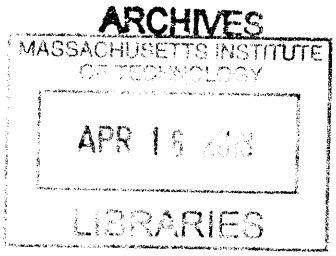
# How Predictable

modeling rates of change in individuals and populations

by
Coco Krumme

MS, MIT 2010
BS, Yale 2005




Submitted to the Program in Media Arts and Sciences
School of Architecture and Planning
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2013



Author: _____

Coco Krumme
December 19, 2012

Certified by: _____

Alex "Sandy" Pentland
Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Accepted by: _____

Prof. Patricia Maes
Associate Academic Head
Program in Media Arts and Sciences

# How Predictable

modeling rates of change in individuals and populations

by
Coco Krumme

Submitted to the Program in Media Arts and Sciences
School of Architecture and Planning
on December 19, 2013
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

## Abstract

This thesis develops methodologies to measure rates of change in individual human behavior, and to capture statistical regularities in change at the population level, in three pieces: i) a model of individual rate of change as a function of search and finite resources, ii) a structural model of population level change in urban economies, and iii) a statistical test for the deviation from a null model of rank churn of items in a distribution.

First, two new measures of human mobility and search behavior are defined: exploration and turnover. Exploration is the rate at which new locations are searched by an individual, and turnover is the rate at which his portfolio of visited locations changes. Contrary to expectation, exploration is open-ended for almost all individuals. A present a baseline model is developed for change (or churn) in human systems, relating rate of exploration to rate of turnover. This model recasts the neutral or random drift mechanism for population-level behavior, and distinguishes exploration due to optimization, from exploration due to a taste for variety. A relationship between the latter and income is shown.

Second, there exist regular relationships in the economic structure of cities, with important similarities to ecosystems. Third, a new statistical test is developed for distinguishing random from directed churn in rank ordered systems. With a better understanding of rates of change, we can better predict where people will go, the probability of their meeting, and the expected change of a system over time. More broadly, these findings propose a new way of thinking about individual and system-level behavior: as characterized by predictable rates of innovation and change.

Thesis supervisor: Alex "Sandy" Pentland
Professor of Media Arts and Sciences

# How Predictable

modeling rates of change in individuals and populations

by

Coco Krumme

Thesis Reader: _____

László Barabási
Professor of Physics
Northeastern University

# How Predictable

modeling rates of change in individuals and populations

by

Coco Krumme

Thesis Reader: _____

Hal Varian
Chief Economist
Google

thank you

at MIT and environs: Sandy Pentland, Nicole Freedman, Yves-Alexandre de Montjoye, Nadav Aharony, Manuel Cebrián, Wei Pan, László Barabási, Chaoming Song, Hal Varian, Thomas Pfeiffer, David Rand, Lily Tran, Nathan Eagle, César Hidalgo, Erik Ross, Dan Ariely, Micah Eckhardt, Santiago Alfaro, Taya Leary

at SFI: Luís Bettencourt, Hyejin Youn, Marcus Hamilton, Geoffrey West, Anne Kandler, James O'Dwyer, Charles Perrault

collaborators: Sandy Pentland, Manuel Cebrián, and Galen Pickard (section two: online market), Sandy Pentland, Alejandro Llorante, Esteban Moro, and Manuel Cebrián (section two: predictability of consumer trajectories), Luís Bettencourt (section five: rank churn)

of course: Maman, Dad, Sasha

# Contents


| | | |
|---|---|---|
| 1 | Introduction and contributions | 12 |
| 2 | Data and background | 16 |
| 3 | Individual rates of change as a function of exploration and resources | 30 |
| 4 | Rates of change in the structure of urban economies | 43 |
| 5 | A statistical test for churn in distributions of ranked items | 55 |
| 6 | Discussion: summary of findings, summary of contributions, and future work | 66 |
| 7 | Appendix: MusicLab polya model | 71 |

# List of Figures


2.1 Probability of visiting a merchant, as a function of merchant visit rank, aggregated across all individuals. Dashed line corresponds to power law fits $P(r) \sim r^{-\alpha}$ to the initial part of the probability distribution with $\alpha = 1.13$ for the European and $\alpha = 0.80$ for the American database. . . . . 25

2.2 Entropy distributions for the American and European time series. Temporally-uncorrelated entropy distributions are slightly higher for both populations. . . . . 25

2.3 Sequence-dependent entropy for a number of artificially sorted sequences. For each window size over which the time series is sorted, we measure the sequence-dependent entropy for the population. . . . . 26

2.4 Markov model results for various temporal windows in training and test. The solid red line indicates hit percentage for Markov model, dashed line exhibits accuracy for the naive model and the dotted line indicates results for the Global Markov model. . . . . 27

3.1 Exploration is the ratio of new store visits to total store visits in a time window $t$. Turnover is the rate at which locations in a basket of top-$n$ most frequent locations are dropped or churn . . . . . 38

3.2 Exploration is open-ended, and rises linearly with respect to the size of the time window $t$ (top). Distribution of Turnover for basket sizes $n = 1$ and $n = 10$ (bottom) 39

3.3 Relationship between Exploration and Turnover in the data, and predictions of the neutral and modified neutral models. The relationship is nearly linear for basket size n=1, but the effect of Exploration on Turnover saturates when n > 2. . 40

3.4 Relationship between T and n for two individuals (top). We can extract two motivations for search: optimization of a portfolio, $O$, and $W$, a taste for variety. Relationship between $W$ and income (bottom). . . . . 41

3.5 Schematic of the neutral and modified neutral models. In the neutral model, an individual replaces visits according to his exploration parameter $E$. He then fills the remaining visits with types chosen from the distribution of visits in $t-1$. In the modified neutral model, he again fills remaining visits in $t$ by choosing from a *weighted* distribution of $t-1$ visits. . . . . 42

4.1 For a simple tree, at each node the actual branch length a, cumulative branch length c, bifurcations b, and abundance n (with each green box representing one firm) . . . . . 50

4.2 Symmetry and evenness measures in hierarchical trees. The regularity of the relationship between a and c gives a sense of the scale-invariance of the tree. The slope gives a sense of the tree's symmetry. . . . . 51

4.3 Number of firms varies linearly with population. Each 2-digit code is a sector: 22 Utilities / 42 Wholesale Trade / 44-45 Retail Trade / 48-49 Transportation and Warehousing / 51 Information / 52 Finance and Insurance / 53 Real Estate and Rental and Leasing / 56 Administrative and Support and Waste Management and Remediation Services / 72 Accommodation and Food Services . . . . . 52

4.4 In small cities there is poor scaling in the economic tree. In large cities, the scaling is stronger. Small cities are lopsided: sectors see either over- (such as a small mining town) or under- (for example, a village without all the shades of retailers you see in a big city) specialization . . . . . 53

4.5 In small cities, the relationship between bifurcations and firm abundance is poor. In large cities, the relationship is stronger in most sectors . . . . . 53

4.6 Evenness as a function of tree shape. Each 2-digit code is a sector: 22 Utilities / 42 Wholesale Trade / 44-45 Retail Trade / 48-49 Transportation and Warehousing / 51 Information / 52 Finance and Insurance / 53 Real Estate and Rental and Leasing / 56 Administrative and Support and Waste Management and Remediation Services / 72 Accommodation and Food Services . . . . . 54

5.1 Samples $N_i$ and $N_{i+1}$ are drawn from global distribution $G$. What is the probability that with the addition of $N_{i+1}$, the rank of the orange type rises from 2 to 3? . . . . . 62

5.2 Most populous U.S. cities, 1900 and 1990 . . . . . 63

5.3 Frequency of top 20 words in the King James Bible . . . . . 63

5.4 Simulation of expected rank churn, given parameters $N_i$, $N_i$ / $N_T$, and $p_i$. Distribution is more narrow for $N_T = 1000$ versus $N_T = 100$ . . . . . 64

5.5 Results of rank churn test for top 3, top 10, for all cities from 1980-1990, and all cities at all times. . . . . 65

5.6 Expected and observed churn for Buffalo, Cambridge, and Los Angeles . . . . . 65

7.1 Quality and appeal are independent. Values are shown for quality and appeal corresponding to the 48 songs in Experiment 2, independent condition $R^2 = 0.012$ 77

7.2 Availability in the independent world of Experiments 1 (A, top) and 2 (B, bottom), indexed to 1. The availability of a position n describes the likelihood that a song in that position will be sampled (where n=1 is the top left corner in Experiment 1, and the topmost position in Experiment 2, and n=48 is the bottom right corner in Experiment 1 and the bottom of the column in Experiment 2). Availability serves as a multiplier in calculating the total probability of a song being sampled, given its position-independent appeal, and its position at a given time in the market. In Experiment , songs on the left side of the grid are more likely to be sampled, all else equal, than songs on the right. In Experiment 2, songs at the top of the column, as well as the final song, are more likely to be sampled. . . . . 78

7.3 Song selection as a two-step process. A listen first selects which song(s) he will listen to, and after listening, decides whether or not to download the song. The first decision is made based on the appeal of a song; the second based on its quality. If a listener listens to more than one song, this process is repeated. . . . . 79

7.4 Inequality (top) and unpredictability (bottom) over the course of the market, with alpha = 900. Inequality is shown for Experiment 1, world 3. RMSE of simulated market's unpredictability is = 0.0017, and average of inequality is = 0.093 . . . . 79

7.5 Inequality (top) and unpredictability (bottom) over the course of the market, with alpha = 200. Inequality is shown for Experiment 2, world 5. RMSE of simulated market's unpredictability is = 0.0012, and average of inequality is = 0.0516 . . . 80

# 1. Introduction and contributions

This thesis develops measures for rates of change in individual- and population- level behavior, and proposes a test for the valance of change in rank-ordered systems.

This research is comprised of three parts: a model for individual rate of change as a function of search and finite resources, a structural model of population level change in urban economies, and a statistical test for the deviation of a distribution from a null model of change in rank.

At the individual level, we often treat choice as occurring in a fixed environment. An individual will find his favorite locations and routes, after which he'll only adapt to exogenous changes in the environment. Absent environmental changes, a person at some point will stand still. Economic theory assumes this bound on an individual's rate of change: once a local optimum is fixed, a person's habits only adapt to exogenous changes.

To consider these assumptions, We draw on one of the largest datasets of individual behavior studied to date. We show that the relationship between exploration rate and rate of behavioral change is approximated by a modified random drift model, in which innovations and copying of past behavior together explain present behavior. Further, we can differentiate two drivers of this innovation: individuals change based on their rates of exploration due to optimization and due to a taste for variety. The latter but not the former is related to individual resources, measured by income.

At the population level, we may be interested in how a system of items changes over time (just as we are interested in how a person's "portfolio" of behaviors changes). We find that patterns of urban-level economic diversity are consistent across cities, and well-predicted by an ecological model of species abundances. This suggests that cities might evolve according to rules similar to those by which ecosystems evolve.

When we observe change at the individual or population level, it is important to be able to sort out the processes that cause it. By testing against a null model of expected churn, the rate of change in a system (such as a set of cities or word frequencies in a corpus) due to random drift (for example, new people being randomly assigned to cities) can be systematically differentiated from rate of change due to a quality endemic to the system (for example, a single city developing a new industry and thus growing more quickly than the null model would predict). We develop a statistical test that relates the rates of individual and population change, and which provides a measure for the level of deviation of a system from expected churn in the null model.

**Why change counts**

Change is the only constant. Classical physics has developed experiments to measure the rate of change of an object: the first derivative, the distance between two points and the time needed to traverse it.

Change may be constant, but it's not consistent. In spite of the sensitivity of many complex systems, such as the dynamics of financial markets or the spread of disease, to continuous human influence, we lack good models for rates of change in human behavior. On the one hand, economic theory holds that individuals optimize over a set of preferences and constraints [37]: rates of behavior change are limited by time scale at which constraints change and force reevaluation of our habits. But we observe many choices -trying a new restaurant without intention to return, for example- that aren't in immediate service of optimization: that is, they don't have any effect on the next month's choices. Utility maximization as a motivation for search is unable to account for choices that don't improve a current set of places and paths.

On the other hand, theories of learning and habit predict that individuals will not simple optimize and settle, but continue to change over time, by sampling new behaviors and selectively copying past ones [3]. William James likens habit to a current, which once it has "traversed a path, it should traverse it more readily still a second time" [28], and more modern neuroscience has confirmed that behavior catalyzed by an arbitrary first step and then reinforced becomes more automatic [21]. Animals form foraging routines as a response to constraints in the environment, often retracing paths even when a superior option becomes available [59]. While such theories account for observed rates of change, they are unable to explain why we search in the first place.

Elective human behavior appears a combination of these two motivations: to optimize for existing needs, and to try new things. The result is that individuals change perceptibly over time, with important impacts on economic and social systems: a person may switch his favorite lunch spot, find a different accountant, move to a new city, upgrade his car, marry and divorce.

At the level of urban systems, rates of change are typically based on counts and comparisons of city features, such as population, industrial diversity, or productivity. Little attention has been paid to how individual-level measures aggregate to population-level ones, or to whether there exist regularities in the rates of change of the structure of a population.

In fact, the presence of economic diversity in cities is typically attributed to competitive factors such as differentiation and economies of scale [32], or to features of demand, such as consumer preferences and taste for variety [29]. Yet the structure of multiple interacting types (such as species or firm categories) is often critical to understanding how an ecosystem or economy might have developed, or why demand is expressed in a certain way.

If change is constant, what are its drivers? In particular, can we distinguish random "noise" from the more fundamental drivers of change? This question is of particu-

lar interest in systems that show a consistent distribution (for example, a Zipf-like or power law distribution) of types over time, but whose types and items-within-type are continually churning. A regularity has been observed in such disparate distributions as individual wealth among people, population across cities, particle sizes, and word frequencies in a corpus [31].

While these overall distribution may remain stable over time, there can exist dynamics between items at different ranks. For example, Madison, WI was the 97th most populous U.S. city in 2000, and the 89th in 2010. And, churn is often more frequent at the tail of the distribution: New York has remained the largest American city since at least 1790 (the first US Census), but Las Vegas, NV moved from the 51st to the 31st in just 10 years. Similar dynamics are observed in word usage statistics, individual wealth, and international city sizes, for example.

Currently, there exists scant methodology to determine how much of this churn results from random fluctuations over time, or how much represents other, potentially important processes, e.g. expressing a systematic advantage of certain types of individuals over others at larger or longer scales. Was it inevitable, statistically speaking, that New York persisted at the top of the distribution of US cities for more than two centuries? Was there something about Las Vegas that let it rise so quickly, or could any city have done the same? We lack the statistical machinery to connect individual and population level rates of change.

This thesis proposes a structured way to look at rates of change in human systems. Each person is a collection of his behaviors, and how quickly he explores or innovates impacts how quickly he changes. Similarly, growth in a city's economic structure is governed by a few common rules. In both individuals and populations, it's important to distinguish the random from the real: a statistical test helps us do so.

**Contributions**

This thesis makes the following contributions:

1. Two new measures of human mobility behavior: exploration and turnover. Exploration is the rate at which new locations are searched by an individual, and turnover is the rate at which his portfolio of visited locations changes. Contrary to expectation, exploration is open-ended for almost all individuals.

2. A baseline model for change or churn in human systems, which relates the rate of exploration to the rate of turnover. With better predictions about rates of change, we can better predict, if they will meet, and quickly the system will evolve over time.

3. This model modifies the neutral or random drift explanation for population-level behavior with an explicit weighting for "habits" that explains observed patterns.

4. The definition of two motivations for of Exploration, $O$ or exploration due to optimization, and $W$ or exploration due to a taste for variety. These can be decomposed,

and $W$ relates to individual resources or income.

5. An analysis of growth and structure of city economic features, finding analogies to growth in ecosystems

6. A new statistical test for distinguishing random from directed churn in rank ordered systems.

At a broader level, the findings in this thesis contribute to a new way of thinking about how behaviors move through time. Individuals and populations have a characteristic rate of change, which can be related to inputs of search or innovation, and which can be systematically differentiated from random effects of sampling.

**Organization**

The following section outlines background literature, the two principle datasets used for analysis, and two studies to introduce, first, the idea of predictability in a particular human system and second, the dynamics that can emerge at the interface of individual- and population- level behavior.

Section 3 introduces the idea of exploration and turnover, and presents a model that describes the relationship between these two properties of individual behavior as a function of individual resources.

Section 4 explores the structure of urban economies and the ways in which they change over time.

Section 5 presents a statistical test to distinguish exogenously-driven churn from the null model of sampling-driven churn.

The thesis concludes with a summary of results, contributions, and suggestions for future work. Finally, there is an appendix and a list of references.

**Introduction**
Data and Background
Individual rates of change as a function of exploration and resources
Rates of change in the structure of urban economies
A statistical test for churn in distributions of ranked items
Discussion
Appendix
Bibliography

## 2. Data and background

The previous section introduced the idea of measuring rates of change in individuals and populations. In this section, the two data main datasets are presented, as well as background analysis on individual predictability. We look at how individual decisions aggregate to population-level urban features, and briefly describe a model for how individual choices might produce the complex outcomes seen in an online market. This latter study is described in more detail in the appendix, as an example of connecting individual- and population- level rates of change in search and behavior.


Introduction
**Data and Background**
- **Financial institution data**
- **US census data**
- **Predictability in human systems**
- **Connecting the individual and population levels**

Individual rates of change as a function of exploration and resources
Rates of change in the structure of urban economies
A statistical test for churn in distributions of ranked items
Discussion
Appendix
Bibliography


#### Financial institution data

We use a dataset of some 80m de-identified credit and debit card accounts, including continuous purchases over a time period of 5+ years.

The analyses in this thesis draw from a relevant sub-sample of transaction records drawn from the database of a major consumer bank. Activity is available dating to 2005, and includes information on transaction date, amount, channel (e.g. check, debit, credit), merchant, merchant category code (MCC; described below), and whether the transaction took place on- or offline. Customers are identified by zip code, join date, and year of birth, and are associated with any linked (e.g. joint) accounts.

Transactions total about 30-35 billion per month and thus can represent significant

flows in the US economy. For the metropolitan areas we consider, a range of 28% to 79% of residents hold accounts with this financial institution.

Individual income can be inferred using inflows to an account. To prevent returned purchases and other debits from being counted as income, only those inflows coming tagged with identifiers for employer direct deposit, annuity or disability payments, and Social Security income are considered.

"Income", or regular deposits into an account, actually captures a reasonable lower bound on true income. It is possible that not all of an individual's true income is captured by our measure: for example, if a person's earnings are primarily in the form of cash or personal check, or if he deposits only a portion of his salary into his account with this Bank, and routes the remainder to a retirement or stock market account, a spouse's account, or a personal account at a separate bank.The effect is stronger for wealthier individuals, who tend to have multiple accounts and are generally more sophisticated financially. Therefore we expect these estimates to exhibit amplified dampening as income rises. For the analysis in Section 3, individuals are divided into observed income quartiles, with the lowest being <$15,000 and the highest >$70,000 per year.

Other sources of sampling bias arise from the 10 million American households (the "unbanked") without bank accounts. This absent slice tends to include recent immigrants to the United States as well as residents of very rural areas and urban centers. The present sample comes from a bank with relatively even distribution across all other income categories, although the wealthiest American consumers tend to be underrepresented by this financial institution. We also expect that in filtering for accounts with electronic inflows (of any amount), we are biasing our sample against individuals who are paid exclusively in cash.

Information on merchants is provided in the form of a string, which often includes store name and (in the case of chain retailers) number, and occasionally information on location. Some of these strings have been hand-coded and standardized: the aggregate business name is also listed in a separate column.

The MCC codes established by MasterCard and Visa are used to categorize merchants. The distribution of codes is heavily skewed in three categories - there are about 150 codes for individual airlines and 200 for individual hotels, for example - to remedy this we create three new aggregate categories to comprise (1) all airlines, (2) all hotels, and (3) all rental car purchases.

**US census data**

The US Economic Census [11] is a detailed report compiled every 5 years by the Census Department, and includes information on the number of businesses, categorized by industrial sector and sub-sectors. The classification hierarchy is based on the North American Industrial Classification System, or NAICS. Data are used from the 1997, 2002 and 2007 censuses, for some 942 metropolitan areas in the United States (by

census-defined Metropolitan or Micropolitan Statistical Areas, together called Core-Based Statistical Areas or CBSAs).

The NAICS classification system is organized hierarchically with six levels of structure, with more digits representing increasing resolution. At the coarsest, 2-digit level, codes describe industrial sectors such as Manufacturing, Transportation, Finance, or Education. The codes are increasingly refined at the 3-, 4-, 5-, and 6- digit levels. For example, one branch is:

45 Retail Trade
451 Sporting Goods, Hobby, Book, and Music Stores
4512 Book, Periodical, and Music Stores
45121 Book Stores and News Dealers
451212 News Dealers and Newsstands

From this tree, we extract snapshots for 942 individual urban areas in the United States, using the U.S. Census-defined Metropolitan Statistical Areas (MSAs). Each snapshot represents a particular expression of the tree, in which only a portion of possible firm types are present. If a firm type has at least one establishment, it is expressed in that city. We also collect information about the number of individual establishments present for each firm type, and the population of each MSA.

In the same way that biological classification systems are subject to the definition of a species and differential ability to find and identify various creatures [13], the NAICS classification is an artifact of its human classifiers. We require, however, for sufficient resolution in codes to ensure that tree expression is not limited by city size. We check for saturation of codes in the largest city, New York, and find, reassuringly, only 90% of types expressed. Others have show via expansion of the codes that true diversity can be estimated, and is constrained by factors other than the limitations of the coding system.

**Predictability in human systems**

Above, two main datasets are summarized. Next we look at basic features of the first dataset of financial transaction time series, and an analysis of the predictability of individual-level patterns


Introduction
Data and Background
- Financial institution data
- US census data
**- Predictability in human systems**
- Connecting the individual and population levels
Individual rates of change as a function of exploration and resources
Rates of change in the structure of urban economies
A statistical test for churn in distributions of ranked items
Discussion
Appendix
Bibliography


We might use the bank dataset to ask: how predictable are individuals in their daily and weekly mobility patterns? Treating as a sensor network, where each swipe is considered a node. We find that individual mobility patterns can be quantified, to a point. The following section details collaborative work.

Economic models of consumption incorporate constraint and choice to varying degrees. In part, shopping is driven by basic needs and constraints, with demand shaped by price, information and accessibility [44, 6]. At the same time, it is believed that shoppers will opt for greater variety if possible [29], although empirical work finds that behaviors such as choice aversion [26] and brand-loyalty can limit search [27, 43, 8]. How do choice and constraint connect? Investigations with mobile phone data find that individual trajectories are largely predictable [10, 20, 2, 36]. Yet these models say little about the motivation for movement. At the same time, models of small-scale decision-making [46, 60, 1, 42] leave open the question of how individual heuristics might form large-scale patterns.

Here, we draw on a unique set of individual shopping data, with tens of thousands of individual time series representing a set of uniquely identified merchant locations, to examine how choice and necessity determine the predictability of human behavior. Data from a wealth of sensors might be captured at some arbitrary waypoint in an individual's daily trajectory, but a store is a destination, and ultimately, a nexus for human social and economic activity.

We use time series of de-identified credit card accounts from two major financial institutions, one of them North American and the other European. Each account corresponds to a single individual's chronologically ordered time series of purchases, reveal-

ing not only how much money he spends, but how he allocates his time across multiple merchants. In the first case we represent purchases made by over 50 million accounts over a 6-month window in 2010-2011; in the second, 4 million accounts in an 11-month window. Data from transactions included timestamps with down-to-the-second resolution.

We filter each sample to best capture actual shoppers' accounts, to have sufficient data to train the Markov models with time series that span the entire time window, and to exclude corporate or infrequently used cards. We filter for time series in which the shopper visits at least 10 but no more than 50 unique stores in every month, and makes at least 50 but no more than 120 purchases per month. We test the robustness of this filter by comparing to a set of time series with an average of only one transaction per day (a much less restrictive filter), and find similar distributions of entropy for both filters.

The median and 25th/75th percentile merchants per customer in the filtered time series are 64, 46, and 87 in the North American (6 months) and 101, 69 and 131 in the European (11 months) dataset.

To quantify the predictability of shopping patterns, we compare individuals using two measures. First, we consider static predictability, using temporally-uncorrelated (TU) entropy to theoretically bound, and a frequentist model to predict where a person will be. Second, we consider a person's dynamics, by taking into account the sequence in which he visits stores. Here we use an estimate of sequence-dependent (SD) entropy to measure, and a set of Markov Chain models to predict location. Both entropy measures, and predictive model, are defined fully below.

At longer time scales, shopping behavior is constrained by some of the same features that have been seen to govern human mobility patterns. We find that despite varied individual preferences, shoppers are on the whole very similar in their overall statistical patterns, and return to stores with remarkable regularity: a Zipf's law $P(r) \sim r^{-\alpha}$ (with exponent $\alpha$ equal to 0.80 and 1.13 for the North American and European datasets respectively) describes the frequency with which a customer visits a store at rank $r$ (where $r = 3$ is his third most-frequented store, for example), independent of the total number of stores visited in a three-month period, see Figure 2.1. This holds true despite cultural differences between the North American and Europe in consumption patterns and the use of credit cards. While our main focus is not the defense of any particular functional form or generative model of visitation patterns, our results support those of other studies showing the (power law) distribution of human and animal visitation to a set of sites [59, 4, 23, 15].

A universal measure of individual predictability would be useful in quantifying the relative regularity of a shopper. How much information is in a shopper's time series of consecutive stores?

Informational entropy [52] is commonly used to characterize the overall predictability of a system from which we have a time series of observations. It has also been used to show similarities and differences across individuals in a population [16].

We consider two measures of entropy:

(i) The temporally-uncorrelated (TU) entropy for any individual $i$ is equal to $S^{\alpha}_{TU} = -\sum_{i \in M_\alpha} p_{\alpha,i} \log(p_{\alpha,i})$ where $p_{\alpha,i}$ is the probability that user $\alpha$ visited location $i$. Note this measure is computed using only visitation frequencies, neglecting the specific ordering of these visits.

(ii) The sequence-dependent (SD) entropy, which incorporates compressibility of the sequence of stores visited, is calculated using the Kolmogorov complexity estimate [33, 35].

Kolmogorov entropy is a measure of the quantity of information needed to compress a given time series by coding its component subchains. For instance, if a subchain appears several times within the series, it can be coded with the same symbol. The more repeated subchains exist, the less information is need to encode the series.

One of the most widely used methods to estimate Kolmogorov entropy is the Lempel-Ziv algorithm [33, 35], which measures SD entropy as:

$$S^{\alpha}_{SD} \approx \frac{\log N}{\langle L(w) \rangle} \tag{2.1}$$

where $\langle L(w) \rangle$ is the average over the lengths of the encoded words.

We can apply the algorithm to observed transitions between locations. A person with a smaller SD entropy is considered more predictable, as he is more constrained to the same sub-paths in the same order.

We find a narrow distribution of TU and SD entropies across each population.

Another dataset, using cell phone traces [55], also finds a narrow distribution of entropies. This is not surprising, given the similarity of the two measures of individual trajectories across space. Yet we find a striking difference between the credit card and the cell phone data. In the shopping data, adding the sequence of stores (to obtain the SD entropy) has only a minor effect of the distribution, suggesting that individual choices are dynamic at the daily or weekly level. By contrast, cell phone data shows a larger difference. Why does this discrepancy occur? A possible explanation is that shoppers spread their visitation patterns more evenly across multiple locations than do callers. Even though visitation patterns from callers and from shoppers follow a Zipf's law, callers are more likely to be found at a few most visited locations than are shoppers. This is true, but to a point. Consumers visit their single top location approximately 13% (North American) and 22% (European) of the time, while data from callers indicates more frequent visitation to top location. Yet shoppers' patterns follow the same Zipf distribution seen in the cell phone data, and the narrow distribution of temporally-uncorrelated entropy indicates that shoppers are relatively homogenous in their behaviors.

An alternative explanation for our observed closeness of temporally-uncorrelated and sequence-dependent entropy distributions is the presence of small-scale interleaving and a dependence on temporal measurement. Over the course of a week a shopper

might go first to the supermarket and then the post office, but he could just as well reverse this order. The ability to compare individuals is thus limited by the choice of an appropriate level of temporal resolution (not necessarily the same for each dataset) to sample the time series. With the large-scale mobility patterns inferred from cell phones, an individual is unable to change many routines: he drives to the office after dropping off the kids at school, while vice versa would not be possible. In the more finite world of merchants and credit card swipes, there is space for routines to vary slightly over the course of a day or week.

To test the extent to which the second hypothesis explains the discrepancy between shoppers and callers, we simulate the effect of novel orderings by randomizing shopping sequence within a 24-hour period, for every day in our sample, and find little change in the measure of SD entropy. In other words, the re-ordering of shops on a daily basis does little to increase the predictability of shoppers, likely because the common instances of order swapping (e.g. coffee before rather than after lunch) are already represented in the data. We then increase the sorting window from a single day to two days, to three days, and so forth.

Yet when we sort the order of shops visited over weekly intervals, thus imposing artificial regularity on shopping sequence, the true entropy is reduced significantly. If we order over a sufficiently long time period, we approach the values seen in mobile phone data. Thus entropy is a sampling-dependent measure which changes for an individual across time, depending on the chosen window. While consumers' patterns converge to very regular distributions over the long term, at the small scale shoppers are continually innovating by creating new paths between stores.

In order to measure the predictability of an individual's sequence of visits, we train a set of first order Markov chain models. These models are based on the transition probabilities between different states, with the order of stores partially summarized in the first-order transition matrix. It is thus related to the SD entropy measure.

While we could use a number of more complicated models (including higher order Markov models), our goal here is not to optimize for predictive accuracy, but rather to show the degree to when the sequence of stores matters to an individual's predictability.

Markov chains are used to model temporal stochastic processes, in which the present state depends only on the previous one(s). Mathematically, let $X_t$ be a sequence of random variables such that

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, X_{t-3} = x_{t-3}, ...) = P(X_t = x_t | X_{t-1}) \qquad (2.2)$$

then $\{X_t\}$ is said to be a Markov process of first order. This process is summarized with transition matrix $P = (p_{ij})$ where $p_{ij} = P(X_t = x_j | X_{t-1} = X_i)$. Markov chains can be considered an extension of a simple frequentist model in which

$$P(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, X_{t-3} = x_{t-3}, ...) = P(X_t = x_t) \qquad (2.3)$$

applied on every observed state.

If the present transaction location depends in some part on the previous one, a 1st order Markov model would be able to predict the location with greater accuracy than a

simple frequentist model. We measure the probability of being at store $x$ at time $t+1$ as $Pr(X_{t+1} = x|X_t = x_t)$ and compare the prediction values to the observed values for each individual. We build several models, varying the range of training data from 1 to 6 months of data for each individual, and compare the model output to test data range of 1 to 4 subsequent months.

We additionally compare the results of the Markov models to the simplest naive model, in which the expectation is that an individual will chose his next store based on his distribution of visitation patterns, e.g. he will always go to one of the top two stores he visited most frequently in the training window (recall that for most people this store visitation frequency is on average just 20-35%). Since this is a simply frequentist approach to the next-place prediction problem, it is strongly related to TU entropy which is computed using the probability that a consumer visits a set of stores.

Comparing the match between model and observed data, we find that using additional months of training does not produce significantly better results. Moreover, results show some seasonal dependency (summertime and December have lower prediction accuracy, for example). For fewer than three months of training data, the frequentist model does significantly better than the Markov model. This suggests the existence of a slow rate of environmental change or exploration that would slowly undermine the model's accuracy.

For each of the two (EU and NA) populations, we next test a global Markov model, in which all consumers' transition probabilities are aggregated to train the model. We find that such a model produces slightly better accuracy that either the naive or the individual-based models (with accuracy $\approx 25-27\%$). To test the sensitivity of this result we take ten global Markov models trained with 5% of time series, selected randomly. We find the standard deviation of the accuracy on these ten models increases to 3.6% (from 0.3% using all data), with similar mean accuracy. Thus the global Markov model depends on the sample of individuals chosen (for example, a city of connected individuals versus individuals chosen from 100 random small towns all over the world), but does in some cases add predictive power.

As previous work has indicated [12], mobility patterns can be predicted with greater accuracy if we consider the traces of individuals with related behaviors. In our case, even though we have no information about the social network of the customers, we can set a relationship between two people by analyzing the shared merchants they frequent. The global Markov model adds information about the plausible space of merchants that an individual can reach, by analyzing the transitions of other customers that have visited the same places, thus assigning a non-zero probability to places that might next be visited by a customer.

Yet in almost every case, we find that people are in fact less predictable that a model based exclusively on their past behavior, or even that of their peers, would predict. In other words, people continue to innovate in the trajectories they elect between stores, above and beyond what a simple rate of new store exploration would predict.

Colloquially, an unpredictable person can exhibit one of several patterns: he may be hard to pin down, reliably late, or merely spontaneous. As a more formal measure for human behavior, however, information-theoretic entropy conflates several of these

notions. A person who discovers new shops and impulsively swipes his card presents a different case than the one who routinely distributes his purchases between his five favorite shops, yet both time series show a high TU entropy. Similarly, an estimate of the SD entropy can conflate a person who has high regularity at one level of resolution (for example, on a weekly basis) with one who is predictable at another.

As example, take person A, who has the same schedule every week, going grocery shopping Monday evening and buying gas Friday morning. The only variation in A's routine is that he eats lunch at a different restaurant every day. On the other hand, person B sometimes buys groceries on Tuesdays, and sometimes on Sundays, and sometimes goes two weeks without a trip to the grocer. But every day, he goes to the local deli for lunch, after which he buys a coffee at the cafe next door. These individuals are predictable at different time-scales, but a global measure of entropy might confuse them as equally routine.

Entropy remains a useful metric for comparisons between individuals and datasets (such as in the present and cited studies), but further work is needed to tease out the correlates of predictability using measures aligned with observed behavior. Because of its dependence on sampling window and time intervals, we argue for moving beyond entropy as a measure of universal or even of relative predictability. As our results suggest, models using entropy to measure predictability are not appropriate for the small scale, that is, their individual patterns of consumption.

Shopping is the expression of both choice and necessity: we buy for fun and for fuel. The element of choice reduces an individual's predictability. In examining the solitary footprints that together comprise the invisible hand, we find that shopping is a highly predictable behavior at longer time scales. However, there exists substantial unpredictability in the sequence of shopping events over short and long time scales. We show that under certain conditions, even perfect observation of an individual's transition probabilities does no better than the simplistic assumption that he will go where he goes most often.

These findings suggest that individual patterns can be bounded, but full predictability is elusive. Moreover, because we're unable to predict where someone will go next, the important question is how quickly this baseline will change. This has important implications for how we think about traffic or consumption or disease modeling.

The difference between this analysis and the main thesis results is that here, we've taken behavior as a static snapshot and asked: how do people compare in their predictability? How much are transition chains repeated? We see a slight decline in predictability as we try to predict months further into the future. In section 3, we will quantify how this predictability is expected to decrease over time, and why.

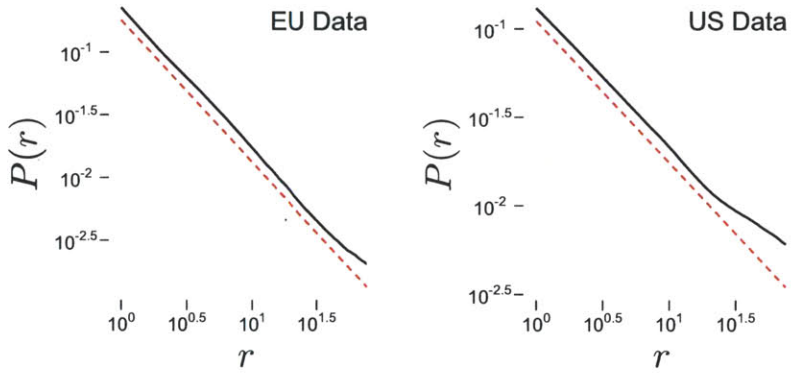



Figure 2.1: Probability of visiting a merchant, as a function of merchant visit rank, aggregated across all individuals. Dashed line corresponds to power law fits $P(r) \sim r^{-\alpha}$ to the initial part of the probability distribution with $\alpha = 1.13$ for the European and $\alpha = 0.80$ for the American database.

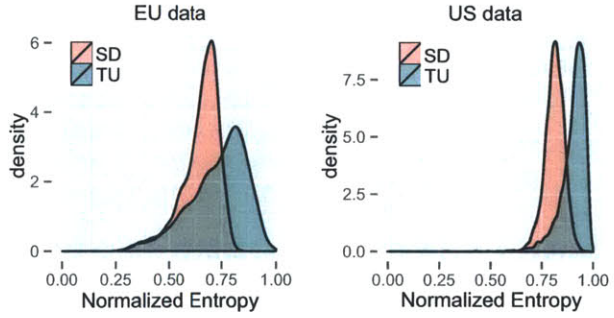



Figure 2.2: Entropy distributions for the American and European time series. Temporally-uncorrelated entropy distributions are slightly higher for both populations.

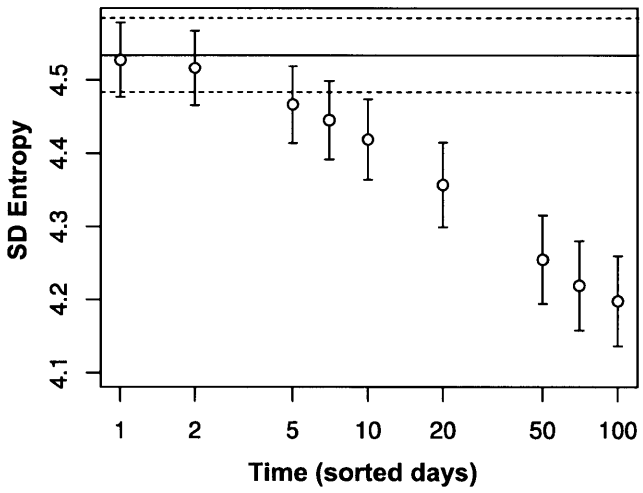



Figure 2.3: Sequence-dependent entropy for a number of artificially sorted sequences. For each window size over which the time series is sorted, we measure the sequence-dependent entropy for the population.

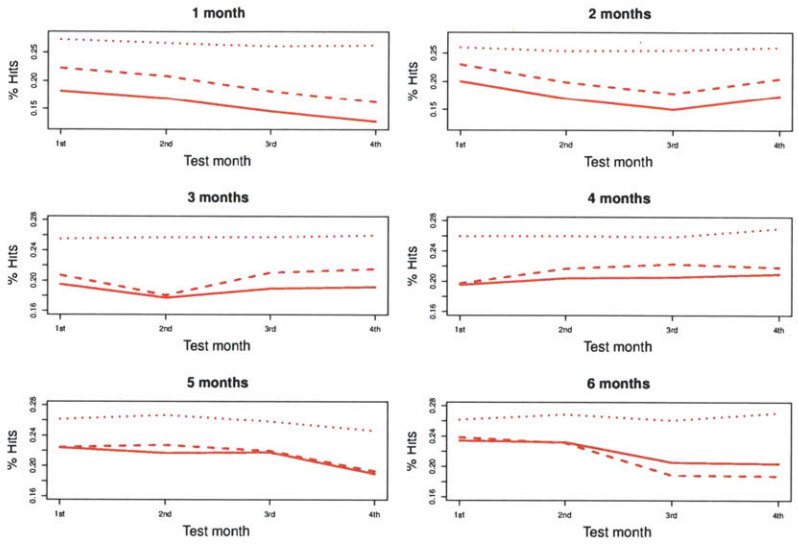



Figure 2.4: Markov model results for various temporal windows in training and test. The solid red line indicates hit percentage for Markov model, dashed line exhibits accuracy for the naive model and the dotted line indicates results for the Global Markov model.

**Connecting the individual and population levels**

The previous sections described the datasets used, and presented analyses of the financial institution dataset. Here, we take a small detour to look at one way in which individual decisions and population-level outcomes might connect. This model draws on a dataset of some 14,000 online decisions, in an online marketplace for downloading music. Using a relatively simple model of the feedback on decisions generated by social influence, it is possible to model the seemingly-complex outcomes of this market. In this section, we briefly explore the implications of the result. The full results and discussion are detailed in the appendix.


Introduction
Data and Background
- Financial institution data
- US census data
- Predictability in human systems
- Individual rates of change aggregated at the city level
**- Connecting the individual and population levels**

Individual rates of change as a function of exploration and resources
Rates of change in the structure of urban economies
A statistical test for churn in distributions of ranked items
Discussion
Appendix
Bibliography


There is an analogy between how we think of rate of change in individual and in populations or systems. In an individual "system", a person has a number of locations (types) he frequents. To each location, he assigns a particular number of visits (items): the overall result is a distribution of visitations. In a population, for example in a set of cities (types), individuals (items) are assigned to each place, according to a distribution. The most populous city can change from year to year, just as a person's most frequented store can change from month to month.

Yet there can be important feedback effects that influence how individual rates of change convert to population level rates of change. For example, social influence may impact the rate of change of a person's top location. We examine the impact of this kind of social influence in going from individual preferences to a system which changes dynamically. Here, as with most systems, it's not merely a matter of "scaling up" individual choices: rather, feedback effects propel the market to unexpected outcomes. Surprisingly, we find that even such complex dynamics can be modeled with a few simple rules.

A recent experimental study [48] found that the addition of social influence to a cultural market increased the unpredictability as well as the inequality of the market share of individual products. However, the study did not propose a model to describe how

such social forces might operate. Here, we present a parsimonious model that sheds light on social behavior in this market. Our model does not rely on assumptions about heterogeneous preferences [54] or solely on the generic notion of herd behavior [7] to explain the outcomes of anonymous social influence: rather, we treat social influence as a variable whose effect grows as the market matures.

In this research we consider the static picture of how individual decisions lead to population-level features. These results show that simple dynamics -namely, a two-step decision process in which social influence in material to the first step only- can model observed data well. An important question for future work is how these connect when we look at the -dynamics- of human behavior, that is, the rates of change of individual behavior to the rate of change the system as a whole.

In this background section, we have considered the bounds of predictability of individuals as well as the static connection between individual and population rates of change. In the next section, we will consider how an individuals' portfolio of behaviors changes slowly over time.

## 3. Individual rates of change as a function of exploration and resources


Introduction
Data and Background
**Individual rates of change as a function of exploration and resources**
- **Introduction**
- **Exploration is open-ended**
- **Rates of change as a function of exploration**
- **Exploration as a function of financial resources**
- **Discussion**

Rates of change in the structure of urban economies
A statistical test for churn in distributions of ranked items
Discussion
Appendix
Bibliography


It is a common lament that a person's behavior is "so predictable"; at the same time, humans are a notably inventive species. Although rates of change in behaviors such as consumption, mobility, and migration have undeniable effect on a number of socio-technical systems, there has been scant focus on modeling and measuring such rates. In this section we study tens of millions of credit card transactions, representing real-time decisions on where to go and what to buy. We find that rates of change, or churn, are largely predicted by an individual's rate of sampling new locations, paired with random copying of his past behaviors. However, for less habitual locations, the effect on churn of additional exploration saturates, suggesting distinct motivations for search combine to effect rates of change. We present a model in which a person explores (i) to optimize his routines and (ii) out of a taste for novelty, and which predicts observed individual rates of change. In addition, we find that taste for novelty increases with financial resources, paralleling adaptations in animal search patterns to allocations of resources in the environment.

### Introduction

Despite the sensitivity of many complex systems, such as financial markets and disease dynamics, to continuous human influence, we lack good models for rates of change in human behavior. On the one hand, economic theory holds that individuals optimize

over a set of preferences and constraints: rates of behavior change are limited by time scale at which constraints change and force reevaluation of where to go. Although behavioral economics has highlighted decision-making biases [19], at root, utility maximization as a motivation for search is unable to account for choices that don't contribute to local optimization.

On the other hand, theories of learning and habit predict that individuals will continue to change over time, based on sampling new behaviors and selectively copying past ones [45, 3]. William James likens habit to a current, which once it has "traversed a path, it should traverse it more readily still a second time" [28], and more modern neuroscience has confirmed that behavior catalyzed by an arbitrary first step and then reinforced becomes more automatic [21]. Animals form foraging routines as a response to constraints in the environment , often retracing paths even when a superior option becomes available [59]. While such theories account for observed rates of change, they are unable to explain why we search in the first place.

Elective human behavior appears a combination of these two motivations: to optimize for existing needs, and to try new things . The result is that individuals change perceptibly over time, with important impacts on economic and social systems: a person may switch his favorite lunch spot, find a different accountant, move to a new city, upgrade his car, marry and divorce.

To understand the how search (measured as the rate of sampling new locations, and dually motivated by optimization and novelty-seeking) affects rates of change in individuals (measured as turnover in an individual's portfolio of store visits), we explore one of the richest time series of economic decision-making studied to date, representing hundreds of millions of economic decisions in the wild.

We first show that consumers' rate of search is open-ended: that is, it does not taper off with time, as economic theory would predict. We then present a model for how people change, as a function of this open-ended search, coupled with copying one's own past choices. A simple process of random drift predicts how often an individual will switch out his most habitually-visited stores (the 1-2 locations he visits most often). For all but the most frequented stores, however, this simple model is insufficient. Instead, the effect on turnover rates saturates after a given level of search, with additional search contributing only marginally to rates of change.

Although this saturation effect varies as a function of how frequently a store is visited (or how "habitual" it is), we show that the relationship between search and turnover can be modeled with a simple adaptation of the random drift mechanism. Moreover, individual resources, measured in terms of income, predict individual rates of search, allowing us to decompose the effect of search or rate of change into two constituents. Search is motivated in part by the optimization of current locations, which does not depend on income, and in part by a taste for novelty, which is facilitated by greater income.

Our aim is to model rates of change as function of search patterns. How does search - that is, the exploration of new locations - vary across the population? We rank all stores that an individual visits in a time period $t$ by the number of visits $f_n$, and define the basket depth $n$ as the rank of a store based on its frequency. Thus the most frequented

store has depth $n = 1$, the second most frequented $n = 2$, and so forth.

We find that although each person develops a unique portfolio of stores to visits, rates $f_n$ are remarkably similar across the population, especially for low $n$. And the most frequented (top 1-3) stores tend to represent only a handful of categories, such as grocery stores and gasoline stations. Rates of visitation follow a Zipf distribution for almost all individuals, while rates of search (unique stores in a time window $t$) vary considerably across the population, from close to 0 to almost 90% of total visits.

In other words, in this a static picture of behavior where the identities of stores are stripped, individuals behave remarkably homogeneously. Below, we seek to enrich this static picture in three ways:

- First, we'd like to understand how rate of search $E$ varies with time window $t$. That is, does E stay constant between multiple windows, or taper off?
- Second, we'd like to relate the rate of an individual's search $E$ to his rate of change $T$ for a given basket depth $n$.
- Finally, we'd like to understand the relationship between rate of search $E$ and individual resources or income.

**Exploration is open-ended**

We define a person's exploration rate $E_i$ as the number of unique stores $U_i$ over number of total visits $V_i$ within a time window $t$:

$$E_i = \frac{U_{i,t}}{V_{i,t}}$$

If a person visits stores a single time, $U_{i,t} = V_{i,t}$ and $E_i = 1$. Conversely if he visits one store all of the time, his rate of exploration will be low, $E_i = 1/V_{i,t}$

Turnover $T_{n,b}$ is the rate of change of a basket $b$ of stores, at basket depth $n$. We measure $T$ by looking at visitation pattern to the store at depth $n$ in two consecutive time windows $t-1$ and $t$, where a change occurs if the store moves to depth $> n$, that is, it becomes less frequent:

$$T_{n,b} = \frac{1}{t_{max}} \Sigma_t d(1 - (B_{n,t-1} \bigcap B_{n,t})$$

Here $b$ is the individual's total basket size (number of total stores $n_{max}$ visited in $t$), $B_{n,t}$ is the portfolio of stores visited in time window $t$ at or above rank $n$ (having frequency $> f_n$) and $d$ is the percentage of the basket explained, or $\frac{n}{b_t}$.

In the data we find a wide range of exploration rates $E$ across the population. Is exploration rate constant over successive time windows for a single person? If the role

of exploration is to help optimize a portfolio of stores, we would expect exploration to decline for any individual over time as he converges on a set of merchants that fits his needs. Likewise, we'd expect rates of change $T$ to be very low, representing only big breaks when an individual changes all of his merchants, for example after a move to a different city.

To the contrary, we find that $E_i$, or $\frac{U_{i,t}}{V_{i,t}}$ is a linear function with $t$ for almost all individuals: in other words, exploration is open and not closed. That $E$ does not vary with $t$ suggests that consumers are continually searching, although each consumer explores at his own characteristic pace. We can thus fix $t$ for further analysis ($t = 30 days$).

Why would the rate of search be open-ended? Exploration may help to optimize a portfolio of stores. In this case, an individual who explores more will presumably be better able to optimize his stores, and rate of search will correlate with rate of churn.

Or, exploration may simply be "discretionary": more exploration will not carry over to the next time period to change the stores that show up in the top ranks, which are visited most frequently.

We call the first motivation for exploration, related to optimization, $O_i$, and the second, related to a taste for variety, $W_i$. These motivations contribute in some proportion to exploration for every individual, and the first has an effect on T, or rate of change. Later, we will specify this relationship and show that $W$, but not $O$, is a function of individual resources.

**Rates of change as a function of exploration**

We now ask how $E$ is related to $T$. We have seen that exploration is open-ended, and also that an individual's total number of visits in time $t$, $V_{i,t}$ varies little across time, and we see a Zipf-like distribution of frequencies assuming sufficiently wide time windows ($t >= 20$).

To what extent do the rates of change $T$ vary across the population? Surprisingly, we find that even the most frequented store at depth $n = 1$ turns over with some probability ($\mu = 0.042$) or approximately one of twenty people's top store will be dropped from one month to the next). The likelihood that a less frequent store $n > 1$ will turn over is even greater.

We can now begin to specify our model for the effect of exploration rate $E$ on turnover rate $T$. Each individual is characterized by his rate of exploration $E_i$. In every time window $t$ he has a certain number of visit slots $V$ to fill. Based on his exploration rate $E_i$, he chooses to fill $V_i$ with $U$ unique stores, and then fills the remainder $V - U$ visits across the stores with a previous visit from period $t - 1$ , selected randomly, in proportion to visitation frequency in $t - 1$.

$E_i = \frac{U_i}{V_i}$

The new stores visited in $t$ convert with some probability to stores he will visit more often, and thus to his turnover $T_i$. For example, a consumer may visit a new grocery store "Grover's" once in time $t-1$ and find it better than his existing grocer, and then return to Grover's sufficient times in the next period $t$ for it to become his 4th most frequented store. However, he may also try a restaurant "Randy's", to which he never returns again.

In order to parameterize this model, we measure how search $E$ is related to rate of change $T$. We start with basket depth $n = 1$, or the rate of change for the single most-frequented store. Surprisingly, a linear relationship (slope = 0.11, $R^2 = 0.97$) explains the population-wide link between $E$ and $T$ remarkably well. That is, the more a person engages in search as a percentage of his total visits, his rate of change increases at a regular rate. If a person has 50% exploration rate (somewhat higher than average), his rate of change in the top store will be about 6% per month (also about 40% greater than average).

We run simulations of the model of model of innovation and copying described above, and find remarkable concord with observed values of E and T in this population. Random drift-type models, such as this one in which innovation, mutation or exploration drives churn, have been proposed to explain population level dynamics, but never for the individual level. It has been argued, in fact, that an individual doesn't simple copy himself, but instead performs a biased random walk []. Here, surprisingly, we show this accepted wisdom to be false in the case of consumption: in fact, an individual has high fidelity to his past behavior, although not necessarily to a particular sequence of transitions between locations []. In the rate of churn in our most frequented store, we are purely defined by habit, with a parameter for the degree to which we explore new places (representing optimization or $O$.

We now test whether the predictions of a random drift model fit observed churn for basket depths $n > 1$. Naturally, we expect a different conversation rate (slope) for the relationship between E and T at different basket depths $n$. Visitation patterns are approximately Zipfian, which is characterized by a function with exponent $-1-(1/s)$ where $s >= 1$, we have the relative frequency $f_n$ of item at rank n versus rank 1: $\frac{f_1}{f_n} = \frac{1}{n^s}$

If the rate of change $T$ in less habitual stores $n > 1$ is affected by search $E$ in the same way as for $n = 1$, we'd expect the slope of similar linear relationship between $E$ and $T_n$ to change proportional to frequencies for $n > 1$, as exploration more easily affects stores with lower visitation frequencies.

Contrary to this expectation, for lower frequency stores, we see something quite different from the random model's prediction: a linear fit is poor. The model breaks down for higher levels of exploration, and greater exploration does not convert to churn at the same rate for high $E$. As exploration rises past a certain point, its effect on $T$ begins to saturate. This suggests that as $n \rightarrow n_{max}$, exploration is motivated by something other than simply copying for optimization of the portfolio.

Looking at the level of the individual, we see how the effect of basket size is different for a "high" vs "low" explorer. A person with high exploration sees little change between n=1 and n=10, while turnover changes significantly individuals with low ex-

ploration.

We now modify the random drift model with an additional variable $S$, describing the distribution of weights placed on sampling from last period according to frequency. By varying the range of $S$ we can recapture the relationship between $E$ and $T$ for the entire population for all basket levels $n >= 1$. We can think of $S$ as capturing the degree to which most-frequented stores are weighted against replacement in the copying stage of the model. When $W$ is stronger, taste for variety is stronger, contributing to a greater offset on search for optimization $O$. We find $S$ described by an exponential.

**Exploration as a function of financial resources**

We have shown that $T$ can be described as a function of of $E$ and $n$ with an adaptation to the random drift model. It remains to understand what factors contribute to an individual's level of exploration.

Recall that rate of search is motivated by $O$, representing search in order to optimize, and $W$, representing a fulfillment of taste for variety. Can we decompose these two motivations? Other models have proposed that the taste for variety should increase with income [] In foraging behavior, animals are observed to change patterns as response to differential resources []. In turn, we might expect the relative amounts of $O$ and $W$ to depend on resources. For each individual, we use income as a proxy for resources.

We have already shown a relationship between $E$ and damping at higher $E$'s for the entire population. We now look at individuals' relationships between $T$ and $n$. We find that as $T$ gets larger (corresponding generally to larger $E$), the slope of $T$ in $n$ decreases. This is in line with the saturation observed as $n > n_{max}$. Moreover, the relationship is linear in most cases, and we can thus decompose the elements of T into optimization-related $O$, which depends on the basket depth, and $W$ the remainder or intercept.

We first find that exploration differs significantly between highest and lowest income individuals, a signal that $E$ is related to income.

For each person, $O$ is the slope of $T$ in $n$ and corresponds to the weighting $S$ in the modified random drift model. So comparing the turnover at two points we have:

$T_{10} = 10 * O + W$ and $T_1 = O + W$,

so $O = (T_{10} - T_1)/9$ and $W = T_1 - O$

We plot $W$ as a proportion of total exploration rate $E$ and find that this corresponds closely with income ($R^2 \approx 0.5$)

Rich and poor are alike in their habits, where the dampening effect of $W$ is less pronounced. In this "habitual" regime, a random model predicts change based on exploration level. Overall, individuals with higher incomes have higher exploration rates, so

their rates of churn are also higher. However, the greatest difference between income groups in rates of churn will be expressed at low ranks, when $O$ contributes fully to exploration rate, (because $O$ contributes fully to conversion to churn, while $W$ does not).

As $n \to n_{max}$, high and low income groups maintain their different ratios of $W$ and $O$. In general, those with more resources explore more, so once they optimize their portfolio they are able to devote remaining $E$ to fulfilling taste for variety. At deeper baskets $n$, less wealthy people will be able to "express" taste for variety, and turnover will look more similar between top and bottom income quintiles

**Discussion**

To summarize, we find first that search rate $E$ is open-ended, or independent of time window $t$. Second, the relationship between rate of search $E$ and rate of change $T$ is predicted for individuals by a modified random drift model, in which the parameters are E and the weighting factor $S$. Third, we can decompose two motivations for the effect of exploration $E$ on turnover $T$, namely optimization ($O$) and taste for variety ($W$). The latter is correlated with an individual's resources or income.

To the extent that we can quantify individual rates of change, we can begin to understand how populations might churn over time, or how aggregated individual churn might make firms or populations more or less resilient. A simple model reveals that rates of change are predicted by exploration rate, which incorporates two motivations: directed search, and search for the sake of variety. The rest of the portfolio is simply described by copying one's past behaviors.

That we can compress an individual's behaviors for different basket depths, with a correction for saturation (the dampening effect of $W$), suggests a universality of habitual behavior across a single person's many movements and locations. Individuals also look surprisingly similar to one another in their habits, and are differentiated by their taste for variety, expressed in less routine behaviors. Somewhat paradoxically, it is in non-habits less frequent behaviors that the effect of income on rate of change washes out, and high and low income individuals look more alike. This is the effect of $W$ "creeping down" the saturation curve, as $O$ (which is independent of resources) has less relative importance with lower frequency behaviors. It is thus easier for a "just this once" visit to change an already unusual behavior, while this fun try has less net effect than search dedicated to improving habits. This is likely also related to non-random distribution of visits to different categories of merchants: we leave this for future work.

The economists' model of search is right, to a point: we explore in order to optimize, but we also leave the door open for non-optimizing search. Perhaps we even evolve a taste for variety not only as a way of displaying resources, but also to facilitate exploration of new landscapes that are not just for optimization of a local portfolio of behaviors [38]. The fact that some have sufficient resources to explore for fun also helps drive the growth of "non-essential" industries (which are often quite productive) in an economy [34].

In showing that the rate of change of human behavior can be described in terms of a simple model, based on search and resources, we open the door for better predictions in complex markets, disease dynamics, and modeling of mobility and migration. We've long known that people are heterogeneous in their choices: we show here that individuals are not only heterogenous, but predictable, in their rates of change.

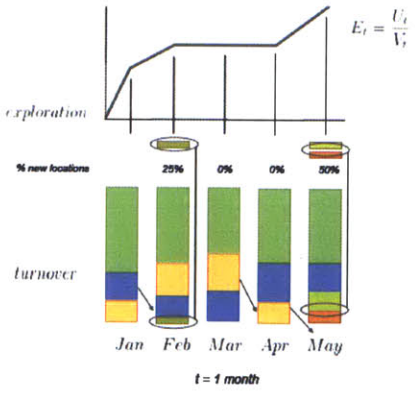



Figure 3.1: Exploration is the ratio of new store visits to total store visits in a time window $t$. Turnover is the rate at which locations in a basket of top-$n$ most frequent locations are dropped or churn

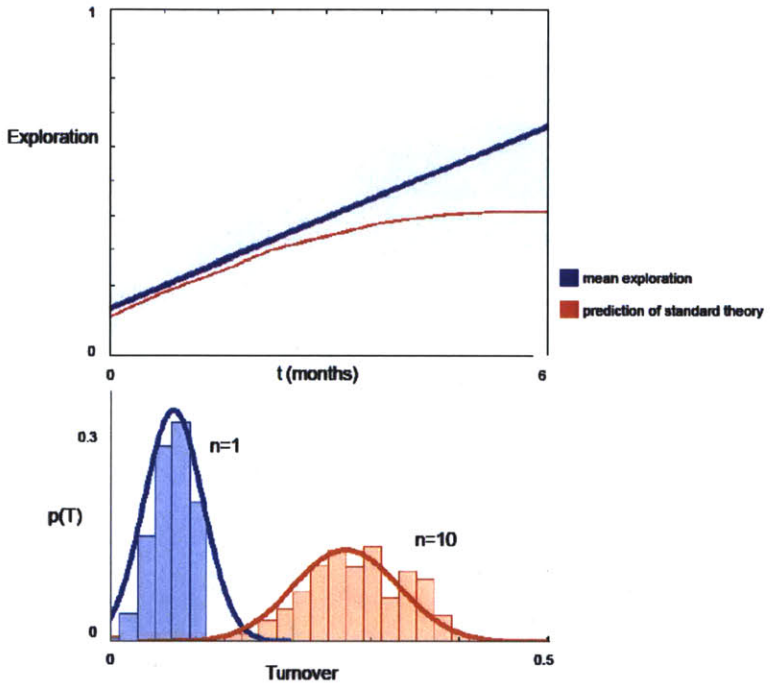



Figure 3.2: Exploration is open-ended, and rises linearly with respect to the size of the time window $t$ (top). Distribution of Turnover for basket sizes $n = 1$ and $n = 10$ (bottom)

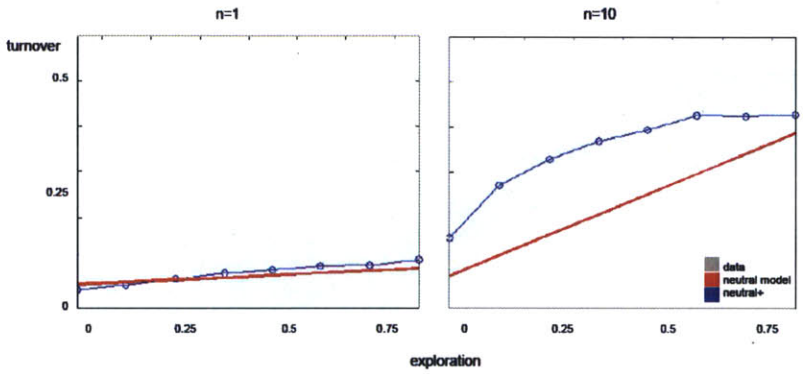



Figure 3.3: Relationship between Exploration and Turnover in the data, and predictions of the neutral and modified neutral models. The relationship is nearly linear for basket size n=1, but the effect of Exploration on Turnover saturates when n > 2.

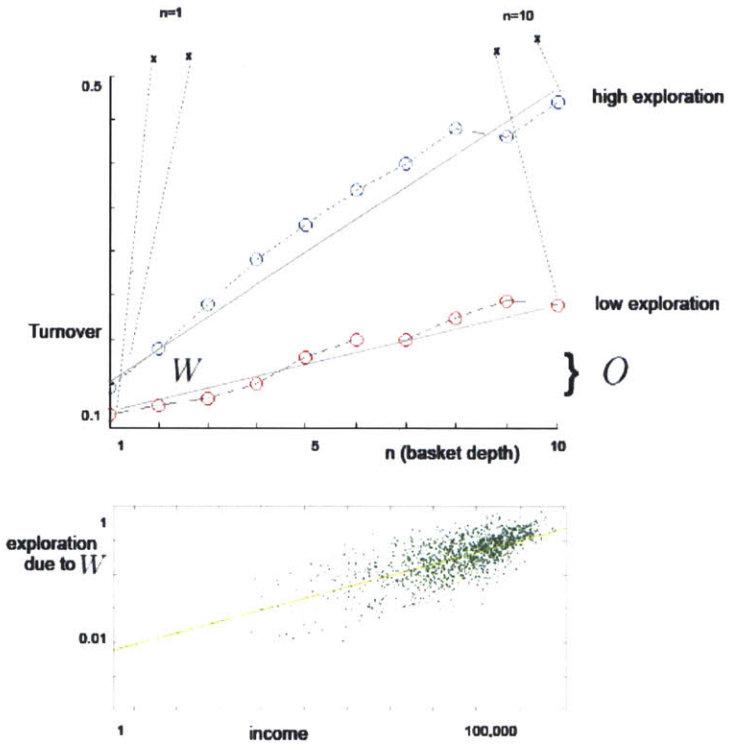



Figure 3.4: Relationship between T and n for two individuals (top). We can extract two motivations for search: optimization of a portfolio, $O$, and $W$, a taste for variety. Relationship between $W$ and income (bottom).

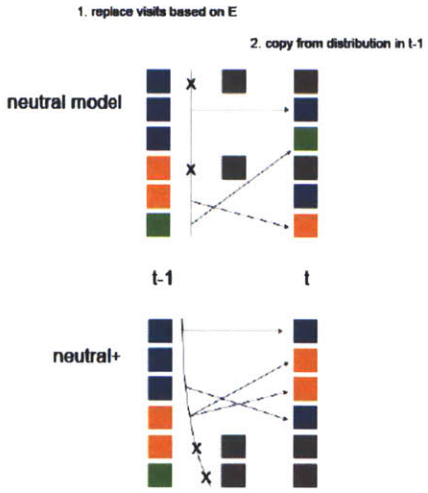



Figure 3.5: Schematic of the neutral and modified neutral models. In the neutral model, an individual replaces visits according to his exploration parameter $E$. He then fills the remaining visits with types chosen from the distribution of visits in $t-1$. In the modified neutral model, he again fills remaining visits in $t$ by choosing from a *weighted* distribution of $t-1$ visits.

## 4. Rates of change in the structure of urban economies


Introduction
Data and Background
Individual rates of change as a function of exploration and resources
**Rates of change in the structure of urban economies**
- **Introduction**
- **Measures of economic structure**
- **Results**
- **Discussion**

A statistical test for churn in distributions of ranked items
Discussion
Appendix
Bibliography


The last section presented a model to connect individual rates of search and rates of change. Here, we zoom out and look at the structure of urban economies, using the US census dataset. We examine a census of business establishments across 942 US metropolitan areas, and find striking regularities in the structure of urban economies. The hierarchical properties of a city's economic organization parallels that found in natural ecosystems, and the distribution of firms in the economy follows predictable patterns. Specifically, (i) as city size grows, the hierarchical tree of economic industries (NAICS) grows more scale-invariant, so that any sub-tree resembles the entire tree. For a given city size, there exists (ii) a relationship between the number of bifurcations and firm abundance at all tree nodes. And (iii) as city population grows, the evenness of firm abundances across nodes decreases, while the tree shape becomes more asymmetric. We present a null model for how firms might be apportioned to different types. While the relationship between species or firm evenness and tree symmetry is positive in ecosystems, (iv) it is only positive for service-based industrial sectors. These results suggest that some of the same generative processes may drive the evolution of structure in economies and ecosystems, and may distinguish characteristic growth patterns of different economic sectors.

### Introduction

The presence of economic diversity in cities is typically attributed to competitive factors such as differentiation and economies of scale [32], or to features of demand, such

as consumer preferences and taste for variety [29]. Yet many of these explanations rely on counts and comparisons of city features, ignoring the effect of economic structure on patterns of diversity therein.

Economies have long been compared to ecosystems [40, 47, 39, 22], and several centuries of ecological research have highlighted simple rules that cut across the tangle of grasses, beasts, and grub. Herrada et al [25] find universal scaling laws in the tree of life, with a similar tree shape present at various levels of resolution. Sugihara et al [56] suggest a relationship between the number of bifurcations in a phylogenetic tree and species abundance. As a tree becomes more symmetric, the evenness of species increases.

Can we derive a coordinating set of rules for urban ecosystems? The structure of multiple interacting types (such as species or firm categories) can reveal how an ecosystem or economy might have developed, or why demand is expressed in a certain way. We might also like to understand how large and small urban economies differ in their growth patterns, and whether there are regularities in the economic tree that cut across cities of different sizes and histories.

Here, we use dataset of more than 7 million business establishments in the United States to understand an urban system's evolving structure. We show that cities, despite differing in their populations and specializations, exhibit common patterns of growth, with striking parallels to those observed in the natural world.

We use the North American Industrial Classification System (NAICS) to describe the "phylogeny" of an urban economy, comparing snapshots of this tree in different cities. The NAICS classification system is organized hierarchically with six levels of structure, with more digits representing increasing resolution. The data is described further in the data section above.

From this tree, we extract snapshots for 942 individual urban areas in the United States, using the U.S. Census-defined Metropolitan Statistical Areas (MSAs). Each snapshot represents a particular expression of the tree, in which only a portion of possible firm types are present. If a firm type $i$ has at least one establishment, it is expressed in that city. We also collect information about the number of individual establishments $n_i$ present for each firm type $i$, and population of each MSA.

In the same way that biological classification systems are subject to the definition of a species and differential ability to find and identify various creatures, the NAICS classification is an artifact of its human classifiers. We require, however, for sufficient resolution in codes to ensure that tree expression is not limited by city size. We check for saturation of codes in the largest city, New York, and find, reassuringly, only 90% of types expressed. Others have show via expansion of the codes that true diversity can be estimated, and is constrained by factors other than the limitations of the coding system.

Importantly, unlike the tree of life, which has only a single expression, here we have multiple expressions of economies of different populations, across which we can make systematic comparisons.

The results of this section are fourfold. First, as city size grows the different levels of an economy (where a level is defined by relative specialization, or the number of digits in the NAICS code) become more similar. That is, the branching patterns at one level (for example the 2-digit level) resemble those at another level (for example the 5-digit level), indicating greater scaling of tree structure.

Second, the relationship between bifurcations and abundance patterns, or the way in which establishments are distributed across levels of the economy, becomes more similar as city size grows, indicating a second form of scaling between levels. Scaling thus strengthens with population in two ways: in branching patterns, and in branching versus firm abundance.

Third, as a city grows, its economy becomes less even, again in two ways. The shape of the economic tree itself becomes more asymmetric, with more diverse branching at a given level. For example, a node $a$ at the 3-digit level may branch into five 4-digit industries, and a neighboring 3-digit node $b$ may only branch into a single 4-digit industry. This pattern is asymmetric, compared to the case in which $a$ and $b$ each branch into three 4-digit industries. We define symmetry more precisely below.

At the same time, the distribution of firm abundances across a given level of resolution becomes more uneven. That is, in larger cities the "rich" industries (nodes with a large number of establishments) get richer. Together these findings suggest that a city specializes in two distinct ways, by (1) allowing more sub-specialties in a given area of expertise, relative to other cities, as well as by (2) generating more firms in a that area.

Finally, we compare the rates at which these two types of unevenness ((1)tree and (2) firm) grow in different sectors. In ecosystems, a regular positive relationship between tree asymmetry and uneveness in species abundances suggests the presence of selection processes [30]. Surprisingly, urban economies conform to this pattern in some but not all sectors. In particular, service-oriented sectors seem to develop in the same way as ecosystems, with tree asymmetry and firm unevenness positively linked, while manufacturing sectors grow in the opposite way.

**Measures of economic structure**

The NAICS classification system includes approximately 1200 industries organized into hierarchical levels. Each level is described by the number of digits in its NAICS code, ranging from level 2 or $l_2$ (such as 45 Retail Trade) to level 6 or $l_6$ (such as 451212 News Dealers and Newsstands). There is a single node at level 1, from which the 2-digit sectors of $l_2$ branch. For each city, we consider only the expressed tree of industries represented by at least one firm. Each node $i$ in the tree corresponds to an industry, such as *4512 Book, Periodical, and Music Stores* as well as to the subtree $S_i$, which includes all of the 5- and 6-digit industries branching from 4512. We further define:

- A *branch* $r_{ij}$ is the single path from any node $i$ to one of its termini $j$ at a lower

level.

- A *sector* $x_i$ is the 2-digit code corresponding to a given $i$. We examine 9 of the NAICS sectors here.

- The abundance $n_i$ is the number of individual establishments in a given industry $i$

- For any *subtree* $S_i$ beginning at node $i$, $a_i$ is the actual branch length, corresponding to its depth in the tree. $c_i$ is the cumulative branch length, or the total sum $\Sigma(a_i)$ of the length of its branches

- The number of *bifurcations* above $i$ in the tree is designated by $b_i$

Two properties of the economic tree describe how an economy grows over time. *Scaling L* describes the degree to which branching in a subtree $S_i$ resembles the branching structure of the entire tree [25].

To illustrate, consider the tree in Figure 4.1. At the finest level of resolution, the actual and cumulative branch lengths $a_i$ and $c_i$ are all 1. At the next level up, on the left side of the tree, $a_i$ is 3, while $c_i$ sums the $a_i$ of the present node, plus the nodes below.

The second property, *symmetry Y*, describes the degree to which a tree's branching patterns are self-similar. A perfectly symmetrical tree will have, at a given level, the same number of branches from each node $i$ at that level. Figure 4.2 illustrates the concepts of symmetry and scaling.

Scaling and symmetry are orthogonal properties: a sub-tree's symmetry is in principle unrelated to the degree of scaling of to entire tree. Scaling concerns the relationship *between* levels, while symmetry describes the pattern *within* levels. Practically, we measure scaling strength as the coefficient of determination of a power-law fit to $a_i$ versus $c_i$ for all $i$ in a city's expressed tree, where $R^2$ would indicate perfect scaling.

So, the relationship between scaling, symmetry and tree features $a_i$ and $c_i$ for every $i$ is

$$c_i = (a_i)^Y$$

with a fit described by $L$

In addition to scaling and symmetry, we can measure how firms are distributed across the various expressed branches. Firm evenness is defined as the degree to which a city's establishments are uniformly distributed across industry codes, described by the Shannon entropy measure:

$$E = \sum_i n_i log(n_i)$$

Because entropy naturally increases as additional types are added, we constrain to only the intersection of industry types existing in cities of 100,000 to 150,000 people and

measure for all cities greater than 100,000 from this benchmark, corrected entropy, which we define as

$E_c = \sum_{i|i\,in\,k} n_i log(n_i)$

Where k is the intersection of industries present in cities of population 100,000 to 150,000.

Just as the emergence of a new industry node $i$ in a city can lead to different changes in symmetry $Y$, the addition of a new establishment to an existing node $i$ changes the abundance $n_i$, and thus the evenness $E$ and $E_c$ of firm distribution.

For each sector, we consider separately the expressed tree of each 2-digit NAICS sector. So the subtree $S_{22}$ beginning at node 22 (Utilities) includes only those industries in the utilities sector. We see how the structure of city-sector pairs changes both across cities and across industrial sectors.

**Results**

As cities grow, they seem to aggregate skills and specialize more effectively [18]. How is this expressed in the structure of the tree? To test for structural trends as a function of size, we characterize each city-sector pair. We first count the abundance or total number of firms for the entire 2-digit sector, and find a near-linear trend, suggesting a universal economic structure that requires a certain number of firms of type $i$ per capita, independent of city size. We note, however, that this trend varies from linear per capita when we examine abundances at the 6-digit industry level.

So larger cities, in terms of assortment of firms, are simply scaled up versions of their smaller counterparts. We might then ask if the economic tree of a small city is also a smaller version of a larger city's. That is, do symmetry and scaling trend with population size? To answer this, we compare the branch length $a_i$ and cumulative branch lengths $c_i$ for every subtree $S_i$. In a city with perfect scaling, branching patterns at the 2- to 3-digit level should resemble branching at the 4- to 5- digit level, and so forth. Closeness to a linear fit to the log-log plot of $C_i$ and $A_i$ for all $i$ indicates the strength of scaling.

We calculate the exponent and fit of $C_i$ versus $A_i$ for all cities. Binning cities by population, we find that for populations less than 50k, we find $R^2$ values in the 0.1 to 0.5 range. As we increase population to cities over 500k, fits rise to the 0.8 to 0.97 range. A moderately sized city sees strong scaling. We will next use the exponent to estimate tree symmetry, for cities with a strong fit ($r^2 > 0.7$).

As we saw above, scaling of sub-trees and tree symmetry are orthogonal properties: strong scaling doesn't require more or less symmetry, and vice versa. Symmetry is the within level measurement of evenness of branching. For all city-sector pairs in which the relationship between $C_i$ and $A_i$ is strong (generally, city sizes > 100,000), we plot the slope of the power-law fit against the population, and find that symmetry

decreases with population. This is in contrast to a random model prediction (Equal-rates Markov model), in which industries branch randomly on the tree, leading to an expected symmetry exponent = 1. Additionally, the random model does not predict a trend with city size. Instead, increasing asymmetry indicates that with population growth a city develops a greater comparative advantage. And because each adds a different set of industrial sectors, each city can be said to specialize in its own way [41]

We would now like to understand how firms are filled in as the tree grows. On average, all firms grow at the same per capita rate. But some specialized 6-digit industries grow much faster (lawyers and software engineers) or slower (turkey farmers and mining) than the linear prediction. To understand tree structure, we would like to study how precise (6-digit) industries grow relative to more general industries (2-digit).

In ecological systems there exists a relationship between the number of bifurcations in the tree and species abundance [56] . Within families of species, organisms with fewer bifurcations tend to be more abundant, and creatures that have speciated many times less abundant. More uneven branching patterns (asymmetry) as well as species abundances are generally associated with the presence of a single ecological factor driving evolution.

We can examine the relationship between bifurcations and firm abundance in the economic tree. A bifurcation $b_i$ is defined as the depth of node $i$, or the number of divisions in the branch above it. Abundance $n_i$ is the number of firms at node $i$. We bin cities by population and compare bifurcations $b_i$ to abundances $n_i$ at each node $i$, and find a scaling relationship similar to the one describing branching at different levels in the tree: above about 150,000, a power-law fit describes the data well.

So, we have some indication that there exists regular scaling in the precision (depth in the tree) of an industry versus volume of firms, at least in cities above a threshold population. To summarize, in the largest cities the economic trees are more strongly scaled in two ways: in tree shape replicating across levels, and in distribution of species abundances scaling between levels.

Next, we would like to understand how specialization changes with city population. In a larger city, we can expect more specialization, perhaps as a result of economies of scale [32], or taste for diversity[29]. Although competing arguments exist for why, people continue to migrate to urban areas despite a clear premium paid for living in a city [18].

But what does this diversity have to do with tree structure, and how does it change as a city grows? We consider two measures of evenness: the evenness $E_c$ across firm abundances, and in the structure of the tree itself, or symmetry $Y$.

We have already shown how the latter changes with city size, using the exponent of fit to $C_i$ versus $A_i$. Now, we plot corrected evenness $E_c$ against city population for every city-sector pair. We find that $E_c$, like tree shape, becomes more uneven as cities grow (the unnormalized entropy $E$ also grows, but here use the more conservative $E_c$).

We compare observed evenness $E$ to evenness generated by a null model in which abundances arise from even splits of a population, beginning with the abundances observed

at the root of the tree. This resembles random fraction models of niche apportionment. We find the observed distributions of evenness are very close to the predictions of the null model.

We can use these expected abundances identify those industries that most differ in abundance from the predictions of the null model. For example, we find more grocery stores than expected from such a model.

In ecosystems, there exists a relationship between tree symmetry $Y$ and evenness of abundances $E$, which is often ascribed to the presence of non-random evolutionary drivers [56]. In economies, what are the relative growth rates versus population? Here, surprisingly, we find that the relationship seen in ecologies holds only in certain sectors: namely, service-oriented industries, figure. $L$ versus $E$ is positively sloped in services, and negatively in other sectors. We might speculate that in service industries, the division of labor and specialization allows for evolutionary drivers of that city to emerge. In manufacturing and agricultural sectors, by contrast, the emergence of specialization is slowed by resource availability and other dampening effects.

**Discussion**

Economic theory is grounded in accounting: for firms, people, and profits. Yet in simply summing up the artifacts of city life, we gloss over the interactions that form the basis of urban experience: small bakeries employing advertising firms, and department stores branching into more specialized sporting goods retailers. Ecologists have long recognized that interactions matter in driving evolutionary outcomes.

By wedding structure to simple counts, we can begin to identify the structural processes that drive growth in cities. We discover several fruitful analogies to ecosystems, suggesting that a large urban economy is able to specialize in two ways. First, as a city grows larger, the economic tree generally becomes more uneven. Second, larger cities see greater unevenness in the distribution of establishments across existing industry types.

We have suggested a baseline methodology for considering economic structure in urban systems. We note that individual cities deviate in interesting ways from the general trend. We leave to future work the question of whether a particular composition of industries can predict such deviations, as well as the identification and testing of mechanisms for growth, and perhaps even for the limits to growth, of industrial sectors.

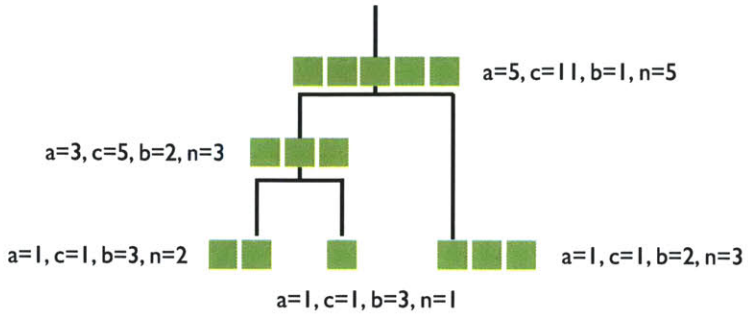



Figure 4.1: For a simple tree, at each node the actual branch length a, cumulative branch length c, bifurcations b, and abundance n (with each green box representing one firm)

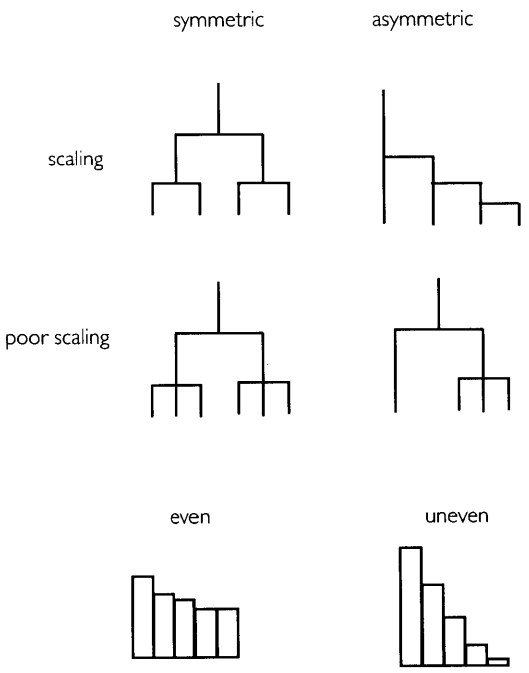



Figure 4.2: Symmetry and evenness measures in hierarchical trees. The regularity of the relationship between a and c gives a sense of the scale-invariance of the tree. The slope gives a sense of the tree's symmetry.

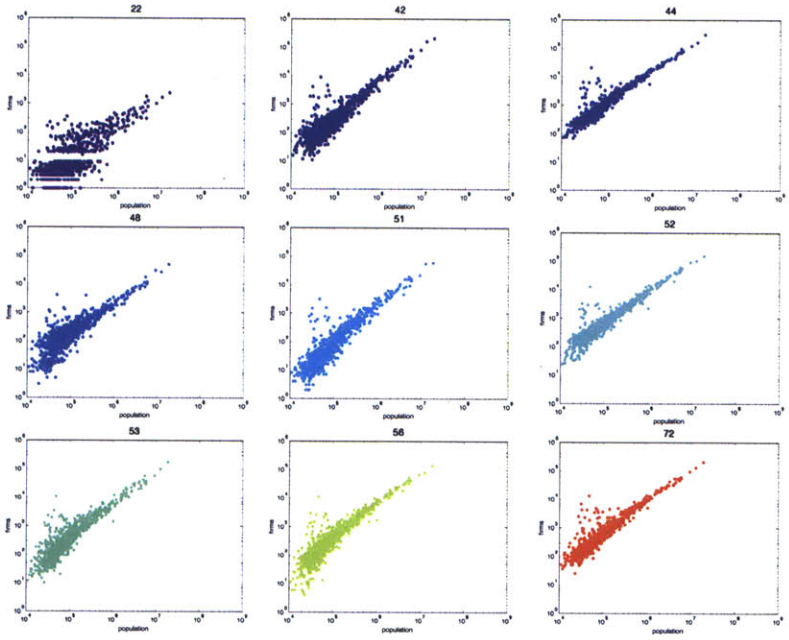



Figure 4.3: Number of firms varies linearly with population. Each 2-digit code is a sector: 22 Utilities / 42 Wholesale Trade / 44-45 Retail Trade / 48-49 Transportation and Warehousing / 51 Information / 52 Finance and Insurance / 53 Real Estate and Rental and Leasing / 56 Administrative and Support and Waste Management and Remediation Services / 72 Accommodation and Food Services

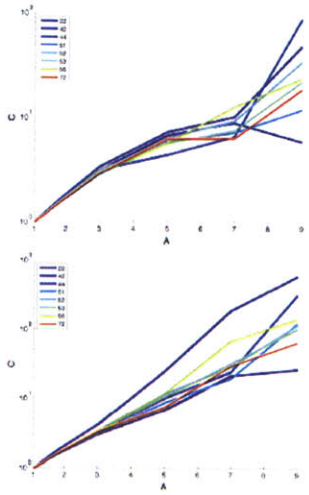


Figure 4.4: In small cities there is poor scaling in the economic tree. In large cities, the scaling is stronger. Small cities are lopsided: sectors see either over- (such as a small mining town) or under- (for example, a village without all the shades of retailers you see in a big city) specialization

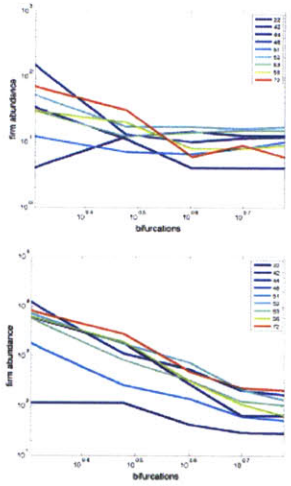



Figure 4.5: In small cities, the relationship between bifurcations and firm abundance is poor. In large cities, the relationship is stronger in most sectors

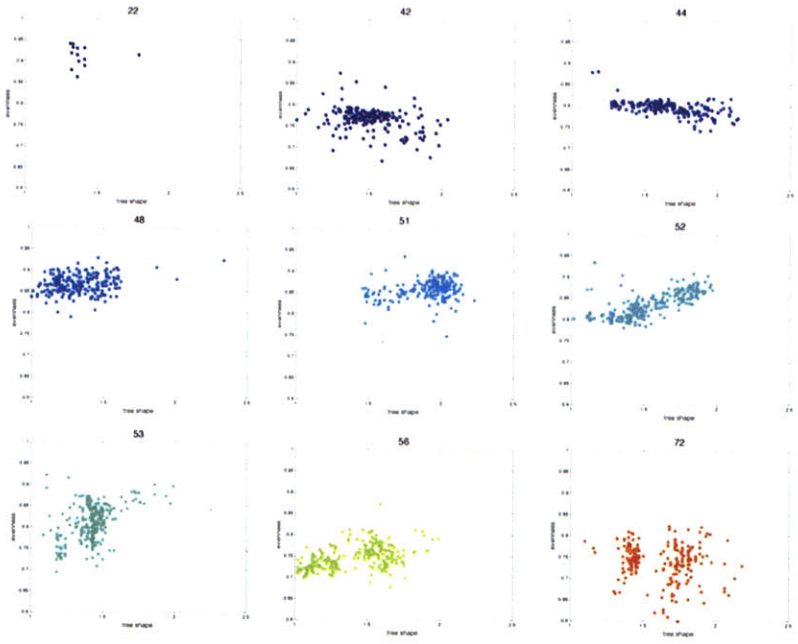



Figure 4.6: Evenness as a function of tree shape. Each 2-digit code is a sector: 22 Utilities / 42 Wholesale Trade / 44-45 Retail Trade / 48-49 Transportation and Warehousing / 51 Information / 52 Finance and Insurance / 53 Real Estate and Rental and Leasing / 56 Administrative and Support and Waste Management and Remediation Services / 72 Accommodation and Food Services

## 5. A statistical test for churn in distributions of ranked items


Introduction
Data and Background
Individual rates of change as a function of exploration and resources
Rates of change in the structure of urban economies
**A statistical test for churn in distributions of ranked items**
- **Introduction**
- **statistical models of rank churn**
- **Binomial statistics**
- **A statistical model**
- **Discussion**

Discussion
Appendix
Bibliography



Rank-size distributions exhibit statistical regularities across a remarkable number of phenomena, from city populations and river lengths, to word frequencies and individual wealth. While the shape of such distributions are often stable over time, this stability masks internal churn: as more data are collected, two items might "switch places" in a rank-ordering, without perturbing the overall distribution. In order to understand the dynamic processes that shape rank-ordered data, it is important to test against a null model of expected churn. To date, no such random process has been described. Here, we present a model to compare observed rank change in system with expected change, given item rank, distribution exponent, and initial and additional sample size. We first establish a simple binomial model, and show that for power-law (but not exponential) distributions, churn can be expected to increase with rank. We next generalize to a multinomial case, and present a statistical test for comparing the predictions of our model to empirical data.


### Introduction

In many complex systems that exhibit a diversity of different types, ranking such types by their frequency can reveal strikingly regular distributions [61]. This regularity has been observed in such disparate domains as individual wealth, city populations, particle sizes, and word usage, which often conform to Zipf, Pareto or related distributions.

However, these regularities may hide important dynamics. It has been noted empiri-

cally that while the overall distribution may remain stable over time, there exists churn between items at different ranks [5, 31]. For example, Madison, WI was the 97th most populous U.S. city in 2000, and the 89th in 2010. In addition, churn often appears to be more frequent at the tail of the distribution: New York has remained the largest American city since at least 1790 (the first US Census), but Las Vegas, NV moved from the 51st to the 31st in just 10 years. Similar dynamics are observed in word usage statistics, for example.

There have been several quantitative descriptions and illustrations of rank churn in different domains, for example in terms of rank clocks, which give an interesting perspective on rank changes over time [5]. In addition, the tendency of individual items to rank shift (or, the state stability of items) has been quantified for various empirically observed systems. Yet to our knowledge there exists no methodology to determine how much of this churn results from random fluctuations over time, or how much represents other, potentially important processes, e.g. expressing a systematic advantage of certain types over others at larger or longer scales. Was it inevitable, statistically speaking, that New York persisted at the top of the distribution of US cities for more than two centuries? Was there something about Las Vegas that let it rise so quickly, or could any city have done the same? We lack the statistical machinery to answer such questions.

#### Statistical models of rank churn

Switches in rank between two types can happen by chance in any finite sample under a stationary process, or they may be the result of the expected rank of a type being sample size dependent.

In addition, the characteristics of the distribution itself are key: at the limit, consider items distributed uniformly versus following a power-law or exponential distribution. In a uniform distribution, adding a single item can perturb the rank ordering.

We consider the simplified, 2-item case in which the measure of distribution is the ratio (and consequently the relative rank) between item-1 to item-2. We define the expected proportion of item-1 as $p$ and of item-2 as $1-p$; we can without changing the model constrain $p$ to $p \geq 1/2$. The pair-wise ratio of arbitrary, consecutive items, then, is simply a modification of the 2-item case, where $p$ is the expected frequency of item-n and $1-p$ of item-$n+1$: taken together these pair-wise ratios form the distribution. While our focus is on the approximately power-law or Zipf distributions that are common in the real world [], we will briefly consider outcomes given alternative distributions.

In Zipf-distributed data, item rank is clearly important: items at the top and bottom of the distribution can be expected to churn differently. Take the example of New York versus Los Angeles, the second largest city with 6 million fewer people. Los Angeles has more "catching up" to do than the 252nd city Albany, GA to the 251st Jackson, MI, where the difference in population is only 3,000. A random process might more easily account for churn in less frequent items, where the frequency difference is reduced. Expected rank churn additionally depends on the amount of data sampled, and specifically on the difference between the sizes of the initial and total samples $N_i$

and $N_T$.At the limit (if $N_i$ is arbitrarily small or large relative to $N_t$), rank churn can be expected to be zero. On the other hand, if $N_i = N_T/2$, the expected churn is maximized for any given item rank and $p$.

Finally, rank churn will depend on $N_t$ itself: a larger sample size will reduce the probability of expected churn. In order to eliminate need for tie-breaking in rank, we restrict our model to unique frequencies, and do not consider the long tail that is often seen in Zipf-distributed empirical data.

#### Binomial statistics

We now consider the simplest 2-item case of rank change. Assuming statistic independence of each trial, the outcomes follow binomial statistics. We denote each sample by $S_i$, characterized by size $N_i$, $1 = 1, 2, \ldots$. The elements of each $S_i$ are drawn from a population $G$ described by the proportion of elements of type 1 and 2 to be $p_i$ and $1 - p_i$, respectively. These probabilities are not necessarily stationary, and may more specifically depend on $N_T = \sum_i N_i$, the total sample size.

In general successive samples $S_i$ might correspond to a measurement in a particular year, as in the case of city populations, or of a continuous corpus, as in the case of word frequencies. In either scenario, the process generating the distribution is assumed to be ongoing, and thus gives the additional sample $S_{i+1}$.

Our aim is to measure the effect of $S_{i+1}$'s addition on the particular rank of each item. Consider the situation that in the initial sample of size $N_i$ there were $k_i$ draws of type 1 (successes). Then $\Delta_i = k_i - (N_i - k_i) = 2k_i - N_i$ is the difference in frequency between the two items at this stage. Additionally we can define the total type asymmetry as $\Delta_i^T = 2k_T - N_T$, where $k_T$ is the total number of successes so far. What is the probability that, given the next sample, the two items change rank?

There are two forms of the problem. In the first, the difference in outcomes in the previous sample $\Delta_i$ is given and we would like to determine the probability that this difference is overcome so that there is churn $c$. We denote this probability $P(c|\Delta_i, N_{i+1}, p_{i+1})$. In the second, more general case we may not know the outcomes of two successive samples, but may know their parameters $N_i, p_i, N_{i+1}, p_{i+1}$. Then we would compute $P(c|N_i, p_i, N_{i+1}, p_{i+1})$.

With these definitions we can now express the discrete probability distribution of churn / no churn ( $c/\bar{c}$) as

$$P(c|N_i, p_i, N_{i+1}, p_{i+1}) = P(\Delta_i > 0)P(\Delta_{i+1} + \Delta_i > 0) + P(\Delta_i < 0)P(\Delta_{i+1} - \Delta_i < 0) \tag{5.1}$$

where the probabilities $P(\Delta_{i+1} + \Delta_i > 0)$, $P(\Delta_{i+1} - \Delta_i < 0)$, with $\Delta_i$ given, describe the problem $P(c|\Delta_i, N_{i+1}, p_{i+1})$, or the probability of churn.

Note that one type or the other being more frequent are mutually exclusive, so that the options are additive. The probability of churn is dependent on the the relative

probability of each type per trial in the sample, $p_i$, and on the size of the two samples $N_i$ and $N_{i+1}$. We now make these probabilities more explicit in terms of particular statistical models.

Note that there is one fundamental difference between the problem in which the outcome of the previous sample is known or not. In the first case $\Delta_{i+1}$ is our statistical variable. In the second, in addition to the $\Delta$'s, the sum and the difference, $\Delta_{i+1} \pm \Delta_i$, that also must be treated as fundamental statistical variables. In the following we derive probabilistic models for all these variables.

We start with the purely binomial case. Taking each trial as independent and distributed as a Bernoulli trial with stationary probability $p_i$ we have

$$P(\Delta_i > 0) = Bi(k_i > \frac{N_i}{2}, N_i) \tag{5.2}$$

and, if we treat $\Delta_i$ (or equivalent $k_i$) as given,

$$P(\Delta_{i+1} + \Delta_i > 0) = Bi(k_{i+1} > \frac{N_i - 2k_i + N_{i+1}}{2}, N_{i+1}), \tag{5.3}$$

$$P(\Delta_{i+1} - \Delta_i < 0) = Bi(k_{i+1} < \frac{2k_i - N_i + N_{i+1}}{2}, N_{i+1}). \tag{5.4}$$

Provided $p_i = p_{i+1}$ (as in the case of Zipf-distributed samples), we obtain a simpler picture with the statistics of $\Delta_i$ approaching a normal distribution with mean $\mu_{\Delta_i}$ and variance $\sigma^2_{\Delta_i}$:

$$\mu_{\Delta_i} = 2\langle k_i \rangle - N_i = (2p_i - 1)N_i = \delta_i N_i. \tag{5.5}$$

and

$$\sigma^2_{\Delta_i} = 4\langle (\delta k_i)^2 \rangle = 4p_i(1-p_i)N = (1-\delta_i^2)N_i. \tag{5.6}$$

with $p_i = 1/2 + \delta_i/2$, $\delta \in [-1, 1]$. We invoke a limit theorem to compute the mean and variance for the sum of samples, which follow from the additive behavior of Gaussians. Then we obtain that

$$\mu_{\Delta_i \pm \Delta_{i+1}} = 2\langle k_i \pm k_{i+1} \rangle - (N_i \pm N_{i+1}) = \delta_i N_i \pm \delta_{i+1} N_{i+1}, \tag{5.7}$$

$$\sigma^2_{\Delta_i \pm \Delta_{i+1}} = (1-\delta_i^2)N_i + (1-\delta_{i+1}^2)N_{i+1}. \tag{5.8}$$

Thus we can write the total probability in terms of Normal CDFs $\Phi(x)$ as

$$P(c|N_i, p_i, N_{i+1}, p_{i+1}) = \Phi(\Delta_i - \Delta_{i+1}) + \Phi(\Delta_i)\left[\Phi(\Delta_{i+1} + \Delta_i) + \Phi(\Delta_i - \Delta_{i+1})\right], \tag{5.9}$$

which is a general function of $N_i, p_i, N_{i+1}, p_{i+1}$. Clearly for $p \to 1/2$ (when the two items have the same underlying frequency) the probability of rank switches remains finite for any sample size. This is because the expectation value of $\mu_{\Delta_i} = 0$ and the variance is maximal, so that in principle there are always new fluctuations that can take the rank ordering in either direction.

In the following section, we present a simple statistical test for the expectation of churn. To do so, we make several assumptions which simplify the general form of the problem. First, we assume $p_i = p_{i+1}$, or that the approximate ratios of frequencies at two neighboring ranks will not change over time. This allows us to treat $P(c|N_i, p_i, N_{i+1}, p_{i+1})$ as $P(c|N_i, p_i, N_{i+1})$. Second, we consider $N_T$ the variable describing the total sample size $\Sigma_i N_i$. Since $N_{i+1}$ can be described as a fraction of $N_T$ we can calculate the probability of churn, given parameters as $P(c|N_i, p_i, N_T)$. We first show the effect on churn of varying $p$, $N_i/N_T$ and $N_T$ in simulation. Churn increases as $p$ increases or decreases from 0.5, and as $N_i/N_T \to 1$.

**A statistical test**

Under the assumption of independence of each sample $S_i$, we can now treat the probability $P(c|\text{parameters})$ as an input to a new binomial distribution. This allows us to perform a simple Binomial test on changes in rank and therefore state with a given level of statistical confidence whether the observations are consistent with a model or should be treated as an anomaly, possibly indicating a change in rank due to interesting new dynamics.

As an example, we apply our method to a dataset of population sizes of the 100 most populous Urban Places in the United States, as defined by the US Census [11]. We use a time series from 1790 to 1990; however, we consider only dates from 1840 forward, because between 1790 and 1840 there were fewer than 100 locations which were considered Urban Places and whose population was tracked.

We extract from the rank-frequency distribution for each year the parameters for each consecutive pair of cities $U_n$ and $U_{n+1}$ in a given year, and $N_i$ is the population in year i minus the population in year $i-1$. $N_T$ is the total population in year $i$.

With these parameters, we simulate 1000 runs of the expected churn for each city. We treat churn only as a movement upward. This ensures independence of results, provided the population of cities does not decrease over time, and the sampling is strictly assigning new items (people) to bins (cities) according to parameter $p_i$. (We could also do the problem in reverse, and look at decreases rather than increases).

For each city $n$ at time $i$ we thus have an expected churn $E(c_{n,i})$, or the probability that the city will have moved up to a lower rank in period $i+1$. From the empirical data, we calculate the observed churn $O(c_{n,i})$

We have as a null model that the observed churn between periods is due to a random process of adding additional samples. That is, that all cities grow according to the same fundamental rules. To test whether to accept or reject this hypothesis, we test whether the difference between $E(c)$ and $O(c)$ is statistically significant.

In the present example, we calculate this significance for the entire dataset, treating each city-year pair as a datapoint. However, one might also do the test separately for each addition of a new sample, i.e. decade by decade in this case.

Note that our values from $O(c)$ are binary (churn or no churn), while values for $E(c)$ are continuous - with the Gaussian approximate these correspond approximately to the distance (in standard deviations) from the mean (defined by the specific parameters). For given $p$, expected churn is maximized when $N_i/N_t \to 0$, and minimized when $N_i/N_t \to 1$. As difference increases the value of $E(c)$ will decrease.

To compare to $O(c)$ we must treat $E(c)$ as a binary variable by assigning a threshold value $\tau$ and setting $E(c) < \tau \to 0$ and $E(c) \geq \tau \to 1$

We find that only with very low thresholds for expected churn (e.g. by including very low probabilities of churning per the random model) and considering only the top cities (which in general, have greater values of $p$, on the order of 0.7, compared with low rank pairs where $p$ is close to 0.5) can we accept the null hypothesis that what we observe is the result of random process. So, we reject the hypothesis and conclude that there exist additional forces changing the rank ordering of cities by population, than mere additional sampling would suggest.

We have presented a statistical test for the entire population. Now, we'd like to answer questions of the form: which items show interesting trends in rank change? For example, we might want to know to what degree the increase in Los Angeles' rank since 1900 is driven by non-random processes of churn such as preferential assignment of population.

To answer such questions, we can compare the expected to the observed churn over time, and assign a probability to each. We consider three cities: Los Angeles, CA; Buffalo, NY; and Cambridge, MA. From 1890 (the first year it appeared in the top 100 places) to 1990, Los Angeles moved from rank 57 to rank 2. Between 1840 and 1990, Buffalo briefly rose from rank 16 to rank 8, and then fell to rank 50 in 1990. Cambridge, MA was the 33rd most populous urban place in 1850; by 1950 it had fallen to 87th, after which it disappeared from the top 100 list.

For each of these cities, we calculate the joint probability of churn over the time period, and compare to observed results. By varying the threshold and calculating the rates of false and true positives and negatives, we can use the AUC (area under the curve) measure to compare the deviation from expected churn of different cities.

**Discussion**

In his important paper in 1955 Herbert Simon [53] showed that rank size statistics result from rules of attachment of new elements to existing types. This depends on the number of their previous occurrences (frequency) in a finite sample, but not necessarily on any intrinsic property of each type. However, the condition of attachment says nothing per se about changes in rank as larger and larger samples are collected.

Thus, even if models that derive the correct rank-size statistics based on only frequency information may be correct at one level, they fail to describe important dynamics by which certain types become systematically more frequent in larger systems, such as

the ascendency of certain cities or economic sectors as nations grow in demographic and economic terms. The capacity to distinguish these important processes from a simple random model is critical to formulating sound theories about growth and change. Such a methodology could also be used to compare multiple datasets, to ask whether generating process occurs across both datasets.

Here, we have developed a test for the deviation from random change in rank of a distribution that is sampled over time. Additionally, we present a test for the probability of any item's change in rank being due to non-random processes. We apply these methods to population data from US cities over more than ten decades, and find that top cities deviate little from expected churn, but that there is significant churn in cities with smaller populations.

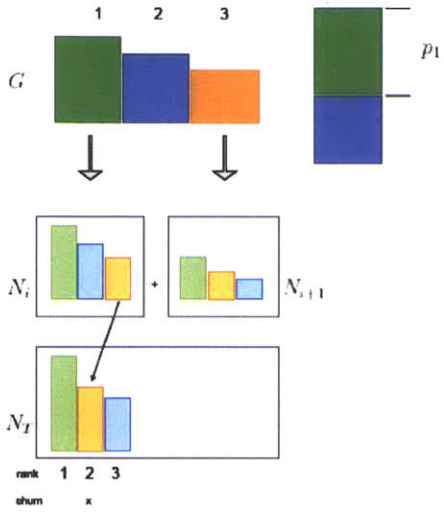



Figure 5.1: Samples $N_i$ and $N_{i+1}$ are drawn from global distribution $G$. What is the probability that with the addition of $N_{i+1}$, the rank of the orange type rises from 2 to 3?

| 1900 | 1990 |
|---|---|
| New York, NY | New York, NY |
| Chicago, IL | Los Angeles, CA |
| Philadelphia, PA | Chicago, IL |
| St Louis, MO | Houston, TX |
| Boston, MA | Philadelphia, PA |
| Baltimore, MD | San Diego, CA |
| Cleveland, OH | Detroit, MI |
| Buffalo, NY | Dallas, TX |
| San Francisco, CA | Phoenix, AZ |
| Cincinnati, OH | San Antonio, TX |

Figure 5.2: Most populous U.S. cities, 1900 and 1990

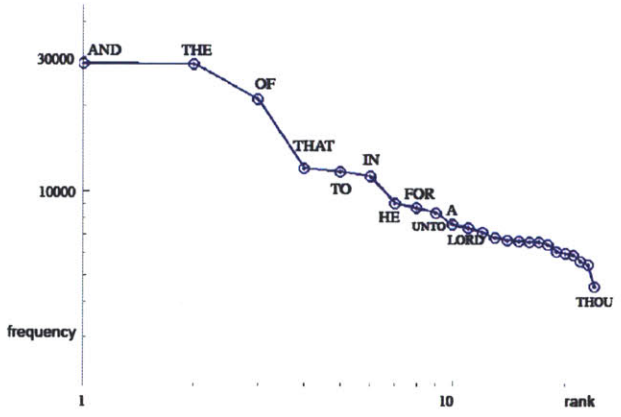



Figure 5.3: Frequency of top 20 words in the King James Bible

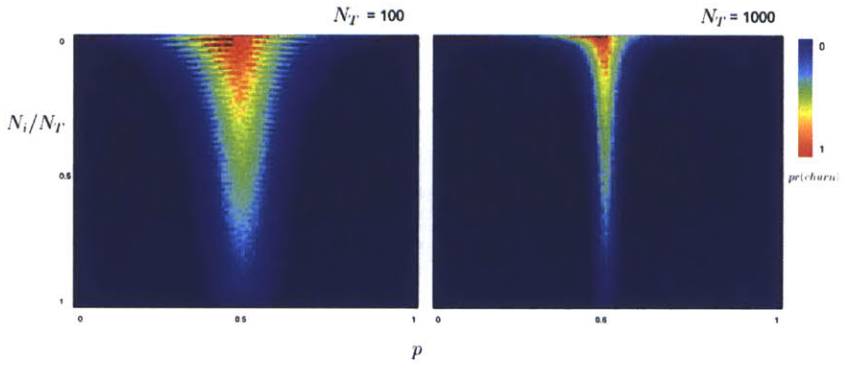



Figure 5.4: Simulation of expected rank churn, given parameters $N_i$, $N_i$ / $N_T$, and $p_i$. Distribution is more narrow for $N_T = 1000$ versus $N_T = 100$

| *tau* | top 3 | top 10 | 1980-90 | all |
|---|---|---|---|---|
| 0.25 | 0.29 | 0.00 | 0.36 | 0.00 |
| 0.50 | 0.29 | 0.00 | 0.00 | 0.00 |
| 1 | 0.01 | 0.00 | 0.00 | 0.00 |

Figure 5.5: Results of rank churn test for top 3, top 10, for all cities from 1980-1990, and all cities at all times.

| | Buffalo | | Cambridge | | L.A. | |
|---|---|---|---|---|---|---|
| | exp | obs | exp | obs | exp | obs |
| 1900 | 0 | 1 | 0.1 | -1 | 0 | -1 |
| 1910 | 0 | 1 | 0.3 | 1 | 0 | 1 |
| 1920 | 0 | -1 | 0.1 | 0 | 0 | 1 |
| 1930 | 0 | -1 | 0.5 | 1 | 0 | 1 |
| 1940 | 0 | -1 | 0.3 | -1 | 0 | 1 |
| 1950 | 0 | -1 | 0 | 0 | 0 | 0 |
| 1960 | 0 | -1 | 0 | -1 | 0 | 1 |
| 1970 | 0 | -1 | 0.4 | -1 | 0 | 1 |
| 1980 | 0.1 | -1 | 0.3 | -1 | 0 | 0 |
| 1990 | 0.4 | -1 | 0.6 | -1 | 1 | 0 |
| 2000 | 0 | -1 | 0 | -1 | 0 | 1 |

Figure 5.6: Expected and observed churn for Buffalo, Cambridge, and Los Angeles

# 6. Discussion: summary of findings, summary of contributions, and future work

*Se hace camino al andar* – Antonio Machado

*If you do not change direction, you may end up where you are heading* – Lao Tzu


Introduction
Data and Background
Individual rates of change as a function of exploration and resources
Rates of change in the structure of urban economies
A statistical test for churn in distributions of ranked items
**Discussion**
- **Summary of findings**
- **Summary of contributions**
- **Directions for future research**

Appendix
Bibliography


**Summary of findings**

This thesis takes three approaches to quantifying the rates of change in human systems. First, we consider individual rates of change. In order to understand how a person's portfolio of behaviors changes over time, it is important to understand the extent to which a static "snapshot" of his behaviors is predictable. Using measures of information entropy, we find that individual predictability (where people go) can be bounded over longer time scales. At the same time, individuals continually innovate in the locations they explore, as well as in the paths they elect between locations.

The rate of exploration is open-ended for almost all individuals, independent of the kinds of stores a person visits, his geography of his resources. This deviates from the economists' prediction that exploration would be closed. In the consumer data, this open-endedness explains why a Markov model of transition probabilities will slowly decrease in accuracy as more months are predicted.

Just as we can measure a person's rate of exploration, we can also quantify the rate at which his aggregate portfolio or genome of habitual stores is changing. The rate of

change of any individual's top store is predicted simply by his rate of exploring new places. This result is in concord with the predictions of a simple random drift model, in which an individual has a characteristic rate of exploration. Based on this rate, he replaces some percentage of total visits with new locations. The remainder of visits are assigned by copying from distribution of the past month. In this way, individual behavior can change slowly over time.

However, if we begin to look at rate of change for a deeper portfolio of behaviors (beyond the top one or two locations), a linear fit is no longer apt, and the simple random drift model must be modified. Turnover is bounded. A saturation of exploration's effect on turnover occurs for those with the highest levels of exploration. A model in which habits are proportionally weighted for replacement captures this saturation effect, not just for the top store but for a basket of arbitrary depth $n$.

These findings suggest a universality of habitual behavior across a single person's many movements and locations. Individuals also look surprisingly similar to one another in their habits, and are differentiated by their taste for variety, expressed in less routine behaviors. The rate of change of a person's behavior can be modeled based on his rate of exploration.

Exploration of new locations can occur for several reasons, among them to optimize one's current locations, as well as for exploration's sake. The latter motivation tends to be tied to resources: the more income a person has, the more he is able to explore just because.

This is a deviation from standard economic theory, which predicts that search is primarily driven by optimization. Indeed in this data, $W$ (exploration due to variety) correlates with income, and has an effect on rate of change. Here, a taste for variety drives more visits to the tail for wealth individuals, but also dampens the effect of change due to optimization on less frequented stores.

The economic model is right, to a point: we search in order to optimize, but we also leave the door open for non-optimizing search. Perhaps we evolve a taste for variety not only as a way of displaying resources, but also to facilitate exploration of new landscapes that are not just for optimization of a local portfolio of behaviors. The fact that some have sufficient resources to explore for fun also helps drive the growth of "non-essential" industries (which are often quite productive) in an economy.

In the next part of the thesis, we seek to connect the individual and population levels. Even for a static case of a market in which the underlying distribution of preferences doe not change, individual-level choices can relate to population-level outcomes in complex ways. Like an individual's distribution of visits to a set of locations, a city has a distribution of individual firms over a portfolio of firm types. Individual visits can be allocated across different locations, and these locations as well as their rank in the frequency distribution can change over time. Similarly, a city can have various individual firms of different types, and the distribution can change over time as new types emerge, or existing types switch ranks.

Institutions in cities are connected by a structure of industrial dependencies, in this case a hierarchical tree (Similarly, individuals are connected by a social structure, but our

data has not permitted us to consider it). Much of economics is grounded in summing up such firms rather than looking at the connections between them. Ecologists have long recognized that interactions matter in driving evolutionary outcomes. By wedding structure to simple counts, we can begin to identify the structural processes that drive growth in cities.

Several analogies to ecosystems emerge from the data, suggesting that a large urban economy is able to specialize in two ways. First, as a city grows larger, the economic tree generally becomes more uneven. Second, larger cities see greater unevenness in the distribution of establishments across existing industry types.

These findings suggest a baseline methodology for considering the rate of change of economic structure in urban systems, as well as identifying important outliers. Individual cities deviate in interesting ways from the general trend. We leave to future work the question of whether a particular composition of industries can predict such deviations, as well as the identification and testing of mechanisms for growth, and perhaps even for the limits to growth, of industrial sectors.

An important question across both individual and population rates of change is whether change due to some intentional intervention can be statistically distinguished from the outcome of a random process. To do so, we consider a system composed of a number of items. Much like an individual and his portfolio, words in a corpus, or people across a set of cities. We assume that there's a global population which can be sampled over time, and so individual things are being added to new types over time.

Ranked items can naturally churn as the outcome of random processes. Or, there can be an additional exogenous force, such as a person purposefully changing his habits or a city taking real steps to attract new people, that causes changes in rank relative to the other stable processes. It is important to distinguish between these two processes.

Models fail to describe important dynamics by which particular types become systematically more frequent in larger systems, such as the ascendency of certain cities or economic sectors as nations grow in demographic and economic terms.

A simple statistical test can offer insight. The capacity to distinguish these important processes from a simple random model is critical to formulating sound theories about growth and change. Similarly, this method could be used to compare multiple datasets, to ask whether generating process is similar.

**Summary of contributions**

To the extent that we can quantify individual and population level rates of change, we can begin to understand how systems might churn over time and become more or less resilient.

At the individual level, showing that the rate of change of human behavior can be described in terms of a simple model, based on search and resources, we open the door

for better predictions in complex markets, disease dynamics, and modeling of mobility and migration. We've long known that people are heterogeneous in their choices: we show here that individuals are not only heterogeneous, but bounded, in their rates of change.

Here, we have presented a new way of modeling human behavior and changed based on turnover in a set of visited locations. Traditional economic models tend to look only at a snapshot of behavior. We have also shown a relationship between income, rate of search, and rate of change.

With a statistical test for rank churn, we can better evaluate policies in city planning, as well as claims in linguistics, economics, and other fields in which ranked, dynamic distributions are important.

At a broader level, this work has implications for how we think about the social sciences. Human behavior is rarely static: rather, our behavior can be characterized by rates of change as a function of inputs, in a dynamic system.

**Directions for future research**

We have considered two examples of how items (individuals in cities, visitation locations for people) are distributed and change across types, as time progresses or more items are sampled. We also develop a statistical test to distinguish random from directed fluctuations in systems.

What are some net steps in understanding rates of change in social and economic systems?

When we look at complex geophysical systems like the earth's climate, and can establish baselines and trending amid the very noisy data. While we are still very far from developing similar "laws" of human nature: bounds, baselines, and trends (rates of change) help us to characterize and find parallels across this system.

The analysis of the relationship between rate of search and rate of change also has applications to a variety of social systems, including economic markets, consumer behavior, disease modeling, and analysis of migration flows or the trajectories of cities or nations.

If we can better understand how behaviors change over time, we can make better predictions about consumer choices. By bounding rate of change based on rate of search, resource distribution, or available sites, we can better predict where people will go and how the system will evolve.

If we can understand the rate at which new strategies are employment, we might better be able to bound the trajectories of economic markets. If we can model how quickly new farming practices and technologies are embraced, we can understand how quickly food production practices will change.

If know how people are expected to move based on their level of resources and the cities in which they live, as well as how often different pockets of the population are likely to interact, to cross paths at a give shop or location, we can better model the spread of diseases and of people across the globe.

An important area for future work is understanding how individual rates of change relate to population level rates of change. For example, the underlying environment is also changing at some rate $X$, as new stores are added to and removed from the landscape. We would like to know the rate $E_i$ relative to $X$, here we take $X$ as fixed for all individuals, because all individuals come from the same city. We can thus compare different individual rates of search $E_i$ across the population. Future work might study relative rates of $E$ and $X$.

Finally, by being able to test statistically for directed churn, in the entire system, and not just an individual prediction, we can identify those systems or items that have undergone real change. Development of this idea will be an important step forward in evaluating policies and making predictions for the future, in everything from city and environmental planning to the modeling of migration and diseases.

If search drives change and change is constant, what remains? This spring in Santa Fe, a visitor encouraged a young scientist on the fence: "Science is merely the process of reducing entropy in the world, or organizing some small corner of human knowledge." And here we are.

## 7. Appendix: MusicLab polya model

Here, we present a simple model that ties together individual choices and market outcomes for an online marketplace in which users can sample and download songs. This study has important implications for the study of individual and population level rates of change.

A recent experimental study [48] found that the addition of social influence to a cultural market increased the unpredictability as well as the inequality of the market share of individual products. However, the study did not propose a model to describe how such social forces might operate. Here, we present a parsimonious model that sheds light on social behavior in this market. Our model does not rely on assumptions about heterogeneous preferences [54] or solely on the generic notion of herd behavior [7] to explain the outcomes of anonymous social influence: rather, we treat social influence as a variable whose effect grows as the market matures.

MusicLab is an online laboratory created in 2004 to evaluate experimentally the role of social influence in the success of cultural products. Researchers invited consumers (about 14,000 in total) to sample 48 previously unknown pop songs via a website, to rate them, and to download whichever of the songs they liked. Songs were arranged on the screen in either a 16x3 grid (Experiment 1) or a single column (Experiment 2).

In each experiment, each visitor was assigned randomly to one of two conditions. In the social influence condition, of which there were eight instances or "worlds", participants received additional information about the number of times each song had been downloaded by his peers, and songs in Experiment 2 were ordered on the screen according to past download count. Songs were shown in random order in the independent condition [48, 49, 50].

Results from the MusicLab experiments suggest that, in this market, information about the behavior of others contributes to greater inequality (differential market share) and unpredictability (variance of possible outcomes), compared to the inequality and unpredictability in the non-social condition.

While Salganik et al. report empirical findings, they do not describe a mechanism for the process of social influence. Others have subsequently proposed theoretical models to explain how a set of individual preferences and responses can create such outcomes. Borghesi and Bouchard model each participant's decision as a multiple-choice situation, and two conditions of "weak" and "strong" herding that fit the empirical data[9]. Hendricks et al develop an equilibrium model to explain how an "anonymous" non-differentiated herd affects low versus high quality products [24]. Our approach differs in several regards. First, we observe the progression of inequality and unpredictability over the course of each experiment, and to compare it to simulation results. Unlike Borghesi and Bouchard, we do not consider decision-making a multiple-choice situa-

tion: we model independent listens rather than listeners, where a listen occurs according to a by-song probability derived from its appeal and a coefficient for social forces.

Social influence exists in non-experimental markets as well, in the form of herding and informational cascades [7]as well as individual decision-making in the presence of complex information[57, 17]. Of course, real markets offer a host of complexities intentionally omitted from the MusicLab experiment in order to test the researchers' hypothesis, such as the possibility for stronger, peer-to-peer social influence and external marketing forces [51, 58]. We discuss below some of the ways in which the experimental setting both resembles and differs from real-world markets. Here, our focus is on parsimoniously modeling the social dynamics of the MusicLab marketplace itself.

To do so, we develop a model for the empirical results that distinguishes between a song's quality and the signal generated by the visible downloads. From the empirical data, we observe that song selection can be modeled as a sequential process in which each song has a probability of being sampled, independent of the other songs a listener chooses, and then an independent probability of being downloaded. Modeling choices are based on empirical observations of user behavior in this market. We describe this process and the model inputs in detail below.

**Music lab**

In the MusicLab experiment, the authors record the choices of participants who enter the market one-by-one. Here, we model song listens rather than market participants, and validate this approximation by examining the consistency of sampling across different participants' propensity to sample more or fewer songs. We find that people who listen to a total of n (where n<40) songs in the system have, on average, the same probability of sampling a particular song i. In fact, over the entire population, the probability song i will be sampled does not depend on the distribution of volume of listens in the population who samples it (r-squared = 0.1). Additionally, the conditional probability of downloading a song (given it was sampled) does not depend on the total number of songs a participant samples .

Again from the empirical data, we observe that there are two stages of decision-making, listening and downloading, that occur according to fixed but independent distributions. The result of the second step (downloading), but not the first (listening), is ultimately visible to future market entrants. We observe that the probability that a user clicks on a song (which we ascribe to the appeal of the song's title) is independent of the conditional probability he downloads the same song, given he listened to it (which we call the song's quality). This finding suggests that in this market, the perception of quality is not subject to social influence.

Sociologists distinguish between the normative and informational facets of social influence [14]: while the former might compel a person to do as others do, the latter acts as a signal of what others like. Because song appeal and propensity to download are independent, we assert that social influence works as a purely informational force in this

market (in other markets, of course, normative influence may be much more relevant).

In both experiments, a song's appeal depends on two factors: first, the inherent attractiveness of its title, and second, its positioning on the screen, which we call availability. Availability is defined as the probability that a song i will be sampled in a given position p:

$$V_i = \sum_p \frac{l_{i,p}}{\sum_i l_{i,p}}$$

We find that in MusicLab, positioning matters: in Experiment 2, participants are more likely than random to click on songs at the top of the list than on those mid-way down. In Experiment 1, the grid interface, the general trend is the same, with a small spike in multiples of three, representing songs positioned on the left side of the screen.

A song's appeal reflects the probability that a participant will want to sample or try it. We can think of appeal as a function of the final listen counts in the independent condition, where:

$$A_i = l_i / \sum_k l_k$$

for $k$ = songs 1 through 48. Here, appeal simply represents the probability of sampling a song, due the attractiveness of the song title in each of the social worlds.

Quality, in turns, measures the conditional probability of download, which we derive directly from the independent world for each of Experiments 1 and 2:

$$q_i = D_i / l_i$$

Finally, we find the total run length of each experimental world, as well as the total download count, which round to an average of 2700 listens in Experiment 1, 2500 listens in Experiment 2, and 1000 downloads in the social conditions of both experiments (with slightly higher variance in total downloads across the eight worlds of Experiment 1).

**Model description: Polya urn**

Using these inputs, we model the dynamic download count of each song i over time, and use the final download counts to compute inequality and unpredictability. The model consists of two steps for each entrance of a listener to the market. These steps are repeated if a listener elects to try more than one song:

1. Select a song to sample, based on its appeal, position, and current download count
2. Choose whether or not to download the song, based on its quality

In the first step, a participant enters the market and chooses a song at random to sample based on a combination of its appeal $A_i$ the availability score of its current position $V_{i,j}$ , and its current download count $D_{i,j}$

The probability that song $i$ is sampled is

$$\frac{V_{i,t}(D_{i,t}+\alpha A_i)}{\sum_j V_{j,t}(D_{j,t}+\alpha A_j)}$$

Here $\alpha$ is a scaling factor, constant across all songs, which captures the strength of the social signal. As $D_i$ grows over the course of the experiment, its value contributes increasingly to the probability a song will be sampled.

In the second step, the user downloads the chosen song with probability $q_i$ , and As $D_i$ is incremented if a download occurs.

While we model *listens* rather than individuals with different listener types, this assumption has little effect on model outcomes. In other words, our model can be said to describe one listener at a time, who listens to at least one song, after which he can choose to repeat these steps, up to a total of 48 times, or to exit the market by not selecting a song.

The decision to listen to a song leaves no signal for others: a song's listen count is invisible to other participants. By contrast, download count is seen by users in the social influence condition (but not by those in the independent condition). So, a user arriving late to the market with social signal receives more information about the songs chosen for download by his peers than does an earlier entrant.

**Inequality and Unpredictability**

In the original experiment, inequality is defined by the Gini coefficient,

$$I = \frac{\sum_{i=1}^{S}\sum_{j=1}^{S}|D_i - D_j|}{2S\sum_{i=1}^{S} D_i}$$

where $D_i$ is the final download count, or market share, of song $i$, and $S$ the total number of songs.

Unpredictability is measured across multiple worlds, with the unpredictability for song $i$

where $m_{i,j}$ is the market share of song i in world j and total unpredictability

$$U = \sum_{j=1}^{S} U_j / S$$

Using the sets of inputs for Experiments 1 and 2, we simulate eight social influence worlds of 2700 listens in Experiment 1, and 2500 listens in Experiment 2 (with resulting download counts ranging from 900-1100) and compute the resulting inequality and unpredictability. To calculate these values for the independent conditions, we run the simulation without the effect of the visibly increasing download count (and its concomitant social effects), so that the probability of sampling song i is simply

$$\frac{V_{i,t}A_i}{\sum_j V_{j,t}A_j}$$

For each experiment, we find, through simulation, the value of alpha that offers the best fit for the values of unpredictability and inequality observed in the original experiment. We are able to replicate the values of inequality and unpredictability over the course of both experiments.

We observe a substantially higher alpha in Experiment 1 (songs displayed in a grid) versus Experiment 2 (songs displayed in a column), suggesting that the impact of a song's appeal is more important in the early stages of the market of Experiment 1. This could be due to the fact that all songs are visible on a single grid, and there is no need to scroll down a long list: a listener employs social information differently to make his choice, compared to the column layout of Experiment 2.

With a frugal model that parallels the decision-making process of the listener (who elects to sample a song based on its inherent appeal, its screen position, and how many others have downloaded it; then decides whether to download it based on its quality), we are able to reproduce the results of the original Experiment 2 with RMSE = 0.0012 for unpredictability and 0.0516 for inequality over the course of the market, and for Experiment 1, RMSE = 0.0017 for unpredictability and 0.093 for inequality.

To summarize the findings described thus far, we first determined, from the experimental data, that the perception of quality, which drives the propensity to download, is not influenced by social forces in this market. Second, with a single scaling factor, we were able to simulate results for inequality and unpredictability over the course of the experiment, suggesting that the dynamics of the market are one of an increasing impact of social factors as the experiment progresses. That is, over time, the weight of the download count grows relative to the appeal of songs in determining a listener's choice of music to sample. Finally, the positioning of songs has an impact, and in particular the screen layouts of Experiments 1 and 2 yield different scaling factors, suggesting that the way in which products are positioned impacts the magnitude of the social forces.

**Long-run dynamics**

In the short run, sampling in the MusicLab market is based largely on initial screen position and on the appeal of songs' titles.

In the longer run, in our model the download to listen ratio increases, suggesting that a larger proportion of higher quality songs are being sampled. Simulating 100,000 listens, the download count to listen count ratio rises significantly, to about 51 downloads per 100 listens in Experiment 2 (in the typical 2500-listen world, this ratio hovers around 39 downloads per listen). Because the number of listens is fixed in the simulation, the higher ratio indicates that a greater number of songs are being downloaded (and that higher quality songs are being sampled more frequently). Of course, in a real market, users may adjust their behavior as market conditions change: for example, they may sample more or fewer songs than earlier entrants.

When social influence is present, unpredictability sinks slightly (to a mean of .0083 with a standard deviation of .00043 on 100 runs after 100,000 listens in Experiment 2), while Gini rises (to a mean of 0.69 with standard deviation 0.033). The unpredictability of the non-social worlds declines significantly (after 100,000 listens in Experiment 2, it reaches a mean of .00005, or about 1% of its value at 2500 listens).

Here we have considered at the static picture of how individual leads to population level features. An important question for future work is how these connect when we look at the -dynamics- of human behavior, that is, the rates of change of individual behavior to the rate of change the system as a whole.

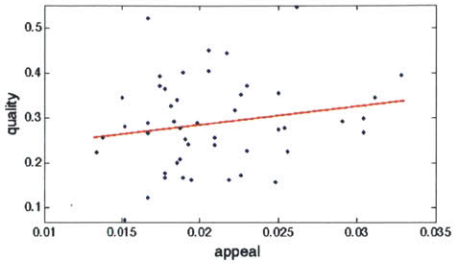



Figure 7.1: Quality and appeal are independent. Values are shown for quality and appeal corresponding to the 48 songs in Experiment 2, independent condition $R^2 = 0.012$

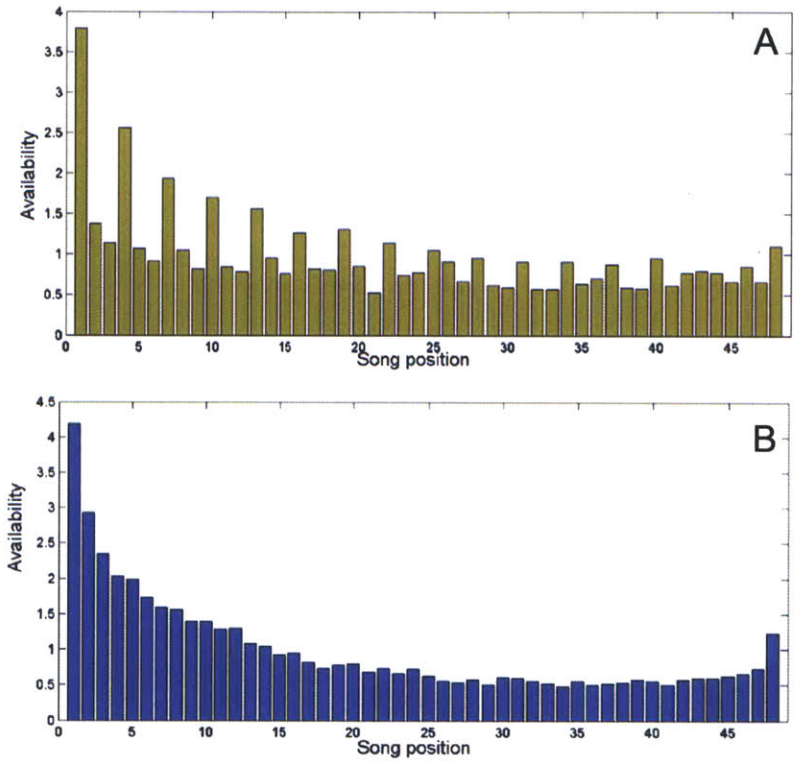



Figure 7.2: Availability in the independent world of Experiments 1 (A, top) and 2 (B, bottom), indexed to 1. The availability of a position n describes the likelihood that a song in that position will be sampled (where n=1 is the top left corner in Experiment 1, and the topmost position in Experiment 2, and n=48 is the bottom right corner in Experiment 1 and the bottom of the column in Experiment 2). Availability serves as a multiplier in calculating the total probability of a song being sampled, given its position-independent appeal, and its position at a given time in the market. In Experiment , songs on the left side of the grid are more likely to be sampled, all else equal, than songs on the right. In Experiment 2, songs at the top of the column, as well as the final song, are more likely to be sampled.

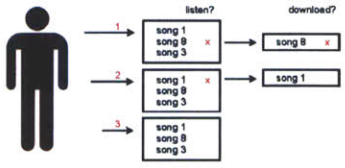



Figure 7.3: Song selection as a two-step process. A listen first selects which song(s) he will listen to, and after listening, decides whether or not to download the song. The first decision is made based on the appeal of a song; the second based on its quality. If a listener listens to more than one song, this process is repeated.

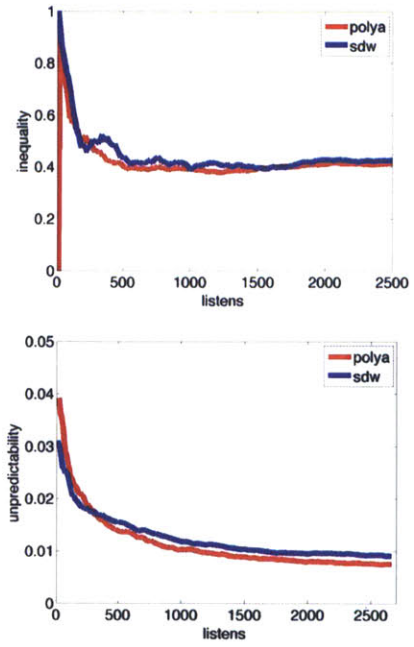



Figure 7.4: Inequality (top) and unpredictability (bottom) over the course of the market, with alpha = 900. Inequality is shown for Experiment 1, world 3. RMSE of simulated market's unpredictability is = 0.0017, and average of inequality is = 0.093

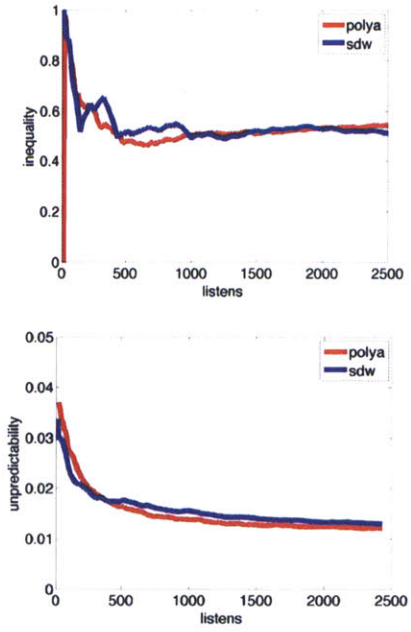



Figure 7.5: Inequality (top) and unpredictability (bottom) over the course of the market, with alpha = 200. Inequality is shown for Experiment 2, world 5. RMSE of simulated market's unpredictability is = 0.0012, and average of inequality is = 0.0516

# Bibliography


[1] I. Ajzen and M. Fishbein. Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin;*, 84(5):888, 1977.

[2] James P. Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PLoS ONE*, 7(5):e37676, 05 2012.

[3] A. Bandura, J.E. Grusec, and F.L. Menlove. Observational learning as a function of symbolization and incentive set. *Child Development*, pages 499–506, 1966.

[4] F. Bartumeus, MGE Da Luz, GM Viswanathan, and J. Catalan. Animal search strategies: a quantitative random-walk analysis. *Ecology*, 86(11):3078–3087, 2005.

[5] Michael Batty. Rank clocks. *Nature*, 444(7119):592–596, 2006.

[6] R.W. Belk. Situational variables and consumer behavior. *Journal of Consumer research*, pages 157–164, 1975.

[7] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

[8] E.H. Bonfield. Attitude, social influence, personal norm, and intention interactions as related to brand purchase behavior. *Journal of Marketing Research*, pages 379–389, 1974.

[9] Christian Borghesi and Jean-Philippe Bouchaud. Of songs and men: a model for multiple choice with herding. 2006.

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

[11] U.S. Economic Census. Establishments by naics and metropolitan area, 2007.

[12] M. De Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. In *Proceedings of the Nokia Mobile Data Challenge Workshop.*, 2012.

[13] K. De Queiroz. Ernst mayr and the modern concept of species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(Suppl 1):6600–6607, 2005.

[14] Morton Deutsch and Harold B. Gerard. A study of normative and informational social influences upon individual judgement. *Journal of Abnormal & Social Psychology*, 51(3):629–36, 1955.

[15] R.O. Doyle. Free will: it's a normal biological property, not a gift or a mystery. *Nature*, 459(7250):1052–1052, 2009.

[16] N. Eagle, M. Macy, and R. Claxton. Network Diversity and Economic Development. *Science*, 328:1029–, May 2010.

[17] Gerd Gigerenzer and Daniel G. Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. 1996.

[18] E.L. Glaeser and D.C. Mare. Cities and skills. Technical report, National Bureau of Economic Research, 1994.

[19] Daniel G. Goldstein and Gerd Gigerenzer. Models of ecological rationality: The recognition heuristic. 2002.

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

[21] A.M. Graybiel. Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.*, 31:359–387, 2008.

[22] Andrew G Haldane and Robert M May. Systemic risk in banking ecosystems. *Nature*, 469(7330):351–355, 2011.

[23] M. Heisenberg. Is free will an illusion? *Nature*, 459(7244):164–165, 2009.

[24] Kenneth Hendricks, Alan Sorensen, and Thomas Wiseman. Observational learning and demand for search goods. *American Economic Journal: Microeconomics*, 4(1):1–31, 2012.

[25] E. Alejandro Herrada, Claudio J. Tessone, Konstantin Klemm, Victor M. Eguiluz, Emilio Hernandez-Garcia, and Carlos M. Duarte. Universal scaling in the branching of the tree of life. *PLoS ONE*, 3(7):e2757, 07 2008.

[26] S.S. Iyengar and M.R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6):995, 2000.

[27] J. Jacoby and R.W. Chestnut. *Brand loyalty measurement and management*. Wiley New York, 1978.

[28] W. James. The principles of psychology. 1952.

[29] Dixit Avinash K. and Stigler George J. Monopolistic competition and optimum product diversity.

[30] Mark Kirkpatrick and Montgomery Slatkin. Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. *Evolution*, 47(4):1171–1181, 1993.

[31] Andras Kornai. How many words are there? 2002.

[32] Paul Krugman. Scale economies , product differentiation , and the pattern of trade. *American Economic Review*, 70(5):950–959, 1980.

[33] A. Lempel and J. Ziv. On the complexity of finite sequences. *Information Theory, IEEE Transactions on*, 22(1):75–81, 1976.

[34] Paolo Leon. *Structural Change and Growth in Capitalism.* John Hopkins University Press, 1964.

[35] M. Li and P.M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications.* Springer-Verlag New York Inc, 2008.

[36] X. Lu, L. Bengtsson, and P. Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 2012.

[37] A. Marshall. Principles of economics, an intorductory volume. 1920.

[38] A. H. Maslow. *Motivation and personality.* Harper and Brothers, 1954.

[39] Robert M May and Nimalan Arinaminpathy. Systemic risk: the dynamics of model banking systems. *Journal of the Royal Society Interface the Royal Society*, 7(46):823–838, 2010.

[40] Robert M. May, Simon A. Levin, and George Sugihara. Complex systems: Ecology for bankers. *Nature*, 451(7181):893–895, 02 2008.

[41] Ernst Mayr. The biological meaning of species*. *Biological Journal of the Linnean Society*, 1(3):311–320, 1969.

[42] P.W. Miniard and J.B. Cohen. Modeling personal and normative influences on behavior. *Journal of Consumer Research*, pages 169–180, 1983.

[43] Albert M. Muniz and Thomas C. OÕGuinn. Brand community. *Journal of consumer research*, 27(4):412–432, 2001.

[44] P. Nelson. Information and consumer behavior. *The Journal of Political Economy*, 78(2):311–329, 1970.

[45] I.P. Pavlov. Lectures on conditioned reflexes. vol. ii. conditioned reflexes and psychiatry. 1941.

[46] M.J. Ryan and E.H. Bonfield. The fishbein extended model and consumer behavior. *Journal of Consumer Research*, pages 118–136, 1975.

[47] Serguei Saavedra, Daniel B. Stouffer, Brian Uzzi, and Jordi Bascompte. Strong contributors to network persistence are the most vulnerable to extinction. *Nature*, 478(7368):233–235, 10 2011.

[48] Matthew J. Salganik, Peter S. Dodds, and Duncan J. Watts. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, February 2006.

[49] Matthew J. Salganik and Duncan J. Watts. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly*, 71(4):338–355, 2008.

[50] Matthew J. Salganik and Duncan J. Watts. Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science*, 1(3):439–468, 2009.

[51] S. Senecal and J. Nantel. The influence of online product recommendations on consumers online choices. *Journal of Retailing*, 2004.

[52] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[53] Herbert A. Simon. On a class of skew distribution functions. 1955.

[54] Lones Smith and Peter Sørensen. Pathological outcomes of observational learning. *ECONOMETRICA*, 68:371–398, 1999.

[55] C. Song, Z. Qu, N. Blumm, and A.L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[56] George Sugihara, Louis-Félix Bersier, T Richard E Southwood, Stuart L Pimm, and Robert M May. Predicted correspondence between species abundances and dendrograms of niche similarities. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5246–5251, 2003.

[57] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, January 1981.

[58] Glen L. Urban, Theresa Carter, and Zofia Mucha. Market share rewards to pioneering brands : an empirical analysis and strategic implications. 1983.

[59] GM Viswanathan, S.V. Buldyrev, S. Havlin, MGE Da Luz, EP Raposo, and H.E. Stanley. Optimizing the success of random searches. *Nature*, 401(6756):911–914, 1999.

[60] D.T. Wilson, H.L. Mathews, and J.W. Harvey. An empirical test of the fishbein behavioral intention model. *Journal of Consumer Research*, pages 39–48, 1975.

[61] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge MA, 1949.