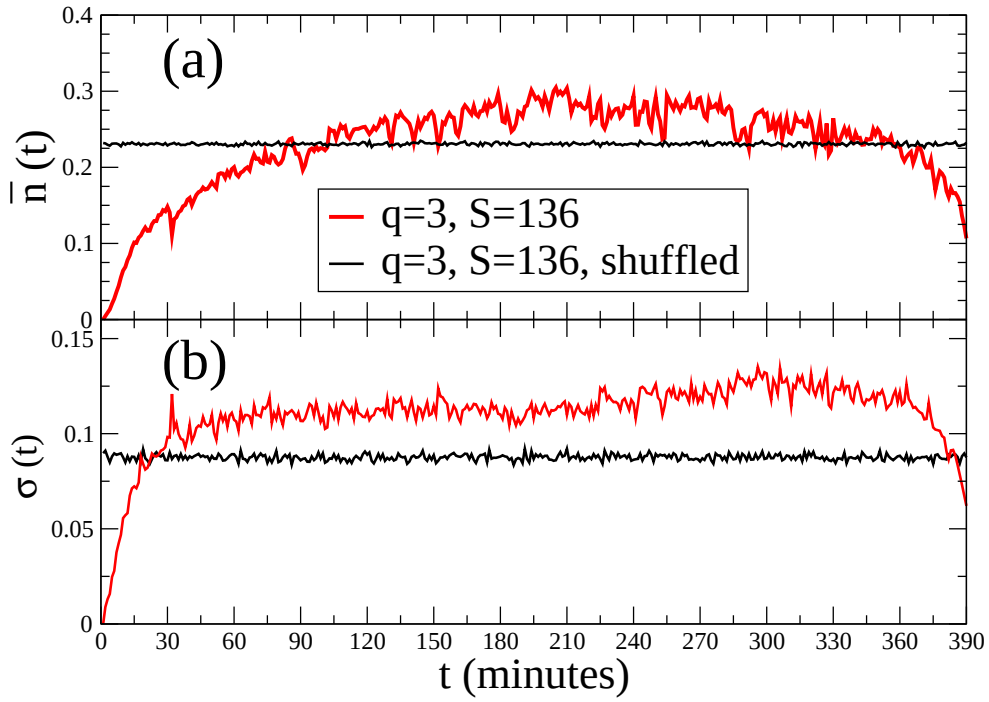

Figure 4.3: The fraction $n(t)$ of the market above the volatility threshold q is non-stationary through the trading day. We plot in (a) the average daily trading pattern $\overline{n}(t)$ for $S = 136$ stocks and in (b) the corresponding standard deviation, to demonstrate the trends we remove in the normalized quantity $n'(t)$. In practice, we use the smoothed average of these curves in order to diminish statistical fluctuations on the minute-to-minute scale. For comparison, we compute $\overline{n}_{sh}(t) \approx 0.23$ and $\sigma_{sh} \approx 0.09$ for shuffled $v_i(t)$. The values of $\overline{n}(t)$ provide an estimate for the background market co-movement that can be attributed to random fluctuations.

power-law response curves are distinguishable, especially if several orders of magnitude in τ is available.

Instead of analyzing $n(|t-T_c|)$, we perform our quantitative analysis on $N(|t-T_c|)$, the cumulative number of events above threshold q at time t minutes, where by definition

$$N(|t-T_c|) = \int_{T_c}^{t} n(|t'-T_c|)dt' \sim \beta(|t-T_c|)^{1-\Omega} \tag{4.6}$$

for market co-movement and

$$N^j(|t-T_c|) = \int_{T_c}^{t} n_j(|t'-T_c|)dt' \sim \beta_j(|t-T_c|)^{1-\Omega_j} \ . \tag{4.7}$$

for the activity of stock j. We perform our regression analysis on $N^j(|t-T_c|)$ because it is less noisy and more monotonic as compared to $n_j(|t-T_c|)$.

Hence, for a given day, we calculate the cumulative time series $N^j(t)$ from $n_j(t)$ for each stock j, where $t = 0$ corresponds to the opening bell at 9:30 AM ET. For comparison, we also analyze the average market response $N(t)$ of the S stocks under consideration, which complements the study of individual stocks.

To demonstrate our approach, in Fig. 4.1 we plot $V(t)$, $N(t)$ and also $N^j(t)$ for four single stocks on 01/11/2002, a day when there was a large market shock corresponding to a publicized comment by the Fed chairman Alan Greenspan concerning economic recovery which occurred at approximately $T_c = 255$ minutes after the opening bell.

In order to compare the dynamics before and after the market shock, we first separate the intraday time series $N(t)$ into two time series $N_b(t|t < T_c)$, and $N_a(t|t > T_c)$. Then, to treat the dynamics symmetrically around T_c, we define the displaced time $\tau = |t - T_c| \geq 1$ as the temporal distance from T_c. As an illustration, we plot in Fig. 4.2 the time series on 01/11/2002 as a function of τ. We then employ a linear fit to find the τ dependence of both $N_b(\tau) \equiv N(T_c) - N(|t - T_c|)$ and $N_a(\tau) \equiv N(t - T_c) - N(T_c)$ on a log-log scale to estimate the Omori power-law exponents Ω_b before the news and Ω_a after the news. By analogy, we define α to be the amplitude $\alpha = \beta(1 - \Omega)$ before T_c as α_b and after the shock as α_a.

4.4 Method for Determining T_c

4.4.1 Calibration using FOMC announcements

We use $n(t)$ to quantitatively determine times T_c in which the market is moving together, possibly in response to a market shock. In Fig. 4.3 we plot the average daily pattern for $\overline{n}(t)$ and the standard deviation $\sigma(t)$. The values of $\overline{n}(t)$ and $\sigma(t)$ are not stationary, so we remove the daily trend from $n(t)$ by defining the detrended quantity $n'(t) \equiv (n(t) - \overline{n}(t))/\sigma(t)$. In order to distinguish significant moments of market co-movement from background fluctuations, we use a significance threshold which we obtain from the distribution of market activity over the entire data set analyzed. Hence, we analyze the quantity $x(t)$ defined as,

$$x(t) \equiv n(t)\frac{n(t) - \overline{n}(t))}{\sigma(t)} \ , \tag{4.8}$$

which is the product of $n(t)$ and $n'(t)$. The value of $n(t)$ quantifies the *size* of the market co-movement, while $n'(t)$ quantifies the *significance* of the market co-movement. Because $\overline{n}(t))$ is not constant during the day, we consider the normalized quantity $n'(t)$ in order to remove the intraday pattern. Then, to restrict our analysis to relatively large market co-movements, we eliminate times toward the beginning and end of each day, when average market activity is lower (significant morning activity is often related to overnight news [54]). We analyze the quantity $x(t)$, which is large only if both $n(t)$ and $n'(t)$ are large. Fig. 4.4 demonstrates how the quantity $x(t)$ is useful for amplifying market co-movement and provides an illustration of a significant shock with substantial preshock and aftershock dynamics.

We analyze the time series $x(t)$ in order to select the set of times $\{t\}$ of the market shocks that are large in the fraction of the market involved (large $n(t)$) as well as significant with respect to the time in which they occur (large $n'(t)$). We determine a significance threshold x_c from the probability density function (pdf) of $x(t)$ as in Fig. 4.5. As a null model, we shuffle the order of each intraday time series $v_j(t)$ and obtain a shuffled market volatility rate $n_{sh}(t)$ for each day. This preserves the empirical pdf of $v_j(t)$ but removes the correlations that exist in the temporal structure of $v_j(t)$. We also plot $\overline{n_{sh}}(t) \approx 0.23$ in Fig. 4.3 which corresponds to the fact that there is a residual 0.23 co-movement due to random fluctuations. We compare

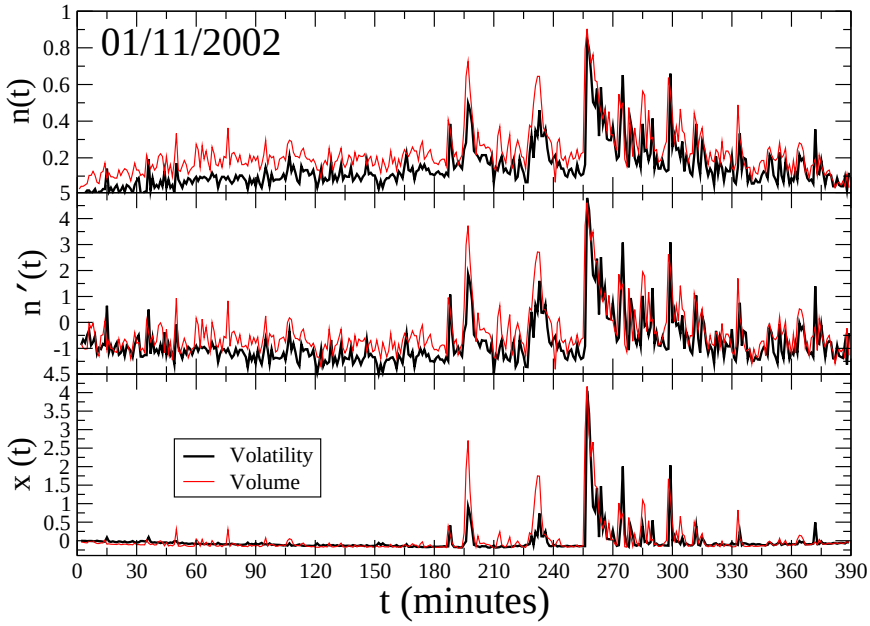

Figure 4.4: Example of market co-movement $n(t)$ in both price volatility and total volume, and the qualitative relationship between the quantities $n(t)$, $n'(t) \equiv (n(t) - \overline{n}(t))/\sigma(t)$, and $x(t) \equiv n(t)n'(t)$. The market shock on 01/11/2002 occurred at $T_c = 256$ in response to a public comment by the Fed chairman A. Greenspan concerning economic recovery [53].

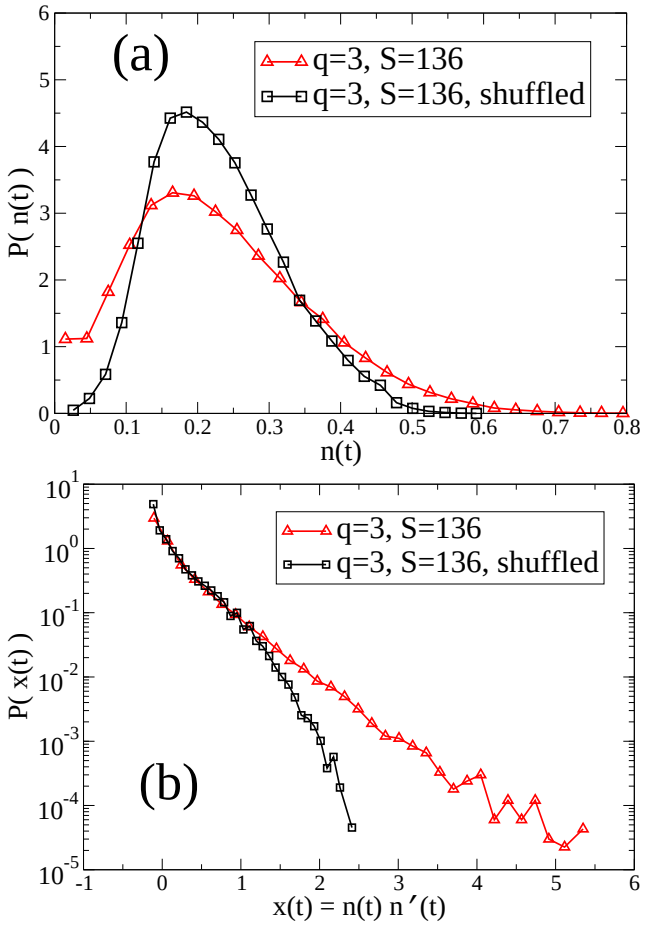

Figure 4.5: Using the volatility threshold $q = 3$ and $S = 136$ stocks, we determine the market co-movement threshold x_c from the pdf of $x(t) \equiv n(t)n'(t)$. (a) The pdf for the 190,000 minutes analyzed of the volatility rate $n(t)$ corresponding to the fraction of the market with volatility $v_i(t) > q$. (b) The pdf of $x(t)$, where in the quantity $x(t)$ we have removed the average daily trend of $n(t)$, so that $x(t)$ is relatively large when market co-movement is large and significant. For comparison, we also plot the pdf of $x_{sh}(t)$ computed from randomly shuffled volatility time series $v_i(t)$. We find a divergence between the pdf of $x(t)$ and of $x_{sh}(t)$ for $x > 1.0$, which we define as the co-movement threshold $x_c \equiv 1$ in our analysis.

the pdfs for $x(t)$ and $x_{sh}(t)$ in Fig. 4.5(b), and observe a significant divergence for $x(t) > 1$.

We calibrate our method for determining T_c from candidate cascades by using the known reported values T corresponding to Fed announcements. We choose the value $x_c = 1.0$ which reproduces with the best accuracy the values of T that we provide for comparison in Table 4.1. The value of $x_c = 1.0$ results in $5,804$ minutes out of $190,000$ minutes analyzed for which $x(t) > x_c$, or roughly 3% of the 2-year period with significant market co-movement.

4.4.2 Determining T_c from candidate cascades

In a typical trading day there are many large fluctuations, for both individual stocks and indices such as the S&P 500 and DOW. This fact is evident in the robust probability density function of volatility which has a stable power-law tail for a wide range of time scales ranging from 1-minute to several days [2, 5, 56, 38]. We select market cascades that are above a "spurious fluctuations" threshold, which we define by randomizing the order $v_i(t)$. We use the corresponding shuffled values $n_{sh}(t)$ as a proxy for background noise.

We find on average approximately 12 minutes per day above the threshold $x_c \equiv 1.0$. So here we develop a method for selecting the most likely time T_c from all candidate times with $x(t) > x_c$. For a given day, we collect all values of $x(t) > x_c$ into a subset $\{x'(t)\}$ of size z. From this subset, we divide the z values into k cascades $\{x'(t)\}_i$, which we define as localized groups using the criterion that a cascade ends when the time between the last x' in cascade i is separated from the first x' in cascade $i+1$ by a time window greater than $\Delta t \equiv 60$ minutes. We next assign to each cascade group $\{x'(t)\}_i$ a weight equal to the sum of the $x'(t)$ values belonging to the given cascade group, and select the cascade group with the largest weight as the most significant cascade. Within the most significant cascade group, we choose the time corresponding to the maximum value of $x'(t)$ as the time T_c of the main shock. We calibrate this method using the reported times for the 19 FOMC interest rate meeting announcements, and find that the values $\Delta t \equiv 60$ and $x_c = 1.0$ best reproduce the known set $\{T\}$, which we provide for comparison in Table 4.1.

Using the parameter values $x_c = 1.0$ and $\Delta t \equiv 60$, we find 373 days with market

FOMC Date	R_{new} (%)	ΔR	$\frac{\Delta R}{R_{old}}$	T	T_c
01/03/01**	6	-0.5	-0.077	210	227
01/31/01	5.5	-0.5	-0.083	285	290
03/20/01	5	-0.5	-0.091	285	286
04/18/01**	4.5	-0.5	-0.100	90	88
05/15/01	4	-0.5	-0.111	285	287
06/27/01	3.75	-0.25	-0.063	285	285
08/21/01	3.5	-0.25	-0.067	285	286
09/17/01**	3	-0.5	-0.143	0	16
10/02/01	2.5	-0.5	-0.167	285	288
11/06/01	2	-0.5	-0.200	285	292
12/11/01	1.75	-0.25	-0.125	285	287
01/30/02	1.75	0	0.00	285	289
03/19/02	1.75	0	0.00	285	293
05/07/02	1.75	0	0.00	285	287
06/26/02	1.75	0	0.00	285	286
08/13/02	1.75	0	0.00	285	291
09/24/02	1.75	0	0.00	285	291
11/06/02	1.25	-0.5	-0.286	285	286
12/10/02	1.25	0	0.00	285	295

Table 4.1: Comparison of announcement times T (as reported in New York Times) with the market clustering times T_c calculated using a threshold $x_c = 1.0$, cascade window $\Delta t = 60$ min., and $S = 136$ stocks. The value of $x(T_c)$ corresponds to the largest value out of all the candidate $\{x\}$ in the most significant cascade of the particular day. Dates of 19 FOMC meetings in the 2-year period between Jan. 2001 – Dec. 2002, where the Federal Funds Target rate (R_{new}) was implemented by the rate change (ΔR) at (T) minutes after the opening bell at 9:30 AM ET. The absolute relative change $|\frac{\Delta R}{R_{old}}| \equiv |\Delta R(t)/R(t-1)|$ has typically filled the range between 0.0 and 0.25. Note: Date** refers to *unscheduled* meetings, in which the announcement time did not correspond to 2:15 PM ET ($T = 285$ minutes).

shocks, out of 495 days studied. If the values of $x'(t)$ were distributed uniformly across all days, then the probability of finding 122 days without one $x'(t)$ is vanishingly small, which confirms that the $x'(t)$ group together forming cascades. We remove all days where T_c is within 90 minutes of opening ($t = 0$) or closing ($t = 390$), and all T_c that occur on half-days (days before or after the 4th of July, Thanksgiving or Christmas), resulting in the data set $\{T_c\}$ constituting 219 individual days.

Furthermore, in order to test the dependence of the data set $\{T_c^{(1)}\}$ found for the time resolution $\delta t = 1$ minute used in this thesis, we also compare the values of $\{T_c^{(5)}\}$ and $\{T_c^{(10)}\}$ found using a volatility series with $\delta t = 5$ min. and $\delta t = 10$ min. resolution, respectively (see Eq. (4.1)). For each of the 219 days with a T_c value we calculate the absolute difference in the time value $T_c^{(\delta t)}$ using two values of δt. We use similar values of x_c for each time resolution so that the number of days with market shocks for each resolution are approximately equal. The difference in $T_c^{(\delta t)}$ depends on the resolution δt and the locality δT_c associated with each market shock. The average of the absolute differences for three values of δt are $|T_c^{(5)} - T_c^{(1)}| = 9$ minutes and $|T_c^{(10)} - T_c^{(1)}| = 15$ minutes. We estimate the standard error for a particular time resolution $\delta T_c^{(\delta t)} \approx 2\delta t$, which implies that $\delta T_c^{(1)} \approx 2$ min. for the 1-min. time resolution. Hence, the use of $T_c \pm \delta T_c$ does not significantly change the results of this section. In the next Section, we analyze the empirical laws that quantify the response dynamics both before and after significant market shocks.

4.5 Results

The analysis performed in this section is largely inspired by the analogies between financial market crashes and earthquakes. A recent study finds significant evidence of Omori power-law relaxation both before and after common FOMC interest rate announcements [52]. The dynamics before the announcements, which are regularly scheduled and pre-announced, are consistent with market anticipated surprise in the Fed news, while the dynamics after the announcements are related to the perceived surprise in the Fed news. We use the relationship between the overnight Effective rate and the U.S. 6-month Treasury Bill to estimate the magnitude of the financial news shock. Here we identify all cascades that meet our significance criterion, and analyze

the dynamics both before and after T_c, using the framework developed in earthquake research [57, 58, 59, 60, 61, 62, 63]. We estimate the magnitude M of each market shock where $M \equiv \log V(T_c)$.

Closely related to the Omori relaxation of aftershocks is the productivity law, which establishes a power-law relationship between the number of aftershocks or preshocks that follow or precede a main shock of magnitude M. To the best of our knowledge, this is the first analysis of financial markets to analyze the productivity law, where we use $M \equiv \log(V(T_c))$. This is analogous to earthquake analysis, where $M = \log E$ and E is the energy associated with the stress released by the main shock. We justify our analogy between market volatility V and earthquake energy E by comparing the cumulative distribution

$$P(V > s) \sim s^{-\eta_V} \tag{4.9}$$

of volatility in financial markets with the cumulative distribution

$$P(E > s) \sim s^{-\eta_E} \tag{4.10}$$

of energy E in seismic earthquakes. Both cumulative distributions are asymptotically power laws, with $\eta_V \approx 3$ [56, 38] and the Gutenberg-Richter law $\eta_E \approx 1$ [60].

For the set of 219 market shocks we analyze, we find a wide range of $V(T_c)$, and hence a wide range of cascade dynamics. In Fig. 4.6 we plot the pdf of Omori parameter values $\Omega_{a,b}$ and $\alpha_{a,b}$ obtained from the power-law fits of $N_b(\tau)$ and $N_a(\tau)$. Figs. 4.6(a) and 4.6(b) show the distribution of parameter values calculated for the average market responses $N_b(\tau)$ and $N_a(\tau)$ corresponding to Eq. (4.6), while 4.6(c) and 4.6(d) show the distribution of parameter values calculated from the individual stock responses $N_b^j(\tau)$ and $N_a^j(\tau)$. The pdfs for individual stock values of Ω and α have a larger dispersion, as the response to each market shock is not uniform across all stocks. For the average market response $N(\tau)$ in Figs. 4.6(a) and 4.6(b) the pdfs of Ω and α are shifted to larger values for $t > T_c$ as compared to $t < T_c$. This is indicative of the stress that can build prior to anticipated announcements and the surprise that is inherent in the news. Larger Ω values correspond to faster relaxation times, while larger α values correspond to higher activity. We also observe $\Omega < 0$, which corresponds to particular time series in which the pre-shocks or after-shocks

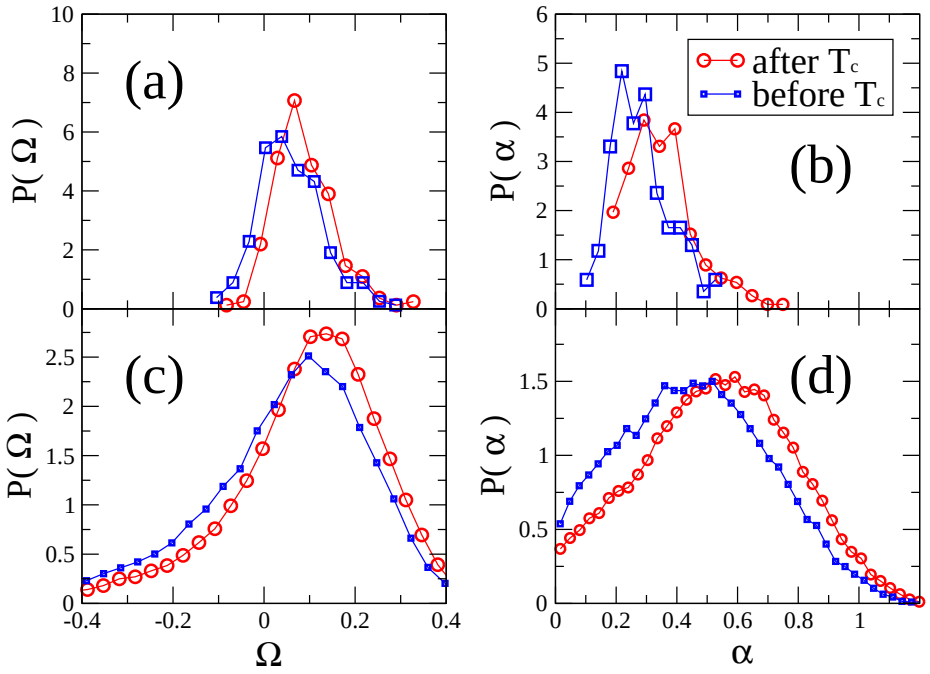

Figure 4.6: (a,b) Comparison of the probability density functions $P(\alpha)$ and $P(\Omega)$ of Omori parameters α and Ω computed from the average market response $N_{a,b}(\tau)$. (c,d) The analogous pdf plots computed from individual stock response $N^j_{a,b}(\tau)$. The average and standard deviation of each data set are (a) $\Omega_a = 0.09 \pm 0.07$, $\Omega_b = 0.06 \pm 0.07$ (b) $\alpha_a = 0.35 \pm 0.11$, $\alpha_b = 0.28 \pm 0.09$ (c) $\Omega_a = 0.08 \pm 0.20$, $\Omega_b = 0.03 \pm 0.22$ and (d) $\alpha_a = 0.53 \pm 0.25$, $\alpha_b = 0.46 \pm 0.24$. Values of both Ω_a and α_a are consistently larger than Ω_b and α_b, indicating that the response time after T_c is shorter than the activation time leading into T_c. However the response cascade after T_c has, generally, a larger amplitude.

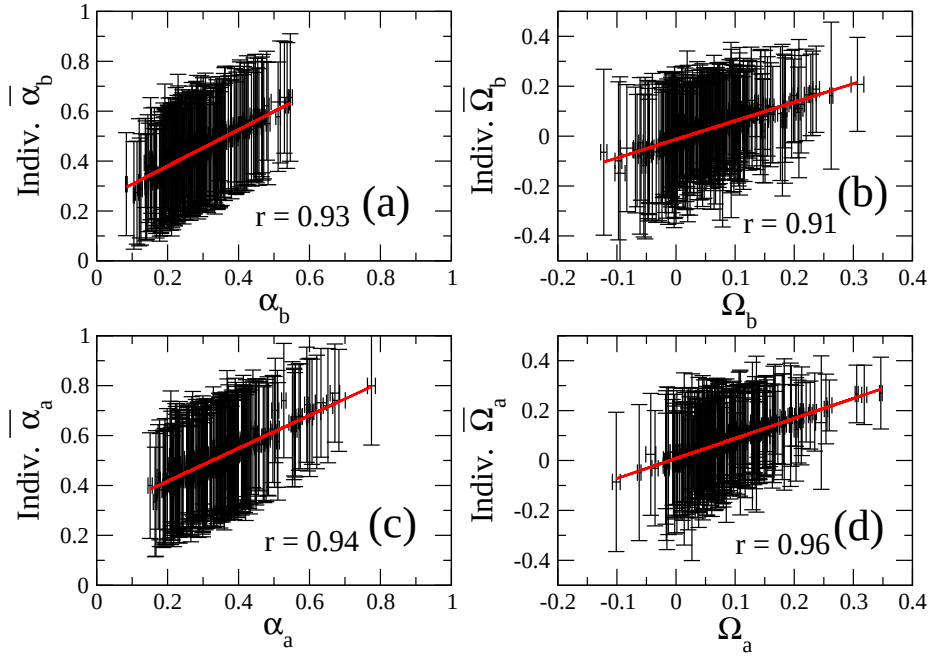

Figure 4.7: In order to account for the dispersion in the pdfs plotted in Figs. 4.6(c) and 4.6(d) for individual stocks, we compare the average values $\overline{\alpha}_{a,b}$ and $\overline{\Omega}_{a,b}$ computed from all $N^j_{a,b}(\tau)$ with the $\alpha_{a,b}$ and $\Omega_{a,b}$ computed from the corresponding average market response $N_{a,b}(\tau)$ for each of the 219 T_c. The visually apparent correlation indicates that the parameters quantifying $N_{a,b}(\tau)$ are a good representation of the average $N^j_{a,b}(\tau)$. The correlation coefficient r for each linear regression is provided in each panel.

farther away from the main shock (for large τ) are dominant over the volatility cascade around T_c. The values of the Omori parameters we find on averageing over all market shocks are given in the figure caption.

Although there is a wide distribution of Omori parameter values when considering all 219 market shocks, there is a strong correlation between the individual stock dynamics for a given market shock. In Fig. 4.7 we relate the values of α and Ω calculated for the average market response to the average and standard deviation of α and Ω calculated for individual stocks for a given T_c. The strong correlation between these quantities over 219 different dates indicates that the dispersion in the values of α and Ω for individual stocks, as demonstrated in Figs.4.6(a) and 4.6(c), result from the broad range of magnitudes of $V(T_c)$, and further, that the dispersion does not result merely from the range of stocks analyzed.

In Fig. 4.8 we plot the relation between the magnitude M of each main shock and the resulting Omori exponents $\Omega_{a,b}$ calculated from both market $N_{a,b}(\tau)$ and individual stock $N^j_{a,b}(\tau)$ response curves. Figs. 4.8(a) and 4.8(c) show a positive relation between M and the decay exponent Ω_a, which indicates that the market responds faster to large shocks on the intraday time scale. Figs. 4.8(b) and 4.8(d) show a significant dispersion across all stocks for a given date. Interestingly, we find a crossover at $M_\times \approx 0.5$ above which $\Omega_{a,b}$ increases sharply to positive values. The values $\Omega \approx 0$ for $M < M_\times$ correspond to a dynamical cascade $n(\tau)$ that is indistinguishable from an exponential decay. Typically, small values of Ω correspond to stocks with relatively low trading activity which are less sensitive to market shocks. For individual stocks, we define M to be the logarithm of the largest volatility within $\Delta t \equiv 3$ minutes of the main shock T_c measured for the average market response $N(\tau)$. This accounts for the possibility of a stock-specific anticipation or delay time in the volatility as a result of the mainshock $V(T_c)$. There is also the possibility that a spurious value of $\Omega \approx 0$ can arise from a stock which has high levels of activity throughout the entire time period analyzed.

In Fig. 4.9 we plot the relation between the magnitude M and the Omori-law amplitude $\alpha_{a,b}$ for both market $N_{a,b}(\tau)$ and individual $N^j_{a,b}(\tau)$ response curves. Interestingly, the relation between α and M is stronger, with less residual error than the relation between Ω and M, even for α_b, indicating a higher information content in

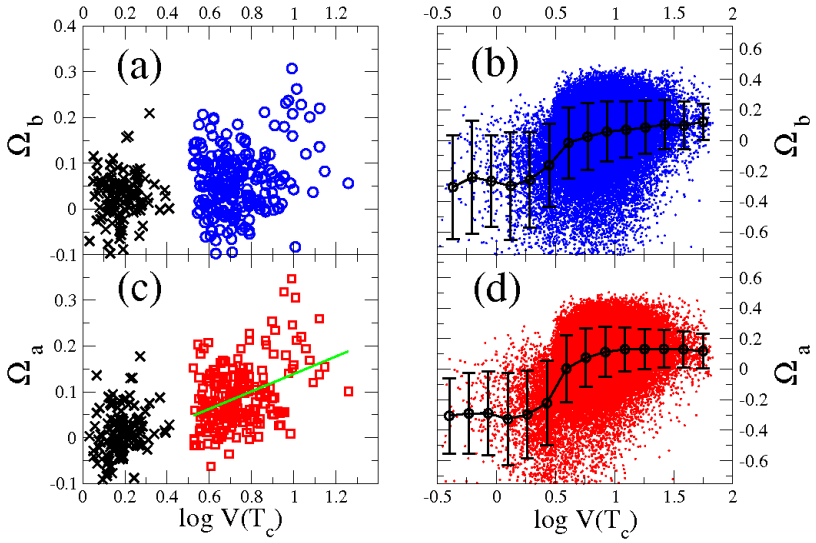

Figure 4.8: The relation between the magnitude $M \equiv \log V(T_c)$ and the Omori exponents $\Omega_{a,b}$. In panels (a) and (c) we compare values calculated from the average market response $N_{a,b}(\tau)$, and in panels (b) and (d) we compare values calculated from individual stock response $N^j_{a,b}(\tau)$. (a) Weak relation before T_c, where we validate the linear regression model at $p = 0.001$ significance level, but with correlation coefficient $r = 0.22$. The dispersion may result from the variability in anticipation preceding the market shock at T_c. (c) The relation between Ω_a and M is stronger after T_c than before T_c, with linear regression significance $p \approx 0$, correlation $r = 0.40$, and regression slope $m = 0.19 \pm 0.03$. The increasing trend demonstrates that a faster response, quantified by larger Ω_a, follows a larger M. Data points in panels (a) and (c) denoted by the symbol x correspond to values of $\Omega_{a,b}$ calculated for randomly selected T_c on those 118 days analyzed without a single value of $x(t) > x_c$. In panels (b) and (d) there is much dispersion in the Ω values of individual stocks for given $V(T_c)$. However, the average trends demonstrate a significant crossover at $M_{\mathrm{x}} \approx 0.5$ from $\Omega_{a,b} < 0$ to $\Omega_{a,b} > 0$. The case of $\Omega < 0$ can occur when there is more volatility clustering for large τ than for small τ, whereas the case of $\Omega > 0$ occurs for large volatility cascading around $\tau \gtrsim 0$. This crossover could result from the difference between anticipated and surprise shocks at T_c.

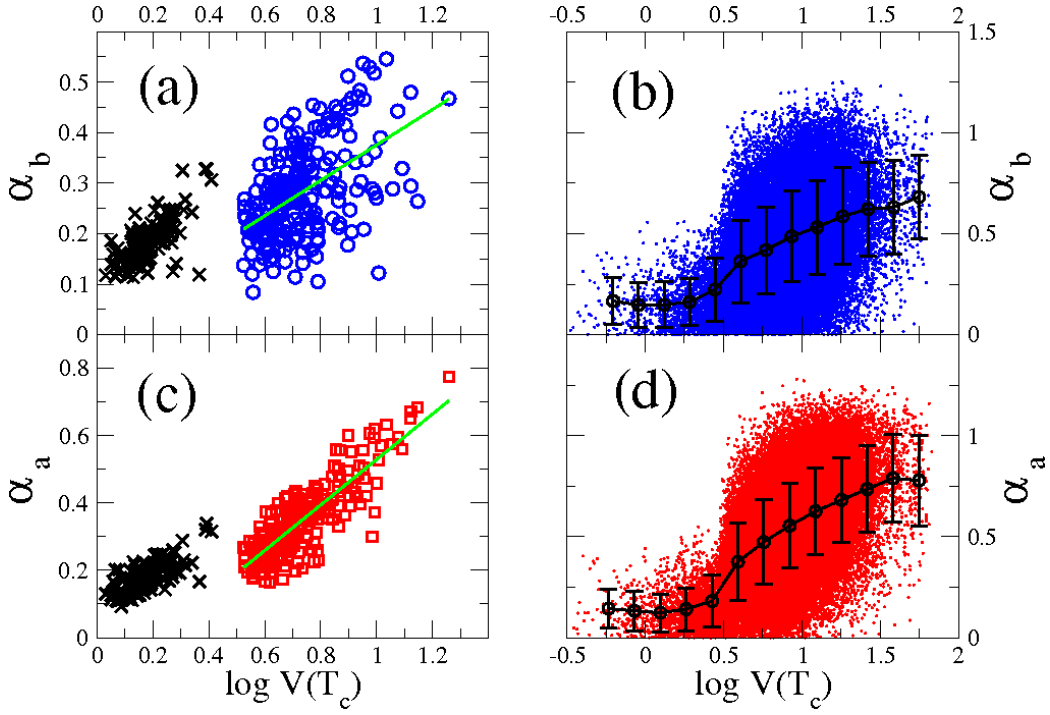

Figure 4.9: The relation between the magnitude $M \equiv \log V(T_c)$ and the Omori amplitudes $\alpha_{a,b}$. In panels (a) and (c) we compare the values calculated from the average market response $N(\tau)$ and in panels (b) and (d) we compare values calculated from individual stock response $N^j_{a,b}(\tau)$. (a) The increasing relation between α_b and M is statistically stronger than the relation between Ω_b and M in Fig. 4.8(a), with significance $p \approx 0$, correlation coefficient $r = 0.52$ and regression slope $m = 0.35 \pm 0.04$. (c) The relation between α_a and M is strong, with significance $p \approx 0$, $r = 0.84$, and regression slope $m = 0.68 \pm 0.03$. Data points in panels (a) and (c) denoted by the symbol × correspond to values of $\alpha_{a,b}$ calculated for randomly selected T_c on those 118 days analyzed without a single value of $x(t) > x_c$. The result that α increases with increasing $V(T_c)$ holds even for random times. In panels (b) and (d) there is much dispersion in the α values of individual stocks for given $V(T_c)$. However the average trends demonstrate a significant crossover at $M_\times \approx 0.5$ from $\alpha_{a,b} \approx 0.2$ for $M < 0.5$ to $\alpha_{a,b} > 0.2$ for $M > 0.5$. This crossover occurs at a similar location as the crossover observed in Figs. 4.8(b) and (d) for $\Omega_{b,a}$. The average amplitude value $\overline{\alpha}$ increases sharply for $M > M_\times$, consistent with first order phase transition behavior.

the amplitude of the Omori law. The strong relation for the average market response suggests that it is possible to identify precursors of market shocks with statistical certainty. However, since often T_c corresponds to anticipated market news, the significant activity prior to the main shock is a natural biproduct of trader anticipation. Interestingly, we also observe a critical threshold for $M_{\times} \approx 0.5$, above which the average response amplitude $\alpha_{a,b}$ increases suddenly, analogous to a first order transition.

In Fig. 4.10 we plot the relation between $V(T_c)$ and the productivity $P_a(\Delta t)$ (or $P_b(\Delta t)$), defined as the cumulative number of aftershocks (or preshocks) greater than the threshold $q \equiv 3$ within $\Delta t \equiv 90$ minutes of T_c. Motivated by the power-law relationship observed for earthquakes we fit the relations $P_a(\Delta t) \sim M^{\Pi_a}$ and $P_b(\Delta t) \sim M^{\Pi_b}$, and find statistically significant values for the market response $\Pi_b = 0.38 \pm 0.07$ and $\Pi_a = 0.48 \pm 0.04$, and for individual stocks $\Pi_b = 0.23 \pm 0.01$ and $\Pi_a = 0.25 \pm 0.01$. For earthquakes, [60] reports a range of $\Pi_a \approx 0.7 - 0.9$ values that are larger than observed here for financial markets, meaning that the productivity of physical earthquakes increases "faster" with main shock magnitude than does the productivity of market shocks. Since for earthquakes $\Pi_a < \eta_E$, this inequality establishes the relative importance of small fluctuations as compared to large fluctuations [60]. In other words, this inequality indicates that small earthquakes play a larger role than large earthquakes in producing the observed number of large earthquake shocks. Using an analogous argument for market volatility, since the cumulative distribution exponent $\eta_V \approx 3$ is robust across many markets [56, 38], then the total number $N_{Tot}(V)$ of aftershocks triggered by a main shock of size V,

$$N_{Tot}(V) = P(V)P_a(\Delta t) \sim 10^{(\Pi_a - \eta_V)\log V} \ , \tag{4.11}$$

is a decreasing function of V. Hence, we also find that aftershock cascades are controlled by the contributions of many smaller V. Thus, the medium-sized market shocks (analyzed here) play a larger role than the large market shocks in producing the observed heavy-tailed distribution of market shocks. We further note that the productivity is a combination of the relationships of both α and Ω with $V(T_c)$, which can be written as

$$P_a(\Delta t) \equiv N_a(\Delta t) \sim (\Delta t)^{1-\Omega_a} \ \alpha_a/(1-\Omega_a) \sim V(T_c)^{\Pi_a} \ , \tag{4.12}$$

with equivalent relation before the shock for $P_b(\Delta t)$.

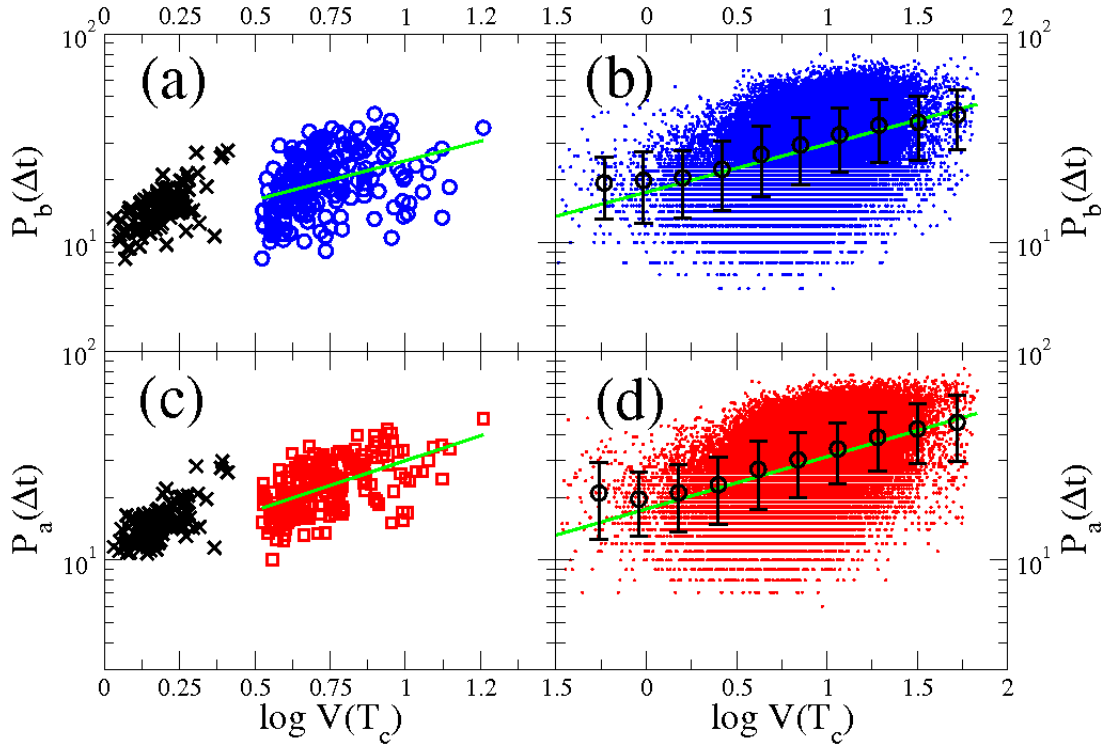

Figure 4.10: The increasing relation between the productivity $P_{a,b}(\Delta t)$ of each market shock and the size of the main shock $M \equiv \log V(T_c)$ with $\Delta t \equiv 90$ min. As is found in earthquakes, we find a power-law relationship between M and $V(T_c)$ described by a productivity exponent Π_b before and exponent Π_a after the market shock. Data points in panels (a) and (c) denoted by the symbol x correspond to values of $P_{a,b}(\Delta t)$ calculated for randomly selected T_c on those 118 days analyzed without a single value of $x(t) > x_c$. The result that $P(\Delta t)$ increases with increasing $V(T_c)$ holds even for random times. For the average market response $N_{b,a}(\Delta t)$, we find (a) $\Pi_b = 0.38 \pm 0.07$ and (c) $\Pi_a = 0.48 \pm 0.04$. For the productivity of individual stocks corresponding to $N^j_{b,a}(\Delta t)$ we find (b) $\Pi_b = 0.23 \pm 0.01$ and (d) $\Pi_a = 0.25 \pm 0.01$. For comparison, the power-law exponent value pertaining to earthquake aftershocks is $\Pi_a \approx 0.7 - 0.9$ [60].

In Fig. 4.11 we plot the values Ω, α, and $P(\Delta t)$, both before and after the main shock at time T_c. Surprisingly, while there is little statistical relation between Ω_b and Ω_a, there is a strong relation between α_b and α_a as well as between $P_b(\Delta t)$ and $P_a(\Delta t)$, for both $\Delta t = 90$ and $\Delta t = 120$ minutes. This result could be of interest for volatility traders and options traders who would like to anticipate the market dynamics *after* an announcement, given the dynamics *before* the announcement.

In Fig. 4.12 we relate the size of the largest shock $V_1 \equiv V(T_c)$ to the sizes of the second largest shock V_2, both before and after T_c. The Bath law parameter B quantifies the relation between V_1 and V_2 as

$$M_1 - M_2 = \log V_1 - \log V_2 = B \ . \tag{4.13}$$

This functional form implies the relation

$$V_2/V_1 = C_B \tag{4.14}$$

and hence $B = -\log C_B$. Fig. 4.12(c) is a scatter plot of V_1 and $V_{2,a}$ which shows a linear relation corresponding to $B_a = -\log(0.90) = 0.046$. Surprisingly, Fig. 4.12(a) also shows a strong relation between V_1 and $V_{2,b}$ with $B_b = -\log(0.81) = 0.092$. Comparing the values of B_b and B_a, the difference between the V_1 and V_2 is smaller after T_c than before T_c. Interestingly, both B_b and B_a are significantly smaller than the value $B_E \approx 1.2$ observed for earthquake aftershocks [59], meaning that the largest preshock and aftershock are of comparable magnitude to the main shock. This significant difference between earthquakes and market shocks is largely due to the relative probabilities of observing first and second-largest events x_1 and x_2. The conditional probability $P(x_1|x_2) = P(x_1 > x_2)$ is given by the corresponding cumulative distribution function. Hence, using Eq. (4.9) and Eq. (4.10), the ratio of the conditional probabilities for E_1 and V_1 is

$$\frac{P(V_1|V_2)}{P(E_1|E_2)} = \frac{P(V_1 > V_2)}{P(E_1 > E_2)} \sim \frac{V_2^{-3}}{E_2^{-1}} \ , \tag{4.15}$$

which roughly explains the 10^2 factor difference $B_E \approx 10^2 B_V$.

We also compare the volatilities V_1 and V_2 for individual stocks in Fig. 4.12(b) before T_c and in Fig. 4.12(d) after T_c. We compute the average value $\langle V_2 \rangle$ for linear bins, and find $V_1 > \langle V_2 \rangle$ for $V_1 > 20$, both before and after T_c. Also, Fig. 4.12 shows

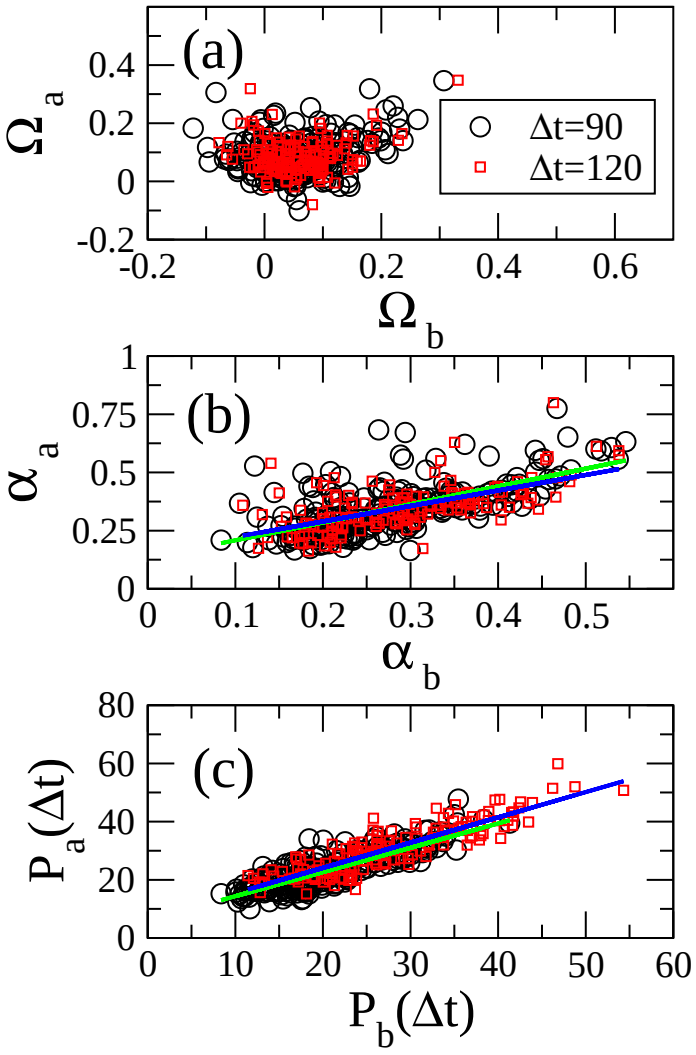

Figure 4.11: A comparison of Omori parameters before and after T_c for $N(\tau)$ and varying Δt indicate that α_b and $P_b(\Delta t)$ are better conditional estimators for the dynamics after T_c. (a) Weak relationship between Ω_b and Ω_a for $\Delta t = 90$ and 120. (b) Strong relationship between α_b and α_a for $\Delta t = 90$ and 120, with both linear regressions passing the ANOVA F-test at the $p < 0.001$ confidence level. (c) Strong relationship between $P_b(\Delta t)$ and $P_a(\Delta t)$ for $\Delta t = 90$ and 120 min. at the $p < 0.001$ confidence level.

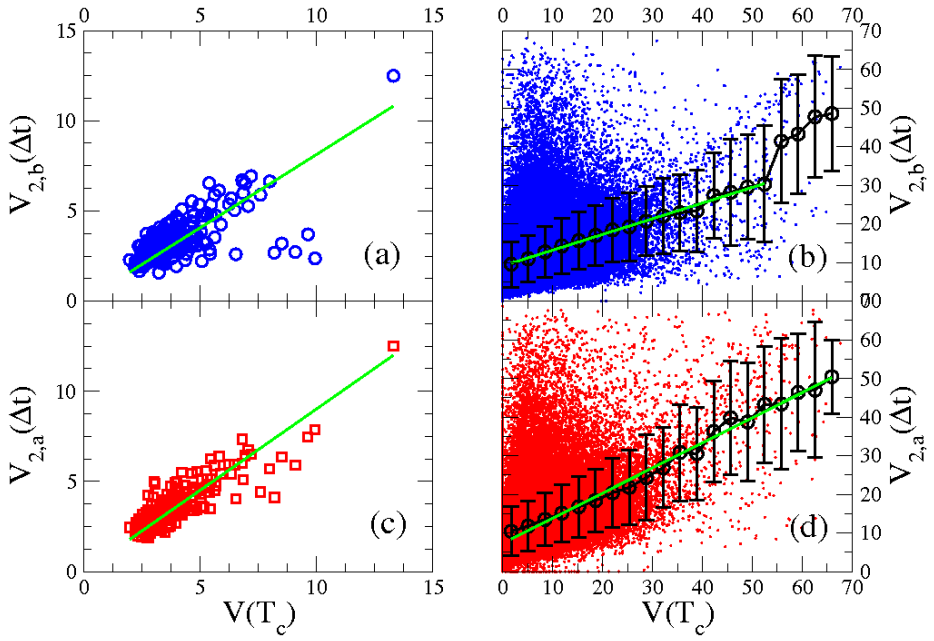

Figure 4.12: The increasing relation between the size of the main shock $V(T_c)$ and the size of the second largest aftershock (or preshock) $V_2(\Delta t)$ within Δt minutes of T_c demonstrates that the volatility of the largest aftershock (or preshock) increases with mainshock volatility. As with the Bath law for earthquakes, we observe a proportional relation $V_{2,a}(\Delta t) \equiv C_B V(T_c)$ which corresponds to a Bath parameter $B = -\log C_B$. For the average market response $N_{b,a}(\Delta t)$ we calculate C_B for (a) the dynamics before, $C_B = 0.81$ with correlation coefficient $r = 0.70$ and $\chi^2 = 212$, and for (c) the dynamics after $C_B = 0.9$ with $r = 0.87$ and $\chi^2 = 109$. For the Bath law corresponding to individual stocks we find that a linear function best fits the relation between $V(T_c)$ and the average value $\overline{V}_2(\Delta t)$ calculated for equal-sized bins as indicated by circles with one standard deviation error bars. We calculate the regression slope for the Bath law (b) before is $m = 0.65 \pm 0.02$ and (d) after is $m = 0.40 \pm 0.01$

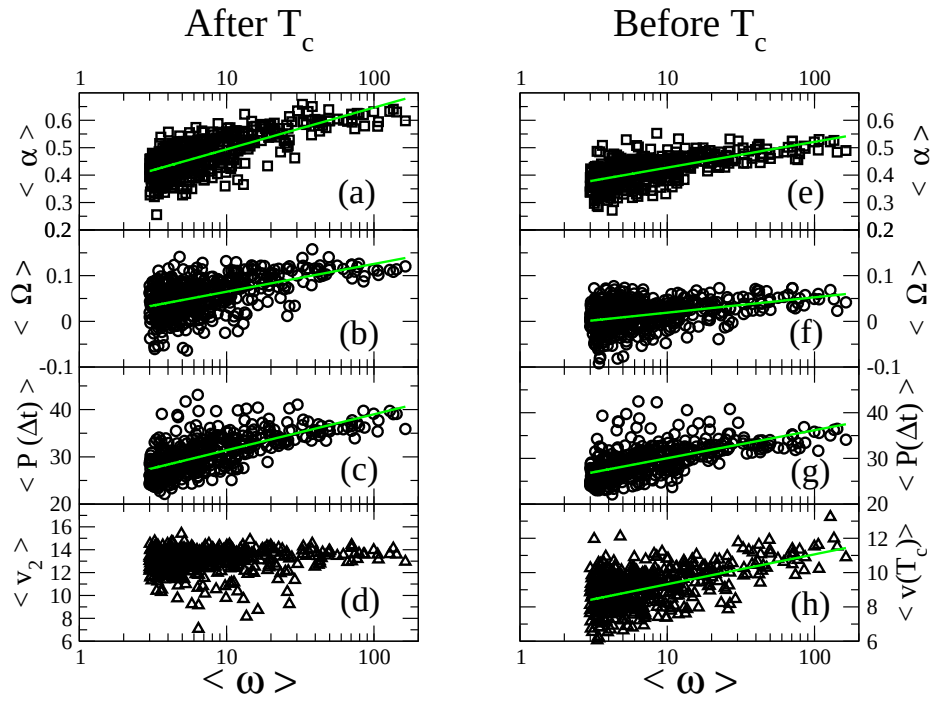

Figure 4.13: Relations between individual stock trading activity and dynamic response parameters (a-d) after T_c and (e-h) before T_c, averaged over all the days with a market shock. We measure the trading activity $\langle\omega\rangle$ for each stock, defined as the average number of transactions per minute over the 2-year period 2001-2002. We find that stocks with large trading activity react both more strongly (larger α and larger $P(\Delta t)$) and quickly (larger Ω) to market shocks. However, panel (d) shows that there is little relation between $\langle\omega\rangle$ and the average size of the largest aftershock $\langle v_2\rangle$.

that $\langle V_{2,a}\rangle > \langle V_{2,b}\rangle$ for most values of $V(T_c)$. Hence, the reaction to surprise causes larger volatility fluctuations than the anticipation of surprise.

We further ask the question, how the response parameters analyzed here depend on the variations between individual stock trading patterns. To answer this question, we quantify the trading capacity of each stock by $\langle\omega\rangle$, the average number of transactions per minute, with $3 \leq \langle\omega\rangle \leq 163$ for the $S = 531$ stocks analyzed. We hypothesize that $\langle\omega\rangle$ is closely related to firm size and market impact. Fig.4.13(a) shows that $\langle\alpha\rangle, \langle\omega\rangle$ and $\langle P(\Delta t)\rangle$ after T_c increase with $\langle\omega\rangle$, indicating that stocks with a large trading base respond to market shocks with large volatility $\langle v(T_c)\rangle$ (shown in Fig.4.13(b)), but also relax more quickly, corresponding to larger Ω values. However, we find no

statistical relation between $\langle\omega\rangle$ and $\langle v_{2,a}\rangle$. Interestingly, Fig. 4.13(b) shows that this positive relation also applies to the dynamic response parameters before T_c.

4.6 Discussion

Cascading avalanche dynamics are a common phenomena in complex systems ranging in scale from solar flares [57, 58] and earthquakes [8, 9, 61, 62] to microscopic vortices in turbulent fluids [64]. Similar bursting phenomena are also observed in human organs, such as the heart [65, 66], lungs [67, 68], and brain [69, 70, 71], and also for common social [35, 72, 73, 74] and economic systems [10, 11, 33, 39, 40, 41, 46, 51, 52]. Neural avalanches in the brain are frequent even in the resting state, and are a signature of healthy brain functioning within the neural network. In fact, the ability to process and disseminate information is largely attributed to the network structure of neuronal correlations which, if inhibited by disease, lead to altered disfunctional states such as in the case of schizophrenia. Extending by analogy, the frequency of cascades in financial markets could also be viewed as a "healthy" optimal state for processing information and eliminating arbitrage among the many the degrees of freedom. Recent work [75] on the switching dynamics around highs and lows in finanancial time series shows further evidence of Omori power-law scaling before and after microtrend extrema, in analogy to the market shocks at T_c developed here. Interestingly, this work on switching dynamics finds cascading trends on time scales ranging from seconds to hundreds of days.

Financial markets are subject to constant information flow, resulting in a large rate of significant events, such as quarterly earnings, splits and dividends announcements, mergers and acquisitions, institutional reports. This information can arrive as "expected" or come as a "surprise". Interestingly, there are precursors extending more than a day in advance of expected announcements such as earnings announcements [40]. Economists have long been interested in the interplay between informed and uninformed traders, and the dissemination of information across a market consisting of rational agents. Early work focusses on the relationship between trading volume and price change, and the relationship between these quantities and the qualitative notions of surprise, importance, and precision of the information [41].

Using methods from statistical physics and geophysics, we analyze the absolute returns of price because of the long-memory property, and the universality of fluctuations in this quantity across diverse markets [37, 38, 38]. Ref.[41] postulates that price changes reflect the average change in market expectations, whereas trading volume reflects idiosyncratic reactions across all traders. Recent work further quantifies trading volume fluctuations and finds that they are similar to price fluctuations, and furthermore, finds significant cross-correlation between volume change and price change [24]. Omori relaxation dynamics are also shown for trading volume in [52]. Here we also observe significant volume cascading as evident in Fig. 4.4. The analysis of volume and transaction dynamics is an avenue of future research, and could highlight the relationship between volume and price fluctuations by studying their correlation around market shocks.

To summarize, we analyzed the cascade dynamics of price volatility, which has potential applications in options pricing and the pricing of other derivatives. The Black-Scholes equation in its simple form assumes that the fluctuations in the price are constant during the duration of the option [76]. However, more sophisticated methods [77] incorporate time dependent price volatility, and are more realistic descriptions of the non-stationarity of financial time series. The results in this section are of potential interest for traders modeling derivatives on short time scales around expected market shocks, e.g earnings reports, etc. The statistical regularity of both market and individual stock behavior before and after a market shock of magnitude $M \equiv \log V(T_c)$ provides information that could be used in hedging, since we observe a crossover in the cascade dynamics for $M \approx 0.5$. Knowledge of the Omori response dynamics provides a time window over which aftershocks can be expected. Similarly, the productivity law provides a more quantitative value for the number of aftershocks to expect. And, finally, the Bath law provides conditional expectation of the largest aftershock, given the size of the main shock, and even the largest preshock. Of particular importance, from the inequality of the productivity law scaling exponents and the pdf scaling exponent for price volatility, we find that the role of small fluctuations is larger than the role of extremely large fluctuations in accounting for the prevalence of aftershocks.

Part III

A Statistical Physics Approach to Quantifying Career Longevity

Chapter 5

Motivation

A typical course on history, art, music, politics, sports, etc. usually focusses on the underlying social and political driving forces that brought about an era, and such a course usually highlights some of the most influential individuals behind the movement. In any type of quantitative approach, such a course of study suffers from extreme statistics bias, since only the relatively rare but monumental events are considered. This is not really a failure of history or science, but rather a problem with data measurement and storage. It is inconceivable that the history of every individual be known, nor is it conceivable that even a tiny fraction every history be recorded. Hence, exhaustive studies of longitudinal careers have been impossible before the information age. With a limited storage capacity, our old technology of written and oral history was doomed to focus on just a few select individuals such as Plato, Da Vinci, Mozart, Napoleon, and Babe Ruth, who were no doubt all stellar in their own right. However, in a new era of data storage and manipulation, a plethora of longitudinal data encompassing individual careers is available.

When data are available on not only the stellar elite, but also an entire labor force, a very simple question arises: What is the distribution of success in a given profession? We are used to thinking of professional athletes as heroic larger-than-life images of refined physical prowess. However, these professional athletes are not playing for free; they are working, just like everybody else. Hence, the same social and economic forces that apply to any given labor market should also apply to professional sports. The same can be said for scientists who also work in a generic competitive arena of

Academia, where "publish-or-perish" is the name of the game.

So how rare were Babe Ruth and Einstein? Phrased in a more quantitative way: Is the distribution of success in such competitive arenas bimodal, where one (smaller) class of individuals corresponds to the greats, and another (larger) class of individuals corresponds to everyone else? A second alternative scenario for the distribution of success is that there are multiple modes of achievement. A third alternative is that there is a non-monotonic decreasing distribution of achievement, resulting in a smooth bridge between the successful "haves" and the unsuccessful "have-nots." Surprisingly, it turns out that this third scenario applies to the two competitive professions we analyze. Such a result is only possible when analyzing the entire labor force, even the relatively insignificant careers that lasted a very short time.

We analyze two competitive professions, professional sports and academia, where individual career data for both longevity and success is easily quantifiable and readily available. As a result, we find statistical scaling laws which quantitatively capture the empirical data. Using the empirical data and the Matthew effect (also known as the "rich-get-richer" effect) concept from sociology, we develop a stochastic model which describes the distribution of career longevity in a competitive labor force. We use master equation methods from statistical physics to solve the a Poisson process model for career progress, resulting in an exactly solvable prediction that depends on only two parameters. Due to the simplicity of the model and the relative freedom of our assumptions, this generic mechanism can conceivably be applied to other professions and even to other systems, such as the dynamics of friendship. And finally, aside from being a "fun" topic, we provide a unifying framework to relate the superstars in professional sports to the "mathletes" of science.

Chapter 6

Case Study: Career Longevity in Major League Baseball 1920–2000

6.1 Summary

Statistical analysis is a major aspect of baseball, from player averages to historical benchmarks and records. Much of baseball fanfare is based around players exceeding the norm, some in a single game and others over a long career. Career statistics serve as a metric for classifying players and establishing their historical legacy. However, the concept of records and benchmarks assumes that the level of competition in baseball is stationary in time. Here we show that power-law probability density functions, a hallmark of many complex systems that are driven by competition, govern career longevity in baseball. We also find similar power laws in the density functions of all major performance metrics for pitchers and batters. The use of performance-enhancing drugs has a dark history, emerging as a problem for both amateur and professional sports. We find statistical evidence consistent with performance-enhancing drugs in the analysis of home runs hit by players in the last 25 years. This is corroborated by the findings of the Mitchell Report [78], a two-year investigation into the use of illegal steroids in major league baseball, which recently revealed that over 5 percent of major league baseball players tested positive for performance-enhancing

drugs in an anonymous 2003 survey.

Baseball is a game of legends, mystique, euphoria and heartbreak. It is also a game of numbers and records. Here we analyze approximately 10,000 players who ended their careers between the years 1920 and 2000, where 1920 is the year widely considered as the beginning of the modern era of baseball. We utilize Sean Lahman's Baseball Archive [79], an exhaustive database consisting of Major League Baseball player statistics dating back to 1871. This database was meticulously constructed, going so far as to extract data from old newspaper reels. We find that baseball players have universal properties despite the distinct eras in which they played. Specifically, we find that the probability density functions of career totals obey scale-free power laws over a large range for all metrics studied. As usual, the probability density function $P(x)$ is defined such that the probability of observing an event in the interval $x + \delta x$ is $P(x)\delta x$. Power law density functions, $P(x) \sim x^{-\alpha}$, arise in many complex systems where competition drives the dynamics [3, 43, 80, 81, 82, 83, 84]. A key feature of the scale-free power law is the disparity between the most probable value and the mean value of the distribution [85]. For a Gaussian distribution, these two values coincide. However, with a power law, the most probable value $x_{mp} = 1$, while the mean value $\langle x \rangle$ diverges for $\alpha \leq 2$. Thus, in power law distributed phenomena, there are rare extreme events that are orders of magnitude greater than the most common events. This leads naturally to the notion of record events and the statistical analysis of sample extremes [86].

6.2 Metrics for career longevity

We begin this section with an analysis of career longevity in American baseball. Because the legacy of a player is based mainly upon his career totals, we also discuss the implications of the power-law behavior found in common career metrics. We conclude with empirical evidence, found in home run statistics, which is consistent with modern performance-enhancing factors including widespread use of performance-enhancing drugs.

In Fig. 6.1 we present the longevity of a player's career measured in at-bats (AB)

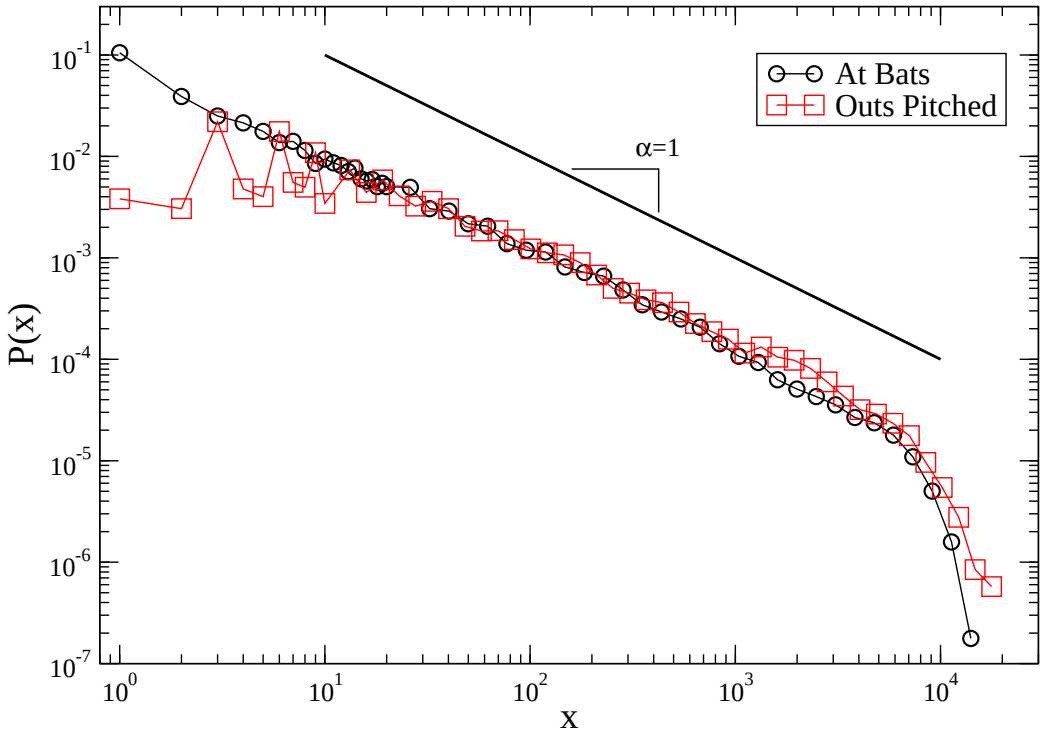

Figure 6.1: Probability density function of player longevity. Longevity is defined as the number of outs-pitched (pitchers), and the number of at-bats (all batters) for players ending their career in the years 1920-2000. Power law extends over more than three orders of magnitude, with $\alpha \approx 1$. For reference: straight line represents the power-law $P(x) \sim x^{-1}$.

and innings-pitched measured in outs (IPO). For these two metrics, we find truncated power-law distributions that range over three decades, marked by a sharp exponential cutoff at a value corresponding to around twenty seasons. It should be noted that unlike a complete power-law distribution with $\alpha \approx 1$, which has a divergent first and second moment, a truncated distribution has a definite mean and second moment. To our surprise, we find that the distributions for career longevity have their maxima around 1 appearance. This implies that most players who make it to the major leagues do not remain for very long, possibly making their professional debut and exit in a single pinch-hit or relief appearance. This leads to a perplexing feature of scale-free power laws, namely that it is just as hard to reach your 10th appearance from your debut appearance as a rookie as it is to reach your 10,000th appearance from your 1000th appearance as a seasoned veteran. In other words, the ratio $P(x_2)/P(x_1) = (x_2/x_1)^{-\alpha}$ depends only on the scale-free ratio of x_2/x_1 and the universal exponent α. This raises a fundamental question addressing longevity in American baseball: How is it possible that the same level of competition can eliminate some players after one appearance while sustaining others for more than two decades? American baseball has a 3-tier farm system, collectively known as the minor leagues. These developmental leagues filter talent up to the major leagues, with only the best players staying at the major league level. Occasionally there are opportunities for minor league players to be promoted to the major leagues for short unguaranteed stints, either if their major league affiliate has a roster vacancy due to injury or if their major league affiliate is not in a position to make the post-season. The long regular season provides ample opportunity for these major league tryouts, thus accounting for the high frequency of short careers.

In Fig. 6.2 we plot the distribution of career batting and pitching totals for all players who ended their careers between the years of 1920-1960 and 1960-2000 (we restrict our analysis to completed careers). Separating players into two subsets allows us to compare careers belonging to each era, where 1961 marks the beginning of the first expansion era in major league baseball. We also find truncated power-law behavior with exponent $\alpha \approx 1$ for all major career metrics. This should not be too surprising since each opportunity (defined in this section as an *at-bat* or *out-pitched*)

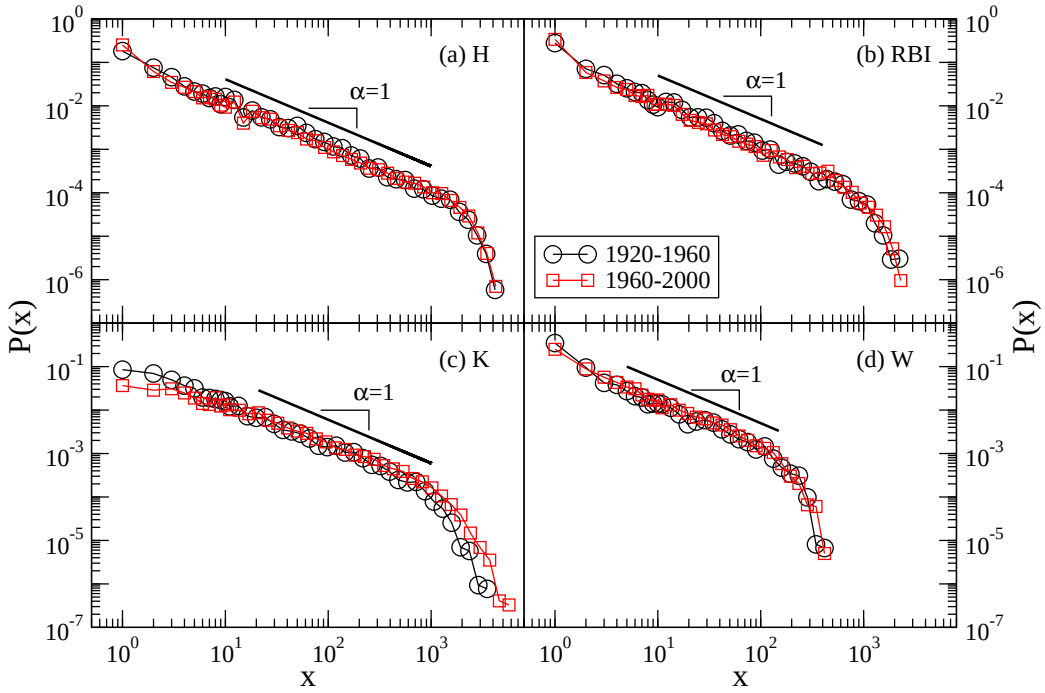

Figure 6.2: Probability density function of career statistics in four categories. (a) Hits, (b) Runs Batted In, (c) K (strikeouts), (d) Wins. Plotted for each statistical metric are the distribution of career totals for players whose career ended in the periods 1920-1960 and 1960-2000. The pairs of distributions are all qualitatively similar, with the exponential cutoffs occurring at the same critical value, indicating that the competition level in baseball has been relatively constant with respect to these career metrics.

is capitalized upon at a player's personal rate (defined in this section as his *prowess*); each success then contributes to the player's career statistical tally. Thus, the exponent from the career longevity power-law should carry over naturally into the density functions of career metrics [87]. In the case of batting statistics, we make no distinction between pitchers and other fielders who are on record for their at-bats. One can also do a statistical analysis on players who do not arise in the pitching database, but the distributions are not qualitatively different. Thus, career longevity measured in at-bats indicates that there is a large disparity between the "iron-horses" and the "one-hit wonders". It is perplexing that there is such a wide range of career lengths despite the typical prowess that distinguishes the upper echelon of baseball talent. It

should also be noted that in the game of baseball there are two classes of pitchers, those that start games, and those that finish games. Pitchers of the first type have routine schedules, pitching once every four or five games in a maintained rotation. Pitchers of the second type pitch more frequently, with game sessions that are shorter, hardly ever exceeding 2 innings (6 outs pitched). Despite these two classes of pitchers, the longevity measure of outs-pitched does not have any evidence of bimodal behavior. One can even notice the fluctuations in the beginning of the distribution for outs-pitched with sharp peaks corresponding to 1 inning (3 outs) and 2 inning (6 outs) stints. Comparing pitchers and batters, there is the remarkable similarity in power-law exponent corresponding to longevity, following from the fact that it is very difficult to reach, and to remain, at the major league level. Moreover, the distributions are nearly equivalent, with the exponential cutoff occurring at approximately the same value. This justifies both the 3000-hit and the 3000-strikeout benchmarks for both batters and pitchers, and suggest that career longevity results from a universal mechanism that is invariant with respect to player type. Baseball relies on precision play, requiring quick physical and mental reflex. The flow of the game is characterized by periods of lull, interlaced with bursts of activity, with both offense and defense capitalizing on sprinting, diving, and sliding plays [88]. In addition, throwing a baseball is very strenuous on the arm. Thus, although not a contact sport in the sense of hockey, rugby, or American football, baseball players are still prone to injury, some of which are career-ending. The perpetual hazard of career-ending replacement or injury provides the main ingredient for explaining the observed power laws. In Ref.[87] we propose a simple stochastic mechanism for career longevity in professional sports which reproduces both the power-law behavior and the exponential cutoff. We follow the explanation of Reed *et al.* [89], which shows that stochastic processes with exponential growth produce power-laws when the process is subject to random stopping times.

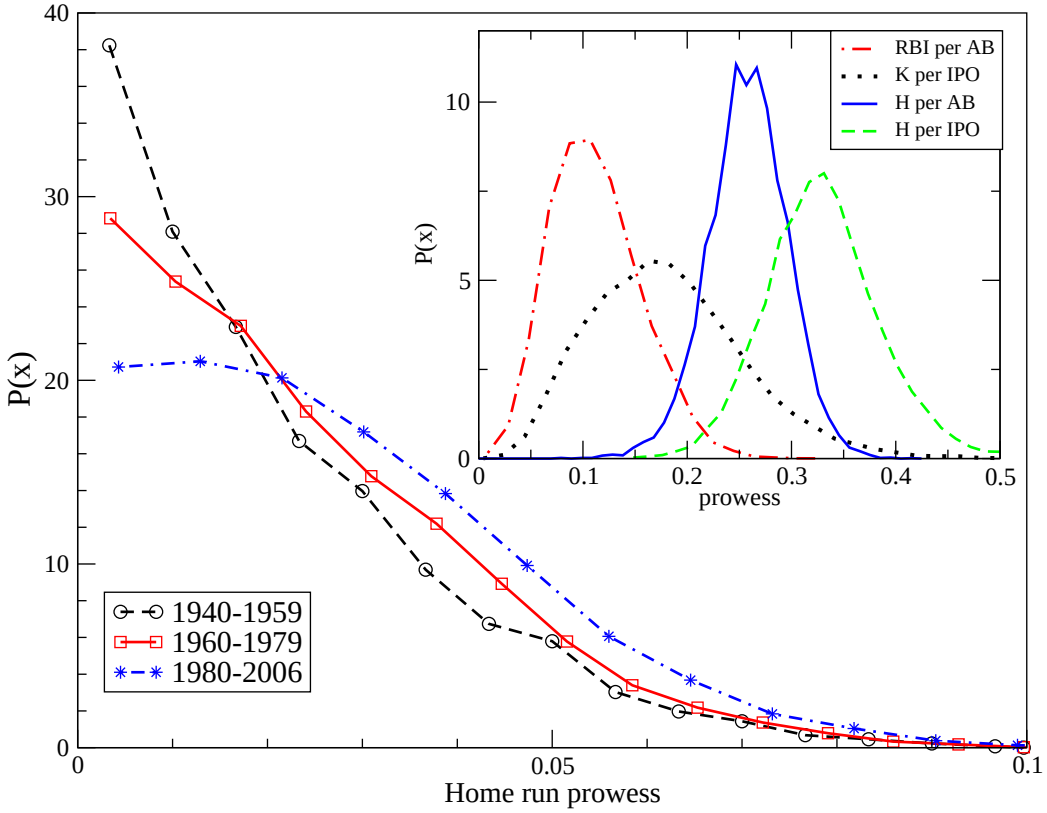

Figure 6.3: Probability density function of seasonal home-run prowess (pitchers are excluded from this analysis). The exponential distribution, representing players in the years 1920-1979, is skewed towards smaller rates, indicating that even at the major league level, the ability to consistently hit home runs is rare. Players from the last 25 years have increased their ability to hit home runs, possibly as a result of modern training regimens, performance-enhancing drugs, expansion-based dilution of talent, and other hypothetical factors. (Inset) Probability density function of seasonal prowess for several key metrics over the seasons 1920-2000. These are centered distributions with a mean $\langle x \rangle$, and standard deviation σ, that define the talent level in the major leagues.

In Fig. 6.3 we analyze seasonal home run prowess, defined as the rate per at-bat in which a particular player hits a home run. Pitchers are excluded from this analysis. We also restrict our analysis to players who exceed N appearances in a given season, and use $N = 100$ to eliminate statistical fluctuations arising from short-lived success. The seasonal prowess distributions for some common batting and pitching

metrics are relatively unskewed, defined by a characteristic standard deviation around a central mean (Fig. 6.3, inset). Thus, there is a typical success rate that defines not only the players, but also the *relative* level of competition between pitcher and batter at the major league level. In contrast, the seasonal prowess distributions for home runs are more exponentially distributed (Fig. 6.3). These distributions are skewed towards small values, indicating that it is rare for players to have prowess that consistently produces home runs. We also note that the distributions for home-run prowess over the past 26 years reveals a shift towards players with higher home-run ability. This increase in home-run prowess could result from modern *natural* weight-training programs with or without the use of performance-enhancing drugs. Other theories suggest that maple bats, a reduced strike-zone, and league expansion all contribute to the increased home run performance of modern players in the "Steroid Era". A recent study by R. Tobin [90] demonstrates that a reasonable increase in a player's muscle-mass, say a 10 percent increase, can produce a significant increase in home run production, ranging from 30-70 percent increase, depending on systemic parameters. Thus, our findings are consistent not only with the factual revelations of the Mitchell Report, but also with the aforementioned Monte-Carlo simulations.

6.3 Empirical evidence for inflation of home-run rates in MLB

It has been known for some time that home run rates over the last two decades have been rising [91], accompanied by home run records falling. In Fig. 6.4 we plot the average prowess of several metrics over 5 year windows from 1900-2005 in order to investigate the evolution of home run prowess. If in a single season player i has prowess $P_i = x_i/y_i$, then we compute the weighted average over all players

$$\langle P \rangle_T = \frac{\sum_i x_i}{\sum_i y_i} = \sum_{i \in T} w_i P_i$$

where

$$w_i = \frac{y_i}{\sum_i y_i}$$

The index i runs over all individual player seasons during the period T, and $\sum_i y_i$ is the total number of events y during the same period. The first era of increasing home-run prowess followed the 1920 revision of the rules (such as the outlaw of the "spitball") which made the batter and pitcher more equally competitive. This was followed by the emergence of sluggers such as Babe Ruth, who popularized the herculean feat of hitting home runs [92]. The first expansion era 1961-1969 saw 8 new teams, accompanied by a decrease in average home-run prowess. It is important to note that expansion within a league has two main effects. On the player level, expansion dilutes the talent in pitching and batting. This allows excellent players to take advantage of their weaker foe, and has been proposed as a possible factor responsible for the increased home run rate during the 1990's [93]. On the team level, the authors of Ref. [94] show that the level of team competition measured in team-versus-team upset probability increases during expansion eras. The second expansion era 1993-1998 saw 4 new teams, accompanied by an increase in average home-run prowess following approximately 20 years of stagnancy.

Because career statistics serve as key metrics for classifying players and establishing their historical legacy, we separate the players in Fig. 6.5 into two subsets, players ending their careers before and after 1980, in order to compare career home run totals. We find that the last 25 years account for many more players with large career home-run tallies. Interestingly, there is similar evidence in the strikeout tallies of pitchers (Fig. 6.2c), which suggests that modern sluggers may be "swinging for the fences" with reckless abandon. We should note that the difference in career statistics for strikeouts is confined to the highly sensitive exponential tail, whereas the differences in the career statistics for home runs extends into the bulk of the distribution. The use of steroids was most recently documented in the Mitchell Report [78], a two-year investigation into the use of performance-enhancing drugs in major league baseball. This report determines a *lower-limit* to the extent of steroid use in major league baseball at 5 percent, the results of a set of anonymous 2003 blood test that confirmed the widespread use of performance-enhancing drugs among major league players. Other Mitchell Report assessments, based on personal accounts, suggest much higher percentages of steroids use in professional baseball. Steroids and other performance-enhancing drugs can be used for two general reasons, to gain strength

and to reduce recovery-time from both workouts and injury. One might expect that performance-enhancing drugs would raise the level of competition across the board, for pitchers as well as batters, since both increased strength and speedy recovery can contribute to high career tallies. However, in our analysis of career statistics, we see evidence for a competitive advantage mainly in the case of home runs. This suggests that the level of competition between pitcher and batter is tipping in the favor of the batter.

Major league baseball is a unique sport, relying on team play while also maintaining a platform for individual play, namely pitcher versus batter. It also has a deep developmental minor league system that filters out the best talent, and serves as a emergency source for randomly depleted team rosters. This provides an explanation for the abundance of hitters and pitchers who experience both their debut and finale in the same game. In [87] we analyze career longevity in Korean baseball, American basketball, and English football. We find the same power-law behavior with exponential cutoff governing career statistics in all these professional sports, and provide a stochastic mechanism to explain the distribution of career length in competitive environments that are subject to random exit times. It should also be noted that performance-enhancing drugs are not a problem unique to American baseball. A separate study of English football also revealed widespread use of performance-enhancing drugs [95]. Moreover, it is not just a problem pertaining to professionals, but amateurs and adolescents as well [96], as performance-enhancing drugs are the core of a pandemic that not only poses personal health risk, but also places the integrity of sports in jeopardy [97, 98]. And finally, crossing over into the academic world, a recent study in the journal *Nature*[99] reveals that cognitive-enhancing drugs are prevalent in the sciences, and pose the same ethical questions that apply to accomplishments in sports [100].

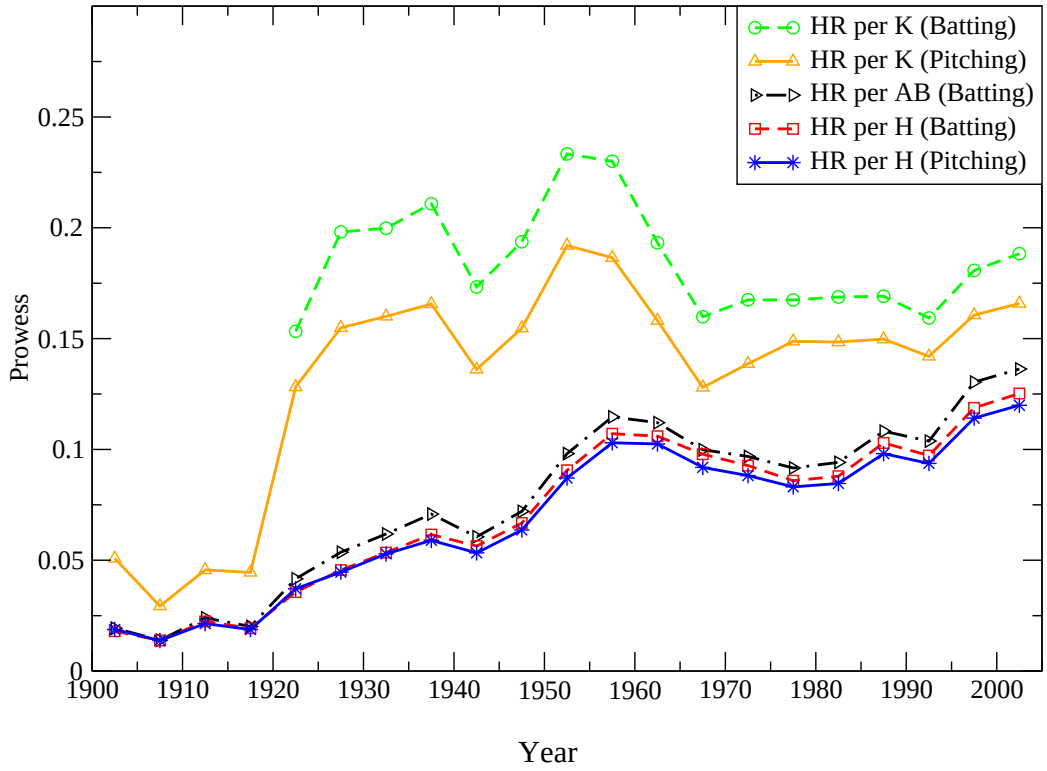

Figure 6.4: Average league prowess $\langle P \rangle$ calculated for 5-year periods starting in 1900. The trendline for HR/AB has been multiplied by a factor of 4 for clarity. The difference in the trendlines for Hr per Hit for batter and pitcher are essentially constant over the years. Following from the form of our weighted average, this suggests that the changes in home run prowess are not due to fluctuations in the talent pool arising from expansion dilution. If more home runs were being hit by veteran batters against poor pitching, then a spread would appear, assuming that poor pitchers don't stay at the major league level very long. Instead, we see that veteran hitters (with a large fraction of the total hits) are hitting more home runs off of veteran pitchers (with a large fraction of hits surrendered, and thus many innings pitched).

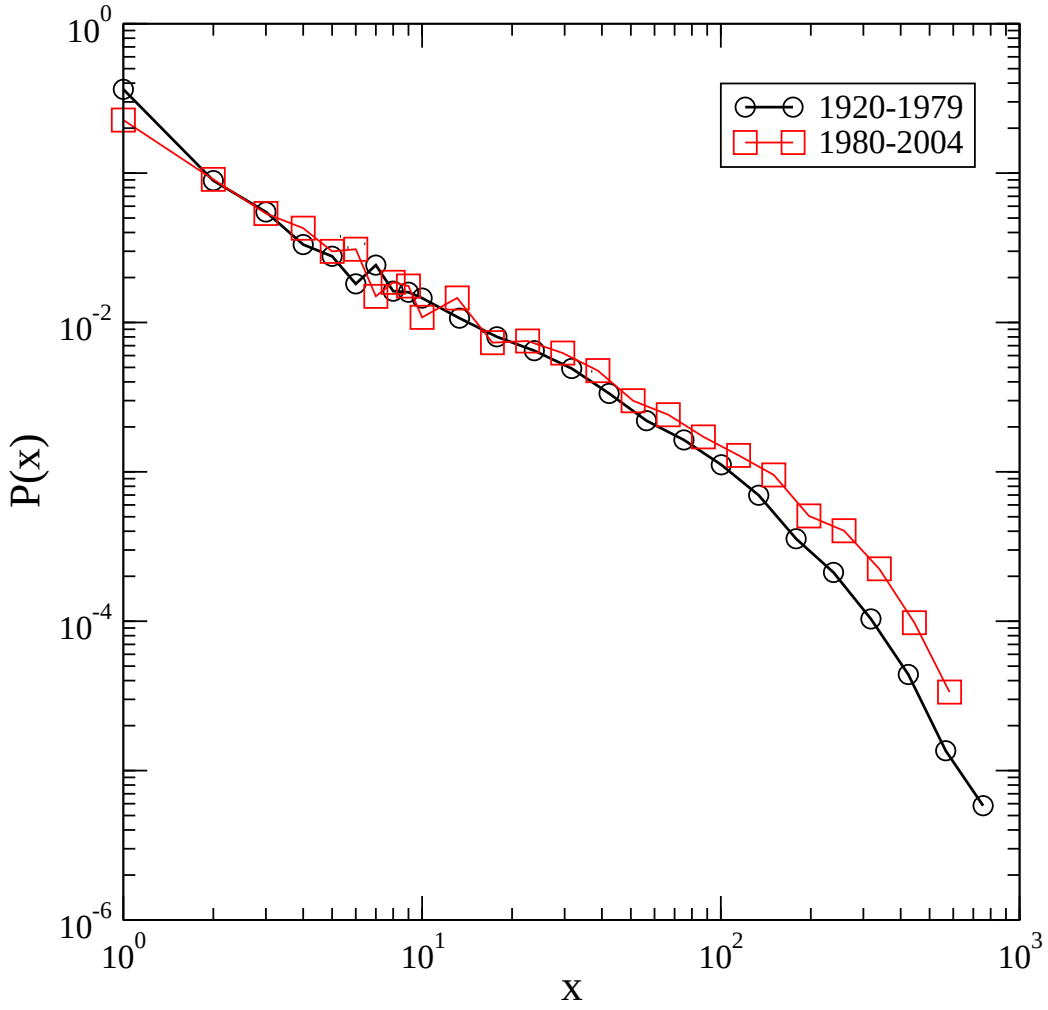

Figure 6.5: Statistical evidence in career home run distributions consistent with performance enhancement drugs. Probability density function of home runs hit over a player's career ending in two different time periods, before and after 1980 (pitchers are excluded). More home runs are being collected in the extreme part of the distribution by individuals ending their careers in the last 25 years, 1980-2004, marked by the "steroids era".

Chapter 7

Quantifying the Matthew Effect: A Model for Career Longevity

7.1 Summary

The Matthew effect refers to the adage written some two-thousand years ago in the Gospel of St. Matthew: "For to all those who have, more will be given." Even two millennia later, this idiom is used by sociologists to qualitatively describe individual progress and the interplay between status and reward. Quantitative studies of professional careers are traditionally limited by the difficulty in measuring progress and the lack of data on individual careers. However, in some professions, there are well-defined metrics that quantify career longevity, success, and prowess, which together contribute to the overall success rating for an individual employee. Here we demonstrate testable evidence, inherent in the remarkable statistical regularity of career longevity distributions, of the age-old Matthew "*rich get richer*" effect, in which longevity and past success lead to a cumulative advantage. We develop an exactly solvable stochastic model that quantitatively incorporates the Matthew effect and validate our model predictions for several competitive professions. These results demonstrate that statistical laws can exist at even the microscopic social level, where the collective behavior of individuals can lead to emergent phenomena. We test our

model on the careers of $400,000$ scientists using data from six high-impact journals. We further confirm our findings by testing the model on the careers of more than $20,000$ athletes in four sports leagues.

The rate of individual progress is fundamental to career development and success. In practice, the rate of progress depends on many random factors. Using a stochastic model, here we find that the relatively small rate of progress at the beginning of the career plays a crucial role in the evolution of the career length. Our quantitative model describes career progression using two fundamental ingredients: (i) random forward progress "up the career ladder", and (ii) random stopping times, terminating a career. This model quantifies the "Matthew effect" by incorporating the everyday property that it is easier to move forward in the career the further along is one in the career. A direct result of the increasing progress rate with career position is the large disparity in the numbers of successful long tenures from unsuccessful short stints. We test this model for both scientific and sports careers, two careers where accomplishments are methodically recorded. We analyze publication careers within six high-impact journals: Nature, Science, the Proceedings of the National Academy of Science (PNAS), Physical Review Letters (PRL), New England Journal of Medicine (NEJM) and CELL. We also analyze sports careers within four distinct leagues: Major League Baseball (MLB), Korean Professional Baseball, the National Basketball Association (NBA), and the English Premier League.

Career longevity is a fundamental metric that influences the overall legacy of an employee because for most individuals the measure of success is intrinsically related, although not perfectly correlated, to his/her career length. Common experience in most professions indicates that time is required for colleagues to gain faith in a newcomer's abilities. Qualitatively, the acquisition of new opportunities mimics a standard positive feedback mechanism (known in various fields as Malthusian growth, cumulative advantage, preferential attachment, the ratchet effect, and the Matthew *"rich get richer"* effect [101, 102, 103, 104, 105, 106]), which endows greater rewards to individuals who are more accomplished than to individuals who are less accomplished. Here we use career position as a proxy for individual accomplishment, so that the positive feedback captured by the Matthew effect is related to increasing career position. There are also other factors that result in selective bias, such as

the "relative age effect", which has been used to explain the skewed birthday distributions in populations of athletes. Several studies find that being born in optimal months provides a competitive advantage, and hence a relatively higher chance of succeeding, at several levels of competitive sports ranging from secondary school to the professional level [107, 108].

In this section we study the everyday topic of career longevity, and reveal surprising complexity arising from the competition within social environments. We develop an exactly solvable stochastic model, which predicts the functional form of the probability density function (pdf) $P(x)$ of career longevity x in competitive professions, where we define career longevity as the final career position corresponding to an effective time duration. The underlying stochastic process depends on only two parameters, α and x_c. The first parameter, α, represents the power-law exponent that emerges from the pdf of career longevity. This parameter is intrinsically related to the rate early in the career during which professionals establish their reputations and secure future opportunities. The second parameter, x_c, is a time scale which distinguishes newcomers from veterans.

7.2 Quantitative Model

In this model, every employee begins his/her career with approximately zero credibility, and must labor through a common learning curve. At each position x in a career, there is an opportunity for progress as well as the possibility for no progress. A new opportunity, corresponding to career position $x + 1$, can refer to a day at work or, more generally, to any assignment given by an employing body. For each particular career, the change in career position Δx has an associated time-frame. Optimally, an individual makes progress by advancing in career position at an equal rate as advancing in time t so that $\Delta x \equiv \Delta t$. However, in practice, an individual makes progress Δx in a subordinate time frame, given here as the career position x. In this framework, career progress is made at a rate that is slower than the passing of work time, representing the possibility of career stagnancy.

As a first step, we postulate that the stochastic process governing career progress is similar to a Poisson process, where progress is made at any given step with some

approximate probability. Each step forward in career position contributes to the employee's resume and reputation. Hence, we refine the process to be a spatial Poisson process, where the probability of progress $g(x)$ depends explicitly on the employee's position x within the career. Career longevity is then defined as the final location along the career ladder at the time of retirement T.

Employees begin their career at the starting career position $x_0 \equiv 1$, and make random forward progress through time to career position $x \geq 1$, as illustrated in Fig. 7.2. Let $P(x|T)$ be the conditional probability that at stopping time T an individual is at career position x. For simplicity, we assume that the progress rate $g(x)$ depends only on x. As a result, $P(x|T)$ assumes the familiar Poisson form, but with the insertion of $g(x)$ as the rate parameter,

$$P(x|T) = \frac{e^{-g(x)T}(g(x)T)^{x-1}}{(x-1)!} \ . \tag{7.1}$$

We derive the spatial Poisson pdf $P(x|T)$ in the following sections.

According to the Matthew effect, it becomes easier for an individual to excel with increased success and reputation. Hence, the choice of $g(x)$ should reflect the fact that newcomers, lacking the familiarity of their peers, have a more difficult time moving forward, while seasoned veterans, following from their experience and reputation, often have an easier time moving forward. For this reason we choose the progress rate $g(x)$ to have the functional form,

$$g(x) \equiv 1 - \exp[-(x/x_c)^{\alpha}] \ . \tag{7.2}$$

This function exhibits the fundamental feature of increasing from approximately zero and asymptotically approaching unity over some time interval x_c. Furthermore, $g(x) \sim x^{\alpha}$ for small $x \ll x_c$. In Fig. 7.3, we plot $g(x)$ for several values of α, with fixed $x_c = 10^3$ in arbitrary units. We will show that the parameter α is the same as the power-law exponent α in the pdf of career longevity $P(x)$ (Fig. 7.3 inset). The random process for forward progress can also be recast into the form of random waiting times, where the average waiting time $\langle \tau(x) \rangle$ between successive steps is the inverse of the forward progress probability, $\langle \tau(x) \rangle \equiv 1/g(x)$.

We now address the fact that not every career is of the same length. Nearly every individual is faced with the constant risk of losing his/her job, possibly as the result of

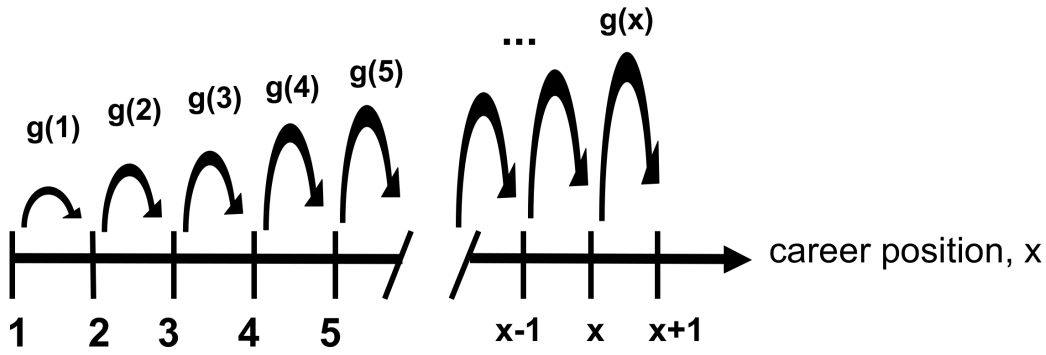

Figure 7.1: Graphical illustration of the stochastic Poisson process quantifying career progress with position-dependent progress rate $g(x)$. A new opportunity, corresponding to career position $x+1$, can refer to a day at work or, even more generally, to any assignment given by an employing body. In this framework, career progress is made at a rate $g(x)$ that is slower than the passing of work-time, representing the possibility of career stagnancy. The traditional Poisson process corresponds to a constant progress rate $g(x) \equiv \lambda$. Here, we use a functional form for $g(x) \equiv 1 - \exp[-(x/x_c)^\alpha]$ that is increasing with career position x, which captures the salient feature of the Matthew effect, that it becomes easier to make progress the further along the career.

poor performance, bad health, economic downturn, or even a change in the business strategy of his/her employer. Survival in the workplace requires that the individual maintain his/her performance level with respect to all possible replacements. In general, career longevity is influenced by many competing random processes which contribute to the random termination time T of a career [109]. The pdf $P(x|T)$ calculated in Eq. [7.1] is the conditional probability that an individual has achieved a career position x by his/her given termination time T. Hence, to obtain an ensemble pdf of career longevity $P(x)$ we must average over the pdf $r(T)$ of random termination times T,

$$P(x) = \int_0^\infty P(x|T) r(T) dT \ . \tag{7.3}$$

We next make a suitable choice for $r(T)$. To this end, we introduce the hazard rate, $H(T)$, which is the Bayesian probability that failure will occur at time $T + \delta T$, given that it has not yet occurred at time T. This is written as $H(T) = r(T)/S(T) = -\frac{\partial}{\partial T} \ln S(T)$, where $S(T) \equiv 1 - \int_0^T r(t)\ dt$ is the probability of a career surviving until time T. The exponential pdf of termination times,

$$r(T) = {x_c}^{-1} \exp[-(T/x_c)] \ , \tag{7.4}$$

has a constant hazard rate $H(T) = \frac{1}{x_c}$, and thus assumes that hazards are equally distributed over time in competitive professions. Substituting Eq. [7.4] into Eq. [7.3], we obtain

$$P(x) \ = \ \frac{g(x)^{x-1}}{x_c(\frac{1}{x_c} + g(x))^x} \approx \frac{1}{g(x) x_c}\ e^{-\frac{x}{g(x) x_c}} \ . \tag{7.5}$$

The theoretical prediction given by Eq. [7.5] is much different than the null model (in which there is no Matthew effect) where the progress rate $g(x) \equiv \lambda$ is a constant value for each individual. In the case of constant progress rate, the pdf $P(x)$ is exponential with a characteristic career length $l_c = \lambda x_c$. Later we further consider the null model where the constant progress rate λ_i of individual i is distributed over a given range. We find again that $P(x)$ is exponential, which is quite different than the prediction given by Eq. [7.5].

In order to account for aging effects, another variation of this model could include a time-dependent $H(T)$. To incorporate a non-constant $H(T)$ one can use a more

general Weibull distribution for the pdf of termination times

$$r(T) \equiv \frac{\gamma}{x_c}(\frac{T}{x_c})^{\gamma-1}\exp[-(\frac{T}{x_c})^{\gamma}] \ , \tag{7.6}$$

where $\gamma = 1$ corresponds to the exponential case [110]. In general, the hazard rate of the Weibull distribution is $H(T) \propto T^{\gamma-1}$, where $\gamma > 1$ corresponds to an increasing hazard rate, and $\gamma < 1$ corresponds to a decreasing hazard rate. We note that the timescale x_c appears both in the definition of $g(x)$ in Eq. [7.2] as a crossover between early and advanced career progress rates, and also as the timescale over which the probability of survival $S(T)$ approaches 0 in the case of $\gamma \geq 1$ as in the case of Eq. [7.4]. It is the appearance of the quantity x_c in the definition of $S(T)$ that results in a finite exponential cutoff to the longevity distributions. Although the timescales defined in $g(x)$ and $S(T)$ could be different, we observe only one timescale in the emprical data. Hence we assume here for simplicity that the two time scales are approximately equal.

From the curves plotted in the inset of Fig. 7.3, one observes that $\alpha_c = 1$ is a special crossover value for $P(x)$, between a bimodal $P(x)$ for $\alpha > 1$, and a monotonically decreasing $P(x)$ for $\alpha < 1$. This crossover is due to the small x behavior of the progress rate $g(x) \approx x^{\alpha}$ for $x < x_c$, which serves as a "potential barrier" that a young career must overcome. The width x_w of the potential barrier, defined such that $g(x_w) = 1/x_c$, scales as $x_w/x_c \approx x_c^{-1/\alpha}$. Hence, the value $\alpha_c = 1$ separates convex progress ($\alpha > 1$) from concave progress ($\alpha < 1$) in early career development. In the case $\alpha > 1$, one class of careers is stunted by the barrier, while the other class of careers excels, resulting in a bimodal $P(x)$. In the case $\alpha < 1$, it is relatively easier to make progress in the beginning of the career, resulting in a remarkable statistical regularity which bridges the gap between very short and very long careers. This statistical regularity for $\alpha < 1$ can be approximated in two regimes,

$$P(x) \propto \begin{cases} x^{-\alpha} & x \lesssim x_c \\ e^{-(x/x_c)} & x \gtrsim x_c \ . \end{cases} \tag{7.7}$$

It has been shown [89] that random stopping times can explain power law behavior in many stochastic systems that arise in the natural and social sciences, with predicted exponent values $\alpha \geq 1$. Our results provide a mechanism which describes systems

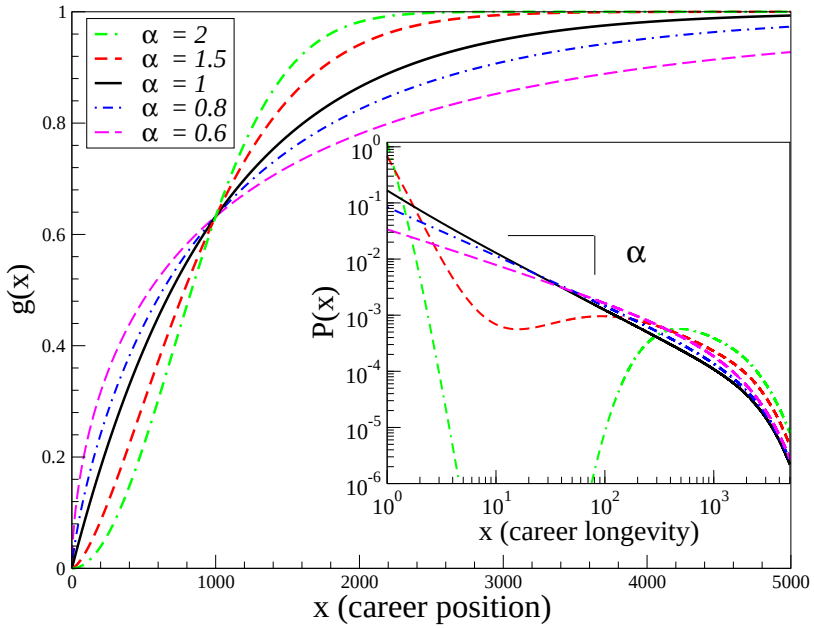

Figure 7.2: A demonstration of the fundamental relationship between the progress rate $g(x)$ and the career longevity pdf $P(x)$. The progress rate $g(x)$ represents the probability of moving forward in the career to position $x+1$ from position x. The small value of $g(x)$ for small x captures the difficulty in making progress at the beginning of a career. The progress rate increases with career position x, capturing the role of the Matthew effect. We plot five $g(x)$ curves with fixed $x_c = 10^3$ and different values of the parameter α. The parameter α emerges from the small-x behavior in $g(x)$ as the power-law exponent in $P(x)$. *(Inset)* Probability density functions $P(x)$ resulting from inserting $g(x)$ with varying α into Eq. [7.5]. The value $\alpha_c \equiv 1$ separates two distinct types of longevity distributions. The distributions resulting from concave career development $\alpha < 1$ exhibit monotonic statistical regularity over the entire range, with an analytic form approximated by the Gamma distribution $Gamma(x; \alpha, x_c)$. The distributions resulting from convex career development $\alpha > 1$ exhibit bimodal behavior. One class of careers is stunted by the difficulty in making progress at the beginning of the career, analogous to a "potential" barrier . The second class of careers forges beyond the barrier and is approximately centered around the crossover x_c on a log-scale.

with $\alpha \leq 1$. Moreover, our model provides a quantitative meaning for the power-law exponent α characterizing the pdf of career longevity.

7.3 Empirical Evidence

The two essential ingredients of our stochastic model, namely random forward progress and random termination times, are general and should apply in principle to many competitive professions. The individuals, some who are championed as legends and stars, are judged by their performances, usually on the basis of measurable metrics for longevity and success, which vary between professions.

In scientific arenas, and in general, the metric for career position is difficult to define, even though there are many conceivable metrics for career longevity and success [111, 112, 113]. We compare author longevity within individual journals, which mimic an arena for competition, each with established review standards that are related to the journal quality. As a first approximation, the career longevity with respect to a particular journal can be roughly measured as the duration between an author's first and last paper in that journal, reflecting his/her ability to produce at the top tiers of science. This metric for longevity should *not* be confused with the career length of the scientist, which is probably longer than the career longevity within any particular journal. Following standard lifetime data analysis methods [114], we collect "completed" careers from our data set which begins at year Y_0 and ends at year Y_f.

For each scientific career i, we calculate $\langle \Delta\tau_i \rangle$, the average time between publications in the particular journal. A journal career which begins with a publication in year $y_{i,0}$ and ends with a publication in year $y_{i,f}$ is considered "complete", if the following two criteria are met: (a) $y_{i,f} \leq Y_f - \langle \Delta\tau_i \rangle$ and (b) $y_{i,0} \geq Y_0 + \langle \Delta\tau_i \rangle$. These criteria eliminate from our analysis incomplete careers which possibly began before Y_0 or ended after Y_f. We then estimate the career length within journal j as $L_{i,j} = y_{i,f} - y_{i,0} + 1$, with a year allotted for publication time, and do not consider careers with $y_{i,f} = y_{i,0}$. This reduces the size of each journal data set by approximately 25% (see Table 5.1). In [115] we further analyze the scientific careers of the authors in these six journal data bases, developing normalized metrics for career success ("citation shares") and productivity ("papers shares"), and we find further evidence of

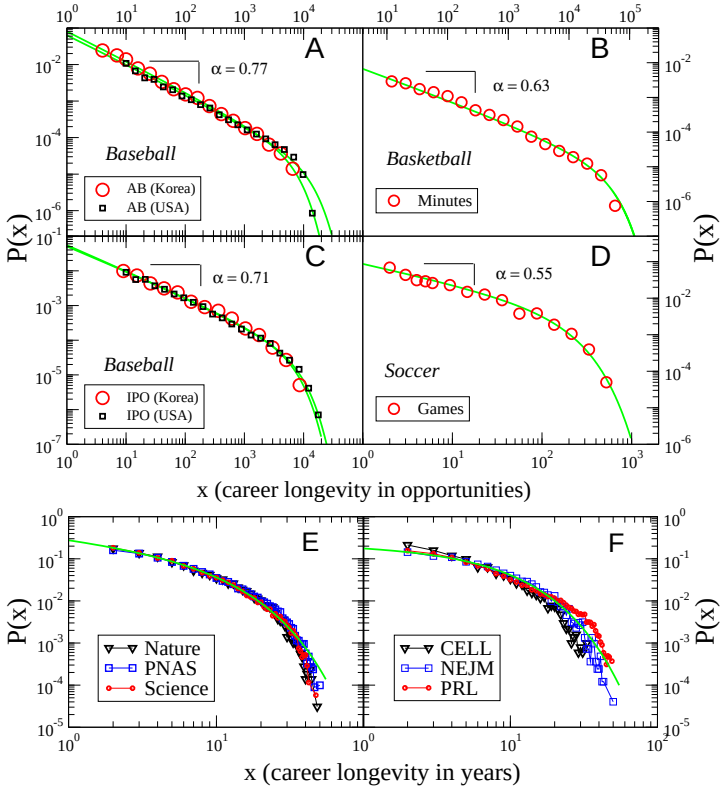

Figure 7.3: Extremely right-skewed pdfs $P(x)$ of career longevity in several high-impact scientific journals and several major sports leagues. Solid curves represent least-squares best-fit functions corresponding to the functional form in Eq. [7.5]. *(A)* Baseball fielder longevity measured in at-bats (pitchers excluded): we find $\alpha \approx 0.77$, $x_c \approx 2500$ (Korea) and $x_c \approx 5000$ (USA). *(B)* Basketball longevity measured in minutes played: we find $\alpha \approx 0.63$, $x_c \approx 21000$ minutes. *(C)* Baseball pitcher longevity measured in innings-pitched measured in outs (IPO): we find $\alpha \approx 0.71$, $x_c \approx 2800$ (Korea), and $x_c \approx 3400$ (USA). *(D)* Soccer longevity measured in games played: we find $\alpha \approx 0.55$, $x_c \approx 140$ games. *(E* and *F)* High-impact journals exhibit similar longevity distributions for the "journal career length" which we define as the duration between an author's first and last paper in a particular journal. Deviations occur for long careers due to data set limitations (for comparison, least-square fits are plotted in panel *(E)* with parameters $\alpha \approx 0.40$, $x_c = 9$ years and in panel *(F)* with parameters $\alpha \approx 0.10$, $x_c = 11$ years).

the Matthew effect in the decreasing inter-publication time $\tau(x)$ with increasing publication x for individual authors within a given journal. Several other metrics for quantifying career success [116, 112], such as the h-index [111] and generalizations [117, 118], along with methods for removing time and discipline-dependent citation factors [119] have been analyzed in the spirit of developing unbiased rating systems for scientific achievement.

In athletic arenas, the metrics for career position, success and success rate are easier to define [120]. In general, a career position in sports can be measured by the cumulative number of in-game opportunities a player has obtained. In baseball, we define an opportunity as an *"at bat"* (AB) for batters, and an *"inning pitched in outs"* (IPO) for pitchers, while in basketball and soccer, we define the metrics for opportunity as *"minutes played"* and *"games played "*, respectively.

In Fig. 7.4 we plot the distributions of career longevity for $20,000$ professional athletes in four distinct leagues and roughly $400,000$ scientific careers in six distinct journals (data is publicly available at [121, 122]). We observe universal statistical regularity corresponding to $\alpha < 1$ in the career longevity distributions for three distinct sports and several high-impact journals (see Table S2 for a summary of least squares parameters). The disparity in career lengths indicates that it is very difficult to sustain a competitive professional career, with most individuals making their debut and finale over a relatively short time interval. The exponential cutoff in $P(x)$ that follows after the crossover value x_c, arises from the finite human lifetime, and is reminiscent of any real system where there are finite-size effects that dominate the asymptotic behavior. The scaling regime is less pronounced in the curves for journal longevity. This results from the granularity of our data set, which records publications by year only. A finer time resolution (e.g. months between first and last publication) would reveal a larger scaling regime. However, regardless of the scale, one observes the salient feature of there being a large disparity between the frequency of long and short careers.

In science, an author's success metric can be quantified by the total number of papers or citations in a particular journal. Publication careers have the important property that the impact of scientific work is time dependent. Where many papers become outdated as the scientific body of knowledge grows, there are instances where

"late-blooming" papers make significant impact a considerable time after publication [123]. In [115] we find that the pdf of total number of normalized citation shares for a particular author in a single journal over his/her entire career follows the asymptotic power law $P(z)dz \sim z^{-2.5}dz$.

In sports, however, career accomplishments do not wax or wane with time. In Fig. 7.4 we plot the pdf $P(z)$ of career success z for common metrics in baseball and basketball. Remarkably, the power-law regime for $P(z)$ is governed by a scaling exponent which is approximately equal to the scaling exponent of the longevity pdf $P(x)$. Later we show that the pdf $P(z)$ of career success z follows directly from a simple Mellin convolution of the pdf $P(x)$ for longevity x and the pdf $P(y)$ of prowess y.

The Gamma pdf $P(x) \equiv Gamma(x;\alpha,x_c) \propto x^{-\alpha}\ e^{-x/x_c}$ is commonly employed in statistical modeling, and can be used as an approximate form of Eq. [7.7]. One advantage to the gamma pdf is that it can be inverted in order to study extreme statistics corresponding to rare stellar careers. In [124], we further analyze the relationship between the extreme statistics of the Gamma pdf and the selection processes for *Hall of Fame* museums. In general, the statistical regularity of these distributions allows one to establish robust milestones, which could be used for setting the corresponding financial rewards and pay scales, within a particular profession. Interestingly, we also find in [124] that the pdfs for career success in MLB are stationary even if we quantitatively remove the time-dependent factors that influence success. This stationarity implies that the right-skewed statistical regularity we observe in $P(z)$ arises from both the intrinsic talent and the longevity of professional athletes, and does not result from changes in technology, economic factors, training improvements, etc.

In summary, a wealth of data recording various facets of social phenomena have become available in recent years, allowing scientists to search for universal laws that emerge from human interactions [125]. Theoretical models of social dynamics, employing methods from statistical physics, have provided significant insight into the various mechanisms that can lead to emergent phenomena [126]. An important lesson from complex system theory is that oftentimes the details of the underlying mechanism do not affect the macroscopic emergent phenomena. For baseball players in Korea and the United States, we observe remarkable similarity between the pdfs of

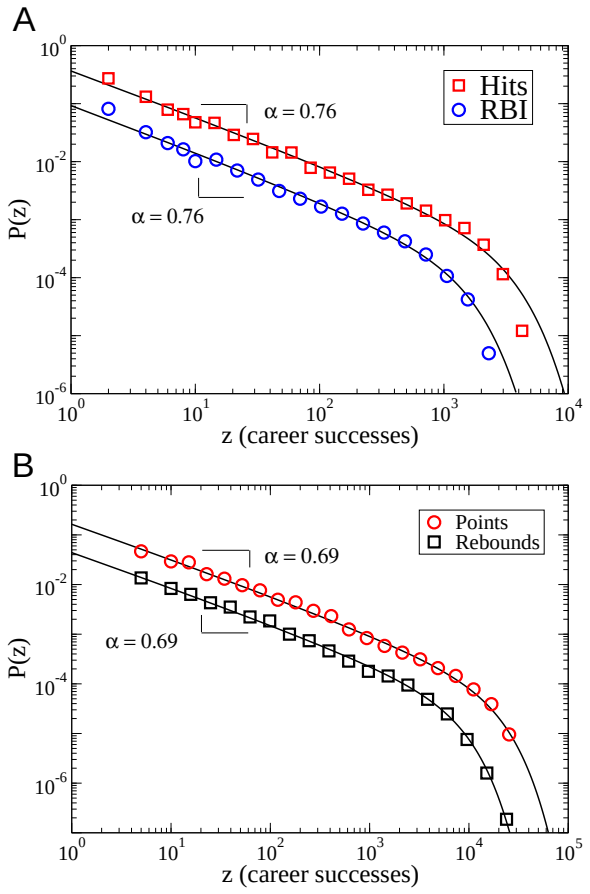

Figure 7.4: Probability density function $P(z)$ of common metrics for career success, z. Solid curves represent best-fit functions corresponding to Eq. [7.5]. *(A)* Career batting statistics in American baseball: $x_c^{Hits} \approx 1200$, $x_c^{RBI} \approx 600$, (RBI = Runs Batted In). *(B)* Career statistics in American basketball: $x_c^{Points} \approx 8000$, $x_c^{Rebounds} \approx 3500$. For clarity, the top set of data in each plot has been multiplied by a constant factor of four in order to separate overlapping data.

career longevity (Fig. 7.4) and the pdfs of prowess (Fig. 7.5), despite these players belonging to completely distinct leagues. This fact is consistent with the hypothesis that universal stochastic forces govern career development in science, professional sports and presumably in a large class of competitive professions.

Here we demonstrate strong empirical evidence for universal statistical laws that describe career progress in competitive professions. Universal power-law behavior also occurs in many other social complex systems [6, 35, 81, 85, 127, 128, 129, 130, 131, 132, 133]. Stemming from the simplicity of the assumptions, the mechanism developed in this section could apply elsewhere in society, such as the duration of both platonic and romantic friendships. Indeed, long relationships are harder to break than short ones, with random factors inevitably terminating them forever. Also, supporting evidence for the applicability of this model can be found in the similar truncated power-law pdfs with $\alpha < 1$, that describe the dynamics of connecting within online social networks [133].

7.4 Derivation of the Spatial Poisson Distribution

The master equation for the evolution of $P(x, N)$ is

$$\begin{aligned} P(x+1, N+1) &- P(x+1, N) \\ = f(x)P(x, N) &- f(x+1)P(x+1, N) \ , \end{aligned} \tag{7.8}$$

with initial condition,

$$P(x+1, 0) = \delta_{x,0} \ . \tag{7.9}$$

Here $f(x)$ represents the probability that an employee obtains another future opportunity given his/her resume at career position x. We next write the discrete-time discrete-space master equation in the continuous-time discrete-space form,

$$\frac{\partial P(x+1,t)}{\partial t} = g(x)P(x,t) - g(x+1)P(x+1,t) \ , \tag{7.10}$$

where $g(x) = f(x)/\delta t$ and $t = N\delta t$ (for an extensive discussion of master equation formalism see Ref. [134]). Taking the Laplace transform of both sides one obtains,

$$\begin{aligned} sP(x+1,s) - P(x+1, t=0) = \\ g(x)P(x,s) - g(x+1)P(x+1,s) \ . \end{aligned} \tag{7.11}$$

From the initial condition in Eq. [7.9] we see that the second term above vanishes for $x \geq 1$. Solving for $P(x+1,s)$ we obtain the recurrence equation

$$P(x+1,s) = \frac{g(x)}{s+g(x+1)} P(x,s) \ . \tag{7.12}$$

If the first derivative $\frac{d}{dx}g(x)$ is relatively small, we can replace $g(x+1)$ with $g(x)$ in the equation above. Then, one can verify the ansatz

$$P(x,s) = \frac{g(x)^{x-1}}{(s+g(x))^x} \ , \tag{7.13}$$

which is the Laplace transform of the spatial Poisson distribution $P(x,t;\lambda = g(x))$ ([135]). The Laplace transform is defined as $L\{f(t)\} = f(s) = \int_0^\infty dt f(t) e^{-st}$. Inverting the transform we obtain

$$P(x,t) = \frac{e^{-g(x)t}(g(x)t)^{x-1}}{(x-1)!} \ . \tag{7.14}$$

The pdf in Eq. [7.14] corresponds to a particular time t that is common to all observations of x. In this section, we define the time $t \equiv T$ to be a conditional stopping time T which characterizes a given subset of careers. We average over a distribution $r(T)$ of stopping times to obtain the empirical longevity pdf $P(x)$ which is comprised of careers with varying T.

7.5 Data and methods

The publication data analyzed in this section was downloaded from *ISI Web of Knowledge* in May 2009. We restrict our analysis to publications termed as "Articles", which excludes reviews, letters to editor, corrections, etc. Each article summary includes a field for the author identification consisting of a last name and first and middle initial (eg. the author name John M. Doe would be stored as "Doe, J" or "Doe, JM" depending on the author's designation). From these fields, we collect the career works of individual authors within a particular journal together, and analyze metrics for career longevity and success.

For author i we combine all articles in journal j for which he/she was listed as coauthor. The total number of papers for author i in journal j over the 50-year period

is n_i. Following methods from lifetime statistics [114], we use a standard method to isolate "completed" careers from our data set which begins at year Y_0 and ends at year Y_f. For each author i, we calculate $\langle \Delta\tau_i \rangle$, the average time $\Delta\tau_i$ between successive publications in a particular journal. A career which begins with the first recorded publication in year $y_{i,0}$ and ends with the final recorded publication in year $y_{i,f}$ is considered "complete", if the following two criteria are met:

(1) $y_{i,f} \leq Y_f - \langle \Delta\tau_i \rangle$

(2) $y_{i,0} \geq Y_0 + \langle \Delta\tau_i \rangle$.

This method estimates that the career begins in year $y_{i,0} - \langle \Delta\tau_i \rangle$ and ends in year $y_{i,f} + \langle \Delta\tau_i \rangle$. If either the estimated beginning or ending year do not lie within the range of the data base, than we discount the career as incomplete to first approximation. Statistically, this means that there is a significant probability that this author published before Y_0 or will publish after Y_f. We then estimate the career length within journal j as $L_{i,j} = y_{i,f} - y_{i,0} + 1$, and do not consider careers with $y_{i,f} = y_{i,0}$. This reduces the size of the data set by approximately 25% (compare the raw data set sizes N to the pruned data set size N^* in Table 7.1).

There are several potential sources of systematic error in the use of this database:

(i) Degenerate names $\rightarrow$ increases career totals. Radicchi *et al.* [112] observe that this method of concatenated author ID leads to a pdf $P(d)$ of degeneracy d which scales as $P(d) \sim d^{-3}$.

(ii) Authors using middle initials in some but not all instances of publication $\rightarrow$ decreases career totals.

(iii) A mid-career change of last name $\rightarrow$ decreases career totals.

(iv) Sampling bias due to finite time period. Recent young careers are biased toward short careers. Long careers located towards the beginning Y_0 or end Y_f of the database are biased towards short careers.

7.6 A Robust Method for Classifying Careers

Professional sports leagues are geared around annual championships that celebrate the accomplishments of teams over a whole season. On a player level, professional sports leagues annually induct retired players into "halls of fame" in order to celebrate and honor stellar careers. Induction immediately secures an eternal legacy for those that are chosen. However, there is no standard method for inducting players into a *Hall of Fame*, with subjective and political factors affecting the induction process. In [124] we quantitatively normalize seasonal statistics so to remove time-dependent factors that influence success. This provides a framework for comparing career statistics across historical eras.

In this section we propose a generic and robust method for measuring careers. We find that the pdf for career longevity can be approximated by the gamma distribution,

$$Gamma(x;\alpha,x_c) = \frac{x^{-\alpha}e^{-x/x_c}}{x_c^{1-\alpha}\Gamma(1-\alpha)} \ , \tag{7.15}$$

with moments $\langle x^n \rangle = x_c^n \frac{\Gamma(1-\alpha+n)}{\Gamma(1-\alpha)}$, where we restrict our considerations to the case of $\alpha \leq 1$, with $x_c >> 1$. This distribution allows us to calculate the extreme value x^* such that only f percentage of players exceed this value according to the pdf $P(x)$,

$$f = \int_{x^*}^{\infty} \frac{x^{-\alpha}e^{-x/x_c}}{x_c^{1-\alpha}\Gamma(1-\alpha)}dx = \frac{\Gamma[1-\alpha,\frac{x^*}{x_c}]}{\Gamma(1-\alpha)} = Q[1-\alpha,\frac{x^*}{x_c}] \ , \tag{7.16}$$

where $\Gamma[1-\alpha,\frac{x^*}{x_c}]$ is the incomplete gamma function and $Q[1-\alpha,\frac{x^*}{x_c}]$ is the regularized gamma function. This function can be easily inverted numerically using computer packages, e.g. *Mathematica*, which results in the statistical benchmark

$$x^* = x_c \ Q^{-1}[1-\alpha,f]. \tag{7.17}$$

In [124] we use the maximum likelihood estimator (MLE) for the Gamma pdf to estimate the parameters α and x_c for each pdf. The values we obtain using MLE are systematically smaller for α values and for x_c values, but the relative differences are negligible.

In Table 7.2 we provide statistical benchmarks x^* corresponding to career longevity and career metrics for several sports. For the calculation of each x^* we use the parameter values α and x_c calculated from a least-squares fit to the empirical pdf $P(x)$

using the functional form of Eq. [7.5], and the significance level value f calculated from historical induction frequencies in the American Baseball Hall of Fame (HOF) in Cooperstown, NY USA. The baseball HOF has inducted 276 players out of the 14,644 players that exist in Sean Lahman's baseball database between the years 1879-2002, which corresponds to a fraction $f \equiv 0.019$. It is interesting to note that the last column, $\frac{x^*}{\sigma} \equiv \beta \approx 3.9$ for all the gamma distributions analyzed. This approximation is a consequence of the universal scaling form of the gamma function $Gamma(x) \equiv U(x/x_c)$, where the standard deviation σ of the Gamma pdf has the simple relation $\sigma = x_c\sqrt{1-\alpha}$. Hence, for a given f and α, the ratio

$$x^*/\sigma = \frac{Q^{-1}[1-\alpha, f]}{\sqrt{1-\alpha}} \tag{7.18}$$

is independent of x_c. Furthermore, this approximation is valid for all statistics in MLB since α is approximately the same for all pdfs analyzed. Thus, the value $x^* \approx 4\sigma$ is a robust approximation for determining if a player's career is stellar at the $f \approx 0.02$ significance level. The highly celebrated milestone of 3,000 hits in baseball corresponds to the value $x^* = 1.26\ \beta\sigma_{hits}$. Only 27 players have exceeded this benchmark in their professional careers, while only 86 have exceeded the arbitrary 2,500 benchmark. Hence, it makes sense to set the benchmark for all milestones at a value of $x^* = \beta\sigma$ corresponding to each distribution of career metrics.

We check for consistency by comparing the extreme threshold value x^* calculated using the gamma distribution with the value x_d^* derived from the database of career statistics. Referring to the actual set of all baseball players from 1871-2006, to achieve a fame value $f_d \approx 0.019$ with respect to hits, one should set the statistical benchmark at $x_d^* \approx 2250$, which account for 146 players (this assumes that approximately half of all baseball players are not pitchers, whom we exclude from this calculation of f_d). The value of $x_d^* \approx 2250$ agrees well with the value calculated from the gamma distribution, $x^* \approx 2366$. Of these 146 players with career hit tallies greater than 2250, 126 players have been eligible for at least one induction round, and 82 of these players have been successfully inducted into the American baseball hall of fame. Thus, a player with a career hit tally above $x^* \approx x_d^*$ has a 65% chance of being accepted, based on just those merits alone. Repeating the same procedure for career strikeouts obtained by

pitchers in baseball we obtain the milestone value $x_d^* \approx 1525$ strikeouts, and for career points in basketball we obtain the value $x_d^* \approx 16,300$ points. Nevertheless, the overall career must be taken into account, which raises the bar, and accounts for the less than perfect success rate of being voted into a hall of fame, given that a player has had a statistically stellar career in one statistical category.

7.7 Career Metrics

In Fig. 7.4 we plot common career metrics for success in American baseball and American basketball. Note that the exponent α for the pdf $P(z)$ of total career successes z is approximately equal to the exponent α for the pdf $P(x)$ of career longevity x (see Table 7.2). In this section, we provide a simple explanation for the similarity between the power law exponent for career longevity (Fig. 7.3) and the power law exponent for career success (Fig. 7.4).

Consider a distribution of longevity that is power law distributed, $P(x) \sim x^{-\alpha}$ for the entire range $1 \leq x \leq x_c < \infty$. The cutoff x_c represents the finiteness of human longevity, accounted for by the exponential decay in Eq. [7.7]. Also, assume that the prowess y has a pdf $P(y)$ which is characterized by a mean and standard deviation, which represent the talent level among professionals (see Ref. [120] for the corresponding prowess distributions in major league baseball). In the first possible case, the distribution is right-skewed and approximately exponential (as in the case of home-runs). In other cases, the distributions are essentially Gaussian. Regardless of the distribution type, the prowess pdfs $P(y)$ are confined to the domain $\delta \leq y \leq 1$, where $\delta > 0$.

Assume that in any given appearance, a person can apply his/her natural prowess towards achieving a success, independent of past success. Although prowess is refined over time, this should not substantially alter our demonstration. Since not all professionals have the same career length, the career totals are in fact a combination of these two distributions as in their product. Then the career success total $z = xy$ has

the distribution,

$$\begin{aligned} P(z=xy) &= \int\int dy\ dx\ P(y)P(x)\delta(xy-z) \\ &= \int\int dy\ dx P(y)P(x)\delta(x(y-z/x)) \\ &= \int dx\ P(\frac{z}{x})P(x)\frac{1}{x}\ . \end{aligned} \tag{7.19}$$

This integral has three domains (Ref. [136]),

$$P(z) \propto \begin{cases} \int_1^{z/\delta} dx\ P(\frac{z}{x})x^{-(\alpha+1)}\ , & \delta < z < 1 \\ \int_z^{z/\delta} dx\ P(\frac{z}{x})x^{-(\alpha+1)}\ , & 1 < z < x_c\delta \\ \int_z^{x_c} dx\ P(\frac{z}{x})x^{-(\alpha+1)}\ , & x_c\delta < z < x_c\ . \end{cases}$$

The first regime $\delta < z < 1$ is irrelevant, and is not observed since z is discrete in the cases analyzed here. For the first case of an exponentially distributed prowess,

$$P(z) \propto \begin{cases} z^{-\alpha}\ , & 1 < z < x_c \quad \delta \\ z^{-\alpha}\ \exp(-z/\lambda x_c)\ , & x_c\delta < z < x_c\ . \end{cases} \tag{7.20}$$

In Ref. [120] we mainly observe the exponential tail in the home-run distribution, as the above form suggests in the regime $x_c\delta < z < x_c$, resulting from $\delta \approx 0$ for the right-skewed home-run prowess distribution. However, in the case for a normally distributed prowess, the power law behavior of the longevity distribution is maintained for large values into the career success distribution $P(z)$, as $x_c\delta > 10^3$.

$$P(z) \propto \begin{cases} z^{-\alpha}\ , & 1 < z < x_c\delta \\ z^{-\alpha}e^{-(\frac{z}{\sigma x_c})^2/2}\ , & x_c\delta < z < x_c\ . \end{cases} \tag{7.21}$$

Thus, the main result of this demonstration is that the distribution $P(z)$ maintains the power law exponent α of the career-longevity distribution, $P(x)$, when the prowess is distributed with a characteristic mean and standard deviation. This result is also demonstrated with the simplification of representing the prowess distribution $P(y)$ as an essentially uniform distribution over a reasonable domain of y, which simplifies the integral in Eq. (7.19) while maintaining the inherent power law structure.

In Fig. 7.5 we plot the prowess distributions that correspond to the career success distributions plotted in Fig. 7.4. It is interesting that the competition level based on the distributions of prowess indicates that Korean and American baseball are nearly equivalent. Also, note that the prowess distributions for rebounds per minute are bimodal, as the positions of players in basketball are more specialized.

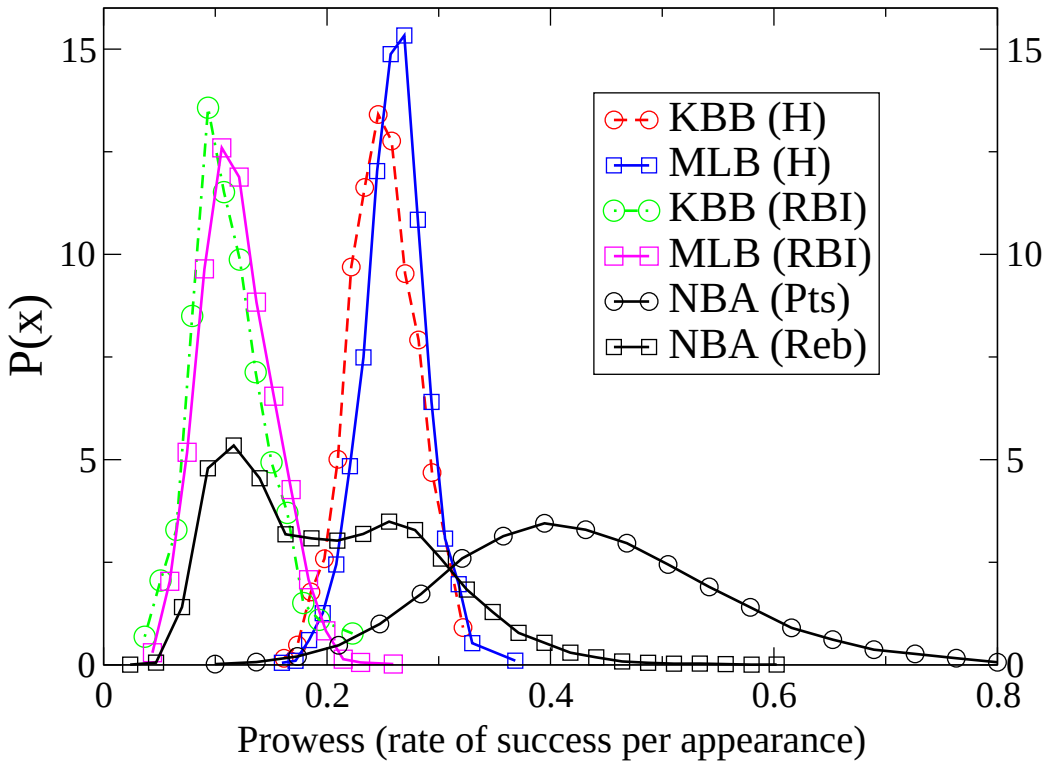

Figure 7.5: Probability density functions of seasonal prowess for several career metrics. Each pdf is normally distributed, except for the bimodal curve for rebound prowess, NBA (Reb.). The bimodal distribution for Rebound prowess reflects the specialization in player positions in the sport of basketball. Furthermore, note the remarkable similarity in the distributions between American (MLB) and Korean (KBB) baseball players.

7.8 A null model without the Matthew effect

In this section, we compare the predictions of our theoretical model with the predictions of a theoretical model which does not incorporate the Matthew effect. Since the Matthew effect implies that the progress rate $g(x)$ increase with career position x, we analyze the simpler model where for each individual i the progress rate $g_i(x)$ is constant,

$$g_i(x) \equiv \lambda_i \ . \tag{7.22}$$

The solution to the conditional longevity pdf $P(x|\lambda_i)$ is still given by Eq. [7.5], taking the form

$$P(x|\lambda_i) = \frac{\lambda_i^{x-1}}{x_c(\frac{1}{x_c}+\lambda_i)^x} \approx \frac{1}{\lambda_i x_c}\ e^{-\frac{x}{\lambda_i x_c}}\ , \tag{7.23}$$

which is an exponential pdf, with a characteristic career length $l_c \equiv \lambda_i x_c$. Hence, this null model corresponds to a career progress mechanism wherein intrinsic ability, which is incorporated into the relative value of λ_i, is the dominant factor. In order to calculate the longevity pdf $P(x)$ which incorporates a distribution of intrinsic abilities across the population of individuals, we average over the conditional pdfs using a pdf $P(\lambda)$ that we assume is well-defined by a mean $\overline{\lambda}$ and standard deviation σ, consistent with what we observe for the seasonal prowess pdfs shown in Fig. 7.5. In the case of $P(\lambda) = Normal(\overline{\lambda}, \sigma)$, then

$$P(x) = \int_0^1 P(\lambda)P(x|\lambda)d\lambda \ \equiv \int_0^1 \frac{e^{-(\lambda-\overline{\lambda})^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} P(x|\lambda)d\lambda \ . \tag{7.24}$$

For the sake of providing an analytic result, we replace $P(\lambda)$ by a uniform distribution,

$$P(\lambda) \approx \begin{cases} 0 & |\lambda-\overline{\lambda}| > 2\sigma \\ \frac{1}{4\sigma} & |\lambda-\overline{\lambda}| \leq 2\sigma\ , \end{cases} \tag{7.25}$$

which does not change the overall result. The integral then becomes,

$$P(x) \approx \frac{1}{4\sigma}\int_{\lambda-2\sigma}^{\lambda+2\sigma} \frac{d\lambda}{\lambda x_c}\ e^{-\frac{x}{\lambda x_c}} = \frac{1}{4\sigma x_c}[\Gamma(0, \frac{x/x_c}{\overline{\lambda}+2\sigma}) - \Gamma(0, \frac{x/x_c}{\overline{\lambda}-2\sigma})] \approx e^{-x/\overline{\lambda}x_c}\ , \tag{7.26}$$

for $1 > \overline{\lambda} > 2\sigma$, where the last approximation corresponds to a relatively small σ. Thus, we find that even with a reasonable dispersion in the constant progress rates λ in a population of individuals, the pdf $P(x)$ is still exponential. Hence, our theoretical model cannot explain the empirical non-exponential form of $P(x)$ unless we incorporate the Matthew effect using $g(x)$ that increase with x.

Journal	Years	Articles	Authors, N	N^*
Nature	1958-2008	65,709	130,596	94,221
Science	1958-2008	48,169	109,519	82,181
PNAS	1958-2008	84,520	182,761	118,757
PRL	1958-2008	85,316	112,660	72,102
CELL	1974-2008	11,078	31,918	23,060
NEJM	1958-2008	17,088	66,834	49,341

Table 7.1: Summary of data sets for each journal. Total number N of unique (but possibly degenerate) name identifications. N^* is the number of unique name identifications after pruning the data set of incomplete careers.

Professional League,	Least-square values		Gamma pdf values				
(success metric)	α	x_c	$\langle x \rangle$	σ	x^*	$\frac{x^*}{\langle x \rangle}$	$\frac{x^*}{\sigma}$
MLB, (H)	0.76 ± 0.02	1240 ± 150	300	610	2400	7.8	3.9
MLB, (RBI)	0.76 ± 0.02	570 ± 80	140	280	1100	7.8	3.9
NBA, (Pts)	0.69 ± 0.02	7840 ± 760	2400	4400	17000	7.0	3.9
NBA, (Reb)	0.69 ± 0.02	3500 ± 130	1100	2000	7600	6.9	3.9

Professional League,	Least-square values		Gamma pdf values				
(opportunities)	α	x_c	$\langle x \rangle$	σ	x^*	$\frac{x^*}{\langle x \rangle}$	$\frac{x^*}{\sigma}$
KBB, (AB)	0.78 ± 0.02	2600 ± 320	580	1200	4700	8.2	3.9
MLB, (AB)	0.77 ± 0.02	5300 ± 870	1200	2500	9700	8.1	3.9
MLB, (IPO)	0.72 ± 0.02	3400 ± 240	950	1800	6900	7.3	3.9
KBB, (IPO)	0.69 ± 0.02	2800 ± 160	840	1500	5900	7.0	3.9
NBA, (Min)	0.64 ± 0.02	20600 ± 1900	7700	12600	48800	6.4	3.9
UK, (G)	0.56 ± 0.02	138 ± 14	61	92	360	5.8	3.9

Academic Journal,	Least-square values	
(career length in years)	α	x_c
Nature	0.38 ± 0.03	9.1 ± 0.2
PNAS	0.30 ± 0.02	9.8 ± 0.2
Science	0.40 ± 0.02	8.7 ± 0.2
CELL	0.36 ± 0.05	6.9 ± 0.2
NEJM	0.10 ± 0.02	10.7 ± 0.2
PRL	0.31 ± 0.04	9.8 ± 0.3

Table 7.2: Data summary for the pdfs of career statistical metrics. The values α and x_c are determined for each career longevity pdf $P(x)$ and each career success pdf $P(z)$ via least-squares method using the functional form given by Eq. [7.5].

Part IV

A Statistical Physics Approach to Quantifying the Role of Zealots in Opinion Formation

Chapter 8

Motivation

Questions concerning the nature of opinion and consensus formation have been studied in the behavioral, social, political, and cognitive sciences. So what does statistical physics have to contribute to the subject? The theory of kinetic spin systems has been very fruitful in describing the evolution of a spin systems, roughly defined as a collection of interacting entities (agents, voters, electron spins, etc.) with dynamic states. The dynamics are determined by the particular rules for the local (or non-local) interaction between spins and the topology of the interaction network between the agents. A key concept is the role of temperature T which determines the significance of fluctuations in the system (interpreted possibly as the role of misinformation or vacillating confidence on the behalf of the agent).

In addition to the application to opinion dynamics, kinetic spin systems have also been used as models for biological dynamics (e.g. the Moran model for population genetics) and for epidemic dynamics (e.g. the SIR, SIS, and other models for infection dynamics in a population of S = susceptible, I = infected, and R = recovered individuals). The power of the statistical physics approach is in the limit of system size $N \to \infty$, so that finite-size effects, boundary and initial conditions are negligible, and so that emergent behavior is directly dependent on the topological features of the interactions and the symmetries (and the corresponding conserved or not-conserved quantities) embedded in the rules for individual spin evolution.

Here we study a simple variant of the well-known Voter model, which is a 2-spin system with the unique property of being analytically solvable in arbitrary dimension.

Interestingly, the 1-d Voter model is equivalent to the critical temperature $T_c = 0$ limit of the 1-d Ising-Glauber model. In higher dimensions, and for $T_c > 0$, the models diverge, where the voter model is characterized by spin-flip rules that obey the local field as a proportional rule, whereas the Ising-Glauber model is characterized by spin-flip rules that obey the majority of the local field.

In some ways, the Voter model has the disadvantage of being overly simple. For example, as a result of this simplicity, the coarsening processes in this Voter model system are defined by coarse domains with no surface tension. Despite the simplicity, the Voter model has been widely regarded as a benchmark spin system which can be modified in numerous ways in order to study the role of modified features. One such modification is to change the topology of the interacting agents so that they interact not through a d-dimensional lattice, but through an arbitrary graph or network. Amazingly, quantities such as the mean consensus time and the exit probability for distinct final states given an initial state are calculable on graphs of arbitrary degree distribution. These results follow from the beautiful property that the voter model can be mapped into a system of coalescing random walkers on a network, therefore opening up a doorway to powerful methods developed in the theory of random walks. Another possible modification to the Voter model is to endow the agents with persistence resulting in a finite opinion memory; this is in contrast to the basic Voter model, where the opinion dynamics of each voter is explicitly Markov, meaning that the transition rates do not depend on the history of the system, meaning that the system has no memory.

The modification to the Voter model that we pursue is to vary the persuasiveness of agents in the voter model between two extremes: some voters are completely "spineless" as in the traditional model, while other voters, here termed "zealots," are completely stubborn. We define stubbornness in the dynamic opinion rules as the inability of a zealot to change opinion, even if this individual is completely surrounded by agents with the opposite opinion. The presence of these voters completely determine the steady-state polarization of the system. Moreover, using methods from electrostatics, it can be shown that the "field of influence" of a single zealot is analogous to the electric field of a single charged particle, such as an electron.

Chapter 9

the Voter Model with Zealotry

9.1 Summary

We study the voter model with a finite density of zealots—voters that never change opinion. For equal numbers of zealots of each species, the distribution of magnetization (opinions) is Gaussian in the mean-field limit as well as in one and two dimensions, with a width that is proportional to $1/\sqrt{Z}$, where Z is the number of zealots, independent of the total number of voters. Thus just a few zealots can prevent consensus or even the formation of a robust majority.

9.2 Introduction

The voter model [137] is one of the simplest examples of cooperative behavior that has been used as a paradigm for the dynamics of opinions in socially interacting populations. In the voter model, each node of a graph is occupied by a voter that has two opinion states, denoted as + and −. Opinions evolve by: (i) picking a random voter; (ii) the selected voter adopts the state of a randomly-chosen neighbor; (iii) repeat these steps *ad infinitum* or until a finite system necessarily reaches consensus. Naively, one can view each voter has having no self confidence and thus takes on the state of one of its neighbors. This evolution resembles that of the Ising model with

zero-temperature Glauber kinetics [138], but with one important difference: in the Ising model, each spin obeys the state of the local majority; in the voter model, a voter chooses a state with a probability that is proportional to the number of neighbors in that state.

There are three basic properties of the voter model that characterize its evolution. The first is the exit probability, namely, the probability that a finite system eventually reaches consensus where all voters are in the + state, $E_+(\rho_0)$, as a function of the initial density ρ_0 of + voters. Because the mean magnetization (averaged over all realizations and all histories) is conserved on any degree-regular graph, and because the only possible final states of a finite system are consensus, $E_+(\rho_0) = \rho_0$ [137].

A second basic property is the mean time T_N to reach consensus in a finite system of N voters. For regular lattices in d dimensions, it is known that T_N scales as N^2 in $d = 1$, as $N \ln N$ in $d = 2$, and as N in $d > 2$ [137, 139]. In contrast, T_N generally scales sublinearly with N on heterogeneous graphs with broad degree distributions [140]. Defining μ_k as the k^{th} moment of the degree distribution, then $T_N \sim N\mu_1^2/\mu_2$, which grows slower than linearly in N for a sufficiently broad degree distribution. Finally, the 2-point correlation function $G_2(r)$, defined as the probability that 2 voters a distance r apart are in the same state, asymptotically decays as r^{2-d} on a regular lattice when the spatial dimension $d > 2$ [134, 139]. This decay is the same as that of the electrostatic potential of a point charge, a correspondence that has proven useful in analyzing the voter model.

In this work, we investigate an extension of the voter model in which a small fraction of the population are zealots—individuals that never change opinion. The role of a single zealot [141] or a small number of zealots [142] on the voter model has been studied previously, and considerable insight has been gained by exploiting the previously-mentioned electrostatic correspondence. One motivation for this work is the obvious fact that consensus is not the asymptotic outcome of repeated elections in democratic societies. An empirical example is the set of US presidential elections [143], where the percentage of votes for the winner has ranged from highs of 61.05% (Johnson over Goldwater 1964) and 60.80% (Roosevelt over Landon 1932) to lows of 47.80% (Harrison minority winner over Cleveland 1888) and 47.92% (Hayes over Tilden 1872). In this compilation, we exclude elections with substantial voting to

candidates outside the top two.

This example, as well as many other election results from democratic countries, show in an obvious way that consensus will never be achieved in large populations. This fact motivates us to to investigate an opinion dynamics model that leads to stationarity due to the influence of zealots. Other natural models that lead to cultural fragmentation and an attendant stasis include the multiple-state Axelrod model [144], the bounded compromise model of Weisbuch et al. [145] and its variants [146], in which lack of consensus is a basic outcome.

The basic question that we wish to address in the voter model with zealots is: what is the nature of the global opinion as a function of the density of zealots? One of our main results is that equal but very small numbers of zealots of both types leads to a steady state with a narrow Gaussian magnetization distribution centered at zero. Here the magnetization is simply the difference in the fraction of voters of each species. Thus a small fraction of zealots is surprisingly effective in maintaining a steady state with only small fluctuations about this state.

In the next section, we define the model. Then in Secs. 9.4 and 9.5, we solve the model in the mean-field limit and on a one-dimensional periodic ring. We then investigate the behavior on the square lattice by numerical simulations in Sec. 9.6 and find behavior that is quantitatively close to that in the mean-field limit. Finally, we conclude and point out some additional interesting features of the role of zealotry on the voter model in Sec. 9.7.

9.3 The model

The population consists of N voters, with a fixed number of zealots that never change opinion, while the remaining voters are susceptible to opinion change. Each voter can be in one of two opinion states, $+1$ or -1 that we term "democrat" and "republican", respectively. Thus the system consists of Z_+ democrat and Z_- republican zealots, as well as N_+ democrat and N_- republican susceptibles. Each type of voter evolves as follows:

1. *Susceptible democrats* can become republicans;

2. *Susceptible republicans* can become democrats;

3. *Zealot democrats* are always democrats;

4. *Zealot republicans* are always republicans.

Each agent, whether a zealot or a susceptible, has the same persuasion strength that we set to 1. That is, after a susceptible voter selects a neighbor, the voter is persuaded to adopt the state of this neighbor with probability 1. Because the total population comprises of agents in one of four possible states, we have $N = N_+ + N_- + Z_+ + Z_-$. Since the number of zealots is fixed, the total number of susceptible individuals $S = N - Z_+ - Z_- = N_+ + N_-$ is also conserved. The dynamics is a direct generalization of the voter model and consists of the following steps:

1. Pick a random voter, if this voter is a zealot nothing happens.

2. If the selected voter is a susceptible, then pick a random neighbor and adopt its state; note that if the selected voter and the neighbor are in the same state, nothing happens in the update.

3. Repeat steps 1 & 2 *ad infinitum* or until consensus is reached.

We will investigate this model on the two natural geometries of the complete graph, a natural realization of the mean-field limit, and regular lattices. For the complete graph, all other voters in the system are nearest-neighbor to any voter. Thus the complete graph has no spatial structure, a feature that allows for a simple solution. In contrast, when the voters live on the sites of a regular lattice, a voter can be directly influenced only by its the nearest neighbors.

9.4 Dynamics on the complete graph

On the complete graph, the state of the population may be characterized by the probability $P(N_+, N_-, t)$ of finding $N_\pm$ susceptible voters at time t. Since $N_- = S - N_+$, we merely need to consider the master equation for $P(N_+, t)$, which reads

$$\begin{aligned} \frac{\partial P(N_+,t)}{\partial t} &= \sum_{\delta=\pm 1} P(N_+ + \delta, t) W(N_+ + \delta \to N_+) \\ &- \sum_{\delta=\pm 1} P(N_+, t) W(N_+ \to N_+ + \delta). \end{aligned} \tag{9.1}$$

The first term accounts for processes in which the number of susceptible democrats after the event equals N_+, while the second term accounts for the complementary loss processes where $N_+ \to N_+ \pm 1$. Here W represents the rate at which transitions occur and is given by

$$\begin{aligned} \delta t\, W(N_+ \to N_+ + 1) &= \frac{N_-(N_+ + Z_+)}{N(N-1)} \\ \delta t\, W(N_+ \to N_+ - 1) &= \frac{N_+(N_- + Z_-)}{N(N-1)}. \end{aligned} \tag{9.2}$$

The first line is the probability of choosing first a republican susceptible and then a democrat (susceptible or zealot), for which a susceptible republican converts to a susceptible democrat in the voter model interaction. We choose $\delta t = N^{-1}$, so that, on average, each agent is selected once at each time step.

While it is usually not possible to solve an equation of the form (9.1), analytical progress can be achieved by considering a continuum $N \to \infty$ limit of the master equation and performing a Taylor expansion [147]. For this purpose, we introduce the rescaled variables $n \equiv N_+/N$, $z_\pm = Z_\pm/N$, and also $s \equiv 1 - z_+ - z_-$ so that $s - n \equiv N_-/N$. In the continuum limit, the reaction rates now become

$$\begin{aligned} W(n \to n + N^{-1}) &= N\,(s-n)(n+z_+) \\ W(n \to n - N^{-1}) &= N\,n(s-n+z_-). \end{aligned} \tag{9.3}$$

Expanding (9.1) to the second order in the variable n, we find the following Fokker-Planck equation [146, 148, 149, 150, 151]:

$$\frac{\partial P(n,t)}{\partial t} = -\frac{\partial}{\partial n}\left[\alpha(n) P(n,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial n^2}\left[\beta(n) P(n,t)\right], \tag{9.4}$$

where (see *e.g.*, Chap. VII of Ref. [147])

$$\begin{aligned} \alpha(n) &= \sum_{\delta n = \pm 1/N} \delta n\, W(n \to n + \delta n) \\ &= [z_+ s - n(1-s)]\,; \\ \beta(n) &= \sum_{\delta n = \pm 1/N} (\delta n)^2\, W(n \to n + \delta n) \\ &= [(n+z_+)(s-n) + n(s+z_- - n)]/N. \end{aligned}$$

The first term on the right-hand-side of Eq. (9.4) leads to the deterministic mean-field rate equation $\dot{n}(t) = \alpha$, with solution

$$n(t) \;=\; \frac{z_+ s}{1-s} + \left[n(0) - \frac{z_+ s}{1-s} \right] e^{-(1-s)t}. \tag{9.5}$$

Thus an initial density of susceptible democrats in an infinite system exponentially relaxes to the steady-state value $n_\infty = z_+ s/(1-s)$. Correspondingly, the magnetization $m = (N_+ + Z_+ - N_- - Z_-)/N$ attains the steady-state value $(z_+ - z_-)/(z_+ + z_-)$. When the number of agents is finite, however, finite-size fluctuations arise from the diffusive second term on the right-hand side of Eq. (9.4). This term leads to a steady-state probability distribution with a finite width that is centered at n_∞. In what follows, we examine these fluctuations around the mean-field steady state when N and $Z_\pm$ are both finite.

9.4.1 Stationary Magnetization Distribution

According to the Fokker-Planck equation (9.4), the stationary distribution $P(n)$ obeys

$$\alpha(n) P(n) - \frac{1}{2} \frac{\partial}{\partial n} \left[\beta(n) P(n,t) \right] = 0, \tag{9.6}$$

whose formal solution is

$$P(n) = \mathcal{Z}\, \frac{\exp\left(2 \int_0^n dn' \, \frac{\alpha(n')}{\beta(n')} \right)}{\beta(n)}. \tag{9.7}$$

Since the density n of agents in the state $+1$ ranges from 0 to s, the normalization constant $\mathcal{Z}$ is obtained by requiring $\int_0^s dn P(n) = 1$. This condition gives

$$\mathcal{Z} = \left[\int_0^s \frac{\exp\left(2 \int_0^n dn' \, \frac{\alpha(n')}{\beta(n')} \right)}{\beta(n)} \, dn \right]^{-1}.$$

We are particularly interested in the distribution of the magnetization $P(m)$ in the continuum limit, which directly follows from (9.7) through a simple change of variables. We first consider the system with the same number of zealots of each type, and then the asymmetric system with unequal numbers of zealots of each type.

9.4.2 Symmetric case: $Z_+ = Z_- = Z$

When the number of zealots of each species is equal, we write $Z_+ = Z_- \equiv Z$. The rate equation (9.5) then gives an equal steady-state density of democrats and republicans, $n^* = n_+ = n_- = s/2$, corresponding to zero average magnetization, $m^* = 0$. We now compute the stationary distribution of magnetization by accounting for finite-size fluctuations. When $Z_+ = Z_- = Z$, $P(n)$ obeys Eq. (9.7) with

$$\begin{aligned} \alpha(n) &= z(1-2z-2n), \\ \beta(n) &= [(2n+z)(1-2z) - 2n^2]/N. \end{aligned}$$

Notice that $\alpha = \frac{Nz}{2}\frac{d\beta}{dn}$, a feature that allows us to solve for the steady-state magnetization distribution easily.

To perform the integral in Eq. (9.7), it is helpful to transform from n to the magnetization $m = (2n-s)/s$ which lies in $[-1,1]$. We therefore find $\exp\left(2\int_0^n dn' \frac{\alpha(n')}{\beta(n')}\right) = \left(1 + \frac{2n(s-n)}{zs}\right)^{Nz}$. According to Eq. (9.7), this leads to the following stationary distribution of susceptible democrats:

$$P(n) = \frac{(zs + 2n(s-n))^{Nz-1}}{\int_0^s dn\ (zs + 2n(s-n))^{Nz-1}}. \tag{9.8}$$

Using the fact that $2n(s-n) = s^2(1-m^2)/2$, we readily obtain the stationary magnetization distribution:

$$P(m) = \frac{(s^{-1} - m^2)^{Z-1}}{\int_{-1}^1 dm\ (s^{-1} - m^2)^{Z-1}}. \tag{9.9}$$

In the limit of large Z, we may then approximate the distribution by the Gaussian $P(m) \propto e^{-m^2/2\sigma^2}$, with $\sigma^2 = /[2s(Z-1)]$.

When zealots are present in equal numbers, the magnetization distribution quickly approaches a symmetric Gaussian, with a width that is inversely proportional to the square-root of the *number* of zealots and not the density. Thus as the system size is increased, the density of zealots needed to keep the magnetization within a fixed range goes to zero. In the limiting case where there is one zealot of each type, the magnetization is uniformly distributed in $[-1,1]$ (Fig. 9.1). Finally notice that the distribution quickly approaches the asymptotic scaling form when $Z \gtrsim 8$ (figure inset).

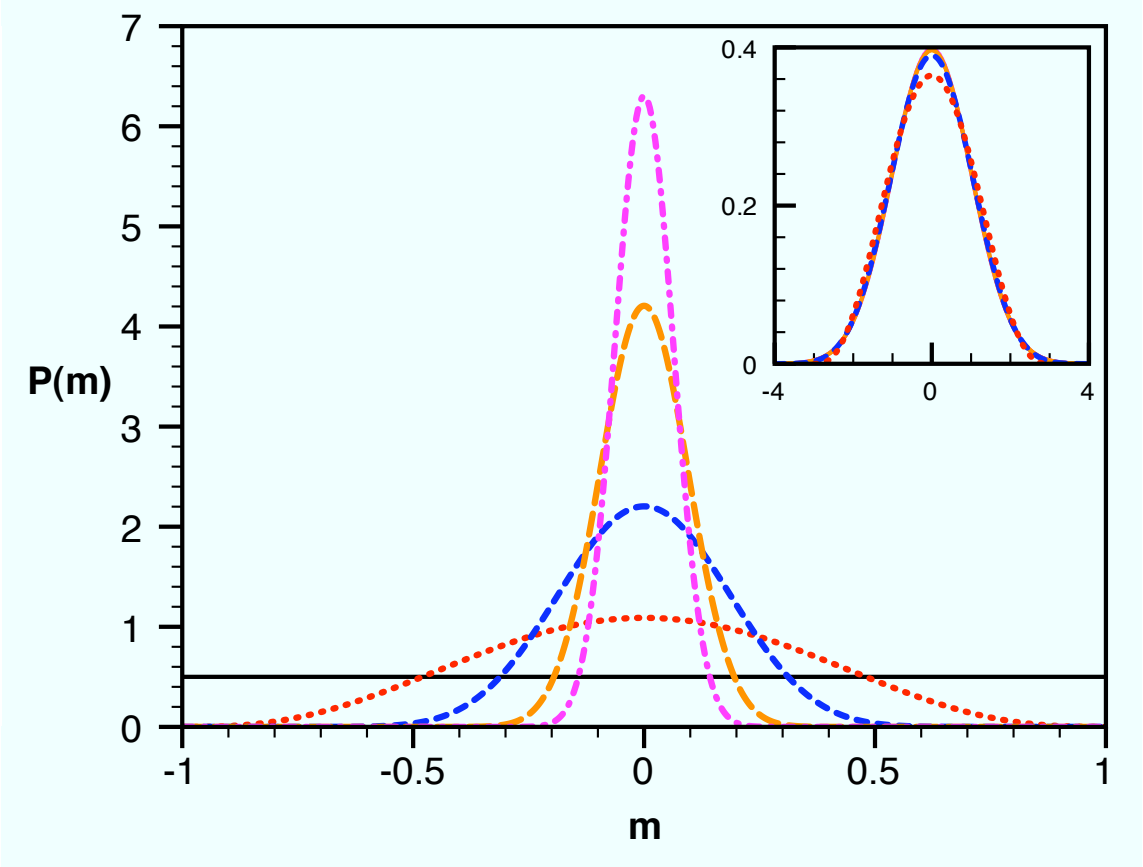

Figure 9.1: Steady-state magnetization distributions for 1000 voters on the complete graph for $Z = 2, 8, 32, 128$, and 512 zealots (progressively steepening curves). The inset shows the scaled form of these distributions for $Z \geq 8$; the case $Z = 8$ slightly deviates from the rest of the distributions that become visually coincident.

9.4.3 Asymmetric case: $Z_+ \neq Z_-$

When the density of zealots of each type are unequal, we now have

$$\alpha(n) = (z_+ + n)s - n, \tag{9.10}$$

$$\beta(n) = [(2n + z_+)(s - n) + nz_-]/N \tag{9.11}$$

in Eq. (9.6). To compute $P(n)$ (and equivalently $P(m)$), it is now convenient to introduce the quantities $\delta \equiv z_+ - z_-$ and $r \equiv \sqrt{\delta^2 + 4s}$. Noticing that one can write $\alpha/\beta = [N(s-1)(d\beta/dn) + \delta(1+s)/4]/\beta$, one can easily compute the integral in Eq. (9.7) and thereby obtain $P(n)$. Transforming from the density to the magnetization by $n = (m+1)s/2$, we obtain the following expression for the stationary magnetization distribution (Fig. 9.2):

$$\begin{aligned} \mathcal{Z}P(m) &= [1 - m(\delta + ms)]^{(Z_+ + Z_- - 2)/2} \\ &\times \left[1 + \frac{r}{ms - \frac{r-\delta}{2}}\right]^{(\delta/2r)\,(2N - Z_+ - Z_-)} . \end{aligned} \tag{9.12}$$

As in the symmetric case, $\mathcal{Z}$ is a normalization constant obtained by requiring that $\int_{-1}^{1} dm\, P(m) = 1$. Notice that $P(m)$ is comprised of two terms. The first term gives a Gaussian contribution (in the limit of large N) and is the analog of Eq. (9.9). The second term is a nontrivial contribution due to the asymmetry that is responsible for the skewness of $P(m)$ which remains peaked around $m^* = \frac{z_+ - z_-}{z_+ + z_-}$. Close to this peak value, there is little asymmetry (*i.e.*, $\delta \ll 1$). Additionally, for a large number of zealots we may approximate the distribution (9.12) by the Gaussian $P(m) \approx e^{-(m-m^*)^2/2\sigma^2}\,[1 + \mathcal{O}((m-m^*)\delta))]$, with $\sigma^2 = [c(Z_+ + Z_- - 2)]^{-1}$.

9.5 One Dimension

We now turn to the one-dimensional system, where the behavior of the classical voter model is quite different from that in the mean-field limit. When zealots are present, however, we generically obtain a Gaussian magnetization distribution, as

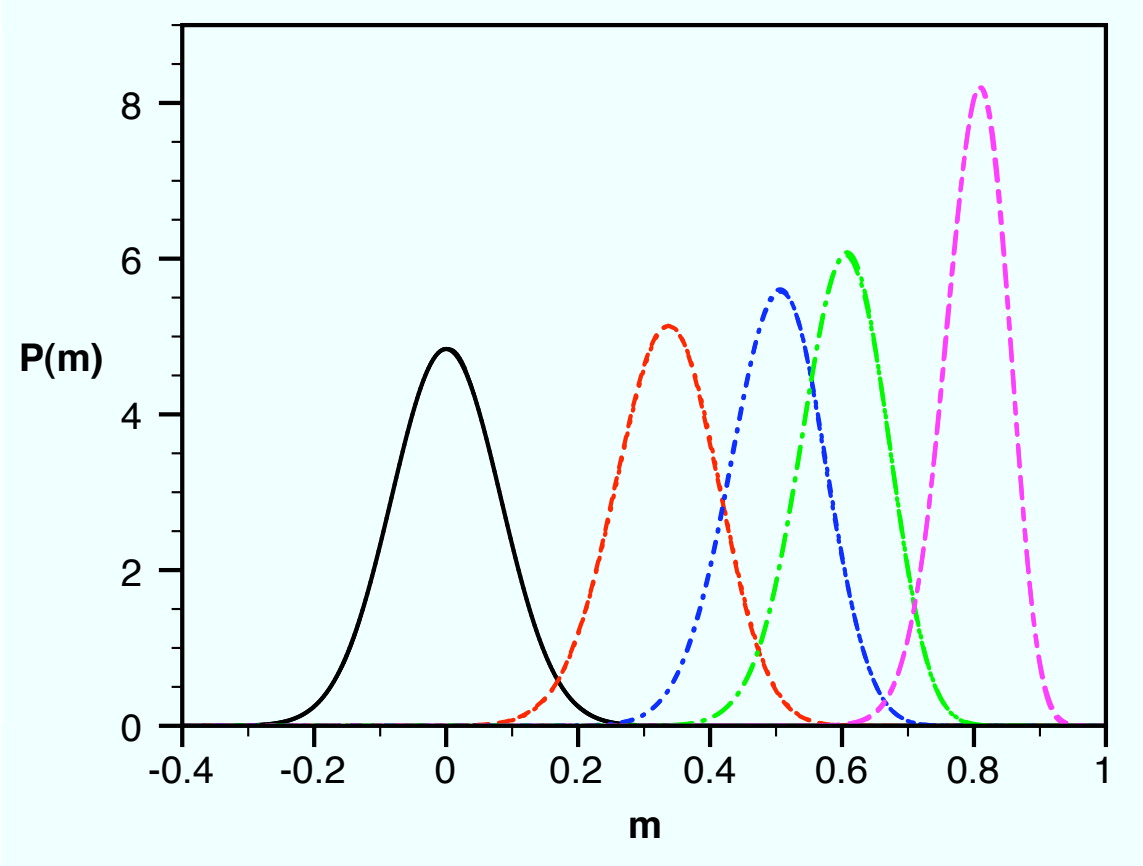

Figure 9.2: Steady-state magnetization distributions on a complete graph of 1000 sites with unequal numbers of zealots. Shown left to right are the cases of $(Z_+, Z_-) = (90, 90)$, $(120, 60)$, $(135, 45)$, $(144, 36)$, $(162, 18)$. The results of voter model simulations and the solution to the master equations are coincident. The mean magnetization of the system equals the magnetization of the zealots: $m = \frac{z_+ - z_-}{z_+ + z_-}$.

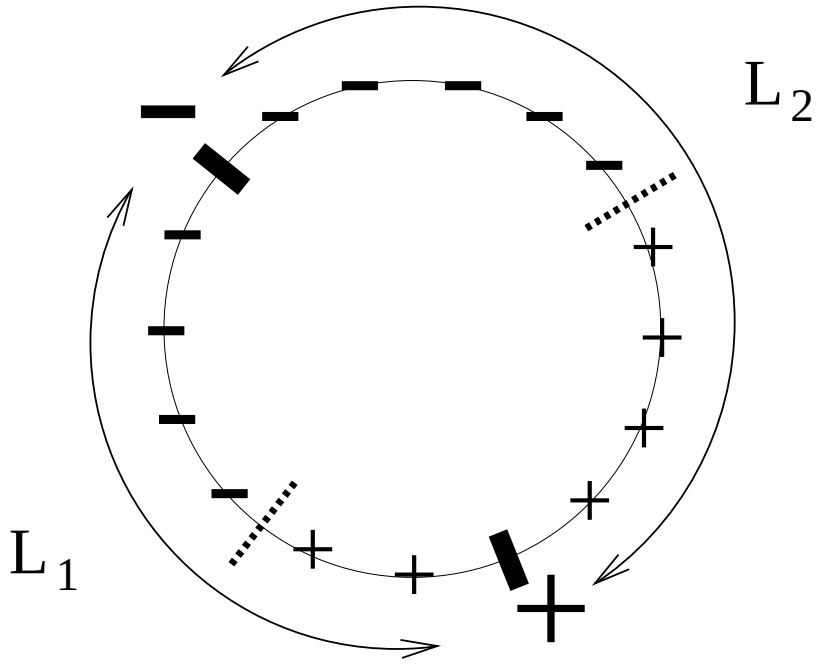

Figure 9.3: A ring divided into two independent segments by oppositely-oriented zealots (thick lines). Also shown is the state of each voter and the domain wall in each segment at long times (dotted lines).

in the mean-field case. We now derive the magnetization distribution—first for two zealots—and then for an arbitrary number of zealots.

9.5.1 Two Zealots

Suppose that two zealots of opposite opinion are randomly placed on a periodic ring of length L. The ring is thus split into two independent segments of lengths L_1 and L_2, with $L = L_1 + L_2 + 2$ (Fig. 9.3). We take the ring to be large so that we may write $L \approx L_1 + L_2$. As shown in Fig. 9.3, the voters in each segment coarsen and eventually there remains one domain of + voters that is separated from one domain of − voters by a single domain wall. Each domain wall performs a free random walk and the walk is reflected upon reaching the end of its segment. A basic fact from the theory of random walks is that each domain wall is equiprobably located within the interval in the long-time limit. We now exploit this property to determine the magnetization distribution.

For interval lengths L_1 and L_2 and respective magnetizations m_1 and m_2, the magnetization m of the entire ring is given by $mL = m_1L_1 + m_2L_2$. Thus a given value of m is achieved if m_1 and m_2 are related by (Fig. 9.4)

$$m_2 = \frac{mL}{L_2} - \frac{m_1L_1}{L_2}. \tag{9.13}$$

Then the probability $P(m|L_1, L_2)$ for a system of two segments with lengths L_1 and L_2 to have magnetization equal to m is proportional to the length of the ray defined by Eq. (9.13) that lies within the unit square in the m_1-m_2 plane. As illustrated in Fig. 9.4, the distribution $P_<(m|L_1, L_2)$, where the subscript $<$ now signifies the range $L_1 < L_2$, increases linearly with m for $-1 < m < (L_1 - L_2)/L$, is constant for $(L_1-L_2)/L < m < (L_2-L_1)/L$, and then decreases linearly with m for $(L_2-L_1)/L < m < 1$.

Using this m dependence of $P_<(m|L_1, L_2)$ and also imposing normalization, we thus find the magnetization distribution for fixed L_1, L_2 with $L_1 < L_2$ to be:

$$P_<(m|L_1, L_2) = \begin{cases} \dfrac{L^2(1+m)}{4L_1L_2} & -1 < m < \frac{L_1-L_2}{L} \\[2ex] \dfrac{L}{2L_2} & |m| < \frac{L_2-L_1}{L} \\[2ex] \dfrac{L^2(1-m)}{4L_1L_2} & \frac{L_2-L_1}{L} < m < 1. \end{cases} \tag{9.14}$$

The complementary distribution $P_>(m|L_1, L_2)$ for $L_1 > L_2$ is obtained from Eq. (9.14) by interchanging the roles of L_1 and L_2.

Now we integrate over all values of L_1 to find the configuration-averaged magnetization distribution $P(m)$. The details of this calculation are a bit tedious and are given in Appendix 9.8. The final result is

$$\begin{aligned} P(m) &= \frac{1}{L}\left[\int_0^{\frac{L}{2}} P_<(m|L_1, L_2)\, dL_1 + \int_{\frac{L}{2}}^{L} P_>(m|L_1, L_2)\, dL_1\right] \\ &= \left(\frac{1-|m|}{2}\right)\ln\left(\frac{1+|m|}{1-|m|}\right) - \ln\left(\frac{1+|m|}{2}\right). \end{aligned} \tag{9.15}$$

As shown in Fig. 9.5, the agreement between Eq. (9.15) and simulations is excellent.

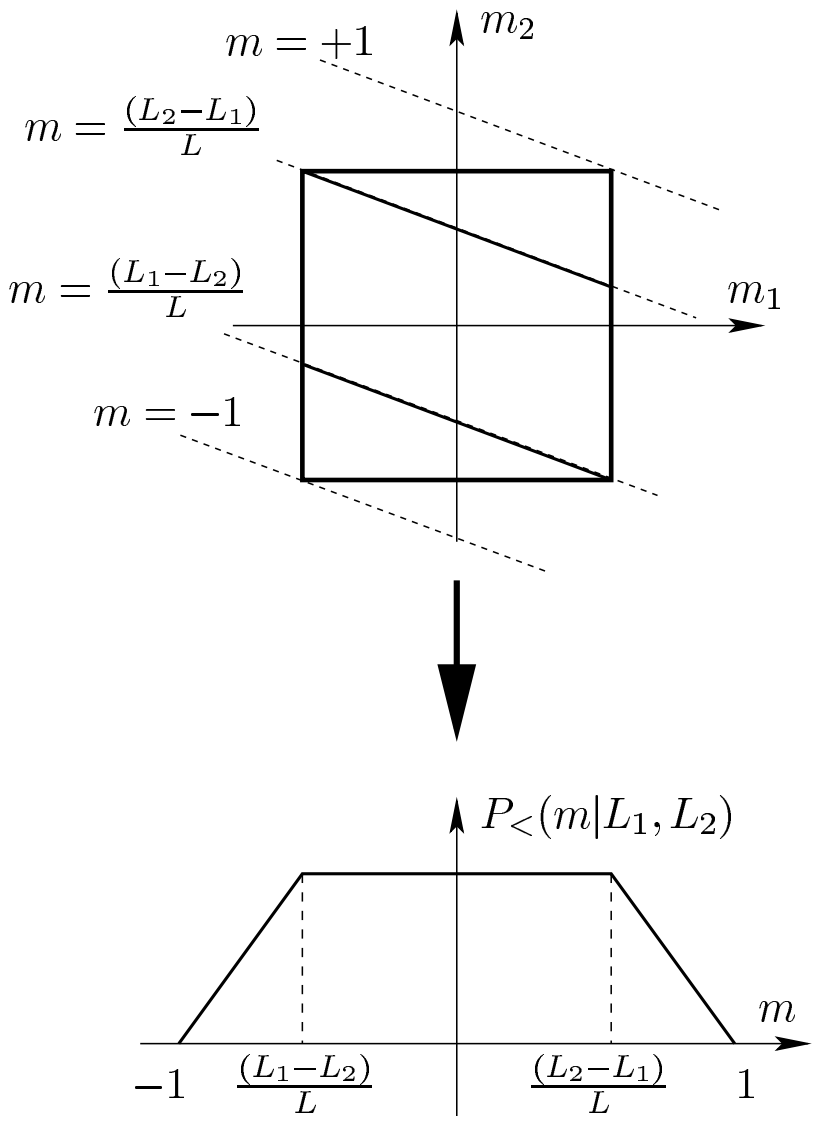

Figure 9.4: (Top) Rays of fixed magnetization (dashed) for the case $L_1 < L_2$. The probability for a given value of m is proportional to the length of the ray corresponding to this m value within the unit square (solid). (Bottom) The resulting magnetization distribution $P_<(m|L_1, L_2)$ for a given L_1 and $L_1 < L_2$.

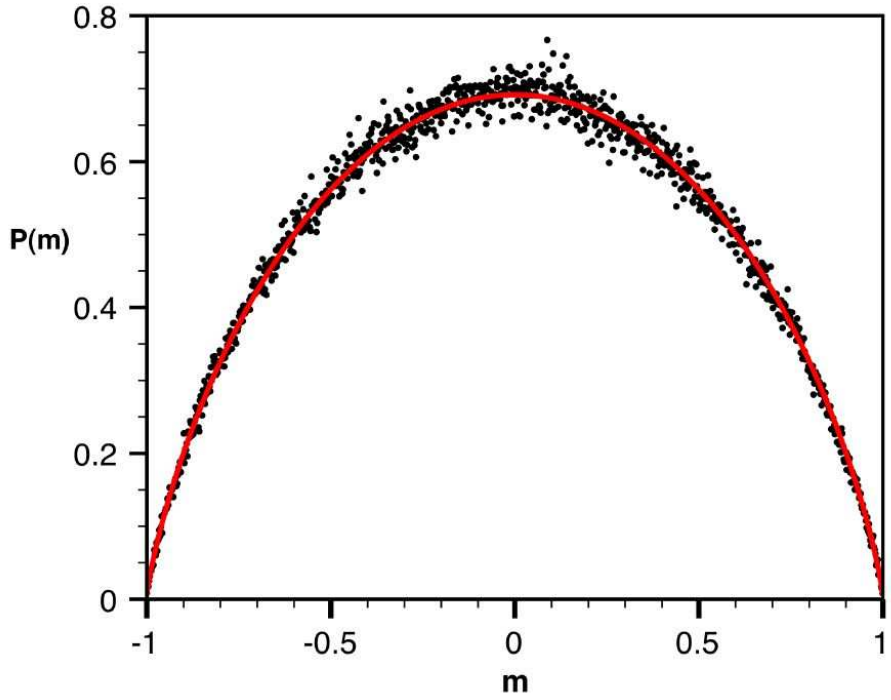

Figure 9.5: Comparison the analytic magnetization distribution for two zealots on the ring (Eq. (9.15)) and simulation results (points).

9.5.2 Many Zealots

We now study the magnetization distribution when many zealots are randomly distributed on the ring, with the restriction of equal numbers of each type of zealot ($Z_+ = Z_- = Z$). Two distinct possibilities can arise:

1. A segment of consecutive susceptible voters is surrounded by two zealots of the same sign. With voter model dynamics, these segments eventually align with the state of the confining zealots so that the segment freezes.

2. A segment of consecutive susceptible voters is surrounded by two zealots of opposite opinion. Eventually a single domain wall remains that diffuses freely within the segment.

We first consider the simpler case where equal numbers of + and − zealots are randomly but alternately placed around the ring so that no frozen segments arise. The segment lengths $\{L_i\}$ with $i = 1, 2, \ldots, Z$, obey the constraint $\sum_i L_i = L$ (ignoring the space occupied by the zealots themselves).

To find the magnetization distribution, we map the state of the voters onto an equivalent random walk as follows. In a segment of length L_i, the difference in the number of + and − voters at long times is uniformly distributed in $[-L_i, L_i]$. We define this difference as the unnormalized magnetization M_i. We now make the following approximations that apply when $L, Z \to \infty$ such that each L_i is also large. In this limit, we may assume that each L_i is independent and identically distributed. As a result, the sum of the unnormalized magnetizations over all intervals is equivalent to the displacement of a random walk of Z steps with each step uniformly distributed in $[-L_i, L_i]$.

To solve this random walk problem, we use the basic fact that the Fourier transform for the probability distribution of the entire walk $\mathcal{P}(k)$ is simply the product of the Fourier transforms of the single-step distributions [134, **?**]. Since the Fourier transform of a uniform single-step distribution over the range $[-L_i, L_i]$ is $\frac{sinkL_i}{kL_i}$, we then have

$$\mathcal{P}(k) = \prod_{i=1}^{Z} \frac{\sin kL_i}{kL_i}. \tag{9.16}$$

Since we are interested in the asymptotic limit where the unnormalized magnetization becomes large, we study the limit of $\mathcal{P}(k)$ for small k. Thus we expand each factor in $\mathcal{P}(k)$ in a Taylor series to first order, and then re-exponentiate to yield

$$\begin{aligned}\mathcal{P}(k) &\approx \prod_{i=1}^{Z}(1-\sum_i k^2 L_i^2/6) \\ &\sim 1-\sum_i^Z k^2 L_i^2/6 \sim e^{-k^2\sum_i L_i^2/6}.\end{aligned}$$

We now invert this Fourier transform to give the distribution of the unnormalized magnetization

$$\begin{aligned}P(M) &= \frac{1}{2\pi}\int e^{-k^2\sum_i L_i^2/6}\, e^{-ikM}\, dk \\ &= \frac{1}{\sqrt{2\pi\sigma_M^2}}\, e^{-M^2/2\sigma_M^2}, \end{aligned} \tag{9.17}$$

with $\sigma_M^2 = \sum_i L_i^2/3$.

What we want, however, is the magnetization distribution; this is related to $P(M)$ by $P(m)\,dm = P(M)\,dM$. We thus find

$$P(m) = \frac{1}{\sqrt{2\pi\sigma_m^2}}\, e^{-m^2/2\sigma_m^2}, \tag{9.18}$$

where $\sigma_m^2 = \sum_i L_i^2/3L^2$. If the number of intervals is large, then each L_i is approximately L/Z, from which we obtain $\sigma_m^2 \approx 1/3Z$. (The result $\sigma_m^2 = 1/3Z$ is exact if all interval lengths are equal.) As in the mean-field limit, the width of the magnetization distribution is controlled by the *number* of zealots and not their concentration, so that a small number of zealots is effective in maintaining the magnetization close to zero.

A similar approach applies in the case where the spatial ordering of the zealots is uncorrelated. In this case, approximately one-half of all segments will be terminated by oppositely-oriented zealots and one-half by zealots of the same species. For the latter type of segments, the unnormalized magnetization will equal $\pm L_i$ equiprobably. Under the assumption that exactly one-half of the segments are frozen and one-half contain a single freely-diffusing domain wall, the analog of Eq. (9.16) is

$$\mathcal{P}(k) = \prod_{i=1}^{Z/2}\frac{\sin kL_i}{kL_i} \quad \prod_{i=1}^{Z/2}\cos kL_i. \tag{9.19}$$

The second product accounts for frozen segments in which the unnormalized magnetization equals $\pm L_i$ equiprobably. For these segments the Fourier transform of the single-step probability for a random walk whose steps length are $\pm L_i$ equals $\cos kL_i$. Following the same steps that led to Eq. (9.18), we again obtain a Gaussian magnetization distribution, but with σ_m^2 given by $\sigma_m^2 = \sum_i 2L_i^2/3L^2 \to 2/3Z$.

9.6 Two Dimensions

In the classical voter model, the two-dimensional system is at the critical dimension so that its behavior deviates from that of the mean-field system by logarithmic corrections. In the presence of zealots, however, the behaviors in two dimensions and in mean field are quite close, as illustrated in Fig. 9.6.

Our results for two dimensions are based on numerical simulations. In our simulations, we pick a random voter and apply the update rules of Sec. 9.3. The unit of time is defined so that a time increment $dt = 1$ corresponds to N update events, so that each voter is updated once on average. The system is initialized with each voter equally likely to be in the + or the − states. From the N_+ voters in the + state, Z_+ of them are designated as zealots, and similarly for voters in the − state. After the system reaches the steady state, we measure steady-state properties at time intervals ΔT. The delay time T to reach the steady state depends on the lattice dimension and the zealot density, while ΔT is the correlation time for the system in the steady state. By making measurements every ΔT steps, we obtain data for effectively uncorrelated systems. Typically, for a given initial condition, we made 100 measurements and then averaged over many configurations of zealots.

The resulting data for the magnetization distribution is typically noisy, and we employ a Gaussian averaging of nearby points to smooth the data. If m_i denotes the i^{th} magnetization value, then the smoothed magnetization distribution at m_i is defined as

$$\overline{P(m_i)} = \frac{1}{\sqrt{\pi d^2}} \sum_{k=-d}^{d} P(m_{i+k})\, e^{-(k/d)^2},$$

where the sum includes the initial point, as well as the d points to the left and to the right of the initial point, with d typically in the range 20–40. Such a smoothed

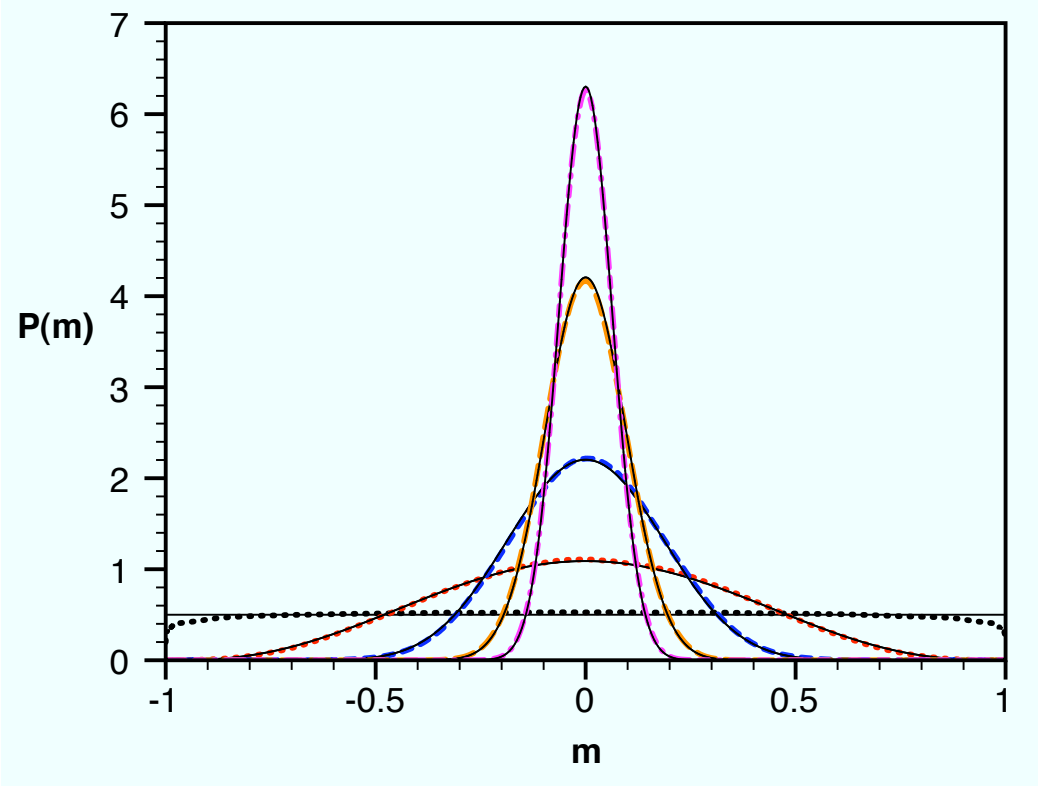

Figure 9.6: Comparison of simulations for the magnetization distribution in two dimensions (dashed) with the mean-field results (solid curves). The simulations are for 1000 voters with $2, 8, 32, 128$ and 512 total zealots, with equal numbers of each type.

distribution is the quantity that is actually plotted in Figs. 9.1, 9.2, 9.6, and in the spatial averaged distribution in Fig. 9.7.

9.7 Discussion

We have shown that a small number of zealots in a population of voters is quite effective in maintaining a steady state in which consensus is never achieved. When there are equal numbers of zealots of each type, the steady-state fraction of democrats and republicans equals 1/2; equivalently, the magnetization equals zero. For unequal densities of the two types of zealots, the steady-state magnetization is simply $m^* = (Z_+ - Z_-)/(Z_+ + Z_-)$, where Z_+ and Z_- are the number of zealots of each type. The magnetization distribution is generically Gaussian, $P(m) \propto e^{-(m-m^*)^2/2\sigma^2}$, with $\sigma \propto 1/\sqrt{Z}$, and $Z = Z_+ + Z_-$ is the total number of zealots. A Gaussian magnetization distribution arises universally in one dimension, on the square lattice (two dimensions), and on the complete graph (mean-field limit). One basic consequence of this distribution is that as the total number of voters N increases, the fraction of zealots needed to keep the magnetization less than a specified level vanishes as $1/\sqrt{N}$.

There are several additional aspects of the influence that zealots have on the voter model that are worth pointing out. Although the time to reach consensus is infinite because this state can never be achieved, one can ask for the time until a specified plurality is first achieved. Equivalently, we can ask for the probability that the magnetization first reaches a value m, when the system is initialized with $m = m_0$. From the above generic Gaussian form of the magnetization distribution, we expect that the mean time for a symmetric system to first reach a magnetization m will thus scale as e^{am^2Z}, where a is a constant of order one. Thus one must wait an extremely long time before the system achieves even a modest deviation away from the zero-magnetization state when the number of zealots becomes appreciable. Perhaps this trivial fact is the underlying reason why so many democratic countries are characterized by small majorities in governance.

Another interesting feature is the role of the zealots' spatial positions on the steady state. For example, if there are only two zealots that are adjacent, one might expect that the effect of this "dipole" would be weaker than that of two separated monopoles.

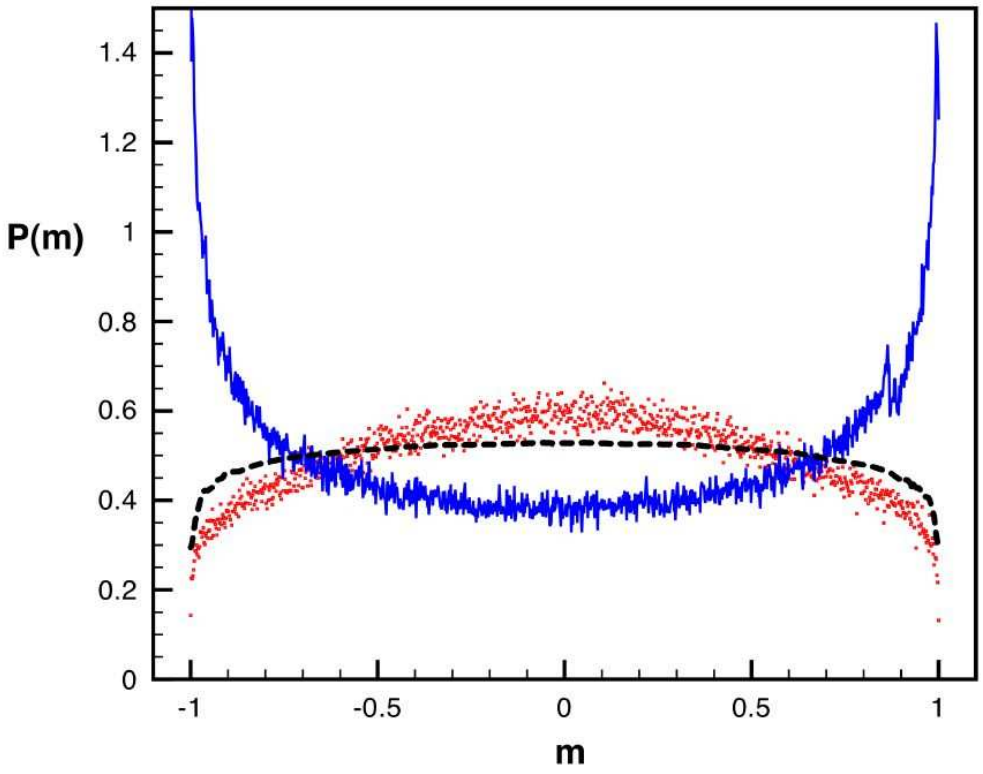

Figure 9.7: Comparison of simulations for the magnetization distribution in two dimensions when the two zealots are adjacent (curve with peaks near ± 1), maximally separated (dots), and averaged over many different zealot configurations (dashed).

This is precisely the effect that is is observed in Fig. 9.7. When the two zealots are adjacent, their effects are substantially screened and the magnetization distribution is peaked near $m = \pm 1$. That is, the voters show a preference for consensus in spite of the zealots. On the other hand, when the zealots are maximally separated, the magnetization distribution is close to the distribution that arises when averaging over possible positions of the two zealots.

Zealots are also quite effective in reducing the total number of opinion changes in the system. If the population is close to zero magnetization, each voter typically has equal numbers of neighbors of each type. If the voters are not strongly correlated, each voter would change its state at a rate that is approximately equal to 1/2. However, simulations on the square lattice show that the flip rate of each susceptible voter is considerably smaller. For example, for 1000 voters with 10 zealots (5 of each type), the rate of opinion changes of the susceptibles is around 1/5 and this rate decreases as the density of zealots decreases.

Finally, a slight embellishment of our model could apply to real voting patterns in a democracy with strong regional differences. Here it is natural to partition a population into enclaves, with an imbalance of one type of zealot over the other in each enclave. Such a spatial distribution would correspond to red (republican) and blue (democrat) states in the parlance of US electoral politics. It would be interesting to study if such an extension can actually account for real voting patterns.

9.8 Magnetization distribution for two zealots

We want to compute the integral

$$P(m) = \frac{1}{L}\left[\int_0^{\frac{L}{2}} P_<(m|L_1, L_2)\, dL_1 + \int_{\frac{L}{2}}^{L} P_>(m|L_1, L_2)\, dL_1\right]. \qquad (9.20)$$

Since $P_<(m|L_1, L_2)$ and $P_>(m|L_1, L_2)$ have different forms in different parts of the interval $[-1, 1]$, each of the above integrals needs to be split into two parts. For $P_<(m|L_1, L_2)$ and assuming that $m > 0$, the linear ramp part of the probability distribution needs to be used for $(L_2 - L_1)/L < m$, which translates for $L_1 > L(1 - m)/2$. Similarly, for $P_>(m|L_1, L_2)$ and again for $m > 0$, the linear ramp must be used

when $(L_1 - L_2)/L < m$, or $L_1 < L(1+m)/2$. Thus the above integral becomes

$$\begin{aligned} P(m) &= \frac{1}{L}\Big[\int_0^{\frac{L}{2}(1-m)} \frac{L\,dL_1}{2(L-L_1)} + \int_{\frac{L}{2}(1-m)}^{\frac{L}{2}} \frac{(1-m)L^2\,dL_1}{4L_1(L-L_1)} \\ &+ \int_{\frac{L}{2}}^{\frac{L}{2}(1+m)} \frac{(1-m)L^2\,dL_1}{4L_1(L-L_1)} + \int_{\frac{L}{2}(1+m)}^{L} \frac{L\,dL_1}{2L_1}\Big]. \end{aligned} \tag{9.21}$$

Each of these integrals is then elementary. We also obtain the result for $m < 0$ by reflecting the result of the above integral about $m = 0$ to give Eq. (9.15).

Chapter 10

Conclusion

This thesis covers work done in three distinct systems where the emergent phenomena is fundamentally related to the underlying social dynamics of humans. Drawing on methods and concepts from statistical physics, we search for statistical patterns that emerge from the complex interactions between individuals in three distinct settings (i) financial markets driven by speculative trading, (ii) competitive professions that are distinguished by productive output measures that are readily quantifiable, and (iii) systems of voters that exchange opinions with their neighbors in an arbitrary contact network.

As a society, we have entered a information age that is marked by the ability to record and store observables that quantify everyday life on both the micro and macro scale. The physical sciences have long benefited from the fundamental time-independent laws of nature, which for several centuries have facilitated the development of novel and repeatable experiments that can either verify or reject existing theories about our physical world. In the last century, economics became a true empirical science with the advent of macroscopic measures like country GDP, and microscopic measures like the price of a stock on a market, which can be used to monitor and quantify the evolution of institutions ranging in size from the nation state to a company, and further discriminate between "good" and "bad" economic theories. However, the social sciences have lagged behind the other empirical sciences because in the past, it has been difficult to obtain human activity data that is both

sufficiently large and sufficiently precise. But we have now entered a new era, and as a result we shall surely see the subfield of quantitative sociology make giant leaps in the years to come.

So what do we expect to find in the new era of quantitative sociology? A common theme in large quantitative studies is the discovery of rare events that occur much more frequently or much less frequently than traditional models predict. For example, the Guttenberg-Richer law quantifies the likelihood of observing an earthquake of a particular magnitude. Before this law was established, there wasn't any formal understanding of the occurrence of earthquakes of a given size, and there was little reason to expect the power-law relation between the frequencies. While seismic activity certainly depends on the geographic region, the earthquake probability distribution can be used to measure the risk of catastrophic disaster, which is critical in many aspects of urban planning. In economics, the standard theories of the last century predict that stock market crashes would not happen, yet large crashes typically occur once a decade. Thus, the empirical probability distributions of price fluctuations in economics can be used to develop investment strategies that incorporate realistic levels of risk. Furthermore, just as in the case of earthquakes, financial markets must incorporate the likelihood of market crashes in the development of financial crisis policy, and in general, policies that can curb dangerous financial activities by banks and large funds.

In this thesis, we also discuss the distributions of career longevity which show a remarkable pattern between the frequencies of long and short careers. Amazingly, we find that most professional sports careers, at the highest level of competition, are terminated within minutes of beginning. In a similar way, most scientists finish their career with just a couple publications in high-impact journals, which presumably everyone aims for, yet there are a few individuals who are able to publish in the high-impact journals repeatedly. Thus, competition which can be a barrier for success, can result in a highly right-skewed distribution of career achievement. It is possible that reinforcement policies could possibly diminish the relative disadvantage of inexperienced individuals so to help young careers reach early career milestones, e.g. young scientist fellowships.

A crucial next step, in addition to data analysis, is to implement policy that in-

corporates our new understanding of social phenomena. This can be difficult, because the emergent patterns can be very robust to details of the underlying system, a tenet of universality in complex systems. As a result, a diagnostic case-by-case approach to change the frequencies of large (or small) events can be time consuming. Thus, a better understanding of the underlying dynamic processes is the next step towards using the vast information that is available to improve the efficiency and capability of modern social systems.

Bibliography

[1] H. Takayasu, H. Miura, T. Hirabayashi, and K. Hamada, Physica A **184**, 127 (1992).

[2] R. N. Mantegna and H. E. Stanley, *Econophysics: An Introduction* (Cambridge University Press, Cambridge England, 1999).

[3] J. P. Bouchaud, M. Potters *Theory of Financial Risk*, (Cambridge University Press, Cambridge England, 2000).

[4] J. P. Bouchaud, Quantitative Finance **1**, 105 (2001).

[5] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, Nature **423**, 267 (2003).

[6] J.D. Farmer, M. Shubik, and E. Smith, Physics Today **58(9)**, 37 (2005).

[7] H. E. Stanley, V. Plerou and X. Gabaix, Physica A **387**, 3967 (2008).

[8] F. Omori, Journal of the College of Science, Imperial University of Tokyo **7**, 111 (1894).

[9] T. Utsu, Geophysical Magazine **30**, 521 (1961).

[10] F. Lillo, R. N. Mantegna, Physical Review E **68**, 016119 (2003).

[11] P. Weber, F. Wang, I. Vodenska-Chitkushev, S. Havlin and H. E. Stanley, Physical Review E **76**, 016109 (2007).

[12] A. Almeida, C. Goodhart and R. Payne, Journal of Financial and Quantitative Analysis **33**, 383 (1998).

[13] T. G. Andersen, T. Bollerslev, F. X. Diebold, and C. Vega, American Economic Review **93**, 38 (2003).

[14] N. Barberis, A. Shleifer, R. Vishny Journal of Financial Economics **49**, 307 (1998).

[15] V. L. Bernard, J. K. Thomas, Journal of Accounting and Economics **13**, 305 (1990).

[16] W. N. Goetzmann, L. Li, K. G. Rouwenhorst, Journal of Business **78**, 1 (2005).

[17] Historical Data for key Federal Reserve Interest Rates:
`http://www.federalreserve.gov/releases/h15/data.htm`
`http://www.federalreserve.gov/fomc/fundsrate.htm`

[18] J. D. Hamilton, O. Jorda, Journal of Political Economy **110**, 1135 (2002).

[19] J. B. Calrson, B. R. Craig and W. R. Melick, Working Paper of the Federal Reserve Bank of Cleveland, 1-50 (2005).

[20] M. Piazzesi, E. Swanson, Working Paper of the National Bureau of Economic Research **10547**, 1 (2002).

[21] K. N. Kuttner, Journal of Monetary Economics **47**, 523 (2001).

[22] S. Kwan, FRBSF Economic Letter **15**, 1 (2007).

[23] J. B. Carlson, B. Craig, P. Higgins and W. R. Melick Economic Commentary of the Federal Reserve Bank of Cleveland, 1-3 (2006).

[24] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley, Proceedings of the National Academy of Sciences USA **106**, 22079 (2009).

[25] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, and H. E. Stanley, Proceedings of the National Academy of Sciences USA **102**, 9424 (2005).

[26] F. Wang, K. Yamasaki, S. Havlin, and H. E. Stanley, Physical Review E **73**, 026117 (2006).

[27] F. Wang, P. Weber, K. Yamasaki, S. Havlin, and H. E. Stanley, European Physics Journal B **55**, 123 (2007).

[28] R. N. Mantegna, European Physics Journal B **11**, 193 (1999).

[29] L. Laloux, P. Cizeau, J. P. Bouchaud and M. Potters, Physical Review Letters **83**, 1467 (1999).

[30] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Physical Review Letters **83**, 1471 (1999).

[31] D. Sornette, A. Helmstetter, Physica A **318**, 577 (2003).

[32] D. Sornette, Y. Malevergne, J. F. Muzy, Risk **16**, 67 (2003).

[33] A. Joulin, A. Lefevre, D. Grunberg, J. P. Bouchaud, Wilmott Magazine **Sept/Oct 46**, 1 (2008). (arXiv:cond-mat/0803.1769).

[34] O. Guedj, J. P. Bouchaud, arXiv:cond-mat/0410079: 1-15 (2004).

[35] R. Crane, D. Sornette, Proceedings of the National Academy of Sciences USA **105**, 15649 (2008).

[36] R. Cont, M. Potters, J. P. Bouchaud, Proceedings of the Les Houches workshop, Les Houches, France, March 10-14, 1997: 1-11.

[37] Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, Physica A **245**, 437 (1999).

[38] P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Physical Review E **60**, 5305 (1999).

[39] A. G. Zawadowsk, G. Andor, J. Kertész, Quantitative Finance **6**, 283 (2006).

[40] D. Morse, Journal of Accounting Research **19**, 374 (1981).

[41] O. Kim and R. E. Verrecchia, Journal of Accounting Research **29**, 302 (1991).

[42] A. M. Petersen, F. Wang, S. Havlin, H. E. Stanley, Physical Review E. **82**, 036114 (2010).

[43] H. E. Stanley, Reviews of Modern Physics **71**: S358 (1999).

[44] Z. Eisler, I. Bartos, J. Kertész, Physical Review E **73**, 046109 (2006).

[45] Z. Eisler, I. Bartos, J. Kertész, Advances in Physics **57**, 89 (2008).

[46] A. Ponzi, F. Lillo, R. N. Mantegna Physical Review E **80**, 016112 (2009).

[47] E. F. Fama, Journal of Business **38**, 34 (1965).

[48] Z. Ding, C.W.J. Granger and R.F. Engle, Journal of Empirical Finance **1**, 83 (1993).

[49] B.B. Mandelbrot, Journal of Business **36**, 394 (1963).

[50] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, Quarterly Journal of Economics **121**, 461 (2006).

[51] S. Ghashgaie, W. Breymann, J. Peinke, P. Talkner, and Y. Dodge, Nature **381**, 767(1996).

[52] A.M. Petersen, F. Wang, S. Havlin, and H.E. Stanley, Physical Review E **81** 066121 (2010).

[53] New York Times article archive, 1981-Present. `http://www.nytimes.com/ref/membercenter/nytarchive.html`

[54] F. Wang, S.-J. Shieh, S. Havlin, and H. E. Stanley, Physical Review E **79**, 056109 (2009).

[55] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Guhr, and H. E. Stanley, Physical Review E **65**, 066126 (2002).

[56] V. Plerou, P. Gopikrishnan, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Physical Review E **60**, 6519 (1999).

[57] L. de Arcangelis, C. Godano, E. Lippiello, and M. Nicodemi, Physical Review Letters **96**, 051102 (2006).

[58] L. de Arcangelis, E. Lippiello, C. Godano, and M. Nicodemi, European Physics Journal B **64**, 551-2008 (2008).

[59] G. Drakatos and J. Latoussakis, Journal of Seismology **5**, 137-145 (2001).

[60] A. Helmstetter, Physical Review Letters **91**, 058501 (2003).

[61] A. Saichev and D. Sornette, Tectonophysics **431**, 7-13 (2007).

[62] A. Saichev and D. Sornette, Physical Review E **70**, 046123 (2004).

[63] M. Bottiglieri, L. de Arcangelis, C. Godano, and E. Lippiello, Physical Review Letters **104**, 158501 (2010).

[64] A. N. Kolmogorov, Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences, **434**, 9 (1991).

[65] P. Ch. Ivanov, L. A. N. Amaral, A. L. Goldberger, S. Havlin, M. G. Rosenblum, Z. Struzik, and H. E. Stanley, Nature **399**, 461-465 (1999).

[66] R. L. Hughson and D.C. Lin, Physical Review Letters **86**, 1650 (2000).

[67] B. Suki, A.-L. Barabasi, Z. Hantos, F. Petak, and H. E. Stanley, Nature **368**, 615 (1994).

[68] A. M. Alencar, S. V. Buldyrev, A. Majumdar, H. E. Stanley, and B. Suki, Physical Review Letters **87**, 088101 (2001).

[69] T. Petermann, T.C. Thiagarajan, M.A. Lebedev, M.A.L. Nicolelis, and D.R. Chialvo, Proceedings of the National Academy of Sciences USA **106**, 15921(2009).

[70] D. Plenz, D.R Chialvo, Arxiv pre-print: (arXiv:0912.5369).

[71] J. Touboul, A. Destexhe, PLoS ONE **5**, e8982 (2010).

[72] A. Vázquez, J. G. Oliveira, Z. Dezso, K.-I. Goh, I. Kondor, and A.-L. Barabási, Physical Review E **73**, 036127 (2006).

[73] K.-L. Goh and A.-L. Barabási, Europhysics Letters **81**, 48002 (2008).

[74] D. Rybski, S. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Proceedings of the National Academy of Sciences USA **106**, 12640 (2009).

[75] T. Preis and H. E. Stanley, Journal of Statistical Physics **138**, 431 (2010).

[76] B.E. Baaquie. *Quantum finance: Path integrals and Hamiltonians for options and interest rates*. (Cambridge University Press, UK, 2004).

[77] J.P. Bouchaud and D. Sornette, Journal de Physique I France **4**, 863 (1994).

[78] G. J. Mitchell, *Report to the Commissioner of Baseball of an Independent Investigation into the Illegal Use of Steroids and other Performance Enhancing Substances by Players in Major League Baseball* (Office of the Commissioner of Baseball, 2007).

[79] Sean Lahman's Baseball Archive: `http://baseball1.com/index.php`

[80] R. Albert, H. Jeong, A.-L. Barabási, Nature **401**, 130 (1999).

[81] F. Liljeros, C. Edling, L. Amaral, H. E. Stanley, Y. Aberg, Nature **411**, 907 (2001).

[82] S. Redner, European Physics Journal B **4**, 131 (1998).

[83] V. Pareto *Cours d'Economie Politique* (Droz, Geneva, 1896).

[84] M. Levy, S. Solomon, Physica A **242**, 90 (1997).

[85] M. E. J. Newman, Contemporary Physics **46**, 323 (2005).

[86] B. Arnold, N. Balakrishnan H. Nagaraja, *Records*, (Wiley-Interscience, New York, 1998).

[87] A. M. Petersen, W.-S. Jung, J.-S. Yang, H. E. Stanley, Proceedings of the National Academy of Sciences USA **108**, 18 (2011).

[88] M. McBeath, D. Shaffer, M. Kaiser, Science **268** 569, (1995).

[89] W. J. Reed, B. D. Hughes, Physical Review E **66**, 067103 (2002).

[90] R. G. Tobin, American Journal of Physics **76**, 15 (2008).

[91] M. R. Yilmaz, S. Chatterjee and M. Habibullah, Journal of Sports Economics **2**, 181 (2001).

[92] G. C. Ward and K. Burns, *Baseball: An Illustrated History* (Knopf, New York, 1994).

[93] J. C. Bradbury, *What Really Ruined Baseball*, New York Times, (April 2, 2007)

[94] E. Ben-Naim, F. Vazquez, S. Redner, Journal of Quantitative Analysis in Sports **2(4)**, 1 (2006).

[95] I. Waddington *et al.*, British Journal of Sports Medicine **39** e18, (2005).

[96] J. J. Koch, Pediatrics in Review, **23** 310, (2002).

[97] C. R. McHenry, Surgery **142**, 785 (2007).

[98] T. D. Noakes, New England Journal of Medicine **351**, 847 (2002).

[99] B. Maher, Nature **452**, 674 (2008).

[100] B. Sahakian, S. Morein-Zamir, Nature **450**, 1157 (2007).

[101] "For to all those who have, more will be given, and they will have an abundance; but from those who have nothing, even what they have will be taken away." Matthew 25:29, New Revised Standard Version.

[102] R. K. Merton, Science **159**, 56-63 (1968).

[103] R. K. Merton, ISIS **79**, 606-623 (1988).

[104] J. H. Walberg, S. Tsai, American Educational Research Journal **20**, 359 (1983).

[105] K. E. Stanovich, Reading Research Quarterly **21**, 360 (1986).

[106] M. Bonitz, E. Bruckner, A. Scharnhorst, Scientometrics **40**, 407 (1997).

[107] W. F. Helsen, J. L. Starkes, J. Van Winckel, American Journal of Human Biology **10**, 791 (1998).

[108] J. Musch, S. Grondin, Developmental Review **21**, 147 (2001) .

[109] M. Segalla, G. Jacobs-Belschak, C. Müller, European Management Journal **19**, 58 (2001).

[110] J. F. Lawless, *Statistical models and methods for lifetime data, 2 ed.* (John Wiley & Sons USA, 2003).

[111] J. E. Hirsch, Proceedings of the National Academy of Sciences USA **102**, 16569 (2005).

[112] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Physical Review E **80**, 056103 (2009).

[113] J. A. Davies, European Physics Journal B **27**, 593 (2002).

[114] J. C. Huber, Scientometrics **45**, 33 (1998).

[115] A. M. Petersen, F. Wang, H. E. Stanley, Physical Review E **81**, 036114 (2010).

[116] A. Sidiropoulos & Y. Manolopoulos, Journal for Systems & Software **79**, 1679 (2006).

[117] A. Sidiropoulos, D. Katsaros & Y. Manolopoulos, Scientometrics **72**, 253 (2007).

[118] P.D. Batista, M.G. Campiteli, A. S. Martinez, Scientometrics **68**, 179 (2006).

[119] F. Radicchi, S. Fortunato & C. Castellano, Proceedings of the National Academy of Sciences USA **105**, 17268 (2007).

[120] A. M. Petersen, W. S. Jung & H. E. Stanley, Europhysics Letters **83**, 50010 (2008).

[121] Sean Lahman's Baseball Archive:
`http://baseball1.com/index.php`
Korean Professional Baseball League:
`http://www.inning.co.kr`
Data Base Sports Basketball Archive:
`http://www.databasebasketball.com/`

Barclays Premier League:
http://www.premierleague.com/

[122] ISI Web of Knowledge:
www.isiknowledge.com/

[123] S. Redner. Physics Today **58**, 49 (2005).

[124] A. M. Petersen, O. Penner, H. E. Stanley, European Physics Journal B **79**, 67 (2011).

[125] D. Lazer, *et al.*, Science **323**, 721 (2009).

[126] C. Castellano, S. Fortunato & V. Loreto, Reviews of Modern Physics **81**, 591 (2009).

[127] B. F. de Blasio, A. Svensson, & F. Liljeros, Proceedings of the National Academy of Sciences USA **104**, 10762 (2007).

[128] J. D. Farmer, M. Shubik & E. Smith, Physics Today **58(9)**, 37 (2005).

[129] A. L. Barabási, Nature **435**, 207 (2005).

[130] D. J. Watts, S. H. Strogatz, Nature **393**, 440 (1998).

[131] M. C. González, C. A. Hidalgo & A. L. Barabási, Nature **453**, 779 (2008).

[132] R. D. Malmgren, *et al.,* Proceedings of the National Academy of Sciences USA **105**, 18153 (2008).

[133] J. Leskovec, *et al.*, Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 462 (2008).

[134] S. Redner, *A guide to first-passage processes.* (Cambridge University Press, UK, 2001).

[135] R. C. Larson, A. R. Odoni, *Urban operations research.* 2nd Ed. (Dynamic Ideas, USA, 2007).

[136] A. Glen, L. Leemis, J. Drew, Computational Statistics & Data Analysis **44**, 451 (2004).

[137] T. M. Liggett, *Stochastic interacting systems: contact, voter, and exclusion processes*, (Springer-Verlag, New York, 1999).

[138] R. J. Glauber, Journal of Mathematical Physics **4**, 294 (1963).

[139] P. L. Krapivsky, Physical Review A **45**, 1067 (1992).

[140] V. Sood, S. Redner, Physical Review Letters **94**, 178701 (2005).

[141] M. Mobilia, Physical Review Letters **91**, 028701 (2003).

[142] M. Mobilia and I. T. Georgiev, Physical Review E **71**, 046102 (2005).

[143] See *e.g.*, `http://www.uselectionatlas.org/RESULTS/index.html`.

[144] R. Axelrod, Journal of Conflict Resolution **41**, 203 (1977); R. Axtell, R. Axelrod, J. Epstein, and M. D. Cohen, Computational and Mathematical Organization Theory **1**, 123 (1996).

[145] G. Weisbuch, G. Deffuant, F. Amblard, and J. P. Nadal J. P., Complexity **7**, 55 (2002); E. Ben-Naim, P. L. Krapivsky, and S. Redner, Physica D **183**, 190 (2003).

[146] F. Vazquez and S. Redner, Journal of Physics A **37**, 8479 (2004).

[147] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, 1985, 2nd edition); N. Van Kampen, *Stochastic processes in physics and chemistry* (North Holland Publishing Company, 1992).

[148] E. Ben-Naim and P. L. Krapivsky, Physical Review E **69**, 046113 (2004).

[149] T. Reichenbach, M. Mobilia and E. Frey, Physical Review E **74**, 051907 (2006).

[150] D. Considine, S. Redner and H. Takayasu Physical Review Letters **63**, 2857 (1989).

[151] A. J. McKane and T. J. Newman, Physical Review Letters **94**, 218102 (2005).