NORTHEASTERN UNIVERSITY
GRADUATE SCHOOL OF COMPUTER SCIENCE
Ph.D. THESIS COMPLETION APPROVAL FORM

THESIS TITLE:
Quantifying the Dynamics of Ranked Systems

AUTHOR: Nicholas Blumm

Ph.D. Thesis Approved to complete all degree requirements for the Ph.D. Degree in Computer Science.

Thesis Advisor Date

Thesis Reader

Thesis Reader

(sent electronically)

Thesis Reader

GRADUATE SCHOOL APPROVAL:

Director, Graduate School

Date

COPY RECEIVED IN GRADUATE SCHOOL OFFICE:

Recipient's Signature

Date

Distribution:
One Copy to Thesis Advisor
One Copy to Each Member of Thesis Committee
One Copy to Director of Graduate School
One Copy to Graduate School Office (with signature approval sheet, including all signatures)
QUANTIFYING THE DYNAMICS OF RANKED SYSTEMS

A dissertation presented by

Nicholas Blumm

to the Faculty of the Graduate School
of the College of Computer and Information Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Northeastern University
Boston, Massachusetts
April, 2012
Abstract

This dissertation uses volatility and spacing to allow one to quantify the dynamics of a wide class of ranked systems. The systems we consider are any set of items, each with an associated score that may change over time. We define volatility as the standard deviation of the score of an item. We define spacing as the distance in score from one item to its neighbor. From these two concepts we construct a model using stochastic differential equations. We measure the model parameters in a variety of ranked systems and use the model to reproduce the salient features observed in the data. We continue by constructing a spacing-volatility diagram that summarizes three unique stability phases and overlay each dataset on this diagram. We end by discussing limitations and extensions to such a model.
# Contents

List of Figures vi

List of Tables vii

Acknowledgements viii

1 Introduction 1

1.1 Motivation ........................................ 1

1.2 Overview ........................................ 2

2 Data 6

2.1 Dataset details .................................... 7

2.2 Spacing $x(r)$ .................................... 8

2.3 Volatility $\sigma(x)$ ................................. 11

3 Model 16

3.1 Stochastic differential equations, briefly ........... 16

3.2 A general model ................................... 18
3.2.1 The steady state solution ........................................... 20
3.3 Score stability (volatility) ............................................. 20
3.4 Rank stability (spacing) ............................................... 25
3.5 Spacing-volatility phase diagram .................................... 26

4 Modelling the data .......................................................... 30
  4.1 Measuring the noise term .............................................. 30
    4.1.1 Universality of scaling exponent $\beta$ ......................... 32
    4.1.2 The noise magnitude $B$ .......................................... 32
    4.1.3 Temporal dependence of the noise .............................. 33
  4.2 Measuring the drift term .............................................. 40
    4.2.1 A simplification for $\alpha$ .................................... 40
    4.2.2 Measuring $\tilde{A}$ .............................................. 42
  4.3 Spacing-volatility phase diagram with data ..................... 44
    4.3.1 Stability phases .................................................. 44
  4.4 Model simulations ..................................................... 47

5 Limits of model assumptions ........................................... 54
  5.1 System size is growing .............................................. 54
  5.2 $p(x)$ not in steady state ........................................... 55
  5.3 $A_i$ not constant .................................................... 58

6 Going further ............................................................... 60
  6.1 Why is a system stable? ............................................. 60

iii
List of Figures

1.1 Example time series ........................................... 3  
1.2 Score stability ................................................. 4  
1.3 Rank stability .................................................. 5  
1.4 Rank stability criteria ........................................ 5  

2.1 $P(x|t)$ ......................................................... 10  
2.2 $x(t)$ .......................................................... 12  
2.3 $r(t)$ .......................................................... 13  
2.4 $\Delta x(x)$ .................................................... 15  

3.1 Deterministic trajectory ....................................... 17  
3.2 Continuum theory of ranking stability. ...................... 24  
3.3 Spacing-volatility phase diagram ............................ 28  
3.4 Rank stability simulations. .................................... 29  

4.1 $\sigma_{\Delta x}(x)$ ............................................. 31  
4.2 $\sigma_{\Delta x}(x|\Delta t)$ ..................................... 34  
4.3 $\beta(\Delta t)$ ................................................... 35
List of Tables

2.1 Summary of datasets ................................................. 9

4.1 Important dataset parameters..................................... 32
Acknowledgements

I’d like to thank my advisor László Barabási for accepting me into his group and supporting me during my studies. He has built the CCNR to be an incredible place to meet a wide variety of people in Network Science. Additional thanks goes to my committee. Zoltán Toroczkai has a breadth of knowledge in dynamical systems, and always welcomed students dropping by. Alan Mislove consistently astonishes me with his enthusiasm and wealth of ideas – a five minute conversation with Alan will provide inspiration for months. Jay Aslam was by far the best teacher I’ve had in graduate school. He has the rare ability to clearly explain any topic using straightforward examples, as well as pose seemingly simple questions that induce deep thought and, ultimately deep insight.

Much thanks goes to the many people who have passed through the CCNR during my time here. They have provided countless conversations, insights and inspirations, especially Yong-Yeol Ahn, Sebastian Ahnert, Andi Asztalos, Jim Bagrow, József Baranyi, Julián Candia, Manuel Cebrian, Marta Gonzales, Natali Gulbahce, Tal Koren, Deok-Sun Lee, Sune Lehman, Yang Liu, Jörg Menche, Juyong Park, Diego Puppin, Filippo Simini, Sameet Srivastana, Tony Strathman, Gabor Szabo, Dashun Wang, and Pu Wang. Also thanks to a number of fellow graduate students both at Notre Dame – Jamie Antonelli, Brian Bucher, Sean Lynch, Matthew Meixner, and Kritsanu Tivakornsasithorn – and at Northeastern – Ahmed Abdelmeged, Keshi
Dai, Shahzad Rajput, and Aaron Turon. Additional thanks goes to my collaborators including Ginestra Bianconi, Jean-Philippe Bouchaud, Nitesh Chawla, Leon Chen, Nicholas Christakis, Darci Davis, Thomas Deisboeck, Zalán Forró, César Hidalgo, Zehui Qu, Max Schich, and Chaoming Song. Thanks also to the present and past CCNR staff who keep the cogs of this machine turning.

Special thanks goes to Gourab Ghoshal with whom I spent a solid two years thinking, discussing, writing and arguing about this project. I have learned much from him. Also to Alec Pawling, for being the first person to recommend CLRS as well as being a good friend. Kathie Newman was instrumental in my transition from non-academia into a graduate career. Naomi Hyun knows what the final days of graduate student life look like. Additionally I thank Pete Avis for the encouragement (and recommendation) to return to graduate school, as well as the many important conversations along the way. I know there will be many more.

Finally the deepest of gratitude goes to my parents, who have wholly supported every decision I have made.
Chapter 1

Introduction

1.1 Motivation

No matter where you look, it is possible to find a ranked list. Perhaps you recently performed an internet search, or scanned the headlines – ranked lists helped you find information you care about. From sports to film to universities, rankings are ubiquitous. Rankings grab your attention: “Did you hear who won the game?” Rankings spur you to action: “What’s next on my todo list?”

Ranked lists also display a wide range of stability. Check the top news stories in the morning, by the evening many will have changed; such ranked lists are volatile on the order of hours. Other lists, bestsellers in books for example, are more stable changing over the course of weeks or months. Yet other lists, like the world’s wealthiest individuals, are stable over the course
of decades.

Many have studied specific rankings, in areas as diverse as news items [34], web pages [20, 46] city populations [1, 47] and sports competitions [2]. Others have explored the volatility in books [19, 52], videos [15] and stock market prices [39, 35, 51].

While individual ranked lists are ubiquitous, relatively little attention has been devoted to a general ranking process. Yet, understanding such mechanisms is crucial from scientometrics [5, 30, 58, 32] and healthcare [28] to economic phenomena [40], raising a number of interesting questions: What distinguishes inherently stable systems, from volatile ones? To what amount is the underlying quality of items incorporated into the rank? We propose to answer these questions using the simple concepts of volatility and spacing.

The rest of this dissertation proceeds as follows: in Section 1.2 we outline our approach; in Chapter 2 we introduce the data used in this study; in Chapter 3 we propose a model; in Chapter 4 measure the model parameters in each dataset and reproduce the qualitative behavior of the data; in Chapter 5 we discuss the limitations of this model; and in Chapter 6 we discuss extensions to this work.

1.2 Overview

The goal of this dissertation is to build a framework to quantify the stability of ranked lists. This framework will allow us to identify stable items in a
system as well as compare stability across systems. To do this we require systems that contain a set of items each with an associated score that can vary over time, Figure 1.1. We define two types of stability: score-stability and rank-stability.

For an item to be score stable, we require it to have a peaked distribution \( p(x) \) over score \( x \), independent of time. If instead an item has some heavy tailed distribution, with \( x \) spanning many orders of magnitude, we consider this score-unstable, Figure 1.2. For an item to be rank-stable it must not only be score-stable but also be unlikely to fall below the score of its neighbor. Thus if we have two score-stable distributions, we want them to be far apart, Figure 1.3. We may then define a rank-stability criteria, \( \Delta > \sigma \), where \( \Delta \) is the distance in score between adjacent ranked items, and \( \sigma \) is the item’s score volatility, Figure 1.4.

These two concepts: \( \Delta \), the \textit{spacing} between items, and \( \sigma \), the item
Figure 1.2: Example of score stable and unstable distributions. Horizontal axis is score, vertical axis is probability.

volatility, summarize our underlying thesis. For infinite $\sigma$ an item is score-unstable; as $\sigma$ decreases we obtain a stable score; and when $\sigma < \Delta$ we achieve rank-stability as well. That dear reader, is our thesis, in the remaining pages we fill in the details.
Figure 1.3: Example of rank stable and unstable distributions. Horizontal axis is score, vertical axis is probability.

Figure 1.4: Our criteria for rank stability is $\Delta > \sigma$, where $\Delta$ is the spacing in score between items, and $\sigma$ is the volatility of an item.
Chapter 2

Data

We restrict our study to data sets meeting the following criteria: a collection of items, with associated scores that may change over time. Let $X_i(t)$ be the score of item $i$ at time $t$. We define a normalized score, or ‘market share’, as

$$x_i = \frac{X_i}{\sum_j X_j}.$$

(2.1)

We use market share so we may control for system growth. For example if all items in a system are doubling in each timestep, $X_i(t + \Delta t) \gg X_i(t)$ but $x_i(t + \Delta t) \approx x_i(t)$. Thus $x_i$ allows us to compare scores even at different timesteps.

We have gathered six diverse data sets meeting this criteria. The temporal resolution of the data varies from hours to years and the time-span from days to centuries. We describe each dataset in detail in the next section, see Table
2.1 for a summary.

2.1 Dataset details

The ngram data came from Google’s “English One Million” [41]. We studied the 10,000 most frequently used words spanning over two centuries, from 1800 to 2008. Here we are ranking words, and $x_i$ represents the fraction of all words published in a given year that were word $i$.

The marketcap data was sourced from the Center for Research in Security Prices (CRSP) [16]. We compiled the daily market capitalization for each of the 11,580 companies listed from Jan 1, 2000 – Dec 31, 2009. Here we are ranking companies; the market share $x_i$ is the fraction of all money in the market belonging to company $i$ at the end of the trading day.

The medicare data came from claims based on MedPAR records of hospitalizations between 1990-1993 as described in [28]. Each record consists of the date of visit, a primary diagnosis and up to 9 secondary diagnoses, all specified by ICD9 codes. For a detailed list of currently used ICD9 codes see [50]. We compiled the number of times each of the 10,045 diseases was diagnosed each month. ICD9 diseases that were not diagnosed during the study period were dropped from our analysis. Here we are ranking diseases, and $x_i$ is the fraction of all diagnoses in a given month that were disease $i$.

The citation data [45, 14, 24] consists of papers published in the Physical Review journals between 1970 and 2009, available upon request from the
American Physical Society [49]. We tracked the annual citation counts of 449,673 papers during this period. Here we rank papers and $x_i$ is the fraction of all citations in a given year pointing to paper $i$.

The *twitter* data [13, 42] consists of all tweets between Aug 14, 2008 and Aug 13, 2009; from which we counted the number of times each hash-tag was used daily. A hash-tag may be thought of as a topic marker, as in

I like #science.

Hashtags appear in search results and are often used in trending topic sites [25]. There were a total of 39,284 hash tags in all tweets during this time period. Thus here we are ranking hash-tags, and $x_i$ is the fraction of all tweets containing hashtag $i$ on a given day.

The *wikipedia* data [12] is available at [57]. We used the hourly English language page views between Feb 5–17, 2011. After limiting to pages that received at least 1,000 page views in total there were a total of 186,603 pages. Thus here we are ranking web pages, with $x_i$ the fraction of all pageviews in a given hour going to page $i$.

### 2.2 Spacing $x(r)$

For our analysis we use the complementary cumulative distribution

$$P(x) = 1 - \int_0^x p(x)dx$$
where \( p(x) \) is the probability density function. By definition, \( p(x) \) tells us where in score-space items are densely packed or relatively rare. Here we ask two questions: (1) what is the shape of \( P(x) \)? and (2) is this shape stable over time? In Figure 2.1 we plot \( P(x) \) for selected timesteps \( t \) spanning the study period. We find that \( x \) spans orders of magnitude. In some systems \( P(x) \) follows a clear power law (ngram, citation, twitter, wikipedia), while in others we observe cutoffs at high \( x \) (marketcap, medicare); \( P(x) \) can be stationary (ngram, marketcap, medicare, wikipedia) or may shift in time (citation, twitter).

\( P(x) \) is also handy as a mapping from scores to ranks [52, 19]. (See [24] and [17] for other approaches for mapping scores to ranks.) Let \( N \) be the number of items in the system. Since \( P(x) \) is a function whose input is a score and whose output is the fraction of items greater than that score, \( NP(x) \) is the number of items with a score greater than \( x \). But this is precisely the

<table>
<thead>
<tr>
<th>item i</th>
<th>score ( X_i )</th>
<th>timespan</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>count in all books (yearly)</td>
<td>200 years</td>
<td>Google</td>
</tr>
<tr>
<td>company</td>
<td>market capitalization (daily)</td>
<td>10 years</td>
<td>CRSP</td>
</tr>
<tr>
<td>disease</td>
<td>diagnoses (monthly)</td>
<td>4 years</td>
<td>Medicare</td>
</tr>
<tr>
<td>paper</td>
<td>citations (yearly)</td>
<td>39 years</td>
<td>Phys. Rev.</td>
</tr>
<tr>
<td>hashtag</td>
<td>count in all tweets (daily)</td>
<td>365 days</td>
<td>Twitter</td>
</tr>
<tr>
<td>article</td>
<td>page-views (hourly)</td>
<td>300 hours</td>
<td>Wikipedia</td>
</tr>
</tbody>
</table>

Table 2.1: The six datasets. The table lists the item ranked, the definition of the score \( X_i \) and its timeframe, the timespan of the study, and the data source.
Figure 2.1: The cumulative distribution $P(x)$ of the market-share $x$ for each studied system at different points in time $t$ indicated by the legend. All six datasets demonstrate a heterogeneous distribution with a time-independent scaling in the tail, which we fit to the form $P(x) \sim x^{-\mu}$. We find that $\mu$ varies between 1 and 2.1.
rank! Thus

\[ r = NP(x). \quad (2.2) \]

Assuming \( P \) is invertible, we also have a mapping from ranks to scores:

\[ x(r) = P^{-1}\left(\frac{r}{N}\right). \quad (2.3) \]

This approach treats \( r \) as a continuous variable; one can round to convert to discrete values. This mapping specifies the spacing in \( x \) between consecutively ranked items:

\[ \Delta(r) = x(r) - x(r + 1). \quad (2.4) \]

Thus the cumulative distribution \( P(x) \) specifies the spacing \( \Delta \) between ranked items.

### 2.3 Volatility \( \sigma(x) \)

To develop some intuition for the volatility we plot \( x(t) \) and \( r(t) \), selected scores and ranks over time, in Figures 2.2 and 2.3. Some rankings, like trends as measured by twitter and wikipedia are intrinsically volatile, changing daily; others, like English word usage, show remarkable stability. These different patterns of ranking stability are illustrated in Figure 2.3, first row,
Figure 2.2: The evolution of score in each system. Selected items were the top five at the middle timestep. **First Row:** Ngram, marketcap and medicare display greater score stability. **Second Row:** Citation, twitter and wikipedia show significant volatility. We plot the corresponding ranks in Figure 2.3.

which indicates that *the* has been the most frequently used English word in printed texts for at least two hundred years and *hypertension* has been the most prevalent disease over 48 months. In contrast, the ranking of research papers based on their citations and wikipedia articles based on their visitations display significant volatility over time (Fig. 2.3, second row).

We now turn to the volatility of items as a function of their score. As we move up in score, do items become more volatile? We begin in by constructing a scatter plot of $\Delta x$ as a function of $x$, where $\Delta x$ is the change in score for a given item over a single timestep, Figure 2.4. We find that for ngram,
Figure 2.3: The evolution of rank in each system. Selected items were the top five at the middle timestep. **First Row:** Ngram, marketcap and medicare display greater score stability. **Second Row:** Citation, twitter and wikipedia show significant volatility. We plot the corresponding scores in Figure 2.2.
medicare and marketcap, $\Delta x$ fluctuates symmetrically around $\Delta x = 0$, so that the score of each item has comparable probability of moving up or down. In contrast citation, twitter and wikipedia are asymmetric, with a tendency to increase at low $x$ and to drop at high $x$. This tendency for low-scoring items to move up while high scoring items fall is the hallmark of an unstable system. While this asymmetry reflects the lack of stability for citations, twitter and wikipedia, it does not explain the origin of the differences between these and the more symmetric systems. We now turn to a model of such systems.
Figure 2.4: Scatter plot of $\Delta x$ as a function of $x$, indicating that for ngram, marketcap and medicare, $\Delta x$ fluctuates symmetrically about $\Delta x = 0$, so that high and low ranked items have comparable probabilities of moving up or down in score. Citation, twitter and wikipedia however, are asymmetric about $\Delta x = 0$, indicating that low scoring items tend to rise in score, while high scoring items tend to fall.
Chapter 3

Model

3.1 Stochastic differential equations, briefly

Imagine we had an entirely deterministic system where the score of an item were known for all time, Figure 3.1(a). We could specify the score as a differential equation of the form $\dot{x} = \mu(x)$, with $\dot{x} = \frac{dx}{dt}$ and initial conditions $x(t = 0) = x_0$. Now imagine we add some noise to this system. Let $\xi(t)$ be a zero mean, unit variance, time independent, Gaussian random variable, i.e. $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \delta(t' - t)$. To be clear, the $t$ indicates that at any timestep, $\xi(t)$ returns a unique, independent draw from distribution $\xi$. This noisy system can be described by an equation of the form

$$\dot{x} = \mu(x) + \sigma(x)\xi,$$
where we omit the $t$ for clarity, see Figure 3.1(b).

We use this as the base of our model since it captures our two important concepts: spacing and volatility. The volatility is controlled by $\sigma$, since the greater $\sigma$, the noisier the trajectory. For spacing, imagine we have two items $x_1$ and $x_2$. And two equations

$$\dot{x}_i = \mu_i(x_i) + \sigma_i(x_i)\xi_i, \quad (3.1)$$

with $i \in \{1, 2\}$. Assuming the system has reached a steady state (independent of initial conditions), the expected distance between $x_1$ and $x_2$ at time $t$ is then controlled by $\mu_1$ and $\mu_2$. Such a stochastic differential equation [43, 23, 29] captures the important concepts of our system. The terms are often referred to as drift $\mu$ and noise $\sigma$ and such an equation is a form of the
Langevin Equation [59].

3.2 A general model

We wish to apply Equation (3.1) to our system. Each item $i$ follows its own trajectory, but as is stands, the value of $x_i$ is unbounded; whereas we have defined $x_i$ to be a normalized market share (2.1). To control for this we use a Lagrange multiplier $\phi$:

$$\dot{x}_i = \mu_i(x_i) + \sigma_i(x_i)\xi_i - \phi x_i,$$

(3.2)

where $\phi$ ensures $x_i$ is properly normalized. The Lagrange multiplier has the form

$$\phi = \phi_0 + \eta,$$

(3.3)

with system drift $\phi_0 = \sum \mu_i(x_i)$, and system noise $\eta = \sum \sigma_i(x_i)\xi_i$ — see 8.1 for a derivation.

The drift term $\mu_i(x_i)$ represents the deterministic mechanisms that drive the score of item $i$, capturing a wide range of system-dependent attributes, from utility (ngram words), to information content (wikipedia) or impact (research papers), together with the relative fitness of each item compared to their peers. The second term $\sigma_i(x_i)\xi_i$ captures the inherent randomness in the system.
Equation (3.2) assumes that the scores of different items do not directly influence each other, apart from the global normalization, hence social [48] and herd effects [4] are absent from our formalism.

We now propose functional forms for the drift and noise terms. First, we postulate that the drift term $\mu_i(x_i)$ can be written as:

$$\mu_i(x_i) = A_ix_i^\alpha,$$

(3.4)

where $0 < \alpha < 1$ is identical for all $i$, whereas the coefficient $A_i$ can be interpreted as the “fitness” of item $i$ [3]. Next we propose that $\sigma_i(x_i)$ is of the form

$$\sigma_i(x_i) = Bx_i^\beta.$$  

(3.5)

We test these assumptions of $\mu$ and $\sigma$ with measurements in Chapter 4.

Our model is now

$$\dot{x}_i = A_ix_i^\alpha + Bx_i^\beta \xi_i - \phi x_i.$$  

(3.6)

For now we assume:

1. we have a closed system, i.e. the number of items $N$ is fixed;

2. $A_i$ is the “fitness” of each item is fixed;

3. $A_i$ is drawn from a heavy tailed distribution: $p(A) \sim A^{-\mu}$.

We revisit each of these assumptions in Chapter 5.
3.2.1 The steady state solution

One advantage of using model (3.6), is that we can calculate steady state distribution \( p(x_i) \) for any given item. Here we apply the Fokker-Planck Equation \([59, 54]\) (also called the Kolmogorov Forward equation \([43, 11, 31]\)) to calculate \( p(x_i) \):

\[
\frac{\partial p(x_i)}{\partial t} = -\frac{\partial}{\partial x_i} [(A_i x_i^\alpha - \phi x_i)p(x_i)] + \frac{1}{2} \frac{\partial^2}{\partial x_i^2} \left( B^2 x_i^{2\beta} p(x_i) \right). \tag{3.7}
\]

Assuming the system evolves to a steady-state with a time-independent \( p(x_i) \) and a constant value for \( \phi = \phi_0 \), the solution of (3.7) is

\[
p(x_i) = C(A_i) x_i^{-2\beta} \exp \left[ \frac{2A_i}{B^2} \left( \frac{x_i^{1+\alpha-2\beta}}{1 + \alpha - 2\beta} - \frac{\phi_0 x_i^{2(1-\beta)}}{2A_i 1 - \beta} \right) \right], \tag{3.8}
\]

with normalization constant \( C(A_i) \) – see Section 8.2 for a derivation.

3.3 Score stability (volatility)

In Section 1.2 we qualitatively defined an item to have a “stable” score if its distribution is peaked. We now examine (3.8) for such conditions. At steady state an item has \( \dot{x}_i = 0 \). In the absence of noise the steady state solution to
Equation (3.6) is

\[ 0 = \dot{x}_i = A_i x_i^\alpha - \phi_0 x_i, \]

or

\[ x_i^* \equiv \left( \frac{A_i}{\phi_0} \right)^{\frac{1}{1-\alpha}}, \tag{3.9} \]

indicating that each item will converge to a score determined by its fitness \( A_i \) and a combination of all other fitnesses via \( \phi_0 = \left( \sum_i A_i^{1-\alpha} \right)^{1-\alpha} \). In the presence of noise, the most probable value of \( x_i \) is the solution of

\[ p'(x_i) = (A x_i^\alpha - \phi_0 x_i) - B^2 \beta x_i^{2\beta-1} = 0, \tag{3.10} \]

where we set the derivative of Equation (3.8) to 0. Therefore the noise term \( B \) shifts the steady-state value of \( x_i \) from its deterministic solution \( x_i^* \) to a new \( x_i^* + \delta_i \).

Let us hold all parameters fixed while varying \( B \). Equation (3.8), says that increasing the noise magnitude \( B \) widens \( p(x) \) from a stable (peaked) distribution to an unstable (heavy tailed) distribution. In Fig. 3.2a we plot \( x_i^* + \delta_i \), the solution of Equation (3.10), as a function of \( B \), finding that \( x_i^* + \delta_i \to 0 \) as \( B \to B_c \). Thus for \( B < B_c \), Equation (3.10) has two solutions, one at \( x_i = x_i^* + \delta_i \) (stable) and the other at \( x_i = 0 \) (unstable).
(Fig. 3.2b). Therefore, \( p(x_i) \) is unimodal with a sharp peak (Fig. 3.2e), indicating that for low noise the market share of item \( i \) will be localized around a value determined by the interplay between the item’s fitness and the noise magnitude.

At \( B = B_c \), however, the non-zero solution disappears (\( x_i = x_i^* + \delta_i \to 0 \), Fig. 3.2b-d). Equation (3.8) says that for \( B > B_c \) the distribution follows a power-law \( p(x_i) \sim x_i^{-2\beta} \) with an exponential cutoff at high \( x_i \). Indeed, if \( 2A_i/B^2 \ll 1 \) then the term in the exponent of Equation (3.8) is small and \( p(x_i) \) is dominated by its pre-factor \( x_i^{-2\beta} \) (Fig. 3.2e-g). This implies that \( x_i \) is no longer confined to the vicinity of \( x_i^* \) but becomes broadly distributed, varying over orders of magnitude.

The consequence of such a noise-induced phase transition is illustrated in Figs. 3.2h–j, where we show the evolution of the ranks for the top five fitness items, as produced by numerical simulations of Equation (3.6). In the stable phase (\( B < B_c \)) the top items maintain their nominal rank, determined by their respective fitnesses; similar to the behavior observed in Fig. 2.3, top row. At the critical point \( B = B_c \) the stable ranking is perturbed by unstable bursts, an intermittent behavior common in dynamical systems at the critical point [8, 21]. Finally for \( B > B_c \) most items become broadly distributed, lacking rank stability, a behavior similar to the one observed in Fig. 2.3, second row.

To further test our model in Fig. 3.2k–m we plot the average change in score \( \langle \Delta x \rangle \) and \( \sigma_{\Delta x} \) as a function of the score \( x \). For \( B < B_c \) the score
variation $\Delta x$ fluctuates symmetrically around $\Delta x = 0$, similar to the behavior observed in stable systems (Fig. 2.4, top row). For $B > B_c$, however there is a systematic downward trend for high $x$, as observed for the unstable systems (Fig. 2.4, bottom row). We see that the choice of the functional form of the drift term in Equation (3.4) is supported by the qualitative agreement between Fig. 2.4 and Fig. 3.2k-m.

To summarize, we have defined a score-critical noise magnitude $B_c$ as the largest noise such that $p(x)$ is still a peaked distribution. Thus

$$B_c \equiv G(A, \alpha, \beta),$$

(3.11)

where $G$ returns the maximum $B$ such that Equation (3.10) has a non-zero solution.
Exponents are fixed at the score ranking unstable. Jumps in rank. For $B < B_c$ only the zero solution remains. The non-zero solution $x^* + \delta$ of Equation (3.10) plotted as a function of the noise coefficient $B$ for a choice of $A = 8 \times 10^{-2}$. $p'(x)$ defined in (3.10) for increasing values of $B$. The point at which $p'(x)$ crosses zero corresponds to a non-zero maxima for $x$ for a given fitness $A$. At $B = B_c$ the non-zero solution disappears and above $B_c$ only the zero solution remains. The score distribution $P(x_i|A_i)$ for the top fitness item plotted in the three regimes. Below $B_c$ we observe a peak around the fitness value $x_i^* + \delta$, at $B = B_c$ the distribution becomes bimodal, while for $B > B_c$ the peak vanishes and the distribution follows a power-law with a cutoff determined by the item’s fitness as shown in Eq. (3.8). The rank evolution for the top fitness items, as predicted by numerical simulations of Eq. (3.6). In the stable phase ($B < B_c$) the top items maintain their rank, the observed dynamical behavior being similar to that seen for the stable systems in Fig. 2.3, first row. At the critical point ($B = B_c$) the stable ranking is perturbed by intermittent jumps in rank. For $B > B_c$ all items lose their rank stability, making the ranking unstable. The average change in score $\langle \Delta x \rangle$ as a function of the score $x$ as predicted by numerical simulations of Eq. (3.6). The scaling exponents are fixed at $\alpha = \beta = 0.7$. 

Figure 3.2: Continuum theory of ranking stability. (a) The non-zero solution $x^* + \delta$ of Equation (3.10) plotted as a function of the noise coefficient $B$ for a choice of $A = 8 \times 10^{-2}$. (b, c, d) $p'(x)$ defined in (3.10) for increasing values of $B$. The point at which $p'(x)$ crosses zero corresponds to a non-zero maxima for $x$ for a given fitness $A$. At $B = B_c$ the non-zero solution disappears and above $B_c$ only the zero solution remains. (e, f, g) The score distribution $P(x_i|A_i)$ for the top fitness item plotted in the three regimes. Below $B_c$ we observe a peak around the fitness value $x_i^* + \delta$, at $B = B_c$ the distribution becomes bimodal, while for $B > B_c$ the peak vanishes and the distribution follows a power-law with a cutoff determined by the item’s fitness as shown in Eq. (3.8). (h, i, j) The rank evolution for the top fitness items, as predicted by numerical simulations of Eq. (3.6). In the stable phase ($B < B_c$) the top items maintain their rank, the observed dynamical behavior being similar to that seen for the stable systems in Fig. 2.3, first row. At the critical point ($B = B_c$) the stable ranking is perturbed by intermittent jumps in rank. For $B > B_c$ all items lose their rank stability, making the ranking unstable. (k, l, m) The average change in score $\langle \Delta x \rangle$ as a function of the score $x$ as predicted by numerical simulations of Eq. (3.6). The scaling exponents are fixed at $\alpha = \beta = 0.7$. 

Unstable phase $B > B_c$.
3.4 Rank stability (spacing)

We now turn to our concept of rank stability from Section 1.1, Figure 1.4. We want an item to be unlikely to swap with its neighbor. Let $\sigma(r)$ be the standard deviation in score for an item with rank $r$. Let $\Delta(r)$ be the distance in score from the item at rank $r$ to the next item at rank $r+1$. Our rank stability condition is

$$\Delta(r) > \sigma(r).$$ \hfill (3.12)

In Section 2.2 we defined $\Delta(r)$ as the distance in score from one rank to the next:

$$\Delta(r) = x(r) - x(r+1),$$ \hfill (3.13)

with $x(r)$ defined in Equation (2.3).

We now turn to $\sigma(r)$. Since $\sigma(r) = \sigma(x(r))$ we must determine $\sigma(x)$. Recall that to be rank stable, an item must first be score stable. Thus $p(x)$ of Equation (3.8) must be peaked. If $p(x)$ is peaked, we can approximate the variance by expanding about the peak, see Section 8.3. We obtain

$$\sigma(x) = \frac{1}{\sqrt{2\phi_0(1 - \alpha)}} B x^\beta,$$ \hfill (3.14)
\[ \sigma(r) = \frac{1}{\sqrt{2\phi_0(1-\alpha)}} B [x(r)]^\beta, \] (3.15)

with \( x(r) \) as defined in (2.3). We can now identify rank stable items using equations (3.13) and (3.15), subject to the condition \( \Delta(r) > \sigma(r) \).

Furthermore, using equations (3.9) and (2.3), rank \( r \) can be written as function of the fitness \( A_i \), and our rank-stability criteria becomes

\[ \Delta(A_i) > \sigma(A_i). \] (3.16)

Thus we may define, \( B_r \), the rank-critical noise magnitude, as

\[ B_r \equiv H(\bar{A}, \alpha, \beta), \] (3.17)

where \( \bar{A} \) is the set of all fitnesses, and \( H \) returns the maximum \( B \) such that Equation (3.16) is satisfied for a given \( A_i \).

### 3.5 Spacing-volatility phase diagram

These results indicate that the stability of a system is best captured by a spacing-volatility phase diagram. The spacing is captured by the distribution of \( A \)'s in the system, the volatility is captured by the noise magnitude \( B \). Given \( \alpha \) and \( \beta \) we construct an \( A-B \) phase diagram in Figure 3.3. The top part \( (B > B_c) \) is the unstable region, where Equation (3.8) states that
the score of each item is broadly distributed; hence neither rank nor score stability is possible, similar to a gas phase where atoms move at random. In the region $B < B_c$ we have score stability, which means that each score fluctuates around a steady state market share $x^*$ determined by its intrinsic fitness. Yet score stability does not necessarily imply rank stability. Hence, there are two distinct phases below $B_c$: a score-stable (liquid) phase between $B_c$ and $B_r$, where each item has a stable score, but the fluctuations around $x^*$ are sufficient for items with comparable score to swap rank. This score stable but not rank stable phase is similar to the notion of a canon in arts and literature, representing “a body of works traditionally considered to be the most significant and therefore the most worthy of study” [9], yet whose relative ranking is ambiguous. For $B$ below $B_r$ we find a solid phase: the noise is sufficiently low that items display not only score but also rank-stability (solid phase). These predictions are supported by Fig. 3.4, where we plot $r(t)$ and $p(x)$ for the top items in each phase.
Figure 3.3: **Spacing-volatility phase diagram.** The A-B phase diagram with A determining item spacing, and B determining volatility. Phases calculated from $B_c$ and $B_r$ – Equations (3.11) and (3.17) – with $\alpha = \beta = 0.7$, $p(A) \sim A^{-2}$, and $\sum A_i = 1$. See also Figure 3.4.
Figure 3.4: Rank stability simulations. Top row: Ranks as a function of time for top-fitness items in each of the three phases. Bottom row: The market share distributions $p(x_i)$ for the same top ranked items. Circle indicates phase as shown in Figure 3.3.
Chapter 4

Modelling the data

So far we have proposed and analyzed a general framework for modelling a system of ranked items. Here we turn to measuring the model parameters in the data.

4.1 Measuring the noise term

Given our model

\[ \dot{x}_i = \mu_i(x_i) + \sigma_i(x_i)\xi_i - \phi x_i, \]

we wish to relate \( \sigma_i \) to something we can measure in the data. Using Equation (3.5) we calculate \( \sigma_{\dot{x}} \) finding

\[ \sigma_{\dot{x}} \approx \sigma_i(x_i) \]

(4.1)

\[ = B x_i^\beta \]

(4.2)
Figure 4.1: \( \sigma_{\Delta x} \) as a function of \( x \), indicating that for each system \( \sigma_{\Delta x} = Bx^\beta \), with \( \beta = 0.76 \pm 0.11 \). Additionally we find that \( B \) is roughly two orders of magnitude larger in unstable systems (bottom row) than in the stable systems (top row). Table 4.1 summarizes these parameters.

(see Section 8.4). Using discrete time we have

\[
\sigma_{\Delta x} \approx Bx_t^\beta, \tag{4.3}
\]

with \( \Delta x = x(t + 1) - x(t) \), the change in score over a single timestep. This \( \Delta x \) is precisely what we measured in Figure 2.4, thus plotting the standard deviation of \( \Delta x(x) \) should give us values for \( B \) and \( \beta \). We do so in Figure 4.1 and, quite surprisingly, find the functional form \( \sigma_{\Delta x} = Bx^\beta \) for all data sets. We summarize the values of \( B \) and \( \beta \) in Table 4.1.
Table 4.1: Summary of important dataset parameters. \( N \) is the number of items in the system. \( \mu \) is the scaling exponent from \( p(x^*) \sim x^{*-\mu} \) as measured in Figure 4.9. \( B \) and \( \beta \) are from \( \sigma_{\Delta x} = B x^\beta \) as measured in Figure 4.1.

<table>
<thead>
<tr>
<th>data set</th>
<th>( N )</th>
<th>( \mu )</th>
<th>( B )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngram</td>
<td>10,000</td>
<td>1.1</td>
<td>( 6 \times 10^{-3} )</td>
<td>0.69</td>
</tr>
<tr>
<td>marketcap</td>
<td>11,580</td>
<td>0.9</td>
<td>( 5 \times 10^{-3} )</td>
<td>0.76</td>
</tr>
<tr>
<td>medicare</td>
<td>10,045</td>
<td>0.8</td>
<td>( 8 \times 10^{-3} )</td>
<td>0.65</td>
</tr>
<tr>
<td>citation</td>
<td>449,673</td>
<td>2.0</td>
<td>( 5 \times 10^{-2} )</td>
<td>0.74</td>
</tr>
<tr>
<td>twitter</td>
<td>39,284</td>
<td>1.0</td>
<td>( 3 \times 10^{-1} )</td>
<td>0.67</td>
</tr>
<tr>
<td>wikipedia</td>
<td>186,603</td>
<td>1.7</td>
<td>( 2 \times 10^{-1} )</td>
<td>0.87</td>
</tr>
</tbody>
</table>

4.1.1 Universality of scaling exponent \( \beta \)

In Figure 4.1 we find the scaling exponent in the range \( 0.67 \leq \beta \leq 0.88 \) (see Table 4.1). The fact that \( \beta < 1 \) implies that relative changes are typically smaller for top-ranked items, which is known in the economic context, where the volatility of large companies is less than that of small companies [39]. While such sub-linear behavior [18, 22] may contribute to the stability of the high ranked items, we do not detect any obvious correlations between \( \beta \) and the observed ranking stability (Table 4.1 and Figure 4.1).

4.1.2 The noise magnitude \( B \)

While the value of \( \beta \) is comparable for all systems, we observe significant differences in the magnitude of the coefficient \( B \): for the three systems with stable ranking we find \( B \approx 10^{-3} \), while for the systems with unstable ranking we observe \( B \approx 10^{-1} \) (Table 4.1). Since the coefficient \( B \) is a direct measure of the noise magnitude, we find that the three unstable systems are affected
by a higher level of noise than the three stable systems. Next we test the
dependence of $B$ and $\beta$ on time.

4.1.3 Temporal dependence of the noise

Here we address two questions:

1. Do the noise parameters depend on timestep size?

2. Are the noise parameters constant over time?

Regarding the first question, one might expect that as the time-interval
(or the window size $\Delta t$) is increased, the noise magnitude $B$ will decrease,
since any short term volatility of the score $\Delta x$ will average out over longer
time-scales $\Delta t \gg 1$. To check this in Fig. 4.2 we plot the standard deviation
$\sigma_{\Delta x}$ in function of $x$ for increasing window sizes, $1 \leq \Delta t \leq T/3$, where $T$ is
the total number of time-steps in the data. For example, $\Delta t = 1$ corresponds
to Fig. 4.1; for $\Delta t = 2$ we collapse two time-steps into one and so on.

In Figure 4.3 we plot $\beta(\Delta t)$, finding the value of the scaling exponent $\beta$
is largely independent of the window size. In Figure 4.4 we plot $B(\Delta t)$. We
find that $B$ is largely constant within our estimated measurements. Hence
the observed differences between the two classes with ranking behavior are
independent of the choice of $\Delta t$.

Regarding the second question posed above – the stability of noise param-
eters over time – we test this by slicing the data into ten time subsets. For
each subset $s$, the first 10th of the timesteps, the 2nd tenth etc., we measure
Figure 4.2: $\sigma_{\Delta x}(x)$ as a function of window size $\Delta t$. For each dataset, we aggregate the data over time-steps $\Delta t$. The colors in the plot correspond to increasing $\Delta t$ with $\Delta t = 1$ being the darkest and $\Delta t = T/3$ being the lightest; $T$ is the total number of time-steps in each dataset. In all six datasets we find both $B$ and $\beta$ are generally independent of $\Delta t$. See Figures 4.3 and 4.4. For each timestep size $\Delta t$ we fit to the form $Bx^\beta \sqrt{\Delta t}$ – the factor of $\sqrt{\Delta t}$ due to the fact that $t$ is discrete [29].
Figure 4.3: Values of $\beta$ for fits to $\sigma_{\Delta t}(x|\Delta t)$ in Figure 4.2. Given error estimates, $\beta$ is roughly independent of window size $\Delta t$. Error bars indicate the 95% confidence intervals for the fit to $\beta$. 

---

35
Figure 4.4: Values of $B$ for fits to $\sigma_{\Delta x}(x|\Delta t)$ in Figure 4.2. Given error estimates, $B$ is roughly independent of window size $\Delta t$. Error bars indicate the 95% confidence intervals for the fit to $B$.

$\sigma_{\Delta x|s} = B x^\beta$. We find $B$ and $\beta$ to be relatively independent of subset $s$. See Figures 4.5, 4.6 and 4.7.
Figure 4.5: Here we test for any drift of the noise parameters $B$ and $\beta$. For each dataset we separate the timesteps into ten subsets. For each subset $s$, the first 10th of the timesteps, the 2nd tenth etc., we measure $\sigma_{\Delta x|s} = B x^\beta$. We find $B$ and $\beta$ to be relatively stable over time. See Figures 4.6 and 4.7.
Figure 4.6: We find $\beta$ relatively stable over time. For each dataset we separate the timesteps into ten subsets. We plot $\beta(s)$ for each subset $s$, where $s$ is the first 10th of the timesteps, the 2nd tenth etc. See Figure 4.5 for additional details.
Figure 4.7: We find $B$ relatively stable over time. For each dataset we separate the timesteps into ten subsets. We plot $B(s)$ for each subset $s$, where $s$ is the first 10th of the timesteps, the 2nd tenth etc. See Figure 4.5 for additional details. For ngram there may be an increase in stability (decrease in $B$) over time, suggesting word usage is becoming more standardized over time.
4.2 Measuring the drift term

We now turn to measuring $A$ and $\alpha$. In a high noise system, the drift term vanishes and it is impossible to measure $A$ and $\alpha$. Only in a low noise system will we be able to measure the fitness. Thus we restrict this discussion to systems where items are score stable.

4.2.1 A simplification for $\alpha$

If an item is score stable there will be some $x^*$ as it’s most probable score.

If the noise is sufficiently small, from (3.10) we have

$$x^* = \left( \frac{A_i}{\phi_0} \right)^{\frac{1}{1-\alpha}}.$$  

And since $\sum x^* = 1$, then $\phi_0 = \left( \sum A_i^{\frac{1}{1-\alpha}} \right)^{1-\alpha}$, thus

$$x_i^* = \frac{A_i^{\frac{1}{1-\alpha}}}{\sum A_j^{\frac{1}{1-\alpha}}}.$$  

At this point, even if we knew $\alpha$ we would only know $A_i$ up to a constant factor. Since $A_i$ would work just as well as $100 \cdot A_i$, we need some constraint on the magnitude of $A_i$. Let’s look again at Equation (3.14),

$$\sigma(x^*) = \frac{1}{\sqrt{2\phi_0 (1-\alpha)}} B x^{\sigma_{\beta}} \quad (4.4)$$
Since we have shown the standard deviation of items is

$$\sigma_{\Delta x} = B x^\beta,$$

then the quantity under the square root in (4.4) must be one. Thus

$$1 = \frac{1}{2\phi_0(1 - \alpha)}$$

$$\phi_0 = \frac{1}{2(1 - \alpha)}$$

$$\left( \sum A_j^{\frac{1}{1-\alpha}} \right)^{1-\alpha} = \frac{1}{2(1 - \alpha)}$$

$$\sum A_j^{\frac{1}{1-\alpha}} = \left[ \frac{1}{2(1 - \alpha)} \right]^{\frac{1}{1-\alpha}}$$

So

$$x_i^* = \frac{A_i^{\frac{1}{1-\alpha}}}{\sum A_j^{\frac{1}{1-\alpha}}}$$

$$= [2(1 - \alpha)A_i]^{\frac{1}{1-\alpha}},$$

or

$$A_i = \frac{x^{*(1-\alpha)}}{2(1 - \alpha)}.$$  \hspace{1cm} (4.5)

Thus given the most probable score $x^*$ and exponent $\alpha$, we can calculate an $A_i$ such that $x^* = \left( \frac{A_i}{\phi_0} \right)^{\frac{1}{1-\alpha}}$ and $\sigma(x^*) = B x^{*\beta}$ hold. Since we still have no
We ran simulations with \( B = 0.01, \beta = 0.7 \) and \( N = 10^3 \) items, for 500 timesteps. We ran three simulations varying only \( \alpha \). \( \bar{A} \) was determined from \( p(A) \sim A^{-2.1} \) subject to the constraint in Equation (4.5) and \( \sum x^* = 1 \). We plot the \( x(t) \) trajectories of five items logarithmically spanning the range of \( A \) values. Independent of \( \alpha \) we find the same results. Thus we choose \( \alpha = 0 \) to simplify the model.

4.2.2 Measuring \( \bar{A} \)

Now that we have a direct mapping from \( x_i^* \) to \( A_i \) in Equation (4.6), we can calculate the distribution of \( p(A) \):

\[
p(A) = p(x^*) \frac{dx^*}{dA}.
\]
Figure 4.9: The cumulative distribution $P(x^*)$. Here we use $x^* \approx \langle x \rangle$. The dashed line indicates a fit of the form $P(x^*) = \left( \frac{x_{\text{min}}}{x^*} \right)^{\mu}$. See Figure 4.15 for a check of the reliability of this measurement.

In Figure 4.9 we plot the cumulative distribution $P(x^*)$. The dashed line indicates a fit of the form $P(x^*) = \left( \frac{x_{\text{min}}}{x^*} \right)^{\mu}$.

Combining this power law cumulative distribution with Equations 4.6 and 4.7 we have

$$p(A) = \left( \frac{A_m}{A} \right)^{\mu},$$

(4.8)

with $A_m = \frac{x_{\text{min}}}{2}$. We find $\mu \approx 1$ for ngram, marketcap, medicare and twitter, and $\mu \approx 2$ for citation and wikipedia. We summarize the measured values of Figure 4.9 in Table 4.1.
If we use a power law \( p(x) = \frac{\mu x^\mu}{x^{\mu+1}} \) with \( \mu > 1 \) we can also calculate \( x_m \). Since \( \sum x_i = 1 \):

\[
1 = \sum x_i = \int_x^1 Nxp(x)dx \approx N\frac{\mu x_m}{\mu - 1},
\]

with \( N \) the number of items and \( x_m \ll 1 \). Thus

\[
x_m \approx \frac{1 - \frac{1}{\mu}}{N}, \tag{4.9}
\]

for \( \mu > 1 \). As a result, the important parameters to measure are only \( N \), the number of items and \( \mu \) the scaling of \( p(x^*) \).

### 4.3 Spacing-volatility phase diagram with data

#### 4.3.1 Stability phases

We now return to score and rank stability and produce an \( AB \) phase diagram as in Figure 3.3 with the data sets included.
Score Stability

For $\alpha = 0$ the Langevin equation (3.6) becomes

$$\dot{x}_i = A_i + B x_i^\beta \xi_i - \phi x_i.$$  \hspace{1cm} (4.10)

The score distribution $p(x_i)$ from Equation 3.8 reduces to

$$p(x_i) = C(A) x_i^{-2\beta} \exp \left[ \frac{2A_i}{B^2} \left( \frac{x_i^1 - 2\beta}{1 - 2\beta} - \frac{\phi}{2A_i} \frac{x_i^{2(1-\beta)}}{1 - \beta} \right) \right].$$ \hspace{1cm} (4.11)

And the location of the peak from Equation (3.10) occurs at the solution to

$$p'(x_i) = A_i - \phi_0 x_i - B^2 \beta x_i^{2\beta-1} = 0.$$ \hspace{1cm} (4.12)

Our definition of score stability is $p'(x) = 0$ with the constraint $x_{\min} \leq x \leq 1$, where $x_{\min}$ is the minimum score in the system. If no solution exists, the item is defined to be score-unstable.

With $\alpha = 0$ our calculation for $B_c$, the score-critical noise magnitude from Equation (3.11) becomes

$$B_c = G(A, \beta).$$
Rank Stability

We now examine rank stability for $\alpha = 0$. Here (3.12) holds as before, we want

$$\Delta(r) > \sigma(r).$$

If $P(x) = \left(\frac{x_m}{x}\right)^{\mu}$, then $x(r) = x_m \left(\frac{N}{r}\right)^{\frac{1}{\mu}}$. And $\Delta(r) = x(r) - x(r + 1)$ or

$$\Delta(r) = x_m N^{\frac{1}{\mu}} \left[x^{-\frac{1}{\mu}} - (r + 1)^{-\frac{1}{\mu}}\right]. \quad (4.13)$$

And we have

$$\sigma(r) = B x^\beta(r), \quad (4.14)$$

or

$$\sigma(r) = B x_m \left(\frac{N}{r}\right)^{\frac{\beta}{\mu}}. \quad (4.15)$$

So

$$\frac{\Delta(r)}{\sigma(r)} = x_m^{1-\beta} N^{1-\beta} B^{-1} \left[r^{-1/\mu} - (r + 1)^{-1/\mu}\right] r^{\beta/\mu}. \quad (4.16)$$

To summarize, with $\alpha = 0$ and $P(x) \sim x^{-\mu}$, from Equation (3.17) we
have

\[ B_r = H(N, \mu, \beta). \]

**Phase diagrams**

We now turn to summarizing these results on a spacing-volatility phase diagram. In Figure 4.10 we overlay the datasets onto our phase diagram. Recall that for \( \alpha = 0 \) we have \( B_c = G(A, \beta) \) and \( B_r = H(N, \mu, \beta) \). For all data sets we use \( \beta = 0.7 \). For ngram, marketcap, medicare and twitter we use \( N = 10^4 \) and \( \mu = 1.1 \). For citation and wikipedia we use \( N = 10^5 \) and \( \mu = 2 \). Since \( B_r(\mu = 1.1) \neq B_r(\mu = 2) \), we make separate figures for each \( \mu \) case. To determine the range \( A_{\min} \) to \( A_{\max} \) in each data set, we convert \( x_{\min}(N, \mu, r = N) \) and \( x_{\max}(N, \mu, r = 1) \) into \( A \) values using Equation (3.9). The value of \( B \) for each data set is tabulated in Table 4.1.

**4.4 Model simulations**

Now that we have measured all the salient parameters in each system (Table 4.1), we can run simulations and compare the results.

In Figure 4.11 we compare the results of our low noise simulations finding \( x(t) \) is qualitatively similar to that observed in the data. Similarly in 4.12 we observe the same symmetry about the \( \Delta x = 0 \) line.

In Figure 4.13 we compare the results of our high noise simulations: we
Figure 4.10: AB phase diagram for $\alpha = 0$. **Left:** We use $N = 10^4$ and $\mu = 1.1$. **Right:** We use $N = 10^5$ and $\mu = 2$. For the low $B$ systems (ngram, marketcap, medicare), we expect to find all items are score stable while top-scoring items are also rank stable. For the high $B$ systems (citation, twitter, wikipedia), we expect only score unstable or score stable items, no rank-stable items.

reproduce the $x(t)$ behavior for each of the systems, including the noisiness of low-scoring items in twitter. Finally in Figure 4.14 we examine $\Delta x(x)$. The model reproduces some of the asymmetry observed in the data. In Chapter 5 we discuss the limits of our model assumptions which may contribute to any lack of asymmetry observed.

At the outset of this chapter we mentioned that we would limit ourselves to low-noise systems in measuring $p(x)$. Yet here we have proceeded to run simulations for high noise systems as well. Is this reasonable? From Equation (3.8) we have that in a sufficiently noisy system, all items will have the same distribution $p(x) \sim x^{-2\beta}$. This would cause all items to have similar average values and thus our measurement of $P(x^*) \sim x^{*-\mu}$ from Figure 4.9 would...
Figure 4.11: Simulations largely reproduce the $x(t)$ stability observed in the data. We simulate each low-noise data set with parameters used in Table 4.1. We plot $x(t)$ for five items logarithmically spaced from the best to the worst scores. **Top row:** $x(t)$ as observed in the data. **Bottom row:** $x(t)$ from corresponding simulations.

have a greater value for $\mu$ – i.e. the $x^*$’s would be more closely packed. In other words the greater the noise $B$, the greater the measured $\hat{\mu}$ and as $B \to 0$, $\hat{\mu} \to \mu$. As a check, in Figure 4.15 we measure $\mu$ from simulations for each system, finding that $\hat{\mu} \approx \mu$, i.e. the measurements of $\mu$ are reasonable for the noise magnitude in the current systems.
Figure 4.12: Simulations reproduce the $\Delta x = 0$ symmetry observed in the data. We simulate each low-noise data set with parameters used in Table 4.1. **Top row**: Scatter plot of $\Delta x(x)$ as observed in the data. **Bottom row**: Scatter plot of $\Delta x(x)$ from corresponding simulations.
Figure 4.13: Simulations largely reproduce the $x(t)$ volatility observed in the data. We simulate each high-noise data set with parameters used in Table 4.1. We plot $x(t)$ for five items logarithmically spaced from the best to the worst scores. Top row: $x(t)$ as observed in the data. Bottom row: $x(t)$ from corresponding simulations.
Figure 4.14: Simulations reproduce some of the asymmetry observed in the data. We simulate each high-noise data set with parameters used in Table 4.1. **Top row:** Scatter plot of $\Delta x(x)$ as observed in the data. **Bottom row:** Scatter plot of $\Delta x(x)$ from corresponding simulations. See Chapter 5 for discussion.
Figure 4.15: Measured $\hat{\mu}$ is close to true $\mu$. In Figure 4.9 we measure the cumulative distribution $P(x^*)$, where we use $x^* \approx \langle x \rangle$. The dashed line indicates a fit of the form $P(x^*) = (\frac{x_{\text{min}}}{x^*})^{\hat{\mu}}$. To test if the measured $\hat{\mu}$ is accurate, we ran simulations with $\mu = 1.1$ (ngram, marketcap, medicare, twitter) and $\mu = 2$ (citation, wikipedia). Here we find comparable values for the measured $\hat{\mu}$. 

53
Chapter 5

Limits of model assumptions

We now return to the assumptions of our general model (Section 3.2).

5.1 System size is growing

In our simulations we keep the number of items $N$ fixed, but in many real systems this number can grow – new words are introduced into the language, previously unknown diseases appear, or new topics are discussed on wikipedia and twitter.

To test whether our studied systems are growing, we plot the number of items $N$ as a function of time on a log-linear scale in Figure 5.1. We find that four of our systems (ngram, marketcap, medicare, wikipedia) do not grow by orders of magnitude during the study period. Two however (citation and twitter) do. This growth can be approximated of the form $N \sim e^{\alpha t}$. This is,
of course expected for citation as new papers are introduced each year and the growth of number of scientists been exponential [44].

Such growth can also be seen by the shifting cumulative distributions of citation and twitter in Figure 2.1. Assuming the cumulative distribution is of the form $P(x) = \left( \frac{x_m}{x} \right)^\mu$ with fixed scaling $\mu$, this shifting $x_m$ is a direct result of system growth, as shown in Equation (4.9),

$$x_m \sim \frac{1}{N}.$$ 

Such growth implies that new items may not have had time to reach a steady state behavior (assuming one exists).

\section{5.2 $p(x)$ not in steady state}

In our analysis we assume items have had time to reach their steady state behavior. In growing systems such as citation and twitter, this may not be the case. In Figure 5.2 we take an example from citation. The paper ‘Self-Consistent Equations’ appears to have reached its steady state behavior, comfortingly obtaining a large market share over the entire study period. The other papers however, all first published around the middle of the study period, gain large market shares in their early years, but settle down to a steady-state score two orders of magnitude below their introduction. Such behavior of items moving from high scores to low contributes to the asymme-
Figure 5.1: The assumption that the number of items \( N \) in the system is not growing by orders of magnitude holds for four of our data sets: ngram, marketcap, medicare and wikipedia. For citation and twitter however, the system growth can be approximated by exponential growth: \( N \sim e^{\gamma t} \). Here \( N \) is measured as the number of items with a non-zero score at time \( t \).

try in the scatter plot of \( \Delta x(x) \) (Figure 5.2, bottom left). In our simulations we measure the system after it has reached its steady state behavior (and introduce all items at the initial timestep), any asymmetry in \( \Delta x(x) \) due to items approaching their steady state will not be observed in our simulations (Figure 5.2, bottom right). As a test, one could introduce new items into the system with random initial conditions, this may result in greater asymmetry in \( \Delta x(x) \) as observed in the data.
Figure 5.2: In systems such as citation, new items (papers) are introduced and take time to reach a steady state distribution. We do not incorporate this behavior in the simulation. **Top:** The paper “Self-Consistent Equations” appears to have reached a steady state distribution for \( p(x) \) as a highly-cited paper. The remaining papers begin with high-scores but reach more stable market shares at much lower values. **Bottom left:** Such behavior contributes to the asymmetry of \( \Delta x(x) \) observed in the data – many top scoring items have a tendency to fall as they approach their steady state distribution. **Bottom right:** Although some asymmetry is observed in \( \Delta x(x) \), it is weaker than that in the data – this may in part be due to no new items introduced during the simulation, thus all items have already reached their steady state values.
5.3 $A_i$ not constant

Additionally our simulations rely on $A_i$ being constant – an item has a preferred fitness and, in the absence of noise, items will be ranked by this fitness. This of course is only a model. A paper may be published, collecting a scant few citations a year, when years later new-found applications may propel its importance to the forefront. Its fitness thus changes over time.

To explore an example observed in our data we look at wikipedia. In Figure 5.3, top, we plot the scores of several top pages over time. During this period both Thomas Edison’s birthday and Valentine’s Day occurred. Accordingly, their scores spiked around the event. Such items are not as ‘fit’ the rest of the year. This behavior also contributes to the volatility of the system contributing to the asymmetry in $\Delta x(x)$ as low scoring items move up and top scoring items fall (Figure 5.3, bottom left). In contrast our model produces $\Delta x(x)$ fairly symmetric about the $\Delta x = 0$ line (Figure 5.3, bottom right).
Figure 5.3: \( A_i \) may not be independent of time. **Top:** Items such as “Valentine’s Day” and “Thomas Edison” have clear spikes in their scores. These are not, however due to randomness, both Valentine’s Day and Edison’s birthday occurred during the study period. On those particular days each item suddenly becomes more “fit” causing it to jump from a low-score to a high-score, and vice-versa after the event. **Bottom left:** Such behavior contributes to the asymmetry of \( \Delta x(x) \). **Bottom right:** In simulations, all items have a fixed fitness \( A_i \), and such asymmetry is not as strongly reproduced.
Chapter 6

Going further

6.1 Why is a system stable?

So far we left an important question unanswered: what determines the value of the noise magnitude $B$? It is a more fundamental question. Why are some systems inherently stable or unstable? And can we quantify the difference?

We hypothesize that $B$ is driven by the amount of attention given to each item. Given sufficient scrutiny, the fitness of all items will be reflected in the ranking resulting in lower noise $B$. Think of this as having sufficiently sampled the items such that we know the true $A_i$ for all items [33]. However if the sampling is too low, many items will have an unknown fitness, hence the outcome remains volatile (high $B$).

We test this hypothesis by measuring the average scrutiny of an item defined as $E/N$, where $E$ is the number of events per unit time, like the
total number of wikipedia visits or disease diagnoses, and $N$ is the number of ranked items. For the three stable systems $B \sim 10^{-3}$ (Table 4.1) while $E/N$ is in the range $10^3$–$10^6$. In contrast, for the three unstable systems $B \sim 10^{-1}$ and $E/N$ is in the range 1–10. We plot $B \left(\frac{E}{N}\right)$ in Figure 6.1(a), clearly showing an inverse relationship between $B$ and $E/N$ – the greater the scrutiny, the lower the noise.

To further test this idea, we reduce the scrutiny, remeasuring $B$ along the way. Specifically, we randomly select a $p$ fraction of the events $E$ to keep. We recalculate $N$ as the number of items having at least one event. We then remeasure $\sigma_{\Delta x} = B \sigma^\beta$ in this less-scrutinized system. If our hypothesis is correct, then $B$ should increase as $E/N$ decreases. We reduce $p$ from $p = 1$ to $p$ sufficiently small that $E/N \approx 1$; in practice this turned out to be $p_{\text{min}}$ in the range $10^{-4}$–$10^{-7}$ depending on the data set. We summarize our findings in Figure 6.1(b).

Thus scrutiny is a possible explanation for the emergence of stable systems. High scrutiny is clearly present in markets, language and diseases, where a relatively small number of items are scrutinized, used or diagnosed millions of times. Stability is less likely to emerge in the environments where the audience is too small to sample all items. Although this is not the only possible explanations for system stability, these results suggest a more rigorous study could prove fruitful.
Figure 6.1: Noise parameter $B$ as a function of the average scrutiny $E/N$, where $E$ is the number of events per timestep and $N$ is the number of ranked items. **Left:** We find low noise (low $B$) systems tend to have higher scrutiny ($E/N$). Error bars indicate the 95% confidence intervals for $B$ from the fit $\sigma_{\Delta x} = Bx^\beta$. **Right:** By randomly reducing the scrutiny, we find even stable systems become unstable. The right most point for each dataset contains all events. As we delete events (reduce $E/N$) we find the noise $B$ remains stable until around 10 events per item, whereupon the noise sharply increases.
6.2 A case for Levy jumps

Here we explore the question of long range jumps for $\Delta x$. Much work has gone into the study of fluctuations in financial markets [7, 55, 51, 40, 38]. A common measurement is the price return $Z = \log \frac{y(t+\Delta t)}{y(t)}$ where $y(t)$ is the price at time $t$ and later time $t+\Delta t$. $Z$ is a measure of the number of orders of magnitude the price has increased or dropped. They find $p(Z) \sim L(Z)$ where $L$ is a Levy stable distribution (generally a distribution in between the cases of the Gaussian and Cauchy illustrated in Figure 6.2) [39, 26, 35]. How does this compare with our study?

In Figure 6.3 we plot $p(Z)$ with $Z = \log \frac{x(t+\Delta t)}{x(t)}$ for each of our datasets. For comparison we plot the two example distributions of Gaussian and Cauchy.
We too find \( p(Z) \) to be much heavier tailed than a Gaussian in several of our data sets. This should be the case for marketcap as we are measuring the same system, but we also find it to be the case for ngram, twitter and wikipedia.

How does \( Z \) relate to the quantity \( \Delta x \) we have focused on so far? Assume for simplicity the tails of \( p(Z) \) are of the form \( p(Z) \sim |Z|^{-\gamma} \). Since \( Z = \log \frac{\Delta x + x}{x} \) we can make a change of variables using \( p(\Delta x) = p(Z) \frac{dz}{dx} \), obtaining

\[
p(\Delta x) \sim \left| \log \frac{\Delta x + x}{x} \right|^{-\gamma} (\Delta x + x)^{-1}
\]
or

\[
p(\Delta x) \sim \begin{cases} 
|\log x|^{-\gamma} \Delta x^{-1}, & \text{for } \Delta x \gg x \\
|\Delta x|^{-\gamma}, & \text{for } \Delta x \ll x.
\end{cases}
\text{(6.1)}
\]

or

\[
p(\Delta x) \sim \begin{cases} 
|\log x|^{-\gamma} \Delta x^{-1}, & \text{for } \Delta x \gg x \\
|\Delta x|^{-\gamma}, & \text{for } \Delta x \ll x.
\end{cases}
\text{(6.2)}
\]

In both cases \( p(\Delta x) \) has heavy tails.

In Figure 6.4 we plot \( p(\Delta x) \) for selected values of \( x \). We are effectively plotting vertical slices from the scatter plot \( \Delta x(x) \) in Figure 2.4. Each \( p(\Delta x|x) \) has an increasingly greater standard deviation \( \sigma = Bx^\beta \), thus using \( \frac{\Delta x}{\sigma} \) as the variable causes all curves to collapse, Figure 6.5. We observe that indeed, if \( p(Z) \) is heavy tailed, so is \( p(\Delta x) \). The asymmetry of the unstable cases in Figure 6.5 is a result of boundary effects due to rescaling and is discussed in Section 8.5.

As an example of such long range jumps, in Figure 6.6 we run two simula-
tions, one with Gaussian jumps and another with Levy jumps. For comparison we plot $x(t)$ for selected items in twitter. Qualitatively we observe similar order of magnitude jumps typical of heavy tailed noise, that are absent in Gaussian noise.

Such distinctions may seem minor but much has been written about the importance of this difference. Indeed estimations of the likelihood of rare events will be drastically different depending on the kind of randomness in the system. Perhaps systems may be classified, not by the noise variance $B$ as we have done here, but by the type of noise observed – some systems subject much heavier tailed jumps than others. Such study is crucial for estimating the likelihood of rare events [6, 36, 53], as different models will produce many orders of magnitude differences in probability estimates, producing statements such as, “The 1987 stock market crash was, according to such models, something that could happen only once in several billion billion years” [37]. Thus such questions are worthy of further study in the area of ranked systems.
Figure 6.3: The log-return distribution $p(Z)$, where $Z = \log \frac{x(t+\Delta t)}{x(t)}$. Solid curve is a Gaussian Fit to the data, dashed curve is a Cauchy distribution. We find four of the data sets (ngram, marketcap, twitter and wikipedia) have tail heavier than a Gaussian. This indicates much greater volatility than a Gaussian model would predict.
Figure 6.4: $p(\Delta x|x)$, the distribution of change in score given $x$. This is equivalent to plotting vertical slices from Figure 2.4. Since we know the width of each distribution scales as $\sigma = B x^\beta$ we plot the rescaled $p\left(\frac{\Delta x}{\sigma}\right)$ in Figure 6.5.
Figure 6.5: $\Delta x$ exhibits heavy tailed jumps. Here we plot $p(\frac{\Delta x}{\sigma})$, the rescaled distribution of $p(\Delta x|x)$ from Figure 6.4. We find that items exhibit long range jumps not used in our current model. The asymmetry of the high noise systems is a function of the rescaling, see Section 8.5.
Figure 6.6: Do some systems exhibit Levy jumps? **Top left:** Simulations using Gaussian noise. **Top right:** Simulations using Levy noise. **Bottom:** Twitter data.
Chapter 7

In Conclusion

In this dissertation we provide a formalism to quantify a wide class of ranked systems using the concepts of spacing and volatility. We propose a model using stochastic differential equations and, despite their diversity, capture the salient features of the systems. Furthermore we condense each system onto a spacing-volatility phase diagram, identifying regimes of score- and rank-stability separated by a noise-induced phase transition. We also examine the underlying causes of system stability by exploring the scrutiny of items, finding that systems with highly scrutinized items tend to be more stable than those with low scrutiny. Finally we look at the possibility of using Levy jumps for the noise term as opposed to Gaussian jumps. With additional embellishments – for example the incorporation of fads, herding and long range jumps [48, 4] – the introduced formalism may be extendible to an even wider range of systems such including biodiversity and linguistics [56, 10].
Yet even as it stands, the proposed model manages to capture the salient features of a wide class of ranked systems.
Appendix

8.1 $\phi$ normalization

To maintain the normalization of $x_i$ we use the Lagrange multiplier $\phi$, thus

$$\dot{x}_i = \mu(x_i) + \sigma(x_i)\xi_i(t) - \phi(t)x_i.$$  \hspace{1cm} (8.1)
\( \phi \) is thus

\[
\begin{align*}
1 &= \sum x_i \quad (8.2) \\
0 &= \sum \dot{x}_i \quad (8.3) \\
0 &= \sum_j [\mu_j(x_j) + \sigma_j(x_j)\xi_j] - \sum_j x_j\phi \quad (8.4)
\end{align*}
\]

\[
\phi \sum_j x_j = \sum_j [\mu_j(x_j) + \sigma_j(x_j)\xi_j] \quad (8.5)
\]

\[
\phi = \sum_j [\mu_j(x_j) + \sigma_j(x_j)\xi_j] \quad (8.6)
\]

\[
\phi = \phi_0 + \eta, \quad (8.7)
\]

where \( \phi_0 = \sum \mu_j(x_j) \), the system drift term, and \( \eta = \sum \sigma_j(x_j)\xi_j \), the system noise term.
8.2 Derivation of $p(x_i)$ from Fokker-Planck equation

Beginning with the Fokker-Planck Equation (3.7) we wish to derive the steady state solution for $p(x_i)$.

We have

$$\frac{\partial p(x_i)}{\partial t} = -\frac{\partial}{\partial x_i} [(A_i x_i^\alpha - \phi x_i)p(x_i)] + \frac{1}{2} \frac{\partial^2}{\partial x_i^2} \left( B^2 x_i^{2\beta} p(x_i) \right).$$

At steady state $\frac{\partial}{\partial t} p(x_i) = 0$, so

$$\frac{\partial}{\partial x_i} [(A_i x_i^\alpha - \phi x_i)p(x_i)] = \frac{1}{2} \frac{\partial^2}{\partial x_i^2} \left( B^2 x_i^{2\beta} p(x_i) \right).$$

Integrating once, assuming the constant term is zero, we obtain

$$(A_i x_i^\alpha - \phi x_i)p(x_i) = \frac{1}{2} \frac{\partial}{\partial x_i} \left( B^2 x_i^{2\beta} p(x_i) \right),$$

which is of the form

$$fp = \frac{\partial}{\partial x_i} (gp),$$
with

\[ f = (A_i x_i^\alpha - \phi x_i) \]
\[ g = \frac{1}{2} B^2 x_i^{2\beta} \]  \hspace{1cm} (8.8)
\[ p = p(x_i). \]

So

\[ fp = \frac{\partial}{\partial x_i} (gp) \]
\[ fp = g'p + gp' \]
\[ p' = \frac{f - g'}{g} p. \]  \hspace{1cm} (8.9)

If \( p \) is of the form

\[ p = e^{m(x)}, \]  \hspace{1cm} (8.10)

then

\[ p' = m'(x)p. \]
Comparing this with (8.9) we have

\[ m'(x) = \frac{f - g'}{g} = \frac{f}{g} - \frac{g'}{g} \]

\[ m(x) = \int \frac{f}{g} dx - \int \frac{g'}{g} dx. \]  

(8.11)

Using (8.8), first integral in (8.11) is

\[ \int \frac{f}{g} dx = \int \frac{A x^{\alpha} - \phi x}{\frac{1}{2} B^2 x^{2\beta}} dx \]

\[ = \frac{2A}{B^2} \frac{x^{\alpha-2\beta+1}}{\alpha - 2\beta + 1} - \frac{\phi}{B^2} \frac{x^{2(1-\beta)}}{1 - \beta}, \]

and the second integral is

\[ \int -\frac{g'}{g} dx = \int \frac{\beta B^2 x^{2\beta-1}}{\frac{1}{2} B^2 x^{2\beta}} dx \]

\[ = \ln x^{-2\beta}. \]

Substituting these results for \( m(x) \) back into (8.10) we obtain

\[ p(x_i) = C(A_i) x_i^{-2\beta} \exp \left[ \frac{2A_i}{B^2} \left( x_i^{1+\alpha-2\beta} \frac{1}{1 + \alpha - 2\beta} - \frac{\phi_0}{2A_i} \frac{x_i^{2(1-\beta)}}{1 - \beta} \right) \right], \]

where we include the normalization constant \( C(A_i) \) such that \( \int p(x_i) dx = 1. \)

We use this result in Equation (3.8).
8.3 Derivation of $\sigma(x^*)$ from $p(x)$

We approximate the standard deviation $\sigma(x)$ of an item given the distribution $p(x)$ following the approach in [27]. We start with $p(x)$ from (3.8), dropping the subscript $i$ for clarity:

$$p(x) = C(A) \ x^{-2\beta} \exp \left[ \frac{2A}{B^2} \left( \frac{x^{1+\alpha-2\beta}}{1 + \alpha - 2\beta} - \frac{\phi_0 \ x^{2(1-\beta)}}{2A \ 1 - \beta} \right) \right].$$

This is of the form

$$p(x) = f(x)e^{g(x)} \quad (8.12)$$

with

$$f(x) = C(A) \ x^{-2\beta}$$

$$g(x) = \frac{2A}{B^2} \left( \frac{x^{1+\alpha-2\beta}}{1 + \alpha - 2\beta} - \frac{\phi_0 \ x^{2(1-\beta)}}{2A \ 1 - \beta} \right).$$

We consider the case when $p(x)$ is peaked about $x^*$. We examine $p(x)$ about this peak by expanding $g(x)$ about $x = x^*$ thus

$$g(x) \approx g(x^*) + g'(x^*)(x - x^*) + \frac{1}{2}g''(x^*)(x - x^*)^2.$$
At the peak the $g'(x^*) = 0$ thus

$$g(x) \approx g(x^*) + \frac{1}{2} g''(x^*)(x-x^*)^2.$$ 

If the exponential is sharply peaked then $f(x) \approx f(x^*)$ thus (8.12) becomes

$$p(x) \approx f(x^*) e^{g(x^*)} e^{\frac{1}{2}g''(x^*)(x-x^*)^2}$$

(8.13)

around $x = x^*$. Comparing (8.13) with a normal distribution,

$$\mathcal{N} = \frac{1}{\sqrt{2\pi} \sigma^2} e^{-\frac{(x-x_0)^2}{2\sigma^2}},$$

and matching the argument in the exponent, we have

$$-\frac{1}{2\sigma^2} = \frac{g''(x^*)}{2},$$

or

$$\sigma^2 = -\frac{1}{g''(x^*)}.$$ 

(8.14)

Now

$$g''(x) = \frac{2A}{B^2} \left[ (\alpha - 2\beta)x^{\alpha-2\beta-1} - \frac{\phi}{A} (1-2\beta)x^{-2\beta} \right]$$

$$= \frac{2A}{B^2} x^{-2\beta} \left[ (\alpha - 2\beta)x^{\alpha-1} - \frac{\phi}{A} (1-2\beta) \right].$$
At the peak \( x^* = \left( \frac{A}{\phi} \right)^{\frac{1}{1-\alpha}} \); substituting in

\[
g''(x^*) = \frac{2A}{B^2} \left( \frac{A}{\phi} \right)^{\frac{-2\beta}{1-\alpha}} \left[ (\alpha - 2\beta) \frac{\phi}{A} - \frac{\phi}{A} (1 - 2\beta) \right]
\]

\[
= \frac{2\phi}{B^2} \left( \frac{A}{\phi} \right)^{\frac{-2\beta}{1-\alpha}} (\alpha - 1).
\]  

(8.15)

Substituting (8.15) into (8.14) we obtain

\[
\sigma^2 = \frac{B^2}{2\phi(1-\alpha)} \left( \frac{A}{\phi} \right)^{\frac{2\beta}{1-\alpha}},
\]

or

\[
\sigma(x^*) = \frac{1}{\sqrt{2\phi(1-\alpha)}} B x^{*\beta}
\]

Which we use in Equation (3.14).
8.4 Derivation of $\sigma_\dot{x}$ from Langevin equation

We wish to calculate the standard deviation of $\sigma_\dot{x}$ given $\dot{x} = \mu + \sigma \xi - x \phi$.

We have

$$\sigma_\dot{x}^2 = \langle \dot{x}^2 \rangle - \langle \dot{x} \rangle^2.$$  \hfill (8.16)

And

$$\langle \dot{x} \rangle = \mu - x \phi_0$$ \hfill (8.17)

with $\phi = \phi_0 + \eta$ as defined in (8.7). Also

$$\langle \dot{x}^2 \rangle = \mu^2 + \sigma^2 + x^2 \langle \phi^2 \rangle - 2\mu x \langle \phi \rangle - 2\sigma x \langle \xi \phi \rangle = \mu^2 + \sigma^2 + x^2 \phi_0^2 - 2\mu x \phi_0 - 2\sigma x \xi \phi$$ \hfill (8.18)

$$= \mu^2 + \sigma^2 + x^2 \phi_0^2 - 2\mu x \phi_0 - 2\sigma x \xi \phi$$ \hfill (8.19)

$$= \mu^2 + \sigma^2 + x^2 \phi_0^2 - 2\mu x \phi_0 - 2\sigma x \xi \phi$$ \hfill (8.20)

$$= \mu^2 + \sigma^2 + x^2 \phi_0^2 - 2\mu x \phi_0 - 2\sigma^2 x.$$ \hfill (8.21)

Thus

$$\sigma_\dot{x}^2 = \langle \dot{x}^2 \rangle - \langle \dot{x} \rangle^2$$ \hfill (8.22)

$$= \sigma^2 - 2\sigma^2 x + x^2 \sum \sigma_j^2.$$ \hfill (8.23)
Since $x_i \ll 1$ we drop the higher order terms and

$$\sigma_i^2 \approx \sigma_i^2$$  \hspace{1cm} (8.24)

(hence our choice of notation $\sigma_i$).
8.5 The Asymmetry of $p\left(\frac{\Delta x}{\sigma}\right)$

In Figure 6.4 we observe the negative tails of $p\left(\frac{\Delta x}{\sigma}\right)$ are tightly bounded for high noise systems. This is a result of the boundary conditions of the system. Let $\Delta x_{max}^-$ be the greatest possible fall in score an item can have. Then $\Delta x_{max}^- = -x_{max}$, a fall from the top score to 0. Let $\sigma$ be the standard deviation for an item at score $x$. Assume $\Delta x_{max}^-$ is $n_{max}^-$ standard deviations, thus $\Delta x_{max}^- = -n_{max}^+\sigma$. So $n_{max}$ is the maximum possible number of standard deviations an item can fall. Since $\sigma = Bx^\beta$ we have

$$n_{max}^- = \frac{\Delta x_{max}^-}{\sigma} = B^{-1}x_{max}^{1-\beta}.$$

Let $x_{max} = 10^{-1}$, and $\beta = 0.7$ as observed in the data. For a low noise system $B = 10^{-3}$ and thus $n \approx 500$. For a high noise system $B = 10^{-1}$ and thus $n \approx 5$. Thus in a low noise system a top scoring item can fall as much as 500 standard deviations, whereas in a high noise system an item that falls more than 5 standard deviations is not observed. As such the left hand side of Figure 6.5 does not exceed 5 standard deviations.

Such a bound is not a factor for $\Delta x > 0$. A similar analysis results in

$$n_{max}^+ = \frac{1 - x_{min}}{Bx^\beta}.$$

If an item has a low score, say $x = 10^{-6}$ a jump of $n = 10^5$ standard deviations
is possible in a low noise system and one of $n = 10^7$ is possible in a high noise system. Thus such bounds are not an issue for $\Delta x > 0$. 
8.6 Publications and Posters During my PhD


Bibliography


