Adaptive Rationality in Communication

by

Charlie Pilgrim

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy in Mathematics of Systems

Mathematics of Real World Systems CDT

June 2023

THE UNIVERSITY OF
WARWICK

The mind is a glass floor.
The mind is the spirit's tear.
The mind is our prior and subsequent ghost.
The mind is the Bullion Express and the blood on the tracks.
The mind is a stone door.
The silver on the backs of mirrors.
The wave that defines the coast.
It's what the drunk grave robbers couldn't stuff in their sacks.
The mind is the sum of all and more.
The spasm between one and zero in the Calendar of Black-Hole Years.
The contract between the lash and the whipping post.
A quilt of dreams stitched with facts.
. . .
The mind is what thought is for.
The parking lot at the Mall of Fears.
The fire-pit for the piggy roast.
What the soul surrendered and won't take back.
The mind is neither either nor or.
The real center of an empty sphere.
— Jim Dodge, Stone Junction

Contents

List of Tables

List of Figures

Acknowledgments

I would foremost like to thank my supervisor Thomas for his enthusiasm, joy, interest, insight and time. I feel much richer for knowing Thomas and I will do my best to live up to his example as an academic.

Throughout the PhD I felt supported and encouraged by the academic community at Warwick and beyond. Especially Paolo Turrini, who believed in me early on and is always a pleasure to be around. I'd also like to thank Weisi Guo, Daniel Sgroi, Adam Sanborn, Colm Connaughton, Stefan Grosskinsky, Magnus Richardson, Yulia Timofeeva, Matthew Turner, Robin Thompson, Ed Brambley, Mirta Galesic, Heather Robson and Jade Perkins.

I am also thankful for the PhD students I met along the way. Especially Eugene Malthouse, with his positive, pragmatic and can-do attitude. And Peter Strong, who helped me through the week to week struggles of PhD life. There really are too many to mention here — all of the people from Warwick, The Alan Turing Institute and University College London.

Finally I would like to thank my family. My Mum, Dad, Michael, Debbie, Kelly, Steve, Leo, Melody and Gabriella. Thank you for being there and supporting me.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy in Mathematics of Systems. It has been composed by myself and has not been submitted in any previous application for any degree. Parts of this thesis have been published or are currently under review:

- Chapter 2 has been published as Pilgrim C, Guo W, Hills T.T, "The Rising Entropy of English in the Attention Economy" on Arxiv. And is under review at a journal.
- Chapter 3 has been published as Pilgrim C, Hills T.T, "Bias in Zipf's Law Estimators" in Nature Scientific Reports.
- Chatper 4 has been published as Pilgrim C, Sanborn A, Malthouse, E, Hills T.T, "Confirmation Bias Emerges from an Approximation to Bayesian Reasoning" on PsyArxiv. And is under review at a journal.

Abstract

Modern telecommunications have transformed the way that people communicate. The situation is dynamic, with a rapidly changing technological and cultural landscape. Furthermore, interactions between this landscape and human behaviour are complex and difficult to predict. The work in this thesis is inspired by the general problem of describing these systems.

We begin with an investigation into historical trends in language, finding that the word entropy of American English has increased steadily since around 1900. We also find differences in word entropy across media categories. These changes are explored in the context of the attention economy, which is the dynamic of increasing competition for human attention in response to a rising abundance of information. A model of information foraging in the attention economy is developed to describe the trends in word entropy.

Word entropy is a property of word distributions, which follow Zipf's law: a power law relationship between the frequency of words and their rank in that frequency distribution. As well as word entropy, we also see trends in changes in Zipf's law over the 20th century. There is difficulty in investigating these trends due to bias in estimators of the Zipf exponent. The source of this bias is explored and shown to be due to inappropriate assumptions in the estimators. The correct estimator is derived but found to be computationally intractable.

Modern advances in information search and social media have been implicated in the creation of separated silos of thought in society and the driving of dangerous political polarisation. This is connected to confirmation bias, which is the tendency for people to search for and consume information in a way that maintains their existing beliefs. A model of confirmation bias is developed that is based on a boundedly rational model of belief updating, which takes into account cognitive limitations.

Chapter 1

Introduction

1.1 Motivation

When was the last time you looked at your phone? In the modern world we are ever-connected, with increasing daily hours of screentime [Rideout et al., 2010] and rising global mobile phone use [Agar, 2013]. This is all driven by recent advances in telecommunications technology [Bawden and Robinson, 2008] that have changed how we communicate and relate to the world [Lacohée et al., 2003], with some users even experiencing their smartphone as an extension of the self [Park and Kaye, 2019]. But does this represent a bright beacon of human progress or are we turned-on, plugged-in and zoned-out?

Modern communication technology provides us with many valuable benefits including entertainment [Vorderer, 2001], education [Hills, 2019] and sense-making [Hills, 2019]. But there is also a "dark side" to communication [Bawden and Robinson, 2008; Hills, 2019] and associated negative effects including addiction [Andreassen et al., 2016; van den Eijnden et al., 2016], anxiety [Bawden and Robinson, 2008; Woods and Scott, 2016; Anderson et al., 1980] and attention disorders [Andreassen et al., 2016; Ra et al., 2018].

At the social level, communication underpins our ability to work together. Effective communication can help us to "put our heads together" to collectively find solutions to problems [Hargadon and Bechky, 2006]. Working together is more important than ever as we face crucial collective challenges including climate change, global conflict and advanced artificial intelligence. But there is a dark side here too, with negative aspects that can work to drive people apart through misinformation [Wang et al., 2019], conspiracy beliefs [Shahsavari et al., 2020], radicalisation [Thompson, 2011], and political polarisation [Hills, 2019].

Why would we build communication systems that do not serve us? For one thing, it can be fundamentally difficult to predict or control these systems, made up as they are of complex interactions between people, culture, ideas and technology. These interactions are not static, and the situation is constantly changing and evolving. Additionally, media producers have mixed incentives that include a struggle to capture human attention [Evans, 2020]. Understandably, human behaviour is not necessarily well suited to flourish in this modern media environment.

Over 50 years ago Simon et al. [1971] pointed out that in the modern world people have access to an abundance of information, in stark contrast to our ancestral evolutionary environment where information was relatively scarce. This trend has only continued in recent times with the advent of digital computing and the internet, and today many of us have instant access to an incredible wealth and diversity of information [Bawden and Robinson, 2008].

An abundance of information creates a scarcity in what information consumes — attention [Simon et al., 1971]. Human attention is valuable and its scarcity creates competition in what has been termed the *attention economy* [Goldhaber, 1997]. Money flows with attention and there is a pressure on media companies to capture both, which is only exacerbated by ad-supported media [Evans, 2020] and the quantification of attention [Terranova, 2012].

In this context, the motivation for the work in this thesis is to make a contribution towards understanding the nature of communication in the modern world. The hope is that a better understanding can help to avoid the negative consequences outlined above. The greater hope is that we can build better communication systems that work *with* human behaviour to improve all of our lives through better media experiences. The even greater hope is that we can unlock human potential through enhanced collective problem solving that can help us meet the pressing challenges facing humanity.

1.2 Chapter Overview

If media is competing for human attention then what are the dimensions of this competition? There is evidence that human attention is attracted to information that is belief-consistent [Hills, 2019; Taber and Lodge, 2006], negative [Hills, 2019; Davis and McLeod, 2003], social [Hills, 2019; Davis and McLeod, 2003], predictive [Hills, 2019] and information dense [Itti and Baldi, 2009; Radach et al., 2003]. It is the last of these, information density, that we will focus on in **Chapter 2, "The Rising Entropy of English in the Attention Economy"**. We find evidence that the information density (or word entropy) of American English is rising, and we develop a model of the attention economy to account for these trends.

Specifically, in Chapter 2 we will ask how word distributions are changing in the attention economy. We consider 3 measures of the lexical diversity of word distributions: word entropy, type token ratio, and Zipf exponent [Bentz et al., 2015]. However, computational measures of the Zipf exponent are biased [Hanel et al., 2017; Corral et al., 2019; Piantadosi, 2014] and it is unclear how to account for these biases. **Chapter 3, "Bias in Zipf's Law Estimators"** investigates the source of this bias and explores potential avenues for building better estimators.

Beyond information density, human attention is also attracted to belief-consistent information [Hills, 2006; Taber et al., 2009; Hart et al., 2009]. This is known as confirmation bias in the selection of sources. More generally, confirmation bias is "the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" [Nickerson, 1998]. Confirmation bias has been connected to political polarisation [Del Vicario et al., 2017], which is a growing social problem that is exacerbated by modern media environments [Settle, 2018, Chapter 4]. **Chapter 4, "Confirmation Bias Emerges from an Approximation to Bayesian Reasoning"**, presents a cognitive model of confirmation bias, including an explanation for the bias for belief-consistent information.

1.3 Approach

Communication is a human problem. When thinking about human (and animal) behaviour we can make a distinction between proximate and ultimate causes [Mayr, 1961; Laland et al., 2011]. A proximate cause answers the "how" question, or describes the mechanisms of behaviour. An ultimate cause is concerned with the evolutionary reason "why" a behaviour exists, or what adaptive benefit it serves. A classic example is the migration of birds [Mayr, 1961], which can be understood both in terms of proximate mechanisms that spur an individual bird to travel, related to the endocrine system and the shortening length of the day; and the ultimate evolutionary cause of the adaptive advantage found in migrating to a place with more food available. While the distinction is useful, a full behavioural explanation will include both proximate and ultimate causes in a complementary way.

One approach that bridges the gap between proximate and ultimate causes is **adaptive rationality**. The view is that the human mind is well adapted to efficiently solve problems that were regularly encountered by our ancestors [Haselton et al., 2009]. To put this in another way, the proximate mechanisms of human cognition can be understood as solutions to the ultimate goal of survival and reproduction in our evolutionary past. Adaptive rationality is similar to bounded rationality [Simon, 1990] and resource rationality [Lieder and Griffiths, 2020]. All share the perspective that cognition has to work within cognitive constraints that include limitations on information and computational resources. The ultimate evolutionary goal of propagation of genes is influenced not only by the quality of cognition but also it's efficiency [Daw et al., 2008].

Adaptive rationality is connected to the idea of **heuristics**, which are cognitive strategies that perform well while being efficient. Fundamentally, the evolutionary function is not on truth seeking or perfect computation but instead on survival and propagation of genes [Friedrich, 1993]. In order to meet a specific evolutionary challenge there may be a variety of potential cognitive and behavioural strategies. Good heuristics have clear adaptive advantages over more computationally demanding forms of cognition — they are quick, efficient and can even outperform more complex strategies [Haselton et al., 2009; Bouskila and Blumstein, 1992].

However, heuristics might not perform as well when taken out of the environment for which they are adapted [Haselton et al., 2009]. We see this in behavioural experiments where researchers need to be careful about overgeneralising findings in artifical settings to behaviour in the real world [Orne, 1962]. This applies not only to behavioural experiments but also the world at large. Behavioural adaptations that are well adapted to the context of the ancestral evolutionary environment may not necessarily generalise well to other contexts (such as the modern world). When studying human behaviour from an adaptionist perpective, we can consider how behaviour observed in the modern world might be expressions of adaptations to the ancestral evolutionary environment.

1.4 Background

This section provides a broad overview of the relevant background to the research presented in the thesis. Additional and more concise context is provided in each chapter.

1.4.1 Information Foraging

Chapter 2, "The Rising Entropy of English in the Attention Economy", asks how people react to the rising abundance of information in the attention economy. To shed light on this question we can turn to information foraging [Pirolli and Card, 1999], a range of models that describe how people manage, search for, and consume information. These models are based on food foraging models from ecology. The justification for using these models in the context of information is that food and information foraging are similar search problems, and that human behaviour is optimised for information foraging in a similar way that animal (and human) behaviour is optimised when foraging to food [Pirolli and Card, 1999]. This is connected to the idea that human beings are informavores, i.e. that our adaptive success is tied to our ability to efficiently process information [Dennett, 2008; Pinker, 2003; Pirolli and Card, 1999]. Pirolli and Card [1999] suggested that human information foraging behaviour might be an evolutionary adaptation that built on existing cognitive mechanisms that govern food foraging. This connection was later supported by comparative biological studies of neural architecture controlling spatial foraging and the cognitive control of attention [Hills, 2006; Hills et al., 2015].

In order to understand information foraging models it is helpful to first describe food foraging models. Optimal foraging theory is a range of models in ecology that make the assumption that food foraging behaviour in animals is optimised to maximise the rate of calorie intake (see e.g. [Stephens and Krebs, 1986]). Given the structure of the food foraging environment, predictions can be made about the expected behaviour of foragers, with good empirical support in a range of contexts in animals [Werner and Mittelbach, 1981; Stephens and Krebs, 1986] and humans [O'Connell and Hawkes, 1981; Kaplan and Hill, 2017; Smith et al., 1983; Winterhalder, 1986]. For example, Charnov's marginal value thoerem [Charnov, 1976] predicts how long foragers will spend in different patches of food (e.g. a raspberry bush) before moving on to the next patch (e.g. when all the easy to reach raspberries have been harvested). Chapter 2 focuses on a different foraging model called the prey choice model [MacArthur and Pianka, 1966], explained below.

The prey choice model asks how an optimal forager should choose which prey items to consume [Stephens and Krebs, 1986; MacArthur and Pianka, 1966]. Each type of prey is assumed to have an expected "handling" time (how long it will take to hunt and eat), as well as an expected utility (in the form of calories), and an expected prevalence (how often these prey items are encountered). From these constraints, we can derive which prey items will be included in the diet of an optimal forager, and which prey will be ignored. The derivation (which is shown in Chapter 2) results in a "diet condition", which states that an optimal forager will only pursue prey with a minimum "profitability" (the expected utility of the prey divided by the expected time to handle the prey). This can help explain animal behaviour such as the fact that oystercatchers, a type of bird whose diet can include mussels, will ignore small mussels (that have little meat) and also ignore large mussels (that take a long time to break into) but instead prefer medium sized mussels (which have some meat and can be eaten quite quickly) — the medium sized mussels are the most “profitable” prey [Meire and Ervynck, 1986]. Another consequence of the prey choice model is that in an abundant prey environment the diet condition is higher, while in a scarce environment the diet condition drops [Stephens and Krebs, 1986], i.e. a starving oystercatcher will be less picky about its mussels.

Following a conceptual review article by Sandstrom [1994], Pirolli and Card [1999] pioneered the use of food foraging models to describe information foraging in human behaviour. Analogously to food foraging, a key assumption is that people optimise their utility rate (or specifically maximise their rate of gaining valuable information). This raises a difficulty in comparision to food foraging — we can easily quantify the calories in a prey item, it is much less clear how to measure the value of information, which is subjective and more multidimensional. We do not necessarily need to measure the sum total of value. Instead, we focus on information rate or density, which can be considered one dimension of the total value of information. And we have reason to believe that human attention will be attracted to high density information from fundamental adaptive arguments as well as behavioural eye tracking experiments [Radach et al., 2003; Itti and Baldi, 2009]. We can measure information density using entropy, a concept from information theory.

1.4.2 Information Theory

The central text in the field of information theory is the article “A Mathematical Theory of Communication” by Shannon [1948], which is concerned with the engineering challenge of transmitting information from a source to a receiver. Shannon provides a definition of a quantity of information that is related to the reduction in uncertainty that the information provides,

$$I = -log(p(x))\,,$$

where $p(x)$ is the prior probability of receiving the message x. This measure has the benefit of being relatively intuitive, is independent of the coding scheme used to transmit information and is additive. For example, if a message can have one of 8 values then the information contained in the message is $log8$ (assuming each of the values is equally likely). This can be encoded in binary as 3 symbols such that each symbol has one of 2 possible values i.e. 0 or 1 (a binary string of length 3 has 8 possible values). The amount of information transmitted in the 3 binary symbols is $3log2 = log8$. Relatedly, a “bit” of information is simply the information transmitted by 1 binary symbol, which is equivalent to the information measure with a logarithmic base of 2. In our example a message with 8 possible values can be conveyed with 3 bits of information.

To quantify the information density of a given coding scheme we can simply take an expectation over the information per symbol,

$$H = -\sum_{x} p(x) log(p(x)) .$$

This gives the **entropy**, H, a name taken from statistical mechanics (the thermodynamic entropy is of the same form). Entropy is a deep concept with applications to disparate fields [Pierce, 2012]. For our purposes we can think of entropy as an information rate or density.

Finally **mutual information** is the amount of information shared between two variables. More precisely, $I(X, Y)$ is the expected reduction in uncertainty we learn from variable X by learning the value of variable Y (or vice-versa). This is an important concept when thinking about how to encode information into symbols for transmission — we might want to maximise the mutual information between the signals in the coding scheme and the message being communicated [MacKay, 2003]. In Chapter 2, "The Rising Entropy of English in the Attention Economy" we will discuss a linguistic model by Cancho and Solé [2003] that considers the mutual information between a) a set of objects that we want to refer to and b) the symbols (or words) that are associated with those objects. This model is discussed further below.

1.4.3 The Entropy of Language

Human language has many proposed functions beyond transmitting information, not limited to facilitating thinking [Bloom and Keil, 2001] and maintaining social relationships in large groups [Dunbar, 2004]. While human communication (and language) is more complicated than the simple model of communication provided by Shannon, the general mathematical constraints do apply (at least approximately). While remembering these limitations, we can carefully apply Shannon's theory to language.

From an information theory perspective, language can be thought of as a coding scheme that consists of a series of symbols (or words) which transmit information between a sender and receiver. We can ask what is the information density (or entropy) of words in language. The calculation of entropy (equation 1.4.2) requires values for the probability of receiving each symbol, $(p(x))$. The probability of a word appearing is highly context dependent and depends on the preceding words and paragraphs. In an engineering environment, the entropy of a source can be calculated exactly if we know the mechanism through which the source generates information [Shannon, 1948]. Human language is generated by cognition and we are not able to interrogate the generative mechanism so easily (although Shannon [1951] used a "guessing game" behavioural experiment to do just this).

Where we don't have access to the data generating mechanism of a source, we can estimate entropy by looking at a long enough sequence of symbols generated by the source [Schürmann and Grassberger, 1996]. In principle, one can estimate the probability of a word appearing given the preceding words. However, in order for this estimator to be accurate one would need a lot of data. For example, given the sentence "the dog barks at the ", we can predict that the next word is likely to be "noise" or "postman" or "cat". But in order to give reliable estimates for these probabilities we would need to find many examples of sentences of this exact format, which would require a very large amount of text. And this only considers the context of the preceding 5 words. Language is more complicated and has long-range correlations between words [Ebeling and Pöschel, 1994]. For example, if the word "post" was mentioned over 100 words ago in the preceding paragraph then we might predict that it is more likely that "the dog barks at the postman".

While we may not be able to accurately estimate the actual entropy of a word, we can use maximum likelihood to estimate a N-gram entropy by taking into account a limited window of previous words and counting the frequencies of a word appearing given the context in the window [Shannon, 1951]. The 1-gram or unigram word entropy considers no context and simply estimates the probabilities of words based on their total observed frequencies in a text sample. The bigram word entropy takes into account one word of context (so a window of two words in total). The trigam word entropy considers a window of three words, and so on. As more words are considered the amount of text that is needed for accurate estimates rises significantly.

In Chapter 2, "The Rising Entropy of English in the Attention Economy", we use the maximum likelihood estimator for the unigram word entropy. The argument for using this estimator is that it correlates with the experienced level of novelty or repetition by a reader. Additionally, it is difficult to find lots of very large text samples, especially when looking at historical trends. In Chapter 2 we used text samples of $N = 2000$ words, and the estimators of higher N-grams have little meaning with text samples of this size.

1.4.4 Zipf's Law

Unigram word entropy is a measure on the word distribution of natural language. In Chapter 2 we also estimate 2 alternate measures: Zipf exponent and type token ratio. Zipf's law describes the empirical relationship of a power law between the frequency (number of occurrences) of words f in natural language and the empirical rank of those words in the frequency distribution r_e,

$$f(r_e) \propto r_e^{-\alpha} . \tag{1.1}$$

The Zipf exponent, α, is usually around 1 although does vary across text samples [Ferrer i. Cancho, 2005; Montemurro and Zanette, 2002]. We investigate this variation in Chapter 2.

A reasonable question to ask is why Zipf's law appears: what is the underlying process from which this statistical pattern emerges? Power laws can emerge from a variety of different processes including e.g. preferential attachment, phase transitions, combinations of exponentials [Newman, 2005], and fractals [Brown et al., 2002]. Accordingly a variety of explanations have been proposed for Zipf's law [Piantadosi, 2014]. At one extreme we have very simple but unrealistic models of Zipf's law such as random typing of letters and spaces to form words [Miller, 1957]. In random typing longer words are less likely to appear than short words and the word distribution forms a power law (this process can also be described by a fractal probability tree [Mandelbrot, 1982]). Zipf [1949] put forward the idea that the pattern emerges from *the principle of least effort* and the aim of communicating with the least amount of work.

The principle of least effort was expanded on more recently [Cancho and Solé, 2003] in a way that incorporates information theory to balance the preferences of speakers and listeners. In this study it is argued that listeners prefer highly informative messages while speakers prefer messages with low entropy because they require less effort to generate. The authors use a model that considers a vocabulary of symbols that can be assigned to refer to a set of objects. Through simulation they assign these symbols in a way that simultaneously a) maximises the mutual information between the objects and signals and b) minimises the entropy of the signals. They found that when these two objective functions are more or less balanced there is a phase transition and a power law in the signal frequency distribution emerges, just as in Zipf's law. Chapter 2 borrows the assumption that listeners prefer highly informative messages, and connects this idea to information foraging models.

1.4.5 Zipf's Law Estimators

If we take logarithms of Zipf's law (Equation 1.1), we find a linear relationship,

$$logf(r_e) = -\alpha log(r_e) + logC\,, \tag{1.2}$$

where C is a constant. If we plot this relationship on a log-log graph then we will find a straight line. Given some data, one approach to fit the Zipf exponent, α, is to first log transform the data and then use ordinary least squares regression to find an estimator for the exponent. However, the assumptions underpinning linear regression do not apply in this case [Clauset et al., 2009]. Specifically, homoscedasticity is unlikely to apply to the errors in the dependent variable following a log transformation. For example, if errors are Gaussian in linear space, they are no longer Gaussian in log space and this assumption no longer holds. As such, ordinary least squares regression of power laws can involve large errors in estimates of the power law exponent [Clauset et al., 2009].

An alternative method to fitting power laws involves maximum likelihood estimation [Clauset et al., 2009]. This overcomes the problems relating to the inappropriate assumptions in ordinary least squares regression. However, the power law estimators also make inappropriate assumptions and have been shown to be biased in the case of Zipf's law [Hanel et al., 2017; Corral et al., 2019]. Specifically, the estimator given by Clauset et al. [2009] assumes that errors in the dependent and independent variable are independent. However, in Zipf's law the power law is between the frequency of words and the rank of words in that frequency distribution. If a word is randomly oversampled and has a higher than expected empirical frequency, it can also move up the empirical frequency ranking, i.e. errors in the frequency and frequency-rank are correlated. This introduces a bias, which is investigated thoroughly in Chapter 3, "Bias in Zipf's Law Estimators".

1.4.6 Beliefs

While information theory is one perspective on human language, it doesn't capture meaning [Shannon, 1948]. Human beings have beliefs that are updated as we receive data about the world through sensory information and communication [L Griffiths et al., 2008]. A common approach to modelling inference behaviour is Bayesian probability theory. Bayes theroem provides the "rational" approach to updating beliefs about whether some hypothesis is true, $P(H)$, given data, D,

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}\,. \tag{1.3}$$

Bayesian updating of beliefs given data is optimal for a rational agent (see e.g. L Griffiths et al. [2008]). We can therefore use this model as a starting point for how we might expect people to make inferences, considering the assumption that historical adaptive pressures drove human cognition towards optimal behaviours [Haselton et al., 2009].

The formalisation in equation 1.3 considers just one belief (or hypothesis) about the world. However, people have lots of beliefs about the world and it is rational to take multiple beliefs into account when making inferences from data [Gershman, 2019]. The canonical example is that if a scientist detects faster than light travel then they might question their beliefs in the quality of their measuring instruments before questioning their belief in the lightspeed limit.

Questioning the validity of measurements is a way of "explaining away" data that is inconsistent with an existing set of beliefs. More generally, people can question the reliability of a source of information. There is evidence that people do hold beliefs about source reliability [Mahoney, 1977; Liberman and Chaiken, 1992; Taber and Lodge, 2006; Lord et al., 1979]. Source reliability, or other beliefs, can be incorporated into Bayes theorem [Merdes et al., 2020; Hahn et al., 2018]. And the interactions between beliefs can be captured by Bayesian networks [Pearl, 2009].

One problem with Bayesian networks as models of human cognition is that the computation of inference in such situations can become intractable [Cooper, 1990]. That is, the computation required to calculate probabilities given data becomes infeasible due to taking too long or using up too much memory space. As such, when modelling human inference it is reasonable to take an adaptive rationality approach [Haselton et al., 2009] and take into account cognitive limitations. Specifically, models should consider how human cognition can efficiently approximate otherwise computationally intractable inferences [Daw et al., 2008]. This is the approach taken in Chapter 4, "Confirmation Bias Emerges from an Approximation to Bayesian Reasoning".

1.4.7 Summary

In summary, the chapters in the thesis all relate to research problems encountered when asking how human behavior interacts with modern communication systems. The main approach to modelling human behaviour is adaptive rationality, which assumes that human cognition tends towards rationality, although limited by cognitive constraints. The chapters that follow were published as academic articles and they retain that structure, which is a suitable way to present the work. This includes abstracts, introductions, discussions and supplementary information for each chapter. Chapter 5, “Discussion”, summarises and brings together the results of the main content chapters, and explores possible future work.

Chapter 2

The Rising Entropy of English in the Attention Economy

2.1 Abstract

We present evidence that the word entropy of American English has been rising steadily since around 1900, contrary to predictions from existing sociolinguistic theories. We also find differences in word entropy between media categories, with short-form media such as news and magazines having higher entropy than long-form media, and social media feeds having higher entropy still. To explain these results we develop an ecological model of the attention economy that combines ideas from Zipf's law and information foraging. In this model, media consumers maximize information utility rate taking into account the costs of information search, while media producers adapt to technologies that reduce search costs, driving them to generate higher entropy content in increasingly shorter formats.

2.2 Introduction

Word entropy is a measure of the amount of repetition (low entropy) or novelty (high entropy) in word distributions. Empirical word distributions typically follow Zipf's law, which describes a power law between a word's observed frequency and that word's rank in the frequency distribution [Zipf, 1949]. This empirical power law is remarkably stable with an exponent around 1 [Bentz et al., 2015; Baixeries et al., 2013; Ferrer i. Cancho, 2005]. The stability of Zipf's law suggests some underlying mechanism, and Zipf himself hypothesised a *principle of least effort* between speakers and listeners. More recently this principle has been expanded to show that power laws in word distributions can emerge from a balance between maximising the benefits of receiving highly informative messages (preferred by listeners) and minimising the costs of generating high word entropy text (preferred by speakers) [Ferrer i. Cancho, 2005].

In recent times this balance between the efforts of listeners and speakers has changed. Modern communication systems have transformed the way that we share and consume information, in particular by increasing the accessibility of information [Hills, 2019]. In the words of Herbert Simon this creates a "poverty of attention" [Simon et al., 1971], such that media producers must compete for the limited resource of human attention [Evans, 2020; Ciampaglia et al., 2015; Terranova, 2012]. This dynamic has been called *the attention economy*, a combination of forces influencing the production and consumption of information, with consequences including a shortening collective attention span [Lorenz-Spreen et al., 2019]. If information adapts to the balance between the preferences of media producers and consumers, then increased competition for attention tips the balance toward the preferences of the consumers. That is, information markets (the distribution of available content) should rise in information density, and specifically, entropy.

We can envision this adaptive process in terms of *information foraging* [Pirolli and Card, 1999; Sandstrom, 1994]. Information foraging describes how people search for and consume information in different environments, including web browsing [Pirolli, 2009b] software debugging [Lawrance et al., 2010a,b; Piorkowski et al., 2013], and the design of information and social environments [Pirolli, 2009b; Piorkowski et al., 2013; Bhowmik et al., 2015]. The basic rationale of this approach is borrowed from ecological models of foraging, which have been shown to be appropriate to a wide range of search problems ranging from spatial foraging to cultural evolution [Hills et al., 2015]. Indeed, handling the exploration versus exploitation trade-off that is common to all of these environments has been proposed to be a defining selective force in the evolution of cognition [Hills, 2006; Todd and Hills, 2020].

In what follows, we first investigate the evolution of information across a wide variety of media sources over the last two centuries, a time marked by increasing media competition. We show how this reveals a characteristic pattern of rising entropy that affects different categories of media in different ways (e.g., books versus news versus social media). We then create a model of the attention economy that expands on existing models of information foraging to incorporate competition for human attention between media producers. This model explains both the general increase in word entropy and the differences in word entropy across categories.

2.3 Materials and Methods

Text Corpora

To investigate the recent history of information evolution we examine a variety of text corpora. The Corpus of Historical American English (COHA) [Davies, 2012] has $116,614$ texts spanning the 1810s to 2000s, balanced between categories of fiction ($n = 11,010$), non-fiction ($n = 2,635$), news ($n = 41,677$) and magazines ($n = 61,292$). The Corpus of Contemporary American English (COCA) has over 150,000 texts from between 1990 to 2008 split between fiction, popular magazines, newspapers, academic journals and spoken word [Davies, 2009]. For our analysis we used a publicly available sample of COCA with $2,362$ texts split between categories of fiction ($n = 275$), academic journals ($n = 266$), news ($n = 872$) and magazines ($n = 949$). The British National Corpus (BNC) contains $8,098$ texts from between 1960 and 1993 including written categories of fiction ($n = 904$), academic prose ($n = 994$), newspapers ($n = 972$), non-academic prose and biography, other published materials and unpublished materials [Burnard, 2007]. Fiction and newspapers are common categories across the corpora. Magazines are a common category between COHA and COCA. We grouped as non-fiction the categories of COHA non-fiction, COCA academic journals and BNC academic prose.

The text sample data was cleaned before analysis in a standard way [Gerlach and Font-Clos, 2020]. COHA and COCA are similar formats and so followed the same procedure. For both:

- Stripped any headers not a part of the main chapter text samples.
- Removed any XML text tags.
- Removed any sentences that contained "@" symbols. COHA and COCA randomly replace words with @ symbol in groups of ten for copyright reasons [Rudnicka, 2018].
- Removed apostrophes and extra whitespace.
- Used python's natural language toolkit (nltk) package to convert text to tokens [Bird et al., 2009].
- Selected the last 2000 tokens (words) of the text sample for processing. This avoids, as much as possible, anomalous text that sometimes appears at the start of text samples such as a contents section.

For the BNC data, python's natural language toolkit package comes with a BNC corpus reader [Bird et al., 2009], which was used to extract tokens. The only other treatment was to remove extra whitespace and apostrophes as with COCA and COHA.

The cleaned datasets had the following surviving sample counts with $N \geqslant 2000$ words:

- COHA total $n = 22,253$. Fiction $n = 8,164$, non-fiction $n = 2,046$, news $n = 725$, magazines $n = 11,318$.
- COCA total $n = 985$. Fiction $n = 167$, non-fiction $n = 166$, news $n = 39$, magazines $n = 133$.
- BNC total $n = 1,319$. Fiction $n = 447$, non-fiction $n = 477$, news $n = 395$.

The COHA dataset was analysed as a timeseries, so requires a large number of samples. The BNC and COCA, being corpora from much narrower time ranges, were analysed as distributions and as such require less samples.

Social Media Data

We also investigated social media. The Twitter dataset consisted of 1.6 million tweets scraped from the twitter API between April and June 2009 [Go et al., 2009] and available online at https://www.kaggle.com/kazanova/sentiment140. To simulate a Twitter feed the tweets were chronologically collated to create $n = 1000$ text samples with $N \geqslant 2000$ words each.

For Reddit, we aimed to capture text samples that were representative of the text a user would see when visiting the site. To achieve this we used Reddit's API to download posts from the Reddit homepage feed at https://oauth.reddit.com/.json. Following Reddit's API rules, we first registered an app and all requests were authenticated with OAuth2. We downloaded 10,000 posts in JSON format in this way. We extracted the text from the posts and combined them to create $n = 90$ text samples with length $N \geqslant 2000$ words each. During processing we found a small number of non-English posts in the feed, which were removed.

The social media data was then cleaned:

- Removed apostrophes and extra whitespace.
- Removed any urls.
- Removed hashtags and usernames i.e. any words containing "@" or "#".

- Used python's natural language toolkit (nltk) package [Bird et al., 2009] to convert the collated samples into a list of tokens, and the last 2000 tokens taken.

Social media statuses are by nature short and are usually much smaller than $N = 2000$ words, and lexical measures of short text samples have little meaning. Our analysis is on the level of the social media feed and we generated large text samples through the collation of posts. This kind of collation will naturally create text samples with high lexical diversity. This isn't a flawed analysis — the high information density of a social media feed is related to the collation of statuses and how people actually consume social media.

Measures of information evolution

Information evolution is measured using unigram word entropy. For robustness we also analysed the type token ratio and Zipf exponent of text samples, which are also measures of lexical diversity [Bentz et al., 2015]. The lexical measures are all sensitive to sample size, so we used truncated text samples to $N = 2000$ words.

Empirical unigram word entropy, H_1, is a function of the relative frequencies of each word, f_i, summed over the set of W unique words in the text sample. We use the maximum likelihood or plug-in estimator, which has the benefit of being simple and well known. And it has been shown to correlate well with more advanced estimators [Bentz et al., 2017].

$$H_1 = -\sum_{i=1}^{W} f_i log_2 f_i \, . \tag{2.1}$$

Type token ratio (TTR) is the number of unique words (types) divided by the total words (tokens) in a text sample.

$$TTR = \frac{\#types}{\#tokens} \, . \tag{2.2}$$

Words in natural language are typically approximately distributed as a power law distribution between type frequency, f_i, and type rank in that frequency distribution, $r(f_i)$ [Clauset et al., 2009]. This power law is parameterised by the Zipf exponent, α, which describes the steepness of the distribution in log space. Maximum likelihood estimation was used to estimate the Zipf exponent [Clauset et al., 2009]. This estimator has the benefit of being widely used and well known. It shows bias (as do all Zipf estimators [Pilgrim and Hills, 2020]), but the bias is systematic so can be ignored for the purpose of comparision of text samples.

$$f_i \propto r(f_i)^{-\alpha} . \tag{2.3}$$

Each of the measures were applied once to the same set of distinct text samples.

Timeseries Breakpoint Analysis

The Corpus of Historical American English (COHA) provides historical text samples across fiction, non-fiction, news and magazines categories. The type token ratio, word entropy and Zipf exponent were calculated for each text sample with over 2000 words.

For each media category and lexical measure, the results were binned into years and the median taken each year. The median was used to reduce the effect of outliers (similar results were found when using the mean). These were plotted on a scatterplot (see Supplementary Information).

Visually, the scatterplots are suggestive of some change in the gradient of the lexical measure in time. In order to estimate the location of these breakpoints, we used python's piecewise-regression package [Pilgrim, 2021] with default settings. The regression fits and locations of breakpoints are shown in the scatterplots in the Supplementary Information.

We ran a similar analysis with the categories combined. In order to combine the categories, we first took means for each year and category and then took the mean across categories for each year. It is more natural to use means than medians when combining categories, and the influence of outliers is smaller as there is more data than in the individual categories. The scatterplot and piecewise-regression fit for the combined word entropy is shown in the Supplementary Information.

Timeseries Trend Analysis

For each category and lexical measure, trend analyses were carried out on the annual median values. This was done between the years 1900 and 2009 (the last year of data). KPSS and MK tests were carried out for each measure and media category in COHA (full results in Supplementary Information).

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test assumes the null hypothesis of a stationary timeseries. p-values below 0.05 mean that we can reject this hypothesis at 5% significance and provide evidence of a trend. The test was applied using python's statsmodels package [Seabold and Perktold, 2010].

The Mann-Kendall (MK) test is a non-parametric trend test [Hussain and Mahmud, 2019]. The test assumes no serial correlation i.e. errors in one observation do not predict errors in other observations [Hussain and Mahmud, 2019]. The text corpora are constructed from independent text samples so this is a reasonable assumption. The null hypothesis is that the data has no trend, and the p-value tells us the probabilty that the data was observed under the null hypothesis. At 5% significance we reject the null hypothesis if $p < 0.05$. The test was carried out using python's pymannkendall package [Hussain and Mahmud, 2019]

In the Supplementary Information we calcaulte Pearson's R between magazine circulation and word entropy.

Timeseries Smoothing

While we included scatterplots for annual binned data in the Supplementary Information, the trends are easier to see visually with a smoothed timeseries. For Figures 2.2 and 2.3 the timeseries was smoothed using a moving average with measures of text samples from ± 5 years. The 95% confidence interval was calculated as the standard error of this mean calculation multiplied by 1.96 (assuming normally distributed errors). For each lexical measure, the mean was plotted for each year with the confidence interval region shaded. We only included years where we had a minimum of 10 data points within the window.

We report the smoothed timeseries for each of the COHA text categories, as well as the categories combined. The timeseries for media categories were combined by taking an average across the timeseries annual means for the media categories that had a value for that year. The 95% confidence interval was again calculated as 1.96 times the standard error. For each year, the standard error of the estimate of the mean, $SE_{\bar{X}}$ was computed based on the delta method,

$$SE_{\bar{X}} = \frac{\sqrt{\sum_{i=1}^{n} SE_i^2}}{n}, \tag{2.4}$$

with n depending on how many media categories had values for the annual mean each year.

Differences Between Media Categories

We looked at the distributions of the lexical measures within media categories in COCA, the BNC and COHA (restricted to 2000-2007 to avoid the effect of historical changes). To test for differences between the groups we carried out ANOVA tests across categories within each corpora separately for each of the lexical measures. At 5% significance, $p < 0.05$ provides evidence that the media categories are drawn from different underlying population distributions. The tests were carried out using python's statsmodels package [Seabold and Perktold, 2010]

For visualisation, the distributions of word entropy for each media category are shown as a kernel density estimate with the bandwidth determined by the Scott rule and the density trimmed to the data range.

US Magazine Circulation

The data for magazine circulation numbers (reported in the Supplementary Information) were taken from Sumner's "The Magazine Century American Magazines Since 1900" [Sumner, 2010] Chapter 1, which are attributed to data originally from the Audit Bureau of Circulation. This data source does not track all US magazines, but does track well-known magazines. The data was plotted without further treatment.

2.4 Results

2.4.1 The Rising Entropy of American English

We analysed the Corpus of Historical American English (COHA), a balanced corpus with text samples from the 1810s to the 2000s categorised into news, magazines, fiction and non-fiction [Davies, 2012]. As discussed in the Methods section, we analysed text samples truncated to $N = 2000$ words. We found a clear trend of rising lexical diversity since approximately 1900 as measured by word entropy, type token ratio and Zipf exponent (Figure 2.1).

The trends in separate media categories follow the same pattern of rising lexical diversity as measured by word entropy (Figure 2.2). We analysed the timeseries of annual averages since 1900 for each media category (fiction, non-fiction, news, magazines) and lexical measure (word entropy, Zipf exponent, type token ratio) using Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and Mann-Kendall (MK) tests on the annual median values (using the annual mean gives similar results). This gives a total of $4 \times 3 \times 2 = 24$ trend tests. All 24 tests show significant evidence of a trend at $p < 0.05$ And 22 out of 24 tests show significant evidence of trends at $p < 0.01$ (the tests for a trend in type token ratio in non-fiction had KPSS $p = 0.015$ and MK $p = 0.013$). Overall there is very strong evidence for a trend of rising lexical diversity in all media categories between 1900 and 2010. For full results and a deeper analysis, see the Supplementary Information.

2.4.2 Higher Entropy in Short-form Media

The historical trend (Figure 2.2) suggests modern differences in entropy between media categories. However, we also know that short-form media has become especially prominent with the recent rise of online platforms for media distribution, such as social media, RSS feeds, and news platforms that present short headlines and snippets that link to long-form articles. To investigate these different media

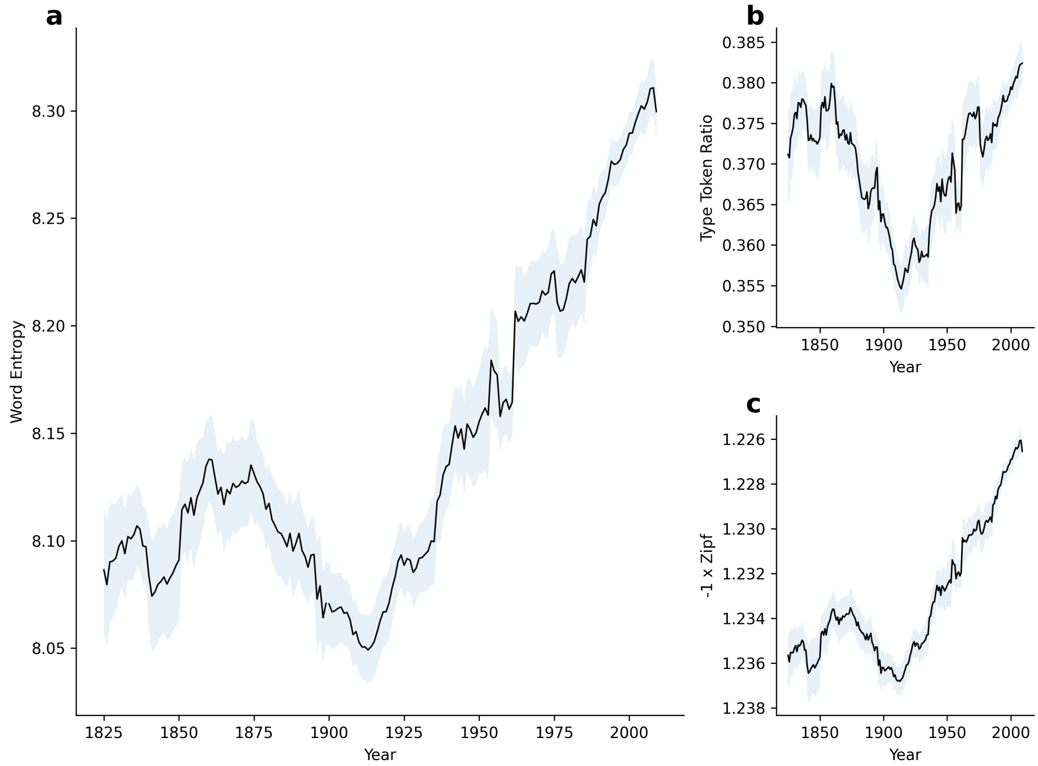


Figure 2.1: Lexical diversity of text samples in the Corpus of Historical American English as measured by a) word entropy, b) type token ratio and c) Zipf exponent. Timeseries are smoothed with a moving average window of ± 5 years, and averaged over media categories. Shaded region shows 95% confidence interval of this average.

categories, we examined the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC), as well as social media data from Twitter and Reddit. Figure 2.3 shows the distribution of word entropy across different media categories. Within COHA (limited to 2000-2007), BNC, and COCA there were significant differences in all lexical measures across media categories (ANOVA tests $p < 0.01$). Full statistical results are in the Supplementary Information. Overall, short-form media categories of news and magazines have higher entropy than long-form media, and social media feeds have the highest entropy of all.

It should be noted that when analysing social media data we collated posts to create text samples with $N = 2000$ words, to match the length of the other media type analyses. Combining posts will naturally lead to high entropy text, with fast switching of contexts and high novelty. This mirrors how people actually consume social media. Essentially, social media platforms generate high entropy information environments in the form of feeds of short messages from different users. This is not

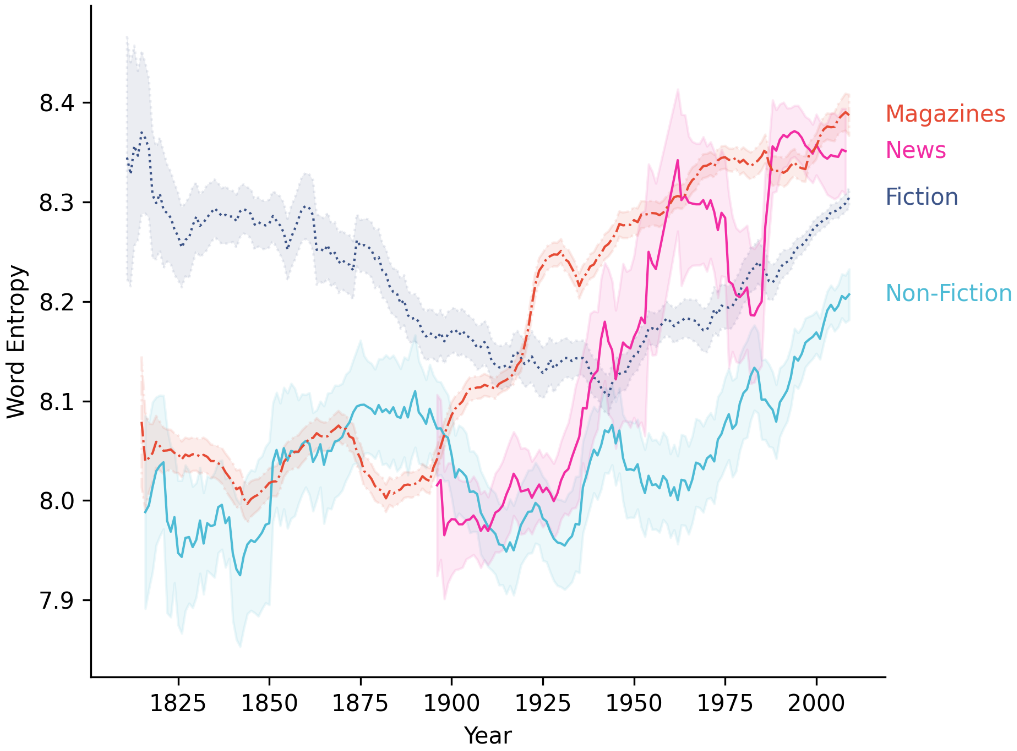


Figure 2.2: Timeseries of word entropy across media categories in the Corpus of Historical American English. For each media category, the timeseries was smoothed using an average over a window of ± 5 years. The shaded regions are 95% confidence intervals of this average. All media categories show an upward trend in word entropy from 1900.

necessarily a linguistic change in how people generate English; it is a change in how people consume English text.

2.4.3 Information Foraging in the Attention Economy

The results are suggestive of a link between competition for attention and word entropy. To explain these results we generate a model of the attention economy based on information foraging. Foraging models relate the consumption of information items with some utility gain to the forager. To bridge utility rates to lexical measures, we borrow the idea of information signal entropy from Shannon [Shannon, 1948]: the entropy of a source of information is a function of the probability of seeing each symbol given the preceding symbols. For our purposes entropy can be thought of as a rate of information. If information foragers gain utility from information then, by definition, an increase in entropy, H, is associated with an

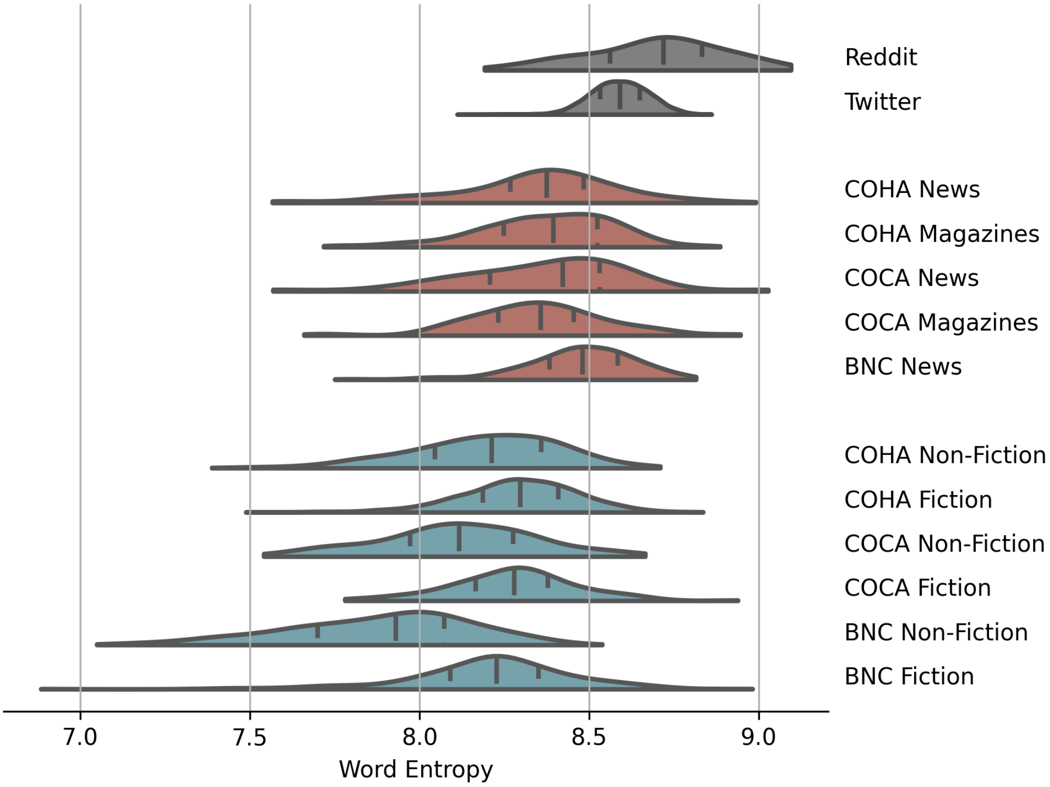


Figure 2.3: Word entropy of very short-form (social) media, short-form (news and magazines) and long-form (fiction and non-fiction) media. For each media category, distributions are kernel density estimates cut to the data range, with quartile positions shown. The COHA data was restricted to 2000-2007 to minimise the effect of historical changes.

increase in utility rate, r. This aligns with Zipf's principle of least effort [Zipf, 1949; Ferrer i. Cancho, 2005].

$$H \propto r\,. \tag{2.5}$$

Animal foragers modulate the selectivity of their diet in response to the environment, becoming more selective in times of abundance [Stephens and Krebs, 1986]. Why waste energy hunting difficult prey when there are plenty of easy calories around? Humans act in the same way when selecting information to consume [Pirolli and Card, 1999; Simon, 1969]. We have all experienced situations where we do not have access to the internet, for example on a plane or train journey, and we become less selective in what we read or watch.

This characterisation of attention corresponds to the prey choice model, which describes which types of prey are worth pursuing and consuming [Stephens and Krebs, 1986]. And this has been applied to information foraging before [Pirolli and Card, 1999]. The derivation of the prey model followed here is exactly analogous to that found in the prey choice model in food foraging. Our contribution will come at the end of this section, where we extend the model to include media competition for attention.

Assume an information forager searches a media environment and encounters information of types, i, at Poisson rates λ_i. If consumed, information provides a benefit u_i in a handling time t_i, during which time the forager is not searching. Alternatively, the forager can choose to ignore information of a certain type and keep searching. The forager's choices to consume or ignore information determine the expected total time spent searching, T_s, and handling, T_h, information, as well as the total utility gain, U. Given these constraints, the forager aims to optimise the expected overall rate of utility of foraging given by

$$R_{media} = \frac{U}{T_s + T_h} \, . \tag{2.6}$$

Here *media* describes the forager's local environment, such as a media platform. Media platforms are analogous to foraging patches in optimal foraging theory. The forager's choices of which information types to consume can be described as an information diet, D. The total expected utility is $U = \sum_D \lambda_i u_i T_s$. Similarly the total expected handling time is $T_h = \sum_D \lambda_i t_i T_s$. Substituting in and cancelling T_s, we can write the expected utility rate given a diet

$$R_{media} = \frac{\sum_D \lambda_i u_i}{1 + \sum_D \lambda_i t_i} \, . \tag{2.7}$$

Consuming an information item carries an expected opportunity cost of not spending that item's handling time looking for other items, equal to $t_i R_{media}$, and an expected utility gain of u_i. To maximise expected utility rate a forager should therefore consume the item if the item utility rate, $r_i = \frac{u_i}{t_i}$, is greater than the overall media platform utility rate, R_{media},

$$r_i \geqslant R_{media} \, . \tag{2.8}$$

This diet threshold condition is a familiar result from foraging theory [Stephens and Krebs, 1986; MacArthur and Pianka, 1966; Pirolli and Card, 1999]. To find the optimal diet, item types can be ranked in order of r_i and added to the diet one by one until this inequality fails [MacArthur and Pianka, 1966]. See the Supplementary

Information for a more thorough derivation.

We can now ask which information types a forager should include in their diet, D, to maximise their expected overall utility rate as a consequence of rising information prevalence, here λ_i. For items with $r_i < R_{media}$, increasing prevalence has no effect as these items are still not included in the diet. For items with $r_i \geqslant R_{media}$, increasing prevalence will mean more time spent handling these items and less time spent searching, so the overall media platform utility rate will increase,

$$\frac{\partial R_{media}}{\partial \lambda_i} \geqslant 0 \quad \forall i \,. \tag{2.9}$$

Combining this with the information diet criterion (Inequality 2.8), increasing information prevalence increases the information utility rate required for diet inclusion: foragers become more selective when prey (or information) is abundant, analogous to the prey model in optimal foraging theory [Stephens and Krebs, 1986].

We now extend traditional foraging theory to information co-evolution by asking how media producers respond to increasing selectivity among information foragers. By assuming there is some cost to media of producing more informative messages — a standard assumption underlying Zipf's principle of least effort [i Cancho and Solé, 2003; Zipf, 1949] — we conclude that an abundance of information creates an adaptive pressure that drives media producers to create information with a higher utility rate. A proxy for utility rate is information density, or word entropy. Figure 2.4 shows a simple simulation of this dynamic.

2.4.4 Competition Between Media Platforms Drives Differences Between Short- and Long-form Media

Information is distributed in media platforms (e.g., newspapers, magazines, books, Twitter, Reddit). The forager has to choose not only which information to consume within a media platform, but also which media platforms to visit. Analogous to the information choice model (Equation 2.8): an optimal information forager will visit a media platform if the expected media utility rate is greater than the background utility rate from foraging in the overall environment (see Supplementary Information for the full model),

$$R_{media} \geqslant R_{env} \,. \tag{2.10}$$

The utility rate of a media platform, R_{media}, is a summation over Poisson processes (Equation 2.7). To simplify this, let $\bar{u}_m$ be the average utility of information items consumed in the media platform, $\bar{t}_m$ the average time spent consuming

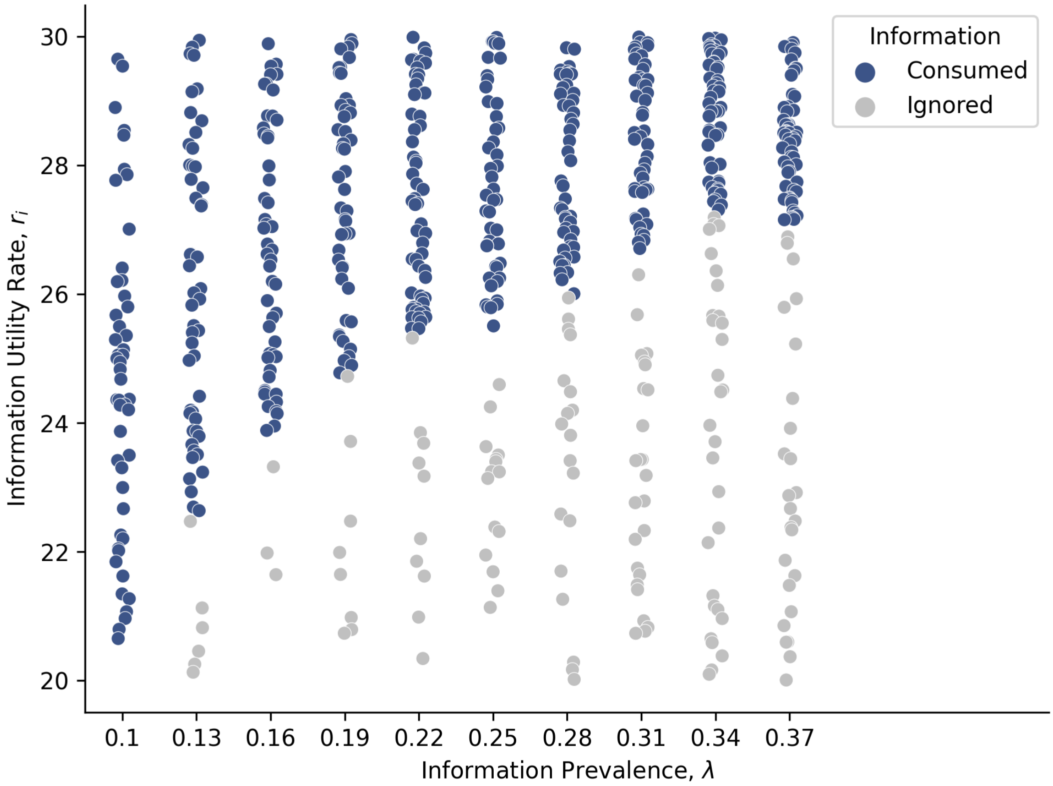


Figure 2.4: Simulation of information foraging in the attention economy. Information items are generated with random utility rates in quantities proportional to the information prevalence. Given the information environment, foragers only consume information items above a minimum information density (blue markers) in order to maximise their foraging rate. Information that is not consumed has less chance of survival (grey markers). Overall the surviving information types have higher utility rates at higher information prevalence.

information items, and λ_m the rate of encounter of any item in the diet. Equation 2.7 then becomes a variation of Holling's disc equation [Holling, 1959] (full derivation in Supplementary Information)

$$R_{media} = \frac{\lambda_m \bar{u}_m}{1 + \lambda_m \bar{t}_m}\,. \tag{2.11}$$

This equation is visualised in Figure 2.5 **a**.

The criteria for inclusion in an information forager's diet is then

$$\frac{1}{\lambda_m \bar{u}_m} + \frac{1}{\bar{r}_m} \leqslant \frac{1}{R_{env}}\,. \tag{2.12}$$

The inclusion of a media platform in the information diet is therefore deter-

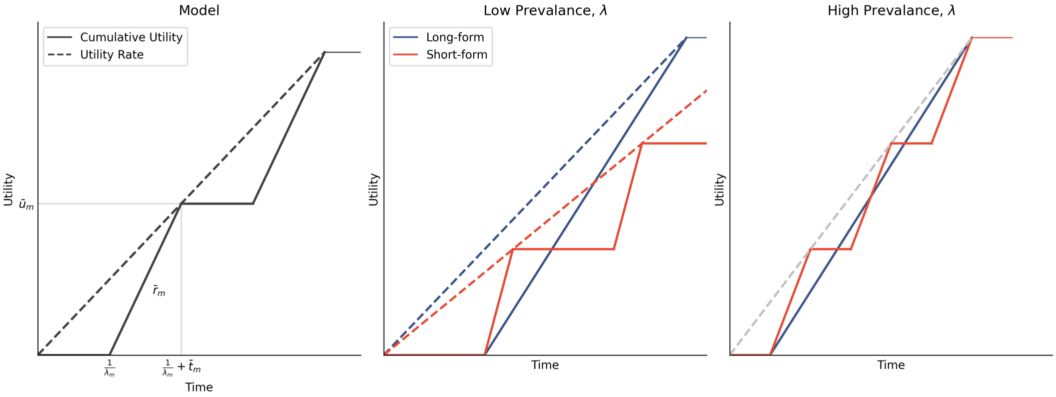


Figure 2.5: The media patch model. a) The expected utility rate of a media patch (dashed line) is determined by the time spent searching for (horizontal solid line) and consuming (diagonal solid line) information items. b) In a low prevalence environment long-form media has an advantage, although at low prevalence foragers are not very selective. c) At high prevalence less time is spent searching between item acquisition. To reach the same overall patch utility rate (dotted grey line), short-form media needs a higher information utility rate (gradient of the solid diagonal red line) than long-form media (gradient of the solid diagonal blue line).

mined by three properties of the information items that it contains and which would be included in the forager's information diet: the average utility (i.e. size) of a item, $\bar{u}_m$; the average item utility rate, $\bar{r}_m$; and the prevalence of items within the media platform, λ_m.

Short-form media platforms such as news and magazines involve more time spent switching (and searching for) articles than long-form media platforms such as books. In order to reach the same overall media platform utility rate, R_{media}, short form media types need to have higher information utility rates (Figure 2.5 **c**). This creates a differential selective pressure on short- and long-form media producers. Given some R_{env}, the short-form media platform needs higher average information utility rates, $\bar{r}_m$, to be accepted in the forager's diet than the long-form media. The long-form media experiences a relaxed selective pressure on information utility rates because there is less time spent switching in these media platforms. This can describe the differences in the observed information utility rates in short- and long-form media as well as the trend towards increased information rates with increasing media prevalence.

2.4.5 Social Media

Inequality 2.12 includes a weaker condition for diet inclusion, $\frac{1}{\lambda_m \bar{u}_m} \leqslant \frac{1}{R_{env}}$. This indicates that information prevalence directly limits the minimal average size of information for diet inclusion. As information prevalence increases, foragers will tolerate media platforms with smaller and smaller information item sizes (Figure 2.6). More intuitively, Twitter only works in a world with instant messages — few people would go to a library in order to check out a single Tweet.

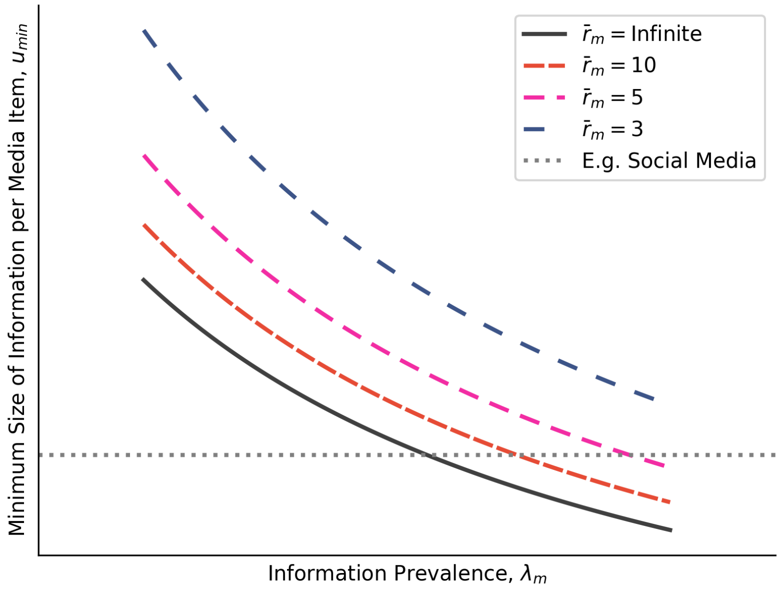


Figure 2.6: Minimum average information size, u_{min}, for media platform diet inclusion for varying levels of information prevalence, λ_m. Increasing average information utility rates, $\bar{r}_m$, can increase this limit only to a point. Very short-form media platforms like social media can only capture attention in a world with high information prevalence.

Finally, our model quantifies the selective forces acting to make media platforms more accessible. If a media platform reduces the expected search time between information encounters, $\frac{1}{\lambda_m}$, then they reduce the left hand side of Inequality 2.12 and become more competitive. This asymmetrically effects utility for short-form media, $\frac{1}{\lambda_m \bar{u}_m}$; for long-form media this term is already small. This could be an explanation for innovations towards minimising time spent searching in short-form media platforms such as infinite scroll and autoplay videos.

2.5 Discussion

We provide evidence that the word entropy of American English has increased over the 20th century. Furthermore, this change is marked by differences across different media categories, with the highest entropy levels found in the shortest media forms. Using a model of the attention economy based on information foraging, we show how a simple model of information selection can drive the observed changes. The attention economy model explains two results: a rise in entropy as information becomes more abundant and a rise in preferences for information dense short-form media.

Our findings offer an interesting contrast to the Linguistic Niche Hypothesis [Lupyan and Dale, 2010], which predicts a loss of complex morphological forms in English due to the influence of second language learners. There is ample evidence that English is undergoing morphological simplification [Michel et al., 2011; Lieberman et al., 2007; Zhu and Lei, 2018], and we might expect this to be associated with a decrease in word entropy (further explored in the Supplementary Information). Our findings show the opposite. Our claim is that the pressure towards information density overcomes the effect of reduced word entropy through linguistic simplification. However it may be that a reduction in morphological complexity and a rise in information entropy are related — in attention markets people may be attracted to both simplicity [Hills and Adelman, 2015] and novelty. Specifically, a loss in morphological complexity may be driven by a pressure towards simplicity and a reduction in the repetition of more difficult to process linguistic forms. That is, the features of the attention economy that drive rising entropy may also drive reduced morphological complexity.

Language evolution has been shown to follow a number of principles governed by human psychology. These principles have, for example, included features of biological and cultural evolution [Smith and Kirby, 2008; Christiansen and Chater, 2008], learning [Hills and Adelman, 2015; Christiansen and Chater, 2008; Lupyan and Dale, 2010], cooling by expansion [Petersen et al., 2012], word formation and distribution [i Cancho and Solé, 2003], and the decay of morphological complexity [Lupyan and Dale, 2010; Lieberman et al., 2007]. Our results extend the psychological consequences on language evolution to word entropy in response to information abundance.

Considering people as information foragers, our model describes observed empirical changes in word entropy of English over time and both within and between media categories in response to increasing information abundance. Empirical findings support the idea that people's attention is attracted to high entropy and high complexity information [Itti and Baldi, 2009; Radach et al., 2003]. Our analysis of historical data shows the entropy of information markets respond predictably to increased competition. The attention economy model offers a simple explanation: humans are, within limits, information rate maximisers responding to rising information abundance and media producers adapt their content to compete for more limited attention.

Humans choices are based on more than entropy. For example, humans respond to social cues and risk [Hills, 2019] just as animals consider factors other than calorie rate such as macro-nutrient content and predators when foraging for food [Stephens and Krebs, 1986]. Moreover, information producers are not only interested in capturing attention, but also in influence [Chen and Stallaert, 2014; Evans, 2020]. Nonetheless, just as animal foraging models have been shown to predict human behaviour in a variety of domains [Winterhalder, 1986; Pirolli and Card, 1999; Pirolli, 2009a; Fu and Pirolli, 2007; Hills et al., 2012], our analyses suggests these models also extend to the shape of information evolution and cultural history, just as the co-evolutionary arguments of Darwin might have predicted [Darwin, 2011].

2.6 Supplementary Information — Linguistic Niche Hypothesis

The finding that word entropy, and lexical diversity, is rising in American English is the opposite of what might be predicted by the Linguistic Niche Hypothesis. That hypothesis makes predictions about the complexity of language morphology (e.g. I ate, la casita) and syntax (e.g. I did eat, la pequeña casa), with the assumption that complexity is balanced between the two. The Linguistic Niche Hypothesis [Lupyan and Dale, 2010] suggests that languages in large, spread out social systems tend to have simpler morphological forms, with the grammatical work instead being done through syntax [Lupyan and Dale, 2010]. The hypothesised mechanism for this is that second language learners prefer simpler forms so that complex morphological forms disappear over time [Lupyan and Dale, 2010]. A global lingua franca like English should therefore be undergoing morphological simplification, and evidence does suggest that this is the case with the regularisation of English past tense verbs [Michel et al., 2011; Lieberman et al., 2007] and a loss of inflectional diversity [Zhu and Lei, 2018]. Further work suggests that this morphological simplification should correlate with a reduction in lexical diversity as measured by type token ratio [Bentz et al., 2015; Kettunen, 2014] (or word entropy) — complex morphological forms are non-repetitive (many unique word types per word token) whilst syntactic grammatical modifiers are repetitive (few unique word types per word token). We find that lexical diversity is instead rising in American English. We suggest some possible explanations:

1. English morphology is overall becoming more complex, against the Linguistic Niche Hypothesis.

2. English morphology is becoming simpler without an increase in syntactic complexity. This would be a further refutation of the already beleaguered [Deutscher et al., 2009; Sampson, 2009] equicomplexity assumption, which states that mature languages have broadly equal grammatical complexity, balanced between morphology and syntax.

3. Lexical diversity (and Type Token Ratio) is not a good measure of morphological complexity. The increase in lexical diversity is instead driven by more concise information and a wider, and faster switching of, contexts in written media.

The third option here is in our opinion at least partly responsible. If people are drawn towards higher utility rate information then that could drive English to be more concise and to switch contexts more quickly.

2.7 Supplementary Information — Historical Analysis of US Magazine Publishing

As a case study we investigated the history of magazine publishing in America. Figure 2.7 shows the historical trend in COHA magazine word entropy alongside magazine circulation figures and important events. Magazine publishers are in a two-sided market where they sell magazines to consumers and attention to advertisers [Evans, 2020], with the majority of revenue from selling attention [Sumner, 2010]. This wasn't always the case in the US — prior to the 1890s most magazine revenue was from sales, with advertising considered undesirable [Sumner, 2010]. Towards the late 19th century a combination of rapidly decreasing printing costs, growth in the literate population, discounts from the US postal service and the ability to target adverts to a niche readership led to a new business model to emerge [Sumner, 2010]. This new model involved selling magazines lower than the price of production, which increased circulation so that those costs could be recouped by advertising revenue [Sumner, 2010]. Before 1893, most magazines sold for 25 cents — until a price war led to the magazines McClure's, Munsey's and Cosmopolitan dropping their prices to 10 cents and subsequently enjoying rises in circulation and advertising revenue [Sumner, 2010]. The 10 cent magazines contributed to a tripling in total magazine readership from 1890 in 1905 [Sumner, 2010], and there was a huge jump in word entropy in the same period (Figure 2.7).

The Audit Bureau of Circulation was created by advertisers in 1914 [Sumner, 2010] to more accurately measure magazine readership numbers. This quantification of attention further increased pressure on magazine publishers to improve their circulation numbers in order to sell advertising. Other changes included moving advertisements from the back of the magazine to alongside the main content — a move that forced copywriters to improve the appeal of the content through adding color and improving graphics [Sumner, 2010].

Word entropy continues to rise throughout the 20th century alongside magazine circulation, with a Pearson's correlation coefficient r= 0.91 ($p < 0.001$), although both rise over time so that confounding factors are not ruled out (Figure 2.7). After the 1890s, the biggest drop in word entropy was during the great depression when magazine circulation also fell. There is a suggestion in the data that things change around the year 2000, as magazine circulation drops but word entropy continues to rise. The rise of digital media around this time is perhaps the biggest change in publishing since the printing press so we would not expect the same trends to necessarily continue — and digital media represents a new competitive pressure.

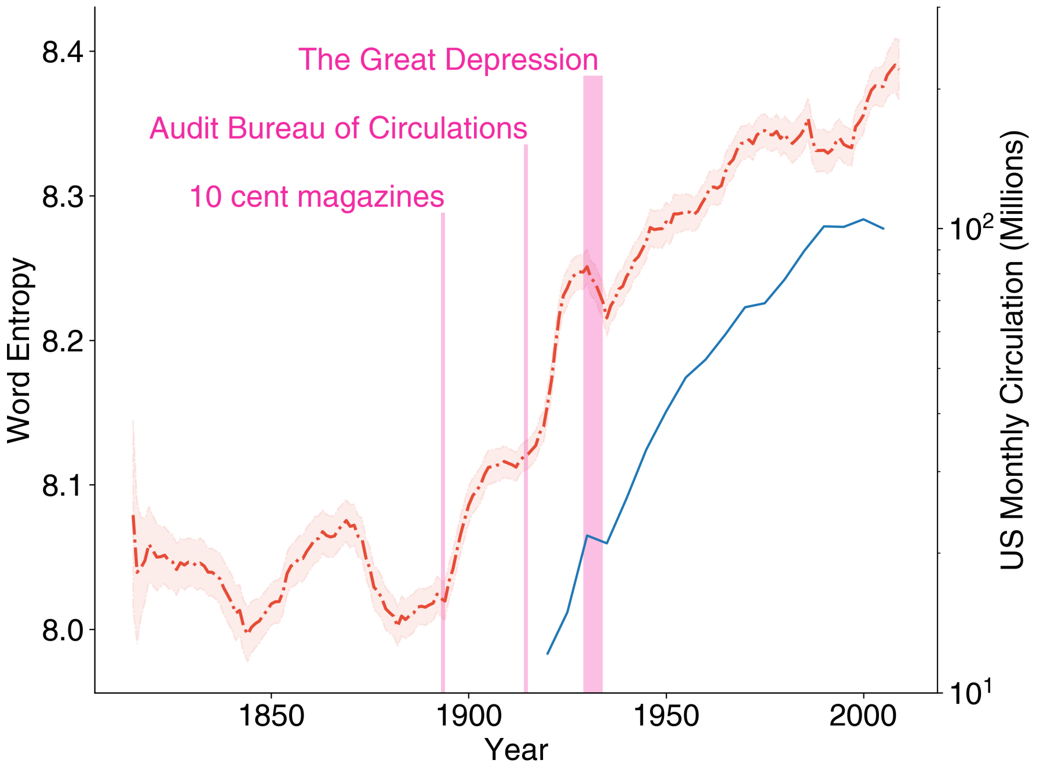


Figure 2.7: Historical analysis of word entropy in magazines (red dotted, timeseries calculated as in previous figure) with key events (pink) and US Monthly Magazine circulation as reported by the Audit Bureau of Circulations (purple).

2.8 Supplementary Information — Prey Choice Model Derivation

In the main chapter text we justify the prey choice algorithm using an argument that considers the opportunity cost of spending time handling a prey versus searching in the environment. Here we derive the same result more rigorously. This is a completely analogous derivation as found in optimal foraging theory [Stephens and Krebs, 1986]. As in the main chapter text, we have information types, i, that are encountered with rates λ_i while searching. Each information item, if consumed, provides a benefit u_i in a handling time t_i, during which the forager is not searching for other items.

In the main chapter text, a media patch expected utility rate is given by,

$$R_{media} = \frac{\sum_D \lambda_i u_i}{1 + \sum_D \lambda_i t_i} \,. \tag{2.13}$$

This assumes that information types are either in the diet, D, in which case they are always consumed upon encounter, or alternatively the items are not in the diet and never consumed. We can generalise this so that forager's have some probability of consuming an information type upon encounter, p_i,

$$R_{media} = \frac{\sum \lambda_i u_i p_i}{1 + \sum \lambda_i t_i p_i} \,. \tag{2.14}$$

The forager can choose the probability of paying attention to each information type, and a forager's strategy can be defined as a vector $\mathbf{p} = [p_1, p_2, ..., p_n]$. These choices are independent. To find the strategy that gives the maximum utility rate we can consider each of these choices, p_j, independently. To find the best strategy we separate p_j from the summations and differentiate

$$\frac{\partial R_{media}}{\partial p_j} = \frac{\lambda_j u_j (1 + p_j \lambda_j t_j + \sum_{i \neq j} p_i \lambda_i t_i) - \lambda_j t_j (p_j \lambda_j u_j + \sum_{i \neq j} p_i \lambda_i u_i)}{(1 + p_j \lambda_j t_j + \sum_{i \neq j} p_i \lambda_i t_i)^2} \,. \tag{2.15}$$

Cancelling like terms

$$\frac{\partial R_{media}}{\partial p_j} = \frac{\lambda_j u_j (1 + \sum_{i \neq j} p_i \lambda_i t_i) - \lambda_j t_j (\sum_{i \neq j} p_i \lambda_i u_i)}{(1 + p_j \lambda_j t_j + \sum_{i \neq j} p_i \lambda_i t_i)^2} \,. \tag{2.16}$$

The sign of this does not depend on p_j. So if $\frac{\partial R}{\partial p_j} > 0$, R_{media} will be maximised with $p_j = 1$, and otherwise with $p_j = 0$. The condition for $p_j = 1$ is

$$\frac{u_j}{t_j} > \frac{\sum_{i \neq j} p_i \lambda_i u_i}{1 + \sum_{i \neq j} p_i \lambda_i t_i)} \,. \tag{2.17}$$

The right hand side is the total expected rate of utility for all items except for item j, $R_{\neg j}$. The item should be included in the diet if the utility rate of the item, $r_i = \frac{u_j}{t_j}$, is greater than the overall rate of foraging without the item.

$$r_j \geqslant R_{\neg j} \,. \tag{2.18}$$

This is equivalent to the diet inclusion criteria given in the main chapter text. To find the optimal diet, one can add items in order of their utility rate until the inequality fails.

2.9 Supplementary Information — Patch Choice Model and Non Constant Patches

The patch choice model considered in the main chapter text is analogous to the information choice model. Patches of each type are randomly encountered in the environment and encountered as a Poisson processes with rates λ_{media}. We also assume that patches have a constant expected rate of utility, R_{media}, and some finite time, T_{media} until the rate drops to zero, which gives each patch a total utility, U_{media}. Foragers can choose to either consume or ignore a patch upon encountering it. This model is identical to the information choice model so that we can follow that derivation and jump to the conclusion that a patch will be included in the diet if the patch utility rate is greater than or equal to the overall rate of foraging in the environment, $R_{media} \geqslant R_{env}$.

Information patches in the real world have non-constant utility rates. Commonly patch marginal utility will decrease with time [Stephens and Krebs, 1986; Charnov, 1976]. This can happen as finite prey are consumed [Bettinger and Grote, 2016; Stephens and Krebs, 1986]. For example, within a patch an optimal forager will consume the most profitable items first if they can, which then makes those items more scarce and reduces the overall utility rate in the patch as time goes on [Bettinger and Grote, 2016]. Examples are collecting raspberries from a bush, or checking your email. Information items themselves may degrade while being consumed, for example news articles often follow an inverted pyramid structure where the most important information is presented first, with extra paragraphs adding marginally diminishing extra information [Pöttker, 2003]. Magazines, fiction and non-fiction have their own styles and utility curves. Overall we can say that utility rates in patches, and information, are not constant.

An optimal forager now has to choose both which patches to consume and how long to spend in those patches. This problem was solved by Charnov's marginal value theorem [Charnov, 1976], which we derive here in the context of information items. We follow the model and derivation given by Stephens and Krebs [Stephens and Krebs, 1986]. We characterise each patch type, k, with an expected utility return rate as a function of time spent within the patch, $g_k(t_k)$. We assume that patches are encountered randomly with rate λ_k as Poisson processes. The forager's decision is now how long to spend in each patch type, with a strategy described as $\mathbf{t} = [t_1, t_2, ..., t_k]$ ($t_i = 0$ meaning the patch is ignored) . We can write the expected patch utility rate as

$$R_{media} = \frac{\sum_k \lambda_k g_k(t_k)}{1 + \sum_k \lambda_k t_k} \,. \tag{2.19}$$

Similarly to the prey choice derivation, we differentiate with respect to the time spent in a patch type, t_j,

$$\frac{\partial R_{media}}{\partial t_j} = \frac{\lambda_j g'_j(t_j)(1 + \sum_k \lambda_k t_k) - \lambda_j(\sum_k \lambda_k g_k(t_k))}{(1 + \sum_k \lambda_k t_k)^2} \,, \tag{2.20}$$

where $g'_j(t_j) = \frac{\partial g_j(t_j))}{\partial t_j}$. Setting this equal to zero, we find the maximum R_{env} when

$$g'_j(t_j) = R_{env} \qquad \forall j \,. \tag{2.21}$$

This is Charnov's marginal value theorem [Charnov, 1976] and states that an optimal forager will leave a patch when the marginal utility rate of the patch equals the overall rate of utility from foraging in the environment. And foragers will not spend any time in a patch if the marginal rate never reaches the environmental rate i.e. $g'_j(t_j) < R_{env} \quad \forall t_j$. This makes sense intuitively — time spent in a patch with rate g_j carries an opportunity cost of time not spent foraging in the wider environment with utility rate R_{env}.

We can find which patches will be visited using the "patches as prey" algorithm [Stephens and Krebs, 1986]. This is a similar algorithm to the diet choice model but with patches ranked in order of their maximum profitability, $\frac{g_k(t^*_k)}{t^*_k}$. patch types are added to the diet one at a time, with the marginal value theorem applied to all included patches after adding each new patch to recalculate the environmental utility rate. This is done with all patch types, or until Inequality 2.21 fails.

How would this model of patches effect the conclusions of the main chapter text? As in the main chapter text, we assume that media producers have an incentive to create information patches that attract and hold attention. People are still driven towards patches with high patch utility rates. If patch degradation occurs through consuming the most attractive items first then then there would still be a selective pressure toward high utility rate information items, as this would make the patch more attractive before degradation and keep foragers in the patch for longer as it degrades. And this pressure would still apply more strongly to short-form media than long-form media (due to more time switching between short-form media). The conclusions in the main chapter text would still follow, although the full model would be more complicated. We are confident that the conclusions would hold under any reasonable model of patch degradation.

2.10 Supplementary Information — The Merged Poisson Process for Patches

Here we justify using average values to describe the expected patch utility rates, instead of summations over information types. We have not seen this derivation before in the foraging literature, but it is relatively straightforward. The result is used without derivation in Pirolli [2009b].

In the main chapter text we write down an equation for the expected patch rate in terms of the characteristics of the information within the patch diet, D,

$$R_{media} = \frac{\sum_{i \in D} \lambda_i u_i}{1 + \sum_{i \in D} \lambda_i t_i} \, . \tag{2.22}$$

In this model, information types are encountered as independent Poisson processes with rates, λ_i, during time spent searching, with total searching time T_s. Items have utilities u_i and handling times t_i. With some simple algebraic manipulation we can write down

$$R_{media} = \frac{(\sum_D \lambda_i) \frac{\sum_D \lambda_i u_i T_s}{\sum_D \lambda_i T_s}}{1 + (\sum_D \lambda_i) \frac{\sum_D \lambda_i t_i T_s}{\sum_D \lambda_i T_s}} \, . \tag{2.23}$$

The rate of a combined Poisson process is equal to the sum of the rate of the independent Poisson processes, $\lambda_p = \sum_D \lambda_i$ [Gallager, 2012].

We define the average utility of items encountered in the patch as the total utility gained divided by the total number of items handled,

$$\bar{u}_p = \frac{\sum_D \lambda_i u_i T_s}{\sum_D \lambda_i T_s} \, . \tag{2.24}$$

Similarly the average time spent handling items encountered is the total time spent handling divided by the number of items handled,

$$\bar{t}_p = \frac{\sum_D \lambda_i t_i T_s}{\sum_D \lambda_i T_s} \, . \tag{2.25}$$

Substituting these relations into equation 2.23,

$$R_{media} = \frac{\lambda_p \bar{u}_p}{1 + \lambda_p \bar{t}_p} \, . \tag{2.26}$$

We can therefore replace the patch rate equation (equation 2.22) with averages taken over the merged Poisson process. This is a variation of Holling's disc equation [Holling, 1959], considering average values.

2.11 Supplementary Information — Full Statistical Results

2.11.1 Timeseries Analysis

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test considers a null hypothesis of no trend. This is a one-sided test. Table 2.1 reports the KPSS statistics and the p-values for each of the analysed categories in the Corpus of Historical American English (COHA). Exact p-values are difficult to calculate below 0.01 and are not provided by python's statsmodels package [Seabold and Perktold, 2010], we have therefore denoted these as < 0.01 where applicable.

The Mann-Kendall test is a non-parametric trend test with the null hypothesis of no trend. This is a two-sided test. We report (Table 2.1) the normalised z-score, the p-value, Kendall's Tau, the Mann-Kendall score and slope. Exact p-values below 0.01 and are not provided by python's statsmodels package [Seabold and Perktold, 2010], we have therefore denoted these as < 0.01 where applicable.

2.11.2 Differences in Media Categories

We ran ANOVA tests to test for differences between media categories in each of the lexcical measures in the British National Corpus (BNC), Corpus of Contemporary American English (COCA), and the Corpous of Historical American English (restricted to 2000-2007). Results are shown in Table 2.2.

2.12 Supplementary Information — Type Token Ratio and Zipf exponent

Figure 2.8 shows the historical trend in type token ratio in COHA. Figure 2.9 shows the trend in Zipf exponent in COHA.

Figure 2.10 shows the distribution of type token ratio in text samples across media categories. Figure 2.11 shows the same for Zipf's law.

Overall these trends support the trends found in word entropy in the main chapter text.

	Word Entropy	
	KPSS (KPSS Statistic, p-value)	Mann-Kendall (z, p-value, Tau, MK score, slope)
news	(1.4725, <**0.01**)	(7.5198, <**0.01**, 0.5157, 2451.0000, 0.0046)
magazines	(1.7361, <**0.01**)	(10.9990, <**0.01**,0.7172, 4144.0000, 0.0027)
fiction	(1.2372, <**0.01**)	(7.5911,<**0.01**, 0.4927, 2900.0000, 0.0017)
non-fiction	(1.4084, <**0.01**)	(5.9100,<**0.01**, 0.3836, 2258.0000, 0.0019)
	Type Token Ratio	
	KPSS (KPSS Statistic, p-value)	Mann-Kendall (z, p-value, Tau, MK score, slope)
news	(1.1982, <**0.01**)	(5.3317, <**0.01**,0.3657, 1738.0000, 0.0005)
magazines	(1.0223, <**0.01**)	(5.9933, <**0.01**, 0.3908, 2258.0000, 0.0002)
fiction	(0.8972, <**0.01**)	(5.9891, <**0.01**, 0.3887, 2288.0000, 0.0003)
non-fiction	(0.6866, 0.0148)	(2.4774, 0.0132,, 0.1609, 947.0000, 0.0001)
	Zipf exponent	
	KPSS (KPSS Statistic, p-value)	Mann-Kendall (z, p-value, Tau, MK score, slope)
news	(1.5085, <**0.01**)	(-7.8083, <**0.01**, -0.5355, -2545.0000, -0.0002)
magazines	(1.7521, <**0.01**)	(-11.4025, <**0.01**, -0.7435, -4296.0000, -0.0001)
fiction	(1.3244, <**0.01**)	(-7.5335, <**0.01**, -0.4890, -2878.0000, -0.0001)
non-fiction	(1.2890, <**0.01**)	(-6.1038, <**0.01**, -0.3962, -2332.0000, -0.0001)

Table 2.1: Timeseries analysis across different categories and measures for text samples from COHA between 1900 and 2009. In each cell, the p-value of a Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test and a Mann Kendall (MK) test are shown respectively. Significant trends at $p < 0.01$ are emboldened. For both tests, p-values below 0.01 mean we can reject the null hypothesis of stationarity at 1% significance.

	Word Entropy ANOVA
COHA (DOF:3)	(F = 86, p = 7.68e-54)
COCA (DOF:3)	(F = 37, p = 8.99e-22)
BNC (DOF:2)	(F = 689, p = 1.76e-205)
	Type Token Ratio ANOVA
COHA (DOF:3)	(F = 34, p = 5.95e-22)
COCA (DOF:3)	(F = 19, p = 5.21e-12)
BNC (DOF:2)	(F = 425, p = 3.63e-143)
	Zipf Exponent ANOVA
COHA (DOF:3)	(F = 92, p = 2.14e-57)
COCA (DOF:3)	(F = 41, p = 3.54e-24)
BNC (DOF:2)	(F = 712, p = 2.67e-210)

Table 2.2: Analysis of differences in word measures across media categories within each text corpus. ANOVA tests are reported. All are significant.

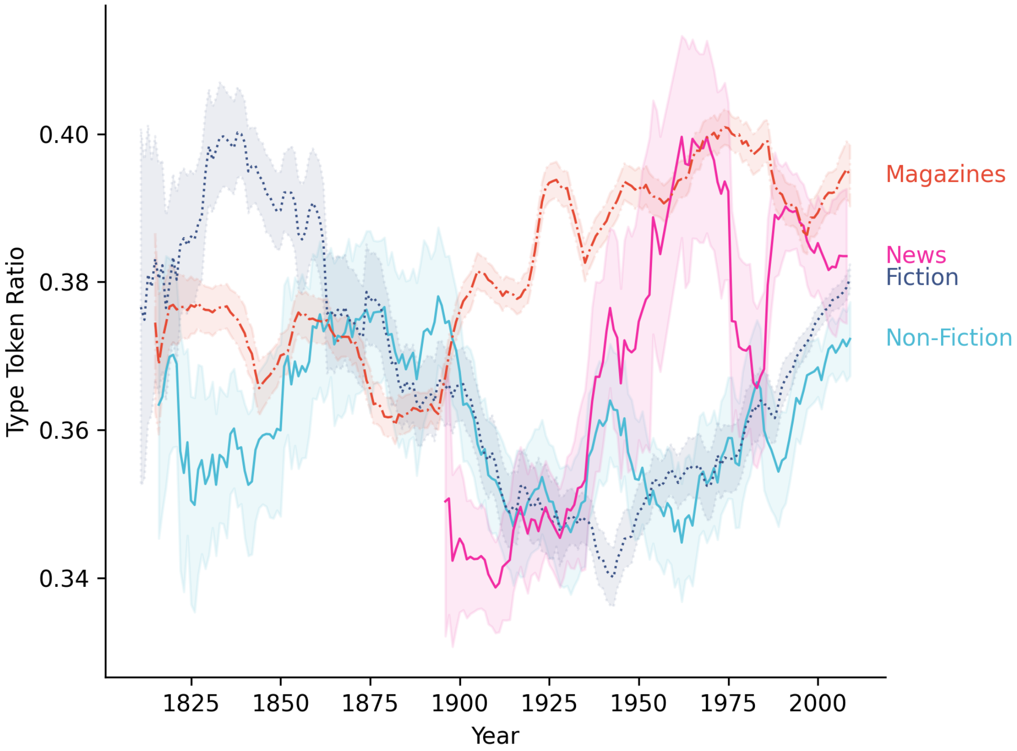


Figure 2.8: Historical timeseries of type token ratio in the Corpus of Historical American English. Type token ratio was calculated for text samples from COHA truncated with $N = 2000$ words. For each media category and year, a moving average of all valid samples with ±5 years was calculated. The shaded region shows a 95% confidence interval for this average.

2.13 Supplementary Information — Timeseries Breakpoint Analysis

As discussed in Methods, we carried out a piecewise-regression analysis on the median annual values for each of the lexical measures and media categories (Figure 2.12). With the type token ration for the News media category, the breakpoint was found close to the edge of the data. If we restrict the position to avoid being close to the edge then the breakpoint is estimated in a similar location as to the Word Entropy and Zipf exponent. The short-form media shows signs of a rise in lexical diversity before long-form media, consistent with the model in the main chapter text.

We ran the same analysis with the media categories collated to give an average mean each year (Figure 2.13). Notably, the confidence interval for the breakpoint

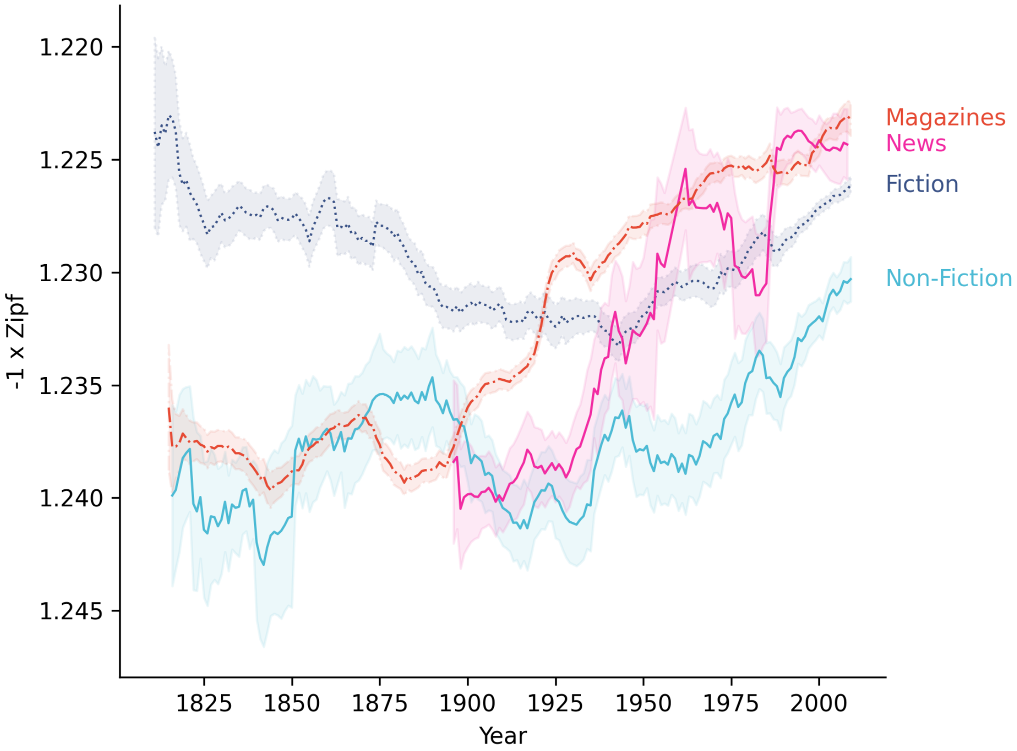


Figure 2.9: Historical timeseries of Zipf exponent in text samples in written media categories in American English. The timeseries was calculated in the same way as in the previous figure.

includes the year 1900.

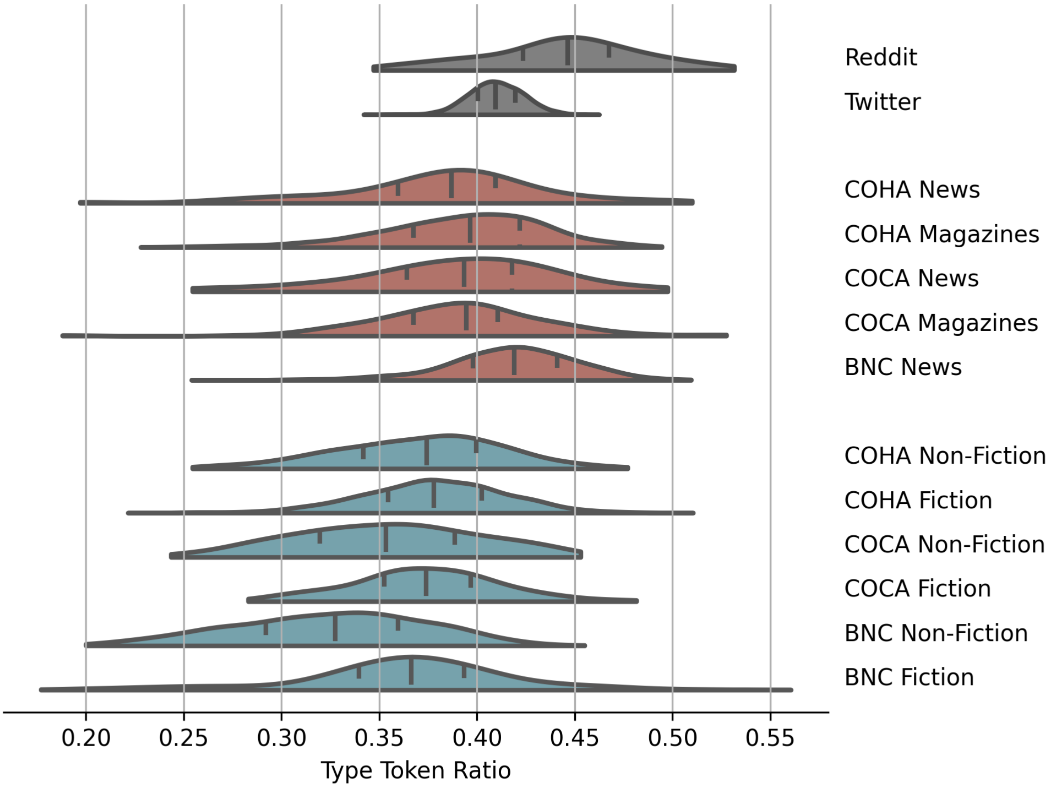


Figure 2.10: Distribution snapshots of type token ratio across different text corpora for text samples with $N = 2000$ words. COHA samples are from the year 2000 onwards only. Social media text samples were collated from status updates.

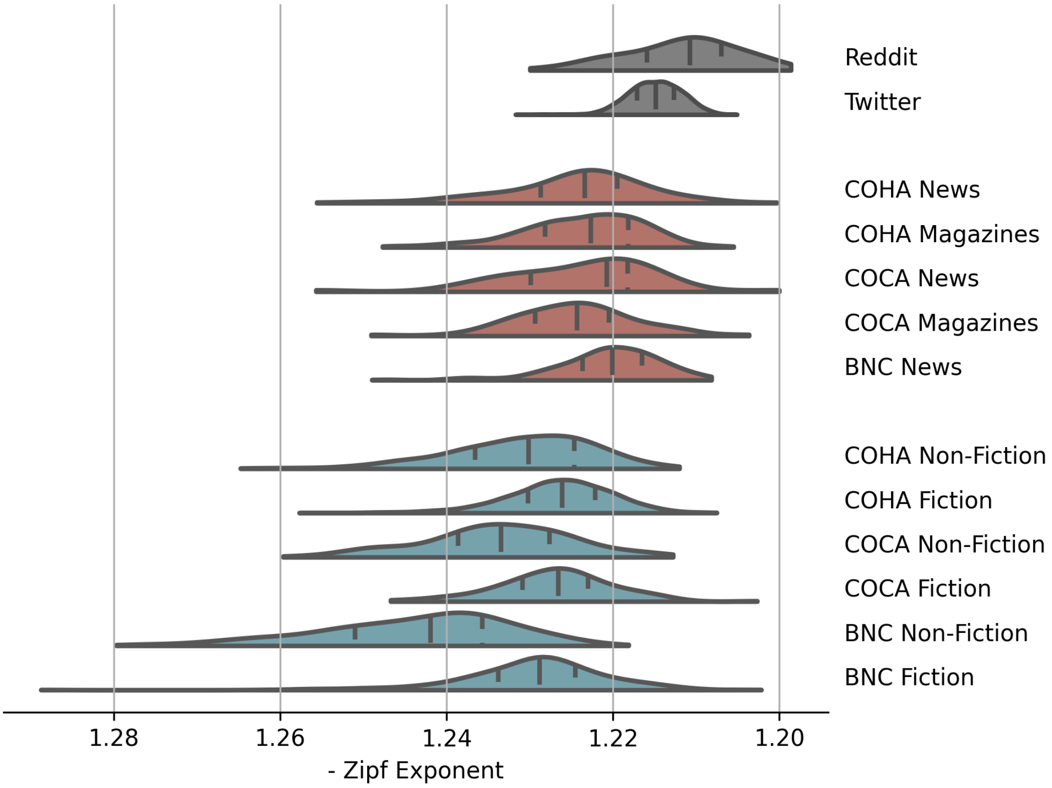


Figure 2.11: Distribution snapshots of the Zipf exponent across different text corpora for text samples with $N = 2000$ words. COHA samples are from the year 2000 onwards only. Social media text samples were collated from status updates.

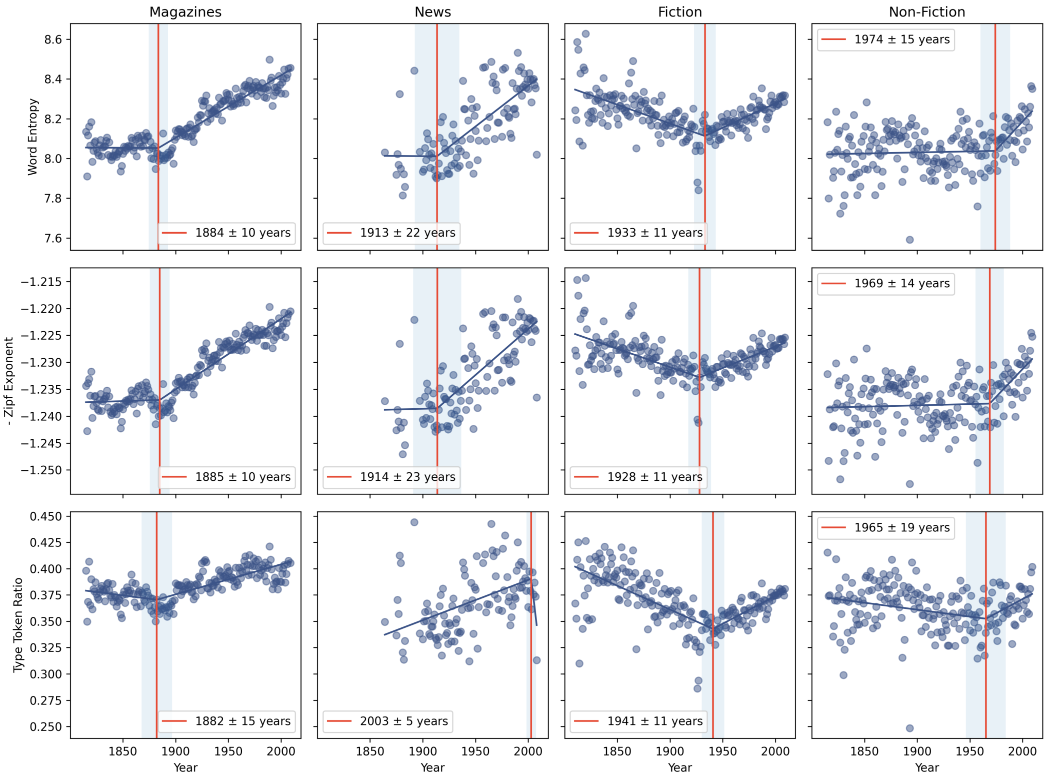


Figure 2.12: Median annual values for each category and lexical measure. The points were fit with a piecewise-regression, with red lines showing the estimated breakpoints. The shaded region shows a 95% confidence interval for those breakpoints.

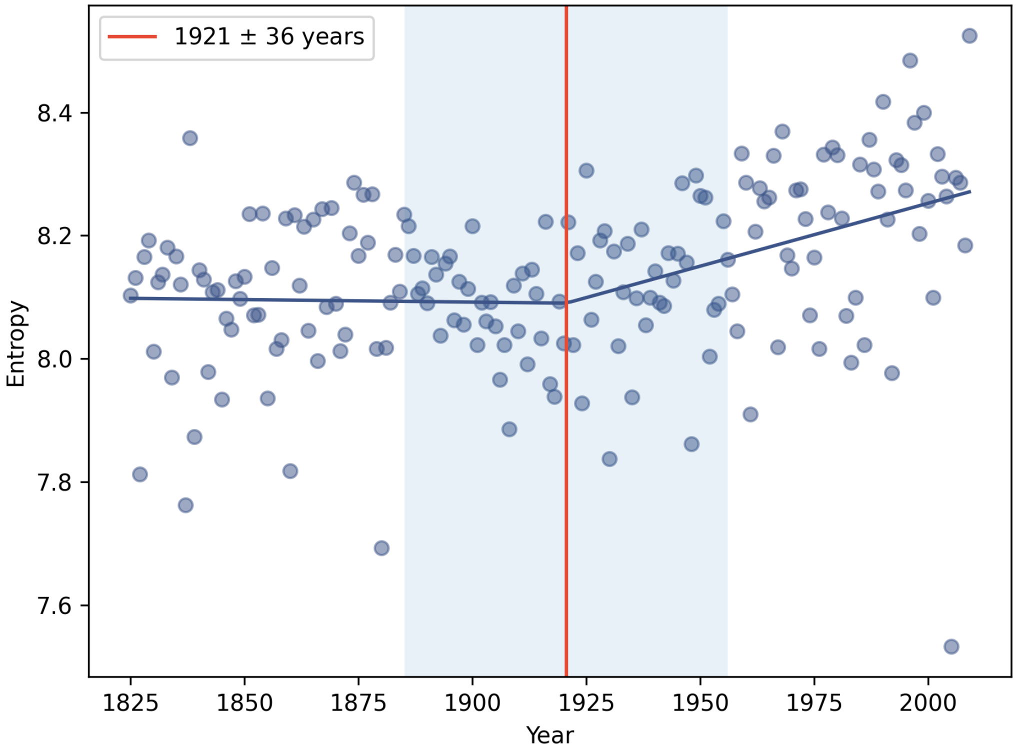


Figure 2.13: Mean annual values for the media categories combined for word entropy. Annual means were first found within each media category, and then averaged over the media categories. The points were fit with a piecewise-regression, with red lines showing the estimated breakpoint. The shaded region shows a 95% confidence interval for that breakpoint.

Chapter 3

Bias in Zipf's Law Estimators

3.1 Abstract

The prevailing maximum likelihood estimators for inferring power law models from rank-frequency data are biased. The source of this bias is an inappropriate likelihood function. The correct likelihood function is derived and shown to be computationally intractable. A more computationally efficient method of approximate Bayesian computation (ABC) is explored. This method is shown to have less bias for data generated from idealised rank-frequency Zipfian distributions. However, the existing estimators and the ABC estimator described here assume that words are drawn from a simple probability distribution, while language is a much more complex process. We show that this false assumption leads to continued biases when applying any of these methods to natural language to estimate Zipf exponents. We recommend that researchers be aware of these biases when investigating power laws in rank-frequency data.

3.2 Introduction

If we take a book and rank each word based on how many times it appears, we will find that the number of occurrences of each word is approximately inversely proportional to its rank [Zipf, 1949]. The second most frequent word will appear approximately $\frac{1}{2}$ as often as the most frequent word, the third around $\frac{1}{3}$ as frequently. This describes a power law relationship between the frequency of a word, n, and the word's rank in terms of its frequency, r_e, with exponent $\gamma \approx 1$ [Piantadosi, 2014],

$$n(r_e) \propto r_e^{-\gamma} . \tag{3.1}$$

This is known as Zipf's law and is consistent, in a general sense, across human communication [Ferrer i. Cancho, 2005; Moreno-Sánchez et al., 2016]. We do not have a satisfactory reason why this is [Piantadosi, 2014] and the exponent, γ, is not always 1 but varies between different speakers [Ferrer i. Cancho, 2005] and texts [Ferrer i. Cancho, 2005; Montemurro and Zanette, 2002]. Sound analytical tools are needed to investigate these research areas.

Equation 3.1 describes an observed empirical relationship. This is sometimes expressed as a relationship between a word's probability of occurrence [Baixeries et al., 2013; Shannon, 1951] and the word's rank in the probability distribution, r_p,

$$p(r_p) \propto {r_p}^{-\lambda}. \tag{3.2}$$

The conflation of equations 3.1 and 3.2 causes the prevailing maximum likelihood estimators to miscalculate λ in equation 3.2 with a positive bias [Corral et al., 2019; Hanel et al., 2017] (Figure 3.1). This bias applies specifically to rank-frequency distributions, where the ranks of events are not known a priori and instead are extracted from the frequency distribution, as is the case in equation 3.1. The existing maximum likelihood estimators make the assumption that the observed empirical frequency rankings of data (r_e in equation 3.1) are equivalent to rankings in an underlying probability distribution (r_p in equation 3.2) [Corral et al., 2019], this is the source of the bias. The nth most frequent word is assumed to be the nth most likely word, which is not necessarily the case.

In the 2000s there were a series of papers [Clauset et al., 2009; Goldstein et al., 2004; Bauke, 2007; Newman, 2005] describing a method of maximum likelihood estimation that gave more accurate (lower bias) estimates for power law exponents than graphical methods [Clauset et al., 2009]. The most influential of these is Clauset et al's paper [Clauset et al., 2009]. The estimators had been derived and presented before [Goldstein et al., 2004] (as early as 1952 in the discrete case [Seal, 1952]) but Clauset et al's paper popularised the idea and provided a clear methodology including techniques to perform goodness of fit tests [Clauset et al., 2009]. In all of these papers, the derivation of the likelihood function assumes that there is some a priori ordering on an independent variable. This works very well for power laws with some natural way to order events, such as the size vs frequency of earthquakes [Clauset et al., 2009]. However, it does not work so well with rank-frequency distributions, where the rank is extracted empirically from the frequency distribution, so that the empirical rank and frequency are correlated variables [Piantadosi, 2014], both dependent on the same underlying mechanism. This difference was not addressed by Clauset et al, who include examples of applying their estimator to rank-frequency

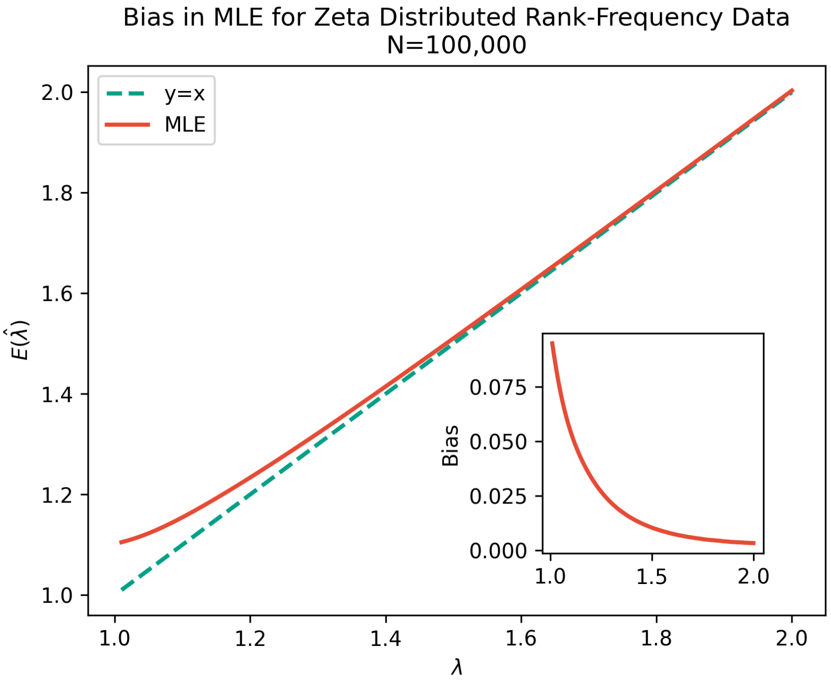


Figure 3.1: Bias in maximum likelihood estimation for rank-frequency data. 100 values of λ between 1 and 2 were investigated. For each λ, samples with $N = 100,000$ were generated from an unbounded power law distribution and Clauset et al's estimator was applied to the empirical rank-frequency distribution. This was repeated 100 times and results averaged. There is a clear and strong positive bias for $\lambda \lessapprox 1.5$.

.

data [Clauset et al., 2009]. The same data can look very different depending on whether we know it's true rank or not, as shown in Figure 3.2.

Recently Clauset et al's estimator has been shown, empirically, to be biased for some rank-frequency distributions [Hanel et al., 2017; Corral et al., 2019]. In particular, Clauset et al's method over-estimates exponents with rank-frequency data generated from known power law probability distributions with exponents below about 1.5 [Hanel et al., 2017] (Figure 3.1). The problem is related to low sampling in the tail [Hanel et al., 2017; Corral et al., 2019], so that the observed empirical ranks tend to "bunch up" above the line of the true probability distribution before decaying sharply at the end of the observed tail (Figure 3.2). To our knowledge this bias has not been adequately explained or solved.

- In 2014 Piantadosi et al [Piantadosi, 2014] suggested splitting a corpora and calculating ranks of words from one part of the split and frequencies from

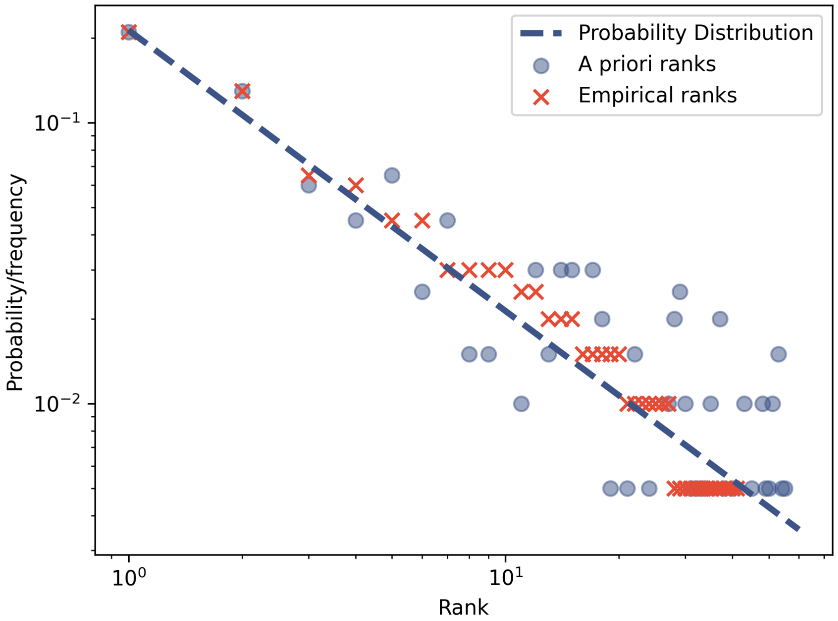


Figure 3.2: Difference between distributions with probability and empirical ranks. Data was generated from an underlying power law probability distribution with exponent $\lambda = 1$, number of possible events $W = 60$ and $N = 200$ samples. The dotted blue line shows the probability distribution. The blue circles show the sampled event frequencies with a priori known probability ranks. The red crosses show the empirical rank-frequency distribution from the same data. There is a significant difference between the two distributions. The current estimators are designed to fit data with a priori known ranks, not empirical ranks.

.

the other, breaking the correlation of errors. However the method does not take into account uncorrelated errors in the ranks. In particular, the empirical ranks of events in the tail will almost certainly be lower than the actual ranks in the probability distribution as many events in the tail will not be observed at all.

- Hanel et al [Hanel et al., 2017] identified the problem and suggested using a finite set of events instead of Clauset et al's unbounded event set [Clauset et al., 2009]. This gives more accurate results in the limited case that the number of possible events, W, is finite and known [Hanel et al., 2017]. Often W is not known and the choice of W can substantially change the results. With Zipf's law in language, W represents the writer's vocabulary and is usually modelled as unbounded [Piantadosi, 2014; Clauset et al., 2009; Bauke, 2007]. This seems appropriate given that Heaps' Law suggests that the number of unique words in a document continues to rise indefinitely as the document length increases [Heaps, 1978].

- In 2019 Corral et al [Corral et al., 2019] examined the problem and explored a technique of transforming the data to a distribution of frequencies representation, $f(n)$, which is also a power law type distribution that they call the Zipf's law for sizes. This distribution does have an a priori known independent variable of frequency sizes, so the bias described here does not apply to this representation. However there is still difficulty in estimating the rank-frequency exponent, as a power law in the rank-frequency distribution, $n(r_e)$, will only approximately map to a power law in the distribution of frequencies, $f(n)$, for real-world sample sizes [Corral et al., 2019].

Overall these ad-hoc methods can remove the bias to some extent but not completely. The methods also introduce a host of somewhat arbitrary choices for the researcher to resolve.

We derive a new maximum likelihood estimator that does not make the false assumption that the empirical ranks, r_e, are equivalent to the probability ranks, r_p. The new estimator considers all the possible ways that the events could be ranked in the underlying probability distribution to generate the observed empirical data. Unfortunately this new likelihood function is computationally intractable for all but the smallest data sets. In order to estimate parameters for larger data sets, we turn to approximate Bayesian computation (ABC), a method that is designed for situations where likelihood functions cannot be computed [Beaumont, 2010]. We show that this method has much lower bias than Clauset et al's estimator for rank-frequency data generated from simple power laws. We further explore two different implementations of ABC and find that they give different results when applied to word distributions in books because ABC and Clauset et al's method both assume an underlying power law probability model, while natural language arises from a more complex model. We suggest that this false assumption means that maximum likelihood estimation with simple models will always have some arbitrary bias when studying rank-frequency data in natural language, including both ABC and Clauset et al's method.

3.3 Model

3.3.1 Likelihood Function - General Case With No A Priori Ordering

A vector of data, $d = [d_1, d_2, ...d_N]$, represents N observations of a random variable X. Each of these observations are one of a discrete set of W events, with no a priori ordinality. An example is words in a book.

We can transform the vector d to counts of each event, ordered from most to least frequent, $n = [n(x_{(1)}), n(x_{(2)}), ..., n(x_{(W)})]$. $n(x_{(r_e)})$ represents the count of the r_eth most common event, where r_e is the event's ranking in the empirical frequency distribution. For ease of notation we will refer to $n(x_{(r_e)})$ as $n(r_e)$.

We assume a simple model where each of these events has some unknown fixed probability of being observed, $p(x_{r_p}) = Pr(X = x_{r_p})$, where r_p is the event's rank in the underlying probability distribution.

The key insight is that given an event's empirical rank, we do not know that event's rank in the underlying probability distribution. We can describe the mapping of events from the data generating probability ranking to the empirical ranking with a vector s, so that $s(r_p) = r_e$. For example $s = [2, 1, 3]$ would mean that the second most probable event was observed empirically the most number of times, the most probable event was seen the second most number of times, and the third most likely seen third most. For any valid mapping, s must be a permutation of the integers from 1 to W. Figure 3.3 shows an example mapping.

We assume that the probability distribution is parameterised by θ. Considering Bayes' rule

$$p(\theta|n) = \frac{p(n|\theta)p(\theta)}{p(n)} \,. \tag{3.3}$$

The likelihood can be written as (ignoring constants of proportionality)

$$p(n|\theta) = \prod_{r_e=1}^{W} p(x_{(r_e)})^{n(r_e)} \,. \tag{3.4}$$

This likelihood equation is in terms of the events' empirical rank, r_e, whereas the underlying probability model is in terms of probability rank, r_p. To convert the likelihood to be in terms of r_p we condition on the mapping vector, s,

$$p(n|\theta, s) = \prod_{r_p=1}^{W} p(x_{r_p})^{n(s(r_p))} \,. \tag{3.5}$$

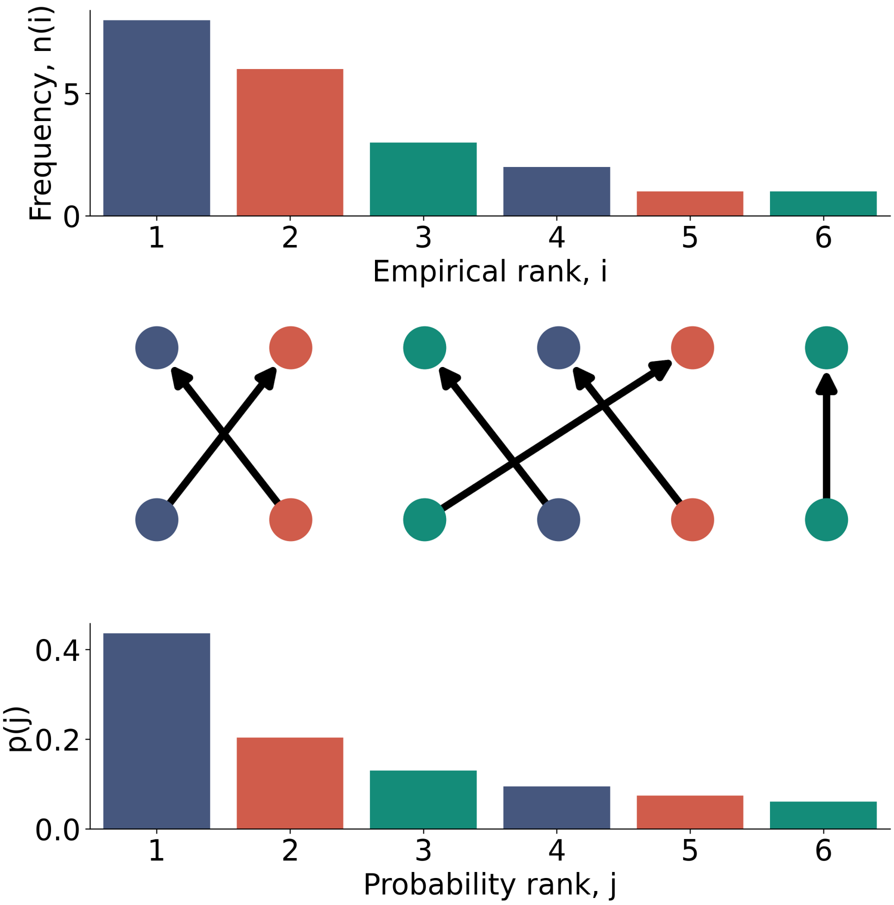


Figure 3.3: An example mapping from probability to empirical ranks. The observed data $n = [8, 6, 3, 2, 1, 1]$ can arise from any valid permutation of events from the probability distribution. Here the permutation is $s = [2, 1, 5, 3, 4, 6]$. The 1st most likely event is observed the second most times ($s[1] = 2$), etc. The likelihood of the data given this permutation is $p(n|s, \theta) = p_1^6 p_2^8 p_3^1 p_4^3 p_5^2 p_6^1$

.

Using the law of total probability we sum over all possible mappings of probability rankings onto empirical rankings. $S(W)$ is the set of all possible permutations of the numbers 1 to W, known as the symmetric group,

$$p(n|\theta) = \sum_{s \in S(W)} \prod_{r_p=1}^{W} p(x_{r_p})^{n(s(r_p))} \,. \tag{3.6}$$

Equation 3.14 is the likelihood for any data that represents observations of discrete events, where the events have no a priori ordering in relation to the underlying model. The equation generalises to $W \to \infty$, suitable to describe models with unbounded event sets, as is the case in many Zipf type models.

3.3.2 Likelihood Function - Power Laws With No A Priori Ordering

A common model applied to rank-frequency distributions is the power law, used by Zipf in his study of words [Zipf, 1949]. A power law probability distribution is of the form

$$p(x_{r_p}) = \frac{r_p^{-\lambda}}{Z_\lambda} \,, \tag{3.7}$$

where λ is the power law exponent, Z_λ is a normalising factor. We use the simplest form of Zipf's law for ease of analysis. The method described here can be used with other models such as the Zipf-Mandelbrot law [Mandelbrot, 1953]. The normalising factor is

$$Z_\lambda = \sum_{r_p=1}^{W} r_p^{-\lambda} \,, \tag{3.8}$$

where W is the number of possible events. In the limit $W \to \infty$, Z_λ becomes the Riemann zeta function, $\zeta(\lambda)$ [Clauset et al., 2009].

Considering equation 3.14, the likelihood can be written as

$$\mathcal{L}(\lambda|n) = \sum_{s \in S(W)} \prod_{r_p}^{W} \left(\frac{r_p^{-\lambda}}{Z_\lambda} \right)^{n(s(r_p))} \,. \tag{3.9}$$

And the differential of the likelihood with respect to λ is

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\lambda|n) = \sum_{s \in S(W)} \left(\left(\frac{N Z_\lambda'}{Z_\lambda} + \sum_{r_p}^{W} n(s(r_p)) ln(r_p) \right) \times \prod_{r_p}^{W} \left(\frac{{r_p}^{-\lambda}}{Z_\lambda} \right)^{n(s(r_p))} \right) \,, \tag{3.10}$$

where Z'_λ is the differential of the normalising factor with respect to λ.

To find the maximum likelihood estimator, we can use numerical methods to either a) maximise equation 3.9 or b) find the root of equation 3.10 (Figure 3.4).

The prevailing estimators from the literature (often implicitly) assume that the empirical ranks match the probability ranks [Piantadosi, 2014; Clauset et al., 2009; Bauke, 2007], so that they only consider the leading term in the main sum in both equations 3.9 and 3.10 (associated with the identity permutation $s_I = [1, 2, ..., W]$). This is the source of the bias in the existing estimators.

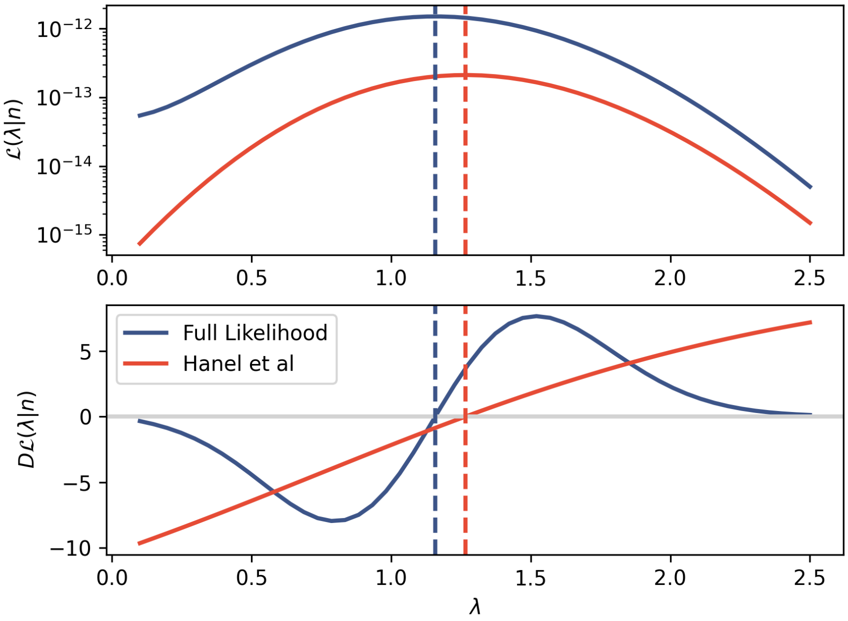


Figure 3.4: Likelihood functions of the full likelihood (blue) and only the leading term (red). Both likelihoods are calculated for the data $n = [10, 3, 3, 2, 1, 1]$. The leading term of the full likelihood is equivalent to the likelihood function as defined by Hanel et al [Hanel et al., 2017], which is adapted for finite event sets from Clauset et al's estimator [Clauset et al., 2009]. The top figure shows the full likelihood compared to Hanel et al's likelihood, with the maximum likelihood estimators shown as dashed lines. The bottom figure shows the differential of the likelihood functions. The form of the differential of the full likelihood is markedly different to only the first term. There is a substantial difference in the maximum likelihood estimator, with the Hanel et al estimator giving $\hat{\lambda} = 1.27$ and the full estimator giving $\hat{\lambda} = 1.16$.

The number of terms in the likelihood function (equation 3.14) scales as $O(W!)$, so that naive computation of the likelihood is impractical even at $W \approx 10$. The computation can be shown to be equivalent to the computation of the permanent of a matrix with entries $a_{ij} = p(x_j)^{n(i)}$. The best known algorithm for exactly computing the permanent of a matrix is Ryser's algorithm [Ryser, 1963; Glynn, 2010] with complexity $O(W2^W)$. This is computationally intractable for real world data sets such as text corpora with vocabularies of $W > 1000$. A more in-depth discussion on the computational complexity can be found in the Supplementary Information.

3.3.3 Approximate Bayesian Computation

Approximate Bayesian computation is a technique for approximating posterior distributions without calculating a likelihood function [Sunnåker et al., 2013; Beaumont et al., 2002; Csilléry et al., 2010]. Instead, we assume a model, $\mathcal{M}$, simulate data, n_i, from possible parameters, λ_i, and observe how close that simulated data is to the empirical data using a distance measure $\rho(n_i, n_{obs})$ [Sunnåker et al., 2013; Csilléry et al., 2010]. The ABC rejection algorithm is based upon the principle that we can approximate the actual posterior by estimating the probability of λ given that the data is within some small tolerance, ϵ, of the observed empirical data [Sunnåker et al., 2013; Sisson et al., 2007]. This assumes that the model, $\mathcal{M}$, is a good representation of the actual data generating process.

$$p(\lambda | n = n_{obs}, \mathcal{M}) \approx p(\lambda | \rho(n, n_{obs}) < \epsilon, \mathcal{M}) \tag{3.11}$$

$$p(\lambda | \rho(n, n_{obs}) < \epsilon, \mathcal{M}) = \frac{p(\rho(n, n_{obs}) < \epsilon | \lambda, \mathcal{M}) p(\lambda | \mathcal{M})}{p(\rho(n, n_{obs}) < \epsilon | \mathcal{M})} \tag{3.12}$$

The ABC rejection algorithm begins by sampling parameter values from the prior. For each of these parameter values, data is then generated from the model and tested on the condition $\rho(n_i, n_{obs}) < \epsilon$ [Sunnåker et al., 2013]. With enough samples, the density of successful parameters will approximate the right hand side of Equation 3.12, and an approximation for the posterior distribution [Sunnåker et al., 2013]. If we use a uniform prior then this will be a proportional estimate to the likelihood.

An ideal distance measure, $\rho(n_i, n_{obs})$, would involve comparing Bayesian sufficient summary statistics from the data [Csilléry et al., 2010]. Usually in practice Bayesian sufficiency cannot be achieved [Csilléry et al., 2010; Sunnåker et al., 2013], and some information will be lost so that the approximation of the posterior includes some error [Sunnåker et al., 2013]. A common technique is to summarise the data sets with summary statistics, $S(n)$, and define the distance as the difference between those, $\rho(n_i, n_{obs}) = S(n_i) - S(n_{obs})$ [Beaumont, 2010; Sunnåker et al., 2013; Csilléry et al., 2010]. Recently the Wasserstein distance, a metric between distributions, has been shown to work well as a distance measure [Bernton et al., 2019]. This is a principled approach that avoids the difficult selection of summary statistics [Bernton et al., 2019], and this is the measure that we use here.

The ABC rejection algorithm requires a small tolerance in order to find a good estimate for the posterior [Sisson et al., 2007]. This in turn requires a high density of samples in order to have enough successful parameters to build the posterior approximation. To sample at a high density across a reasonable parameter space with a uniform prior would be prohibitively computationally expensive. Instead, we use population Monte Carlo to sample from a proposal distribution that focuses on areas of high posterior probability while avoiding areas of negligible probability [Cappé et al., 2004]. At each time step, the results are weighted using principles from importance sampling to account for the fact that we are sampling from the proposal distribution instead of the prior [Cappé et al., 2004]. This algorithm, adapted from [Beaumont et al., 2009], is shown in Algorithm 1 and Figure 3.9 (the 2 parameter algorithm is equivalent, with the variance replaced by a covariance matrix). The parameters in the algorithm were set following trial and error to balance computation time and accuracy.

We also investigated an alternative approximate Bayesian computation approach known as ABC regression. Instead of the Wasserstein distance, we used the mean of the log transformed event counts as a summary statistic with this method. Full details are in the Supplementary Information.

3.4 ABC Results

3.4.1 Approximate Bayesian Computation with Zipf Distributions

Rank-frequency data was generated (N =10,000) from an unbounded power law with exponents ranging from 1 to 2. For each generated data set, the exponent was estimated using a) Clauset et al's estimator and b) ABC-PMC with the Wasserstein distance. This was repeated 100 times to find the mean bias and variance. The ABC method has much lower bias and similar variance to Clauset et al's method, (Figure 3.10).

We also investigated how the bias changes with varying sample size. Rank-frequency data was generated with $\lambda = 1.1$ and varying sample size up to N =1,000,000. Clauset et al's estimator shows positive bias at all values of N, although it decreases with large N. ABC shows much lower bias for all values of N. The variance of ABC is higher for $N \lessapprox 1000$. Overall the variance is still very low, and is insignificant

Algorithm 1: Approximate Bayesian Computation Population Monte Carlo Zipf's Law

```
Input: The observed data $n = [n_1, n_2, \dots, n_W], \theta_{min} \leftarrow 1.001, \theta_{max} \leftarrow 3, survivalFraction \leftarrow 0.4, nParticles \leftarrow 256, nGenerations \leftarrow 10$
Output: Maximum likelihood estimator $\hat{\theta}$
$priorDist \leftarrow uniformDist(\theta_{min}, \theta_{max})$
$nData \leftarrow sum(n)$
$tolerance \leftarrow \infty$
$proposalDist \leftarrow priorDist$
for $g \leftarrow 1$ to $nGenerations$ do
    $\theta s \leftarrow array()$
    $ds \leftarrow array()$
    $weights \leftarrow array()$
    for $i \leftarrow 1$ to $nParticles$ do
        $hit \leftarrow FALSE$
        while $!hit$ do
            $\theta \leftarrow proposalDist.sample()$
            if $\theta_{min} \leqslant \theta \leqslant \theta_{max}$ then
                $z \leftarrow generateData(\theta, nData)$
                $d \leftarrow wassersteinDistance(n, z)$
                if $d \leqslant tolerance$ then
                    $\theta s[i] \leftarrow \theta$
                    $ds[i] \leftarrow d$
                    $weights[i] \leftarrow priorDist.evaluate(\theta)/proposalDist.evaluate(\theta)$
                    $hit \leftarrow TRUE$
    $tolerance \leftarrow getTolerance(ds, survivalFraction)$
    $var \leftarrow weightedVariance(\theta s, weights)$
    $proposalDist \leftarrow KDE(\theta s, weights, bandwidth = sqrt(2 \times var))$
$posterior \leftarrow KDE(\theta s, weights, bandwidth = sqrt(var))$
$\hat{\theta} \leftarrow max(posterior)$
return $\hat{\theta}$
```

compared to the positive bias showed by Clauset et al's estimator (Figure 3.11).

In addition to the results shown here, we explored a variation of the algorithm using ABC rejection with the mean of the logged event counts as a summary statistic. This method has similarly low bias and variance as the results shown here. See the Supplementary Information for full details.

3.4.2 Approximate Bayesian Computation with Zipf-Mandelbrot Model

The Zipf-Mandelbrot law is a modification of Zipf's law derived by Mandelbrot that accounts for a departure from a strict power law in the head of the rank-frequency distribution [Mandelbrot, 1953],

$$p(r_p) \propto (r_p + q)^{-\lambda} \,, \quad q \in [0, 1, 2...] \,. \tag{3.13}$$

We tested the ABC PMC algorithm with this 2 parameter model. The algorithm is of the same form as Algorithm 1, with the variance replaced with a covariance matrix. The algorithm is demonstrated with one generated data set with q =4, λ=1.2 and N =100,000. ABC PMC performs well, with close estimates to the true parameters (see Figure 3.8). The approximated likelihood function gives negligible probability for q =0, suggesting that the algorithm can discriminate between data generated from Zipf's law and the Zipf-Mandelbrot law.

3.4.3 Analysis of Books

Both Clauset et al's method and the approximate Bayesian computation method described here assume a Zipfian data generating model. We have demonstrated that ABC-PMC with the Wasserstein distance works well for data generated from a known power law, with much lower bias than Clasuet et al's method. In the Supplementary Information, we also describe an ABC regression method using the mean log of the word counts that has similar low bias when applied to data from a power law distribution.

It is reasonable to suggest that natural language is a more complex process than drawing words from a power law probability distribution. Indeed, deep learning language models like GPT-3 use billions of parameters [Brown et al., 2020]. As such, models that assume Zipfian data generating models are not necessarily suitable for analysing language. To demonstrate the problem, we analysed books using a) Clauset et al's method, b) ABC-PMC with the Wasserstein distance c) ABC regression with the mean of the log transformed word counts as a summary statistic

(Table 3.1). All of the books were downloaded from Project Gutenberg [Gut, 2020]. Each text sample was first "cleaned" by removing all punctuation, replacing numbers with a # symbol, and converting all text to lowercase. The word frequencies were then counted.

The two forms of ABC give different results, which bracket the results of the Clauset et al estimator. This does not imply that the Clauset et al is the best approximator as we show above that it is biased upwards. What these results indicate is that there is no correct "ground truth" because the assumed underlying models are wrong.

Book	**Clauset et al**	**ABC PMC with Wasserstein**	**ABC regression with mean log**
Moby Dick	1.19	1.25	1.16
A Tale of Two Cities	1.21	1.27	1.17
Alice In Wonderland	1.22	1.25	1.18
Chronicles of London	1.19	1.20	1.15
Ulysses	1.18	1.22	1.14

Table 3.1: Comparision of estimators of Zipf's law in books.

3.5 Discussion

We have demonstrated that the prevailing Zipf's law maximum likelihood estimators for rank-frequency data are biased due to an inappropriate likelihood function. This bias is particularly strong in the range of natural language, with exponents close to 1. The correct likelihood function is intractable. We have presented one approach to overcoming this bias using a likelihood-free method of approximate Bayesian computation. The ABC method is shown to work well with data generated from actual power law distributions, with lower bias than Clasuet et al's estimator.

ABC works well in an idealised situation where the true model is known. However when applied to analysing books, the two ABC approaches that we explored give very different estimates for the Zipf exponents. The Zipfian approaches we investigate all assume a simple bag of words probability model, whereas our results on books indicate that natural language generation is a more complex process– otherwise the two ABC methods would converge. The ABC algorithms are searching a parameter space for the closest model based on the distance measure. This works well when the parameter space includes the true data generating process. But with natural language the assumed simple Zipf model is wrong so there is no "correct" location in the parameter space (or the "correct" location is outside the parameter space). Different distance measures will prejudice different aspects of the observed data and so arrive at different estimates. This bias is arbitrary in nature and there seems to be no reasonable way to decide which distance measure is “correct”. The error lies in the assumption of an incorrect data generating model. This problem applies to ABC and Clauset et al’s estimator, and seems to be inherent in applying maximum likelihood estimation using simple models to describe rank-frequency power laws in natural language.

Zipf’s law for word types [Corral et al., 2019] is an empirical relationship between frequencies of words and ranks in that frequency distribution. The difficulty arises when a probabilistic model is used to describe the mechanism that is generating this relationship, when the actual mechanism is more complex. The main aim of this publication is to clearly show that Clauset et al’s estimator is biased for rank-frequency data. The correct likelihood function provides an unbiased framework that works well when the underlying data generating process is known. This does not appear to be the case for natural language. All Zipf estimators have some bias and the best choice will depend on the specific application. Graphical methods such as ordinary least squares may be more suitable to study Zipf’s law when investigating the empirical relationship between ranks and frequencies (Equation 3.1) and not the probability distribution (Equation 3.2). The bias in rank-frequency estimation provides some support for focusing on the alternative frequency-size representation of word counts and Zipf’s law for sizes [Corral et al., 2019] when studying natural language.

3.6 Supplementary Information — Computational Complexity

The general likelihood for inferring probability distributions from rank-frequency data is given in the main paper as

$$p(n|\theta) = \sum_{s \in S(W)} \prod_{r_p}^{W} p(x_{r_p})^{n(s(r_p))} \,. \tag{3.14}$$

The number of terms in the likelihood function scales as $O(W!)$, so that naive computation of the likelihood is impractical even at $W \approx 10$. When analysing Zipf’s law for words in a book W represents the writer’s vocabulary. Even considering a lower bound for W as the number of unique words in a book, $W > 1000$ so that the likelihood is extremely computationally expensive using a naive algorithm. Here we will explore how to make this computation more efficient.

The full likelihood function (equation 3.14) is equivalent to the calculation of the permanent of a matrix with entries $a_{ij} = p(x_j)^{n(i)}$:

$$A = \begin{bmatrix} p_1^{n(1)} & p_2^{n(1)} & \cdots & p_W^{n(1)} \\ p_1^{n(2)} & p_2^{n(2)} & \cdots & p_W^{n(2)} \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{n(W)} & p_2^{n(W)} & \cdots & p_W^{n(W)} \end{bmatrix}, \tag{3.15}$$

$$\mathcal{L}(\theta|n, M)) = per(A)\,. \tag{3.16}$$

The permanent is similar to the determinant, with the difference that the negative signs in the Laplace expansion formula for the determinant are all positive [Agrawal, 2008]. A well known algorithm for exactly computing the permanent of a matrix is Ryser's algorithm [Ryser, 1963; Glynn, 2010] with complexity $O(W2^W)$. The exact computation of the permanent is thought to be $\#P$-hard [Valiant, 1979; Scott, 2011], so that no polynomial algorithm exists if $P \neq NP$. A polynomial time approximation algorithm for the permanent of a non-negative matrix (as our matrix is), was discovered by Jerrum et al. [2004], with complexity $O(W^{10})$. These algorithms are improvements on the naive case but are still prohibitively computationally expensive for the use case of a text corpora with a vocabulary of $W > 1000$.

We investigated a method of reducing the computational complexity of Ryser's algorithm (in our case) by several orders of magnitude by considering tied empirical ranks, which are equivalent to repeated columns in the matrix A. This can be done but the computation time remains extremely prohibitive. A lower bound to an estimate of the computational complexity using this technique would be $O(F2^F)$, where F is the number of unique empirical counts, as the computation would be at least as complex as computing the permanent of a matrix of the unique columns. This would remain prohibitively computationally expensive for real world data sets. The slim hope that remains is to use the structure and symmetry of the matrix to find some shortcut or a reasonable approximation, we leave this as an open question.

3.7 Supplementary Information — Approximate Bayesian Computation Regression with Mean Log

Approximate Bayesian computation is a technique for approximating posterior distributions without having to calculate a likelihood function [Sunnåker et al., 2013; Beaumont et al., 2002; Csilléry et al., 2010]. Instead, we simulate data, n_i, from possible parameters, λ_i, and observe how close that simulated data is to the empirical data (using a distance measure $\rho(n_i, n_{obs})$). By looking at the behaviour of simulated data with close distances, we can approximate the posterior distribution, $p(\lambda|n_{obs})$.

In order to use ABC to we need a way to measure the "distance" between two data sets. A common technique is to summarise the data sets with a summary statistic, $S(n)$, and define the distance as the difference between those, $\rho(n_i, n_{obs}) = S(n_i) - S(n_{obs})$ [Beaumont, 2010; Sunnåker et al., 2013]. A good summary statistic will capture a lot of information relevant to the likelihood function so that $p(\lambda|n) \sim p(\lambda|S(n))$ []. With rank-frequency distributions, the mean of the logs of the observations is of a similar form to the likelihood function derived in the main paper. Through experiment this statistic was found to be a good candidate summary statistic,

$$S_i = \sum_{r_e=1}^{W} n_i(r_e) log(r_e) \, . \tag{3.17}$$

There are several flavours of ABC [Beaumont, 2010; Csilléry et al., 2010]. Here we use the regression method [Beaumont et al., 2002; Csilléry et al., 2010; Leuenberger and Wegmann, 2010]. We only consider distances within some tolerance, ϵ, of the observed data, i.e. $|S(n_i) - S(n_{obs})| < \epsilon$. The regression method has advantages over the rejection method that it is computationally more efficient and does not require careful tuning of the tolerance [Beaumont et al., 2002]. The key assumption is a linear approximation within the tolerance region:

$$\lambda_i = \beta S(n_i) + \alpha + \phi_i \, . \tag{3.18}$$

Assuming that ϕ has an invariant distribution within this tolerance region, we can find estimates $\hat{\beta}$ and $\hat{\alpha}$ using ordinary least squares regression. To estimate the posterior we are interested in $p(\lambda|S(n_{obs}))$, which can be estimated by translating the data points along the regression line,

$$\lambda_i^* = \lambda_i - \hat{\beta}(S(n_i) - S(n_{obs})) \, . \tag{3.19}$$

The frequency histogram of these translated points will be approximately proportional to the likelihood function. The histogram can be smoothed using a kernel density estimate and the mode taken to find the maximum likelihood estimator. The process is summarised in Figure 3.9.

3.7.1 ABC Regression Results

Rank-frequency data was generated ($N = 10000$) from an unbounded power law with exponents ranging from 1 to 2. For each generated data set, the exponent was estimated using a) Clauset et al's estimator and b) ABC. This was repeated 100 times to find the mean bias and variance. The ABC method has much lower bias and similar variance to Clauset et al's method, (Figure 3.10).

We also looked at changing sample size. Rank-frequency data was generated with $\lambda = 1.1$ and varying sample size up to $N = 1000000$. Clauset et al's estimator shows positive bias at all values of N, although it decreases with large N. ABC regression shows much less bias at all tested values of N. The variance of ABC regression is higher for $N \lessapprox 1000$. Overall the variance is still very low, and is insignificant compared to the positive bias showed by Clauset et al's estimator (Figure 3.11).

Overall ABC regression with the mean log as a summary statistic shows much less bias and similar variance to Clauset et al's estimator, when applied to data generated from a Zipfian probability distribution.

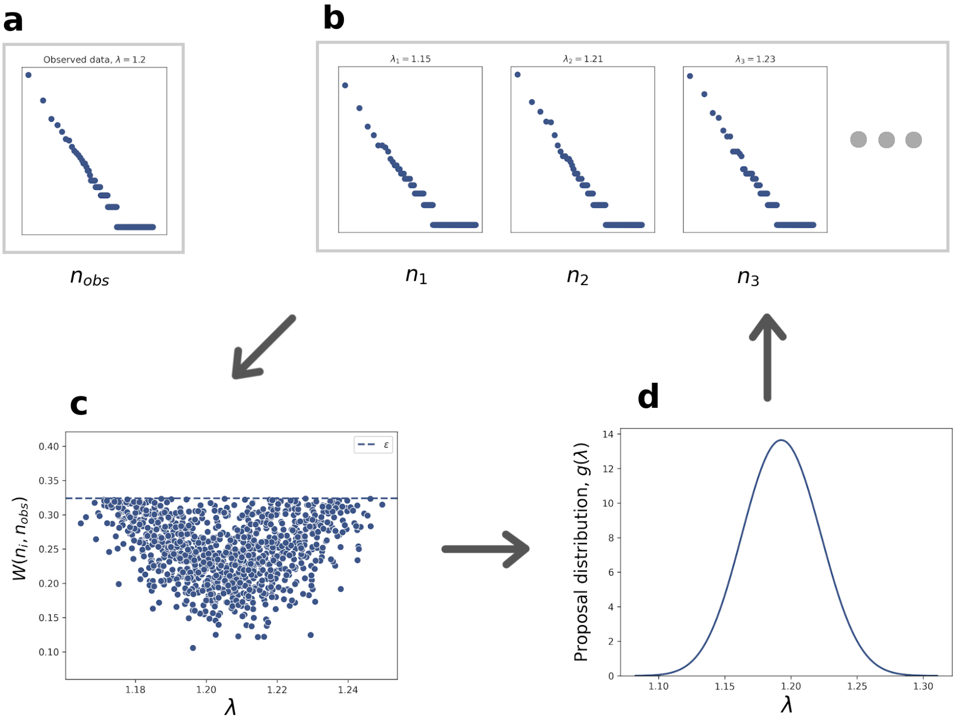


Figure 3.5: Approximate Bayesian computation with population Monte Carlo (ABC-PMC). a) Given the observed data. b) Particles are generated from a proposal distribution and data is simulated for each particle. For each particle, the Wasserstein distance is measured between the simulated data and the observed data. c) This is repeated until $nParticles$ samples are generated with Wasserstein distance within a tolerance ϵ. d) A new proposal distribution is generated by a weighted kernel density estimate on the accepted particles, with a weighting based on importance sampling principles. A new tolerance is set based upon a proportion of $survivalFraction$ particles with the smallest distances found in this time step. This is repeated for a given number of generations. The final successful particles are used to generate an approximation of the posterior distribution using a weighted kernel density estimate. Figure adapted in part from [Sunnåker et al., 2013] and [Csilléry et al., 2010].

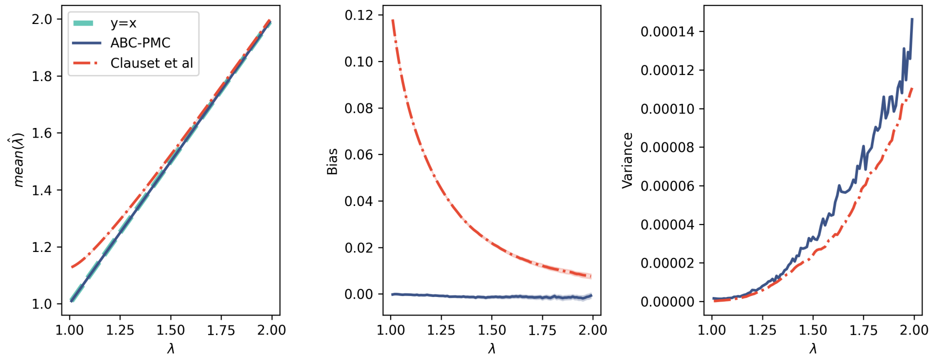


Figure 3.6: Bias in ABC (solid blue) vs Clauset et al's estimator (dashed red) for unbounded power laws. For each of 100 values of λ between 1.01 and 2, rank-frequency data (N =10,000) was generated by sampling an unbounded power law. This was run 100 times. The left figure shows the known λ and the mean estimated λ. The centre figure shows the mean bias, with a 68% confidence interval shaded. The right figure shows the variance of the estimators. The ABC estimator has much lower bias and similar variance to Clauset et al's estimator.

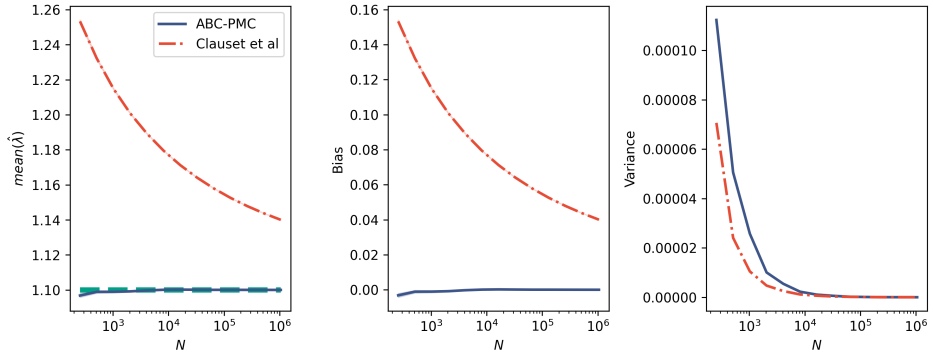


Figure 3.7: Bias in ABC (solid blue) vs Clauset et al's estimator (dashed red) for unbounded power laws. Rank-frequency data was generated for $\lambda = 1.1$ with varying sizes, N. This was run 100 times. The left figure shows the known λ against the mean estimated λ. The centre figure shows the mean bias, with a 68% confidence interval shaded. The right figure shows the variance of the estimators. The bias is much lower with ABC. The ABC estimator has higher variance than Clauset et al at low N, although the variance is still very low.

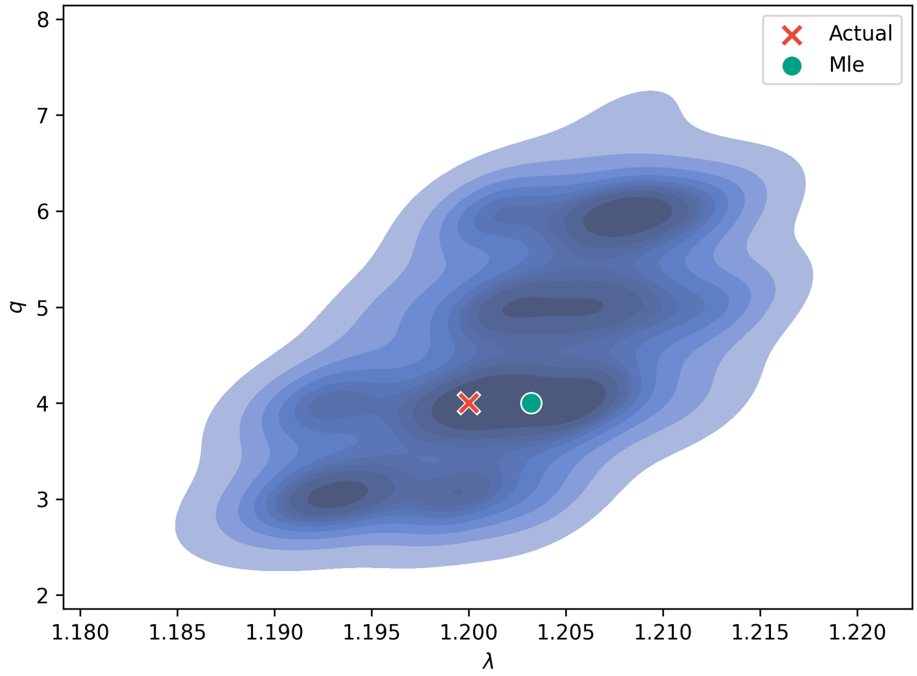


Figure 3.8: Results of ABC-PMC for the Zipf-Mandelbrot law with data generated with known exponent $\lambda = 1.2$ and $q = 4$ (red cross) with N =100,000 words. The likelihood function (darker blue regions have higher likelihood) was approximated using a kernel density estimate. The mode of the KDE gives the maximum likelihood estimate (green circle). The estimator correctly identifies q and is close to the correct exponent λ.

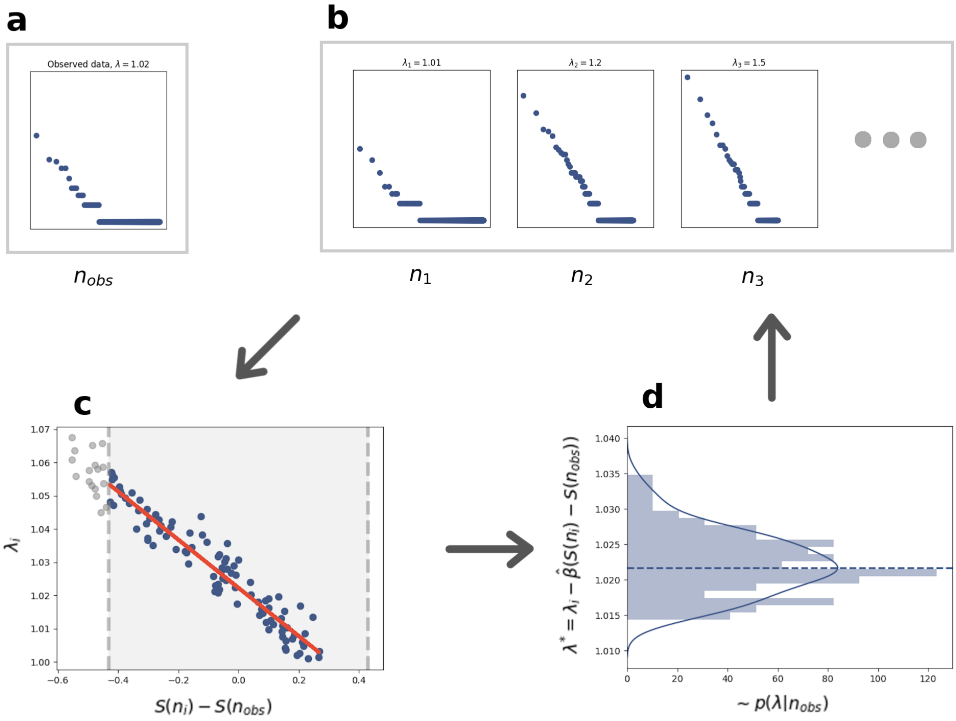


Figure 3.9: Approximate Bayesian computation regression with the mean log. ABC proceeds as shown. a) A summary statistic $S(n)$ is calculated from the observed data. b) Parameters are sampled from a uniform distribution. For each parameter, λ_i a set of data, n_i, is generated, and a summary statistic, $S(n_i)$, is calculated. c) A tolerance is chosen to accept a given proportion, P_ϵ, of the simulations with close summary statistics to the observed data, shown as the shaded region. A linear regression is fit to the accepted simulation results. d) The accepted parameters are adjusted along the regression line to $S(n_i) = S(n_{obs})$. The histogram of these corrected parameter values approximates the likelihood function. A kernel density estimate is used to smooth the likelihood and find the maximum likelihood estimate for λ. Here the initial data was generated with $\lambda = 1.02$ and the maximum likelihood estimator was $\hat{\lambda} = 1.023$, this is a typical result. Figure idea adapted from [Sunnåker et al., 2013] and [Csilléry et al., 2010].

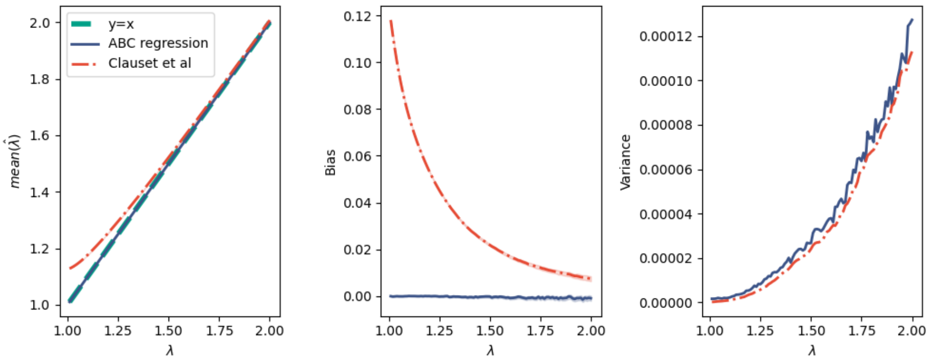


Figure 3.10: Bias in ABC regression (blue solid line) vs Clauset et al's estimator (red dashed line) for unbounded power laws. Rank-frequency data was generated with $N = 10,000$ for 100 values of λ between 1.01 and 2. This was run 100 times. The left figure shows the known λ against the mean estimated $\hat{\lambda}$ over 100 runs. The central figure shows the mean bias (the difference between the mean estimated $\hat{\lambda}$ and λ) with a shaded 68% confidence interval. The right figure shows the variance of the estimators. The ABC estimator has much less bias and similar variance to Clauset et al's estimator.

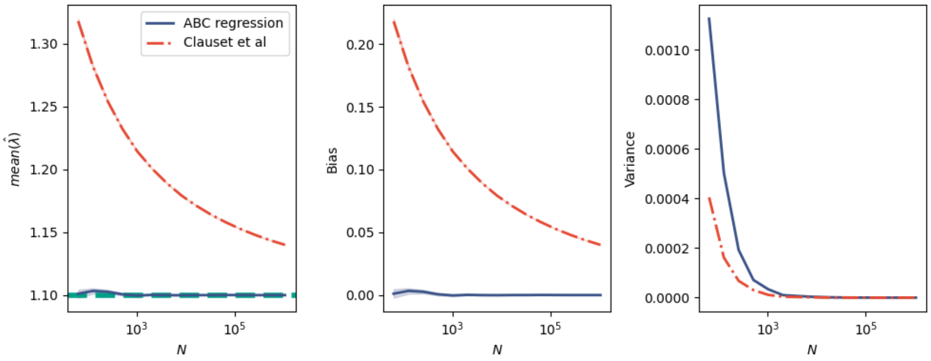


Figure 3.11: Bias in ABC regression (blue solid line) vs Clauset et al's estimator (red dashed line) for unbounded power laws. Rank-frequency data was generated for $\lambda = 1.1$ with varying sizes, N. This was run 100 times. The left figure shows the known λ against the mean estimated $\hat{\lambda}$. The centre figure shows the mean bias, with a 68% confidence interval shaded. The right figure shows the variance of the estimators. The ABC estimator has much smaller bias and similar variance to Clauset et al's estimator.

Chapter 4

Confirmation Bias Emerges from an Approximation to Bayesian Reasoning

4.1 Abstract

Confirmation bias is defined as searching for and assimilating information in a way that favours existing beliefs. We show that confirmation bias emerges as a natural consequence of boundedly rational belief updating by presenting the BIASR model (Bayesian updating with an Independence Approximation and Source Reliability). In this model, an individual's beliefs about a hypothesis and the source reliability form a Bayesian network. Upon receiving information, an individual simultaneously updates beliefs about the hypothesis in question and the reliability of the information source simultaneously. If the individual updates rationally then this introduces numerous dependencies between beliefs, the tracking of which represents an unrealistic demand on memory. We propose that human cognition overcomes this memory limitation by assuming independence between beliefs, evidence for which is provided in prior research. We show how a Bayesian belief updating model incorporating this independence approximation generates many types of confirmation bias, including biased evaluation, biased assimilation, attitude polarisation, belief perseverance and confirmation bias in the selection of sources.

4.2 Introduction

Confirmation bias is the search for and assimilation of information in a way that favors the preservation of prior beliefs [Nickerson, 1998]. It has been described as one of the most pernicious [Nickerson, 1998] of the cognitive biases, with impacts felt in many social domains including religion [Batson, 1975; Nickerson, 1998], politics [Nyhan and Reifler, 2010; Lord et al., 1979; Nickerson, 1998; Taber and Lodge, 2006], climate change [Cook and Lewandowsky, 2016; Hart and Nisbet, 2012], health and medicine [Liberman and Chaiken, 1992; Nickerson, 1998; Malthouse, 2022], justice [Nickerson, 1998], stereotyping [Darley and Gross, 1983], conspiracy theories [McHoskey, 1995], and science [Mahoney, 1977; Nickerson, 1998]. Understanding the underlying cognitive mechanisms that drive confirmation bias is therefore of fundamental theoretical and practical interest.

Confirmation bias encompasses numerous distinct but closely related behaviours [Klayman and Ha, 1987; Friedrich, 1993; Fischhoff and Beyth-Marom, 1983; Nickerson, 1998]. Though many such behaviours have been identified, we focus on five here which have received wide empirical support (see Klayman [1995] for a review): i) *biased evaluation*: judging information that opposes one's views more critically than that which supports them [Koehler, 1993; Lord et al., 1979; Taber and Lodge, 2006; Russo et al., 1996]; ii) *biased assimilation*: whereby people are less influenced by opposing than confirmatory sources [Lord et al., 1979; Taber and Lodge, 2006]; iii) *attitude polarisation*: extreme views both for and against a hypothesis can become more extreme upon seeing the same evidence [Lord et al., 1979; Taber and Lodge, 2006]; iv) *belief perseverance*: the reluctance to change beliefs in the face of disconfirmatory evidence [Anderson et al., 1980; Batson, 1975]; and v) *confirmation bias in the selection of sources*: preferring sources of information that confirm existing beliefs [Taber and Lodge, 2006; Redlawsk, 2002]. [1]

Given confirmation bias' wide prevalence and potential negative impact, a natural question is why confirmation bias exists at all? Sufficiently costly tendencies should be expected to disappear under evolutionary pressures [Nickerson, 1998], unless they are themselves an adaptive solution to a more costly alternative. While confirmation bias may be an impediment to finding the truth, the adaptive force on cognition is primarily towards pragmatic survival and only secondarily concerned

[1] There are many other behaviours that may fall under the umbrella of confirmation bias. Importantly is *positive hypothesis testing*, where people search for evidence in a way that will verify existing hypotheses as opposed to falsifying those hypotheses [Klayman, 1995; Klayman and Ha, 1987; Wason, 1960, 1968]. It has been claimed that this should not be labelled as a confirmation bias and is instead a search heuristic which is useful in many real-world contexts [Klayman and Ha, 1987].

with truth seeking [Friedrich, 1993]. In light of this, we may ask is confirmation bias truly a dysfunction, or does it serve some adaptive purpose?

Explanations for confirmation bias have been put forward at the social [Mercier and Sperber, 2011; Norman, 2016; Peters, 2020], individual [Kunda, 1990; Nickerson, 1998; Friedrich, 1993; Festinger, 1962] and information processing levels [Jern et al., 2014; Cook and Lewandowsky, 2016; Gerber and Green, 1999; Koehler, 1993; Henderson and Gebharter, 2021]. These levels of analysis are qualitatively different but are nonetheless connected. Social behaviour emerges from individual behaviour, and individual behaviour emerges, in part, from information processing. In this paper we present a normative explanation at the information processing level, although our description complements many existing social and individual explanations.

Before we go further, it is helpful to discuss the definition of "bias". In the psychological literature, the word bias is used to mean a variety of things [Hahn and Harris, 2014]. This ranges from the everyday usage of the term as a leaning or tendency in one direction, to the precise use in statistics of a systematic departure from accuracy. Within the context of research on beliefs, bias is usually accepted to mean a departure from a normative model [Hahn and Harris, 2014], which is often Bayesian rationality [Klayman, 1995; Hahn and Harris, 2014]. This definition introduces difficulties because behaviour that is irrational given one belief-updating model may be rational given a different belief-updating model. This has been the case for biased evaluation [Koehler, 1993] and attitude polarisation [Jern et al., 2014; Cook and Lewandowsky, 2016; Henderson and Gebharter, 2021]. We aim to sidestep the issue by not claiming that the behaviours are fundamentally biased under all possible belief-updating models. Rather, we will define behaviours as departures from specific rational belief updating models described in previous literature.

Our contribution is multifold. We present a model of information processing that can generate a large range of empirically confirmed confirmation bias type behaviours, more so than other explanations. In particular, we explore existing Bayesian models of inference in a world with uncertain beliefs and unreliable sources of information [Koehler, 1993; Bovens et al., 2003; Olsson, 2011; Hahn et al., 2018; Merdes et al., 2020]. We argue that maintaining full rationality is impossible for realistic agents due to the high memory demands of remembering dependencies between beliefs. As a consequence, humans are forced to make approximations in order to maintain complex world models. We demonstrate how these approximations to rationality can introduce small biases that magnify as data is processed sequentially over time. In different task domains, these biases encompass the five confirmation bias behaviours we list above.

We will begin with a discussion of information processing models of belief updating in the literature. This will lead to a description of the BIASR model and an interrogation of each of its assumptions. We will then evaluate each of the 5 confirmation bias type behaviours we list above, defining each and showing how the BIASR model can generate the behaviour. We will end with a general discussion of the model and it's position in the literature.

An example may help build intuition. Alice has a neutral belief about vaccine safety. She talks to her new neighbour Bob, who tells her that vaccines have not been thoroughly tested and that they are dangerous. Alice is at first not entirely convinced, but she does become slightly more wary about vaccines. The next week Bob again tells Alice about the dangers of vaccines. Alice is more receptive now as she already has a slight belief that vaccines are dangerous, and she starts to see Bob as reliable because his information matches her slight belief. This continues for several weeks until Alice is convinced that vaccines are dangerous and that Bob is a very reliable source of information. When Alice now hears on the news that vaccines are safe she is not convinced — after all, both she and Bob can't both be wrong, especially considering how knowledgeable Bob is. Alice's beliefs about the reliability of Bob and the dangers of vaccines are correlated. If she forgets this correlation then she does not give enough consideration to the counterfactual world where Bob is wrong and vaccines are safe. In this example Alice exhibits biased evaluation, biased assimilation and belief perseverance.

Table 4.1: Comparison of information processing models of confirmation bias. Checkmarks denote which behaviours have been explained using the different models.

	Simple Version of Bayes' theorem	Biased Evaluation Prior to Assimilation[Gerber and Green, 1999; Lord et al., 1979]	Bayesian Updating Including Source Reliability [Koehler, 1993]	Belief-based Sequential Updating with Source Reliability [Bovens et al., 2003; Olsson, 2011; Merdes et al., 2020; Hahn et al., 2018]	Bayesian Networks [Jern et al., 2014; Henderson and Gebharter, 2021; Cook and Lewandowsky, 2016]	BIASR. Bayesian updating with an Independence Approximation and Source Reliability
Biased Evaluation		51	51	51	51	51
Biased Assimilation		51		51		51
Attitude Polarisation					51	51
Belief Perseverance		51		51		51
Selection of Sources						51

4.3 Models of Information Processing

Bayes' theorem provides the objectively optimal way to update beliefs given new evidence, where beliefs are described in terms of degrees of uncertainty. For human cognition, inference affects behaviour, which in turn affects adaptive success. One could therefore expect that adaptive pressures over our evolutionary history would drive our inference mechanisms towards Bayesian rationality.

The **simple version of Bayes' theorem** is

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}\,. \tag{4.1}$$

If one applies this simple rule to a single hypothesis, H, then data from all sources is treated equally. There is no judgement of evidence quality. Under this model, any unequal treatment of evidence is considered biased evaluation. Indeed, bias as the unequal consideration of evidence is a definition of confirmation bias (often implicitly) used in the literature [Lord et al., 1979; Plous, 1991; Lord et al., 1984; Miller et al., 1993].

The simple version of Bayes' theorem is not adequate in terms of describing either observed or desirable behaviour (see Table 4.1 and below). Indeed, it is reasonable to judge the quality of evidence based on assessments of the reliability of the source [Fischhoff and Beyth-Marom, 1983]. Lord et al. [1979] state,

> "Our subjects' main inferential shortcoming ... did not lie in their inclination to process evidence in a biased manner. Willingness to interpret new evidence in the light of past knowledge and experience is essential for any organism to make sense of, and respond adaptively to, its environment. Rather, their sin lay in their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially "justified" the processing bias."

Here Lord et al. [1979] are suggesting **biased evaluation prior to assimilation**. Data is first evaluated based on prior beliefs, with unlikely data considered as less reliable evidence. And then the result of this evaluation determines the weight of the evidence in updating those same prior beliefs. This idea was formalised mathematically by Gerber and Green [1999], who present a Bayesian model of belief updating combined with biased learning. The biased learning is represented as a weakening of the strength of evidence that disconfirms prior beliefs, before updating those same beliefs within the Bayesian machinery. They provide an example of a politician's supporters considering whether the politician is corrupt or not. In this example, evidence in support of corruption is discounted by a factor, $\alpha < 1$.

Russo et al proposed a similar model involving "predecisional distortion of information" in relation to choice among alternatives [Russo et al., 1996; Russo, 2014, 2018]. Prior preferences influence the evaluation of data, and this evaluation influences how the data is used to update beliefs, generating a bias towards initial preferences [Russo, 2014, 2018]. These ideas can describe biased evaluation and biased assimilation, and can go some way to describing belief perseverance [Carlson et al., 2006] (Table 1). Though useful, these models are not Bayesian and do not have a clear normative basis.

The Bayesian framework does, however, allow us to incorporate evidence evaluation in the form of **Bayesian updating including source reliability**. Koehler [1993] argues that a normative account of belief updating should consider an individual's prior beliefs about source reliability as well as evidence evaluation. Koehler [1993] proposes a rational Bayesian model that includes source reliability and which is able to generate biased evaluation (Table 1). This model supports proposals in the literature that judging evidence based on prior beliefs is not necessarily irrational, such that it can be rational to consider unlikely evidence more critically [Lord et al., 1979; Koehler, 1993; Klayman, 1995; Fischhoff and Beyth-Marom, 1983].

This idea is extended by Bovens et al. [2003] and Olsson [2011] to account for sequential belief updates when receiving data over time. In both these cases, individuals maintain separate beliefs about a central hypothesis and source reliability, and upon receiving information those beliefs are updated simultaneously [Merdes et al., 2020]. This type of Bayesian updating has been described as a **"belief-based" strategy** for inference in a world with unknown source reliability [Hahn et al., 2018; Merdes et al., 2020]. As data is received, beliefs about the central hypothesis and source reliability both change over time, which influence how subsequent data is interpreted. It has been shown that these models demonstrate order effects such that the order in which information is received changes the final belief [Hahn et al., 2018]. This has been described as a form of confirmation bias [Hahn et al., 2018; Merdes et al., 2020] (connected to belief perseverance), and the updating process is a departure from Bayesian rationality in the same way as the model that we present. In this research, the normative argument is that the rational approach would be too much of a cognitive burden, as it would require remembering all data received so far and updating from initial priors whenever data is received [Hahn et al., 2018]. However, we contend that this normative argument is not complete as it does not consider the alternative rational approach of maintaining a joint belief distribution over the central hypothesis and source reliability.

More generally, **Bayesian networks** are graphical representations of causal and information dependencies between variables which can describe how to reason about events that could be caused by multiple factors [Pearl, 2009]. Cook and Lewandowsky [2016] used Bayesian networks to explain attitude polarisation in participants who were given evidence about climate change. They demonstrated that polarisation can be rational given a Bayesian network, because individuals' beliefs about the evidence they observed were influenced not only by whether they believed climate change was true, but also by their worldview and trust in scientists [Cook and Lewandowsky, 2016].

Jern et al. [2014] generalise this idea by describing the set of Bayesian networks that can lead rational agents to attitude polarisation — crucially this set of networks share the property that upon receiving some data, beliefs about more than one hypothesis are updated simultaneously. In order to generate rational attitude polarisation, individuals require differences in prior beliefs about the "central" hypothesis in question, and importantly also some difference in other "auxiliary" prior beliefs [Gerber and Green, 1999; Henderson and Gebharter, 2021]. For example, those with strong views about the dangers of climate change may also believe that scientific evidence is more reliable than those who are less worried about climate change [Cook and Lewandowsky, 2016]. Bayesian networks can also be used to describe biased evaluation [Jern et al., 2014] (Table 1).

So far our discussion of information processing has centered on Bayesian rationality. However, this is not necessarily the appropriate normative standard when modelling human probabilistic reasoning. We must also take into account realistic cognitive constraints [Dasgupta et al., 2020; Klayman, 1995; Daw et al., 2008]. People update many hypotheses simultaneously [Gershman, 2019], which can be computationally demanding [Dasgupta et al., 2020]. Dependencies between hypotheses mean the computational scale of inference can quickly overwhelm any realistic agent, who will be forced to make approximations to optimal Bayesian inference. To understand how human cognition may overcome this limitation we can take inspiration from computer science, a field with much experience in the approximation of computationally expensive Bayesian reasoning [Sanborn, 2017]. This path from computer science to human cognition is a well worn road, and algorithms such as Markov Chain Monte Carlo have shed light on human cognitive processes including behavioural biases [Sanborn, 2017].

4.4 The BIASR Model Assumptions

We present the BIASR model (Bayesian updating with an Independence Approximation and Source Reliability; see Figure 4.1), which rests on the following assumptions:

1. Source reliability. Upon receiving information, we update our beliefs about the reliability of the source.

2. Simultaneous updating. We update our beliefs about source reliability and the central hypotheses at the same time.

3. Independence approximation. Simultaneous updating introduces dependencies between our beliefs about a) central hypotheses and b) source reliabilities. Our model approximates these dependencies away by taking marginal beliefs and assuming independence.

4. Sequential Updating. Data is received and processed sequentially over time. The independence approximation is applied between sequential updates.

Each of these assumptions will be explored in the following sections. The belief updating process under BIASR is visualised in Figure 4.1.

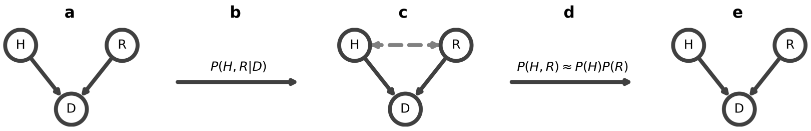


Figure 4.1: The BIASR belief updating model. a) Data received is believed to be causally influenced by both the true value of the hypothesis at hand and the source reliability in a collider Bayesian network. b) When receiving data, beliefs about the hypothesis and source reliability are updated simultaneously. c) This updating introduces information dependencies (grey dotted line) between beliefs. d) These belief correlations can be approximated away by assuming independence. e) This approximation simplifies the Bayesian network structure. The process repeats when more data is received.

4.4.1 Source Reliability

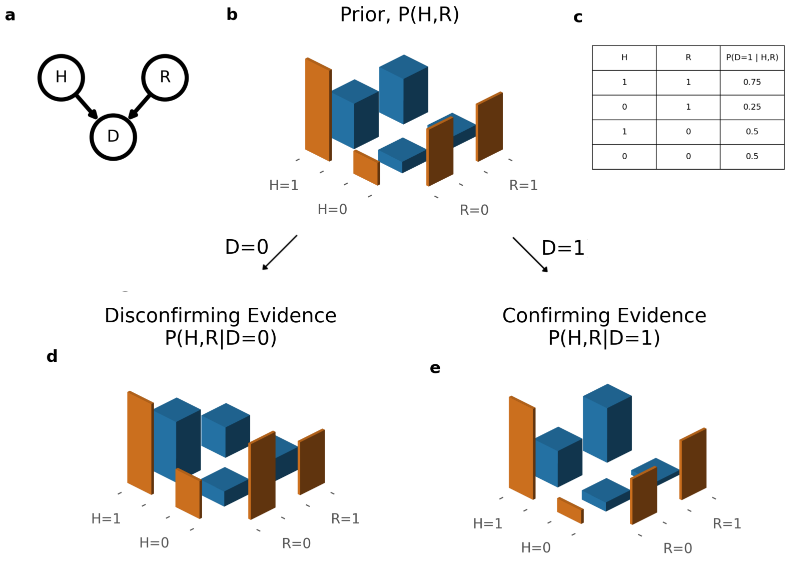
It is rational to hold and update beliefs about the reliability of sources of information Merdes et al. [2020]; Hahn et al. [2018]. Doing so allows us to weigh the quality of evidence based on the source and protects against the influence of unreliable sources that may foster misinformation. The inclusion of source reliability in belief updating models has been suggested as a possible rational basis for the conjunction fallacy [Bovens et al., 2003; Jarvstad and Hahn, 2011; Fischhoff and Beyth-Marom, 1983; Tversky and Kahneman, 1983]. There is also an abundance of empirical evidence that people track source reliability [Mahoney, 1977; Liberman and Chaiken, 1992; Taber and Lodge, 2006; Lord et al., 1979].

4.4.2 Simultaneous Updating of Source Reliability and Central Hypotheses

In the absence of an objective standard of truth, we can judge a source's reliability based on our assessment of the plausibility of the information received [Hahn et al., 2018]. If someone tells us that Elvis Presley is outside, it's a fair guess that we won't believe them. Instead, we are likely to downgrade our belief in them as a reliable source. We update our belief about their source reliability and Elvis simultaneously. Our strong prior belief in the hypothesis that Elvis is not outside is protected by an auxiliary hypothesis in the reliability of the source. In this way our belief in source reliability can absorb disconfirmatory evidence about strongly held central beliefs (Figure 4.2). Empirical evidence shows that people do consider information about source reliability when updating beliefs about a hypothesis, and vice versa [Collins et al., 2018].

The idea that our beliefs are not updated in isolation is known, in the context of scientific epistemology, as the Duhem-Quine thesis [Gershman, 2019]. No hypothesis can be tested in isolation and upon receiving evidence we update a set of beliefs together, sometimes partitioned into central and auxiliary hypotheses, while maintaining overall coherence. The auxiliary hypotheses (e.g., source reliability) can act to absorb disconfirmatory evidence, allowing us to maintain central beliefs (Figure 4.2). This can be rational; if a scientist detects faster than light travel it is sensible to question the accuracy of the measurements [Gershman, 2019; Lord et al., 1979]. Empirically, scientists question whether disconfirmatory evidence is the result of an error before abandoning a central hypothesis [Dunbar, 1995]. And people test hypotheses more extensively when told that discomfirmatory evidence may be in error [Gorman, 1989].

Following existing models in the literature [Koehler, 1993; Merdes et al., 2020], our model assumes that individuals believe evidence, $D \in \{0, 1\}$, is influenced by both the truth value of the central hypothesis, $H \in \{0, 1\}$, and the reliability of the source, $R \in \{0, 1\}$ (Figure 4.1). In our model all sources are less than perfectly reliable, but a "reliable" source, $R = 1$, has less noise and is more likely to report



H	R	P(D=1 | H,R)
1	1	0.75
0	1	0.25
1	0	0.5
0	0	0.5

Figure 4.2: Simultaneous updating of source reliability, R, and the central hypothesis, H. Blue (thick) bars show the joint belief distribution. Orange (thin) bars show the marginal belief distributions. a) The Bayesian network structure describes how the individual believes the data is influenced by the true values of the central hypothesis and source reliability, given by c) the conditional probability distribution. b) Prior beliefs favour the central hypothesis, $P(H = 1) = 0.8$, and are neutral about the source reliability, $P(R = 1) = 0.5$. d) The posterior following disconfirming evidence shows how disconfirmatory data can be explained away as coming from an unreliable source, with only a small impact on belief in the central hypothesis. e) The posterior following confirming evidence is updated towards stronger belief in both the central hypothesis and the reliability of the source.

the true value of the central hypothesis than an “unreliable” source, $R = 0$, which has more noise. These probabilities can be quantified as $P(D = 1|R = 1, H = 1)$ for the reliable source and $P(D = 1|R = 0, H = 1)$ for the unreliable source (and symmetrical in the case that $H = 0$). A quantitative example is given in Figure 4.2. The true values of the central hypothesis and the source reliability are causally independent. Our individual has prior beliefs in the central hypothesis, $P(H)$, and the source reliability, $P(R)$. Assuming initial independence between these beliefs, this is a collider type Bayesian network, $H- > D < -R$. Notably, this is one of the set of Bayesian networks that Jern et al. [2014] proved can lead to attitude polarisation. Beliefs are simultaneously updated by Bayes' rule,

$$P(H, R|D) = \frac{P(D|H, R)P(H, R)}{P(D)}\,. \tag{4.2}$$

Initially, the individual's beliefs in H and R are independent so that $P(H, R) = P(H)P(R)$ and we can simplify the update rule to

$$P(H, R|D) = \frac{P(D|H, R)P(H)P(R)}{P(D)} \,. \tag{4.3}$$

A general property of causal graphs of this structure, including collider type Bayesian networks, is that upon receiving data, beliefs in H and R are no longer independent [Pearl, 2009] (Figure 4.1). Mathematically, the individual's beliefs no longer fulfil the independence relationship and $P(H)P(R) \neq P(H, R)$. If we later learn that Elvis is alive, we should update our belief in our friend's reliability. Moreover, upon receiving subsequent evidence from the same source about the same hypothesis, in order to remain Bayesian rational we can no longer use the simpler update rule (Equation 4.3) and must instead consider the full joint belief distribution (Equation 4.2).

Notably, this formalism is different to previous work on belief-based updating [Merdes et al., 2020; Hahn et al., 2018]. In that work, beliefs are assumed to be stored marginally, i.e. individuals have some belief about the hypothesis in question and a separate belief about the reliability of a source. Given this belief structure, the rational benchmark is to remember the entire history of data received and carrying out full inference on all the data at each timestep [Hahn et al., 2018], which Merdes et al. [2020] argue is unrealistic and not normative. We argue that the more complete rational benchmark also includes the possibility of maintaining a joint belief distribution as described here. However, as we will see in the following section, normative arguments based on cognitive limitations recover the marginal belief structure.

4.4.3 Independence Approximation

In our minimal example (Figure 4.2) we are considering only one central hypothesis, H, and one source reliability, R, each of which can take values of either 1 or 0. In this case, the joint belief distribution is relatively small, with 4 possible combinations, $\{(H = 1, R = 1), (H = 1, R = 0), (H = 0, R = 1), (H = 0, R = 0)\}$, shown as the blue (thick) bars in Figure 4.2. In our actual day to day reasoning we track many more hypotheses, each with many possible values, and evidence from many different sources. If we combine all these then there are many possible combinations to track in the joint belief distribution.

We can ask how the size of the hypothesis space scales as we add more attributes to a world model. In computer science, the amount of computational resources that an algorithm uses is known as the computational complexity. This can be measured in terms of processing time or memory space. Big O notation is a way of comparing the computational complexity of algorithms as the size of inputs to that algorithm grows. As we add new attributes to our world model the memory space requirements scale exponentially, as $\sim O(k^n)$, where n is the number of attributes and k is the number of hypotheses per attribute. This is the curse of dimensionality [Bellman, 2015 (1957]. If we were tracking 300 binary attributes about the world then the joint belief distribution would have size $2^{300} \approx 10^{90}$ — more than the number of atoms in the observable Universe.

This combinatorial explosion in memory space will quickly exhaust any reasonable level of cognitive resources (or for a fixed cognitive resource, will limit the richness and resolution of the agent's world model). A realistic agent would therefore require an approximation to optimal Bayesian reasoning, and this should be included in a normative account. Alternatively, one could remember the entire history of data received from all sources and carry out the full inference with the initial priors each time new data is received, but this also imposes an unrealistic computational burden [Merdes et al., 2020].

Variational approximations are an approach to approximating Bayesian reasoning that can reduce computational requirements associated with large posterior distributions [Ormerod and Wand, 2010]. One option is to take a mean field approximation of the joint posterior distribution by partitioning variables and assuming that the partitions are independent [Ormerod and Wand, 2010; Sanborn and Silva, 2013]. This kind of approximation has been applied before to understanding how human behaviour can emerge as a consequence of cognition overcoming computationally intractable problems, in the realm of associative learning [Sanborn and Silva, 2013; Sanborn, 2017]. At one extreme, if we assume all variables are independent then our memory requirements now scale linearly as $\sim O(kn)$, a vast improvement. We can now track 300 binary attributes with a belief distribution of size 600. If we limit belief partitions to d variables, then computation scales linearly as $\sim O(n(d^k))$. This type of partial or structured mean field approximation [Sanborn, 2017] will preserve dependencies between some variables while avoiding the curse of dimensionality. Figure 4.3 shows an example of this partial approximation with $d = 10$ variables, as compared to no approximation and a full mean field approximation. It should be noted that we are still updating beliefs simultaneously — the mean field approximation disentangles these beliefs following simultaneously updating.

The quality of this compression in terms of loss of information will depend upon the degree to which the attributes being inferred are actually independent,

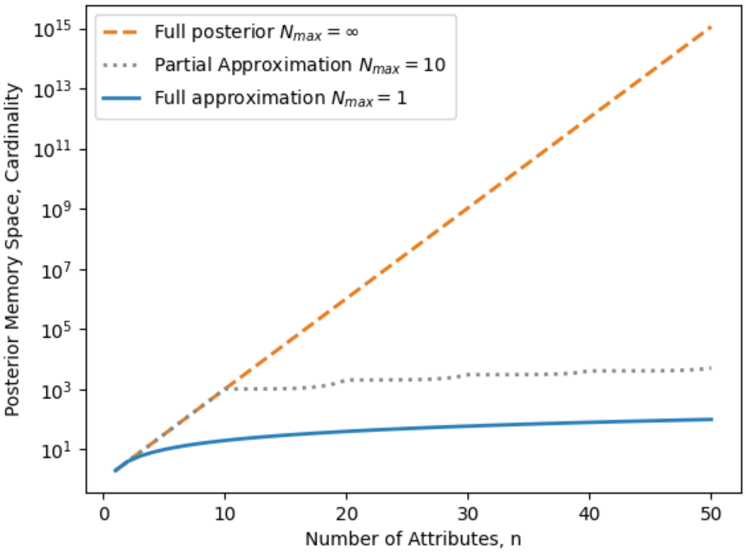


Figure 4.3: Memory space scaling as a function of number of binary attributes. With no approximation (red dashed line) memory requirements scale exponentially (a straight line on this log-linear figure). A full mean field approximation (blue solid line) scales linearly. A partial approximation (grey dotted line) scales exponentially up to d and then linearly after, in this case with $d = 10$.

and the type of approximation we make. A common choice of measure to guide the approximation is to minimise the Kullback-Leibler divergence between the full posterior and the approximated posterior [Sanborn, 2017; Ormerod and Wand, 2010]. This is achieved by taking marginal belief distributions for each attribute. By assuming independence the full joint belief distribution can be approximated from the marginal belief distributions (Figure 4.1),

$$P(H, R) \approx P(H)P(R)\,. \tag{4.4}$$

Human cognition is unlikely to always use a full approximation, such that people would be unable to remember any dependencies between beliefs — the key point is that people are unlikely to remember all the dependencies and will have to make some approximations. Do these approximations include forgetting dependencies between source reliability and more central beliefs? There is evidence that people do not always correctly associate sources of information with their beliefs and instead people can experience source confusion where the belief remains but the source is mis-attributed [Johnson et al., 1993].

4.4.4 Sequential Updating

A common assumption underlying Bayesian inference is exchangeability, i.e. that the order that data is received is irrelevant [Gelman et al., 1995]. Data can be processed in any order, or all at once, and the final beliefs will be the same. This assumption holds in the rational case if the data generating process is static, as in our model. However, exchangeability does not hold under the BIASR model because the independence approximation introduces path dependency, such that biases accumulate over successive steps. Therefore, the order that data is received and sequentially processed influences the final beliefs. A general consequence of sequential updating with approximations to Bayesian inference is the potential loss of exchangeability and the introduction of effects that are dependent on the order of processing [Daw et al., 2008].

There is evidence that people do not process data all at once, but update sequentially. Empirical evidence for this includes the primacy effect [Bruner and Potter, 1964], where the order that data is seen has an influence on final beliefs. In the realm of decision making, Russo [2014] describe a stepwise evolution of preference paradigm. This stepwise updating has been shown empirically in many contexts, with experiments showing that people sequentially update their preferences and their opinion on the diagnosticity of the data [Russo, 2014].

4.5 Evaluation of Five Forms of Confirmation Bias

In this section we evaluate the BIASR model in relation to the five forms of confirmation bias outlined in Table 1. For each form of confirmation bias, we first discuss the literature and empirical evidence. We then define a mathematical requirement for this behaviour in the context of Bayesian rationality. Following this definition, we simulate the behaviour under different models of information processing.

For each form, we simulate how an individual could update their beliefs about whether a central hypothesis is true or false, $H \in \{1, 0\}$, and whether a source is reliable or not, $R \in \{1, 0\}$. If a source is reliable, they transmit the true state of the hypothesis 75% of the time, $P(D = 1|H = 1, R = 1) = 0.75$. If the source is unreliable, they transmit the true value only 50% of the time, $P(D = 1|H = 1, R = 0) = 0.5$. In most cases we use a neutral prior on the source reliability, $P(R) = 0.5$, and a strong prior belief in the central hypothesis, $P(H) = 0.8$. In the case of attitude polarisation, we also include a strong prior belief against the central hypothesis, $P(H) = 0.2$. And in the case of belief perseverance we start with a neutral prior in the central hypothesis, $P(H = 1) = 0.5$. In all the examples, the simulated individual receives multiple datums sequentially from either a single source or two sources. The values used, and the problem setup itself, are intended to minimally demonstrate the behaviours as clearly as possible. The behaviours are robust and emerge under a wide range of parameters.

We consider 3 information processing models:

1. Simple version of Bayes' theorem. Beliefs in source reliability are not updated at all, i.e. the prior belief, in this case $P(R) = 0.5$, remains the same. Beliefs in the central hypothesis are updated according to Bayes' rule,

$$P(H|D) = \frac{P(H)\sum_R P(D|H,R)P(R)}{P(D)}\,. \tag{4.5}$$

2. Rational updating including source reliability. Beliefs about the central hypothesis and source reliability are updated simultaneously. This introduces a dependency in the joint belief distribution, $P(H,R|D)$. This dependency is remembered between successive datums by updating using Bayes' rule over the full joint belief distribution (Equation 4.2). Give the data generating process, this is the rational way to update beliefs. As such, exchangeability holds and this is equivalent to updating on all data received using initial priors.

3. BIASR model (Bayesian updating with an Independence Approximation and Source Reliability). Beliefs are updated as in the rational case, but dependencies between the central hypothesis and source reliability are forgotten between successive datums. We take marginal beliefs

$$P(H) = \sum_R P(H,R) \tag{4.6}$$

   and

$$P(R) = \sum_H P(H,R)\,. \tag{4.7}$$

   We ignore dependencies by using these marginal beliefs in the independent version of Bayes' rule (Equation 4.3).

4.5.1 Biased Evaluation (Biased Assimilation)

In the confirmation bias literature, the terms biased evaluation and biased assimilation are often used interchangeably. We can strictly define evaluation as the judgement of the quality of the evidence and assimilation as concerning the degree of belief change in the central hypothesis at hand. These are separate, but connected, beliefs. In many studies what could strictly be thought of as biased evaluation is sometimes called biased assimilation [Lord et al., 1984; Miller et al., 1993]. This is understandable as the meanings of the two overlap: if a piece of evidence is rated as "more convincing"[Lord et al., 1984] or "more persuasive"[Miller et al., 1993], is that evaluation or assimilation? From a cognitive dissonance perspective, assimilation and evaluation are connected through coherence in beliefs — a disconfirmatory piece of evidence creates a cognitive dissonance that can be resolved through biased evaluation [Kunda, 1990]. Or more simply, contrary evidence is explained away as coming from an unreliable source. An early mention of confirmation bias is found in the writings of Bacon [1620], who also links evaluation and assimilation,

> "Once a human intellect has adopted an opinion (either as something it likes or as something generally accepted), it draws everything else in to confirm and support it. Even if there are more and stronger instances against it than there are in its favor, the intellect either overlooks these or treats them as negligible or does some line-drawing that lets it shift them out of the way and reject them. This involves a great and pernicious prejudgment by means of which the intellect's former conclusions remain inviolate." *Francis Bacon*

There is strong empirical evidence for biased evaluation, some of which also supports biased assimilation. Mahoney [1977] found that scientists judged studies more harshly when the findings disagreed with their own theoretical positions. This was followed by Lord et al. [1979], who ran an experiment with two sets of students — those with strong prior opinions either for or against capital punishment. Both groups were shown the same set of evidence that consisted of studies for and against capital punishment. When the students were asked to rate the quality of the evidence, the studies that agreed with their position were rated higher than those that disagreed. Students were also asked to self-report on their degree of attitude change following reading the studies, finding that the students rated confirmatory studies as having a greater influence. Lord, Ross and Lepper went on to replicate those findings and explore confirmation bias in different contexts in a range of papers, [Vallone et al., 1985; Lord et al., 1984].

Gilovich [1983] recruited volunteer students to gamble on American football games, and found evidence of biased evaluation in the post-match description of losses and wins, with losses more likely to be explained away. They were even able to influence participants' future likelihood of gambling on a match by mentioning that a previous match was decided by a "fluke" play that could have gone either way, and so bringing into question the reliability of the previous match result as a predictor of future results. Liberman and Chaiken [1992] found that caffeine drinkers were more critical of messaging that linked caffeine to health problems [Liberman and Chaiken, 1992]. Koehler [1993] found a bias in scientists evaluating studies that either agreed or disagreed with their prior positions. McHoskey [1995] found that prior beliefs had a strong effect on people's ratings of the persuasiveness of evidence for and against a conspiracy. Malthouse [2022] found biased evaluation in the assessment of evidence for the efficacy of vaccines. These studies represent just some of the empirical evidence for biased evaluation.

It has often been pointed out that judging evidence based on prior beliefs is not irrational, as it can be rational to consider unlikely evidence more critically [Lord et al., 1979; Koehler, 1993; Klayman, 1995; Fischhoff and Beyth-Marom, 1983]. Nevertheless this effect is still often called *biased evaluation* — judging confirmatory sources more favourably, and disconfirmatory sources less favourably. We will follow this naming convention and define a sufficient condition given our minimal model:

$$P(R|D_{for}) > P(R|D_{against}) \,, \tag{4.8}$$

where D_{for} is a set of data that agrees with a prior hypothesis, and $D_{against}$ disagrees to the same extent.

Figure 4.4 shows biased evaluation effects with a strong initial prior belief in the central hypothesis, $P(H) = 0.8$, and a neutral prior belief in source reliability, $P(R) = 0.5$. With the BIASR model, we see biased evaluation as the confirmatory sources (Figure 4.4 **a**) are judged to be more reliable than the disconfirmatory sources (Figure 4.4 **b**). Given our model setup, the message receiver will eventually be persuaded and come to trust the source. This is because at worst an unreliable source is sending only noise. If we instead allowed anti-reliable sources, who consistently lie, then the overall effect would be stronger and it is possible for trust in a source to consistently move towards 0.

Notably, when given confirmatory information the belief in the reliability of the source is lower in the BIASR model than the rational Bayesian network model (Figure 4.4 **a**). This was unexpected and is related to an underestimation of probability mass for the correlated beliefs $P(H = 1, R = 1)$ in the BIASR model. This is explored in the Supplementary Information.

Given our model, a sufficient condition for biased assimilation is if an individual updates their beliefs in the central hypothesis more so than they would do under the rational version of Bayes' theorem with the Bayesian network. In the case of confirmatory evidence

$$P(H|D_{for}) > P(H|D_{for})_{rational} \,. \tag{4.9}$$

And in the case of disconfirmatory evidence,

$$P(H|D_{against}) > P(H|D_{against})_{rational} \,. \tag{4.10}$$

Biased assimilation is simulated in Figure 4.4. The BIASR model shows a stronger posterior belief in the central hypothesis than rational updating under both the Bayesian network and simple models, for both confirmatory and disconfirmatory evidence.

What causes these dynamics? When receiving confirmatory data a positive correlation is induced between H and R — it is more likely that the source is either correct and reliable, $P(H = 1, R = 1)$, or incorrect and unreliable, $P(H = 0, R = 0)$, than the alternatives. With the BIASR model, the agent forgets about this correlation, i.e. the agent forgets that their belief in the central hypothesis is partly due to their belief in the source reliability, and vice versa. One consequence is that the agent does not give enough consideration to the counterfactual world where the central hypothesis is wrong and the source is unreliable. A similar pattern happens with disconfirmatory evidence. During the independence approximation, probability mass is effectively moved away from correlated beliefs where those beliefs go against an indiviudal's priors. This is explored further in the Supplementary Information, where we explore belief updating across the entire joint belief distribution, $P(H, R)$.

We gave an intuitive example of confirmatory evidence in the Introduction with Alice and Bob. Here we give an example in the case of disconfirmatory evidence. Alice has a strong belief that vaccines are dangerous. She meets a new acquaintance, Chris, who tells her that vaccines are actually safe. This goes against Alice's strongly held views and so she naturally questions how reliable Chris is, and only updates her beliefs about vaccines slightly. The next time they meet, Chris again raises points about vaccine safety. This information again goes against Alice's views, and this time she already has question marks over Chris's reliability and is able to dismiss the evidence more easily. Over time, Alice is able to hold onto her belief that vaccines are dangerous and dismiss Chris as an unreliable source. Under the BIASR model, she does not remember the relationship between her belief in Chris' reliability and her beliefs in vaccine safety. As a consequence, she gives little consideration to the possibility that Chris is reliable and vaccines are safe.

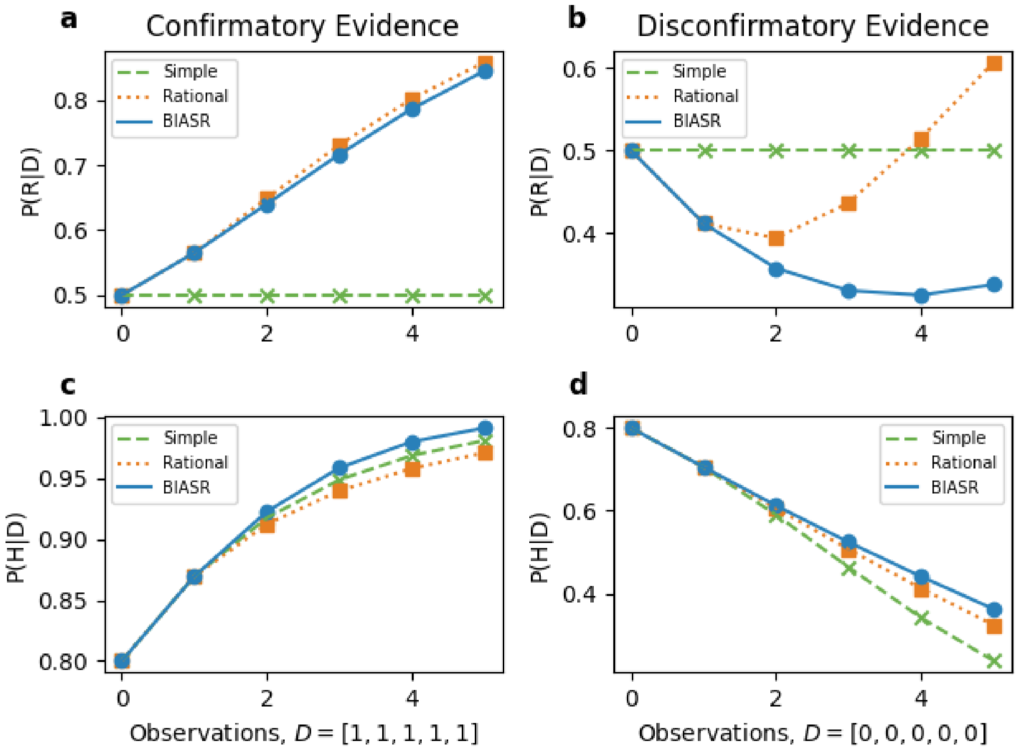


Figure 4.4: Assimilation and evaluation for confirmatory evidence and disconfirmatory evidence for sequential information from the same source. The BIASR model shows biased evaluation with **a**) confirmatory sources judged to be more reliable than **b**) disconfirmatory sources (the blue line is higher in the top-left figure than the top-right figure). The BIASR model shows biased assimilation, with a stronger posterior belief in the central hypothesis than the simple and rational models following both **c**) confirmatory and **d**) disconfirmatory evidence (the blue line is the highest line in both bottom figures).

4.5.2 Attitude Polarisation

In the study that we described above by Lord et al. [1979], people's evaluation of the evidence for and against capital punishment depended on their prior beliefs: Those who were pro-capital punishment self-reported that the evidence swayed them to be more fervent in their beliefs, and those who were against capital punishment stated that they also became more fervent, in the opposite direction — the two groups diverged in their beliefs after seeing the same data. This attitude polarisation has been replicated in the context of climate change [Cook and Lewandowsky, 2016], gun control [Taber and Lodge, 2006], affirmative action [Taber and Lodge, 2006], the Iraq war [Nyhan and Reifler, 2010], the JFK assassination [McHoskey, 1995], homosexual stereotypes [Munro and Ditto, 1997], drug use [Taber et al., 2009], freedom of speech [Taber et al., 2009] and nuclear energy [Plous, 1991].

A sufficient condition for *attitude polarisation*, given our model, is that individuals with different prior beliefs in the central hypothesis update in opposite directions,

$$\begin{cases} P(H|D) > P(H), & \text{if} \quad P(H) > 0.5 \\ P(H|D) < P(H), & \text{if} \quad P(H) < 0.5 \end{cases} . \tag{4.11}$$

D is the same set of evidence shown to both individuals, which can include more than one source and multiple datums from each source, both for and against hypotheses.

When considering a single hypothesis in isolation, it is a property of Bayesian updating that different prior beliefs will converge given the same data (or more precisely, not diverge). However, if we have a more complicated belief structure then it can be rational for individuals to update in opposite directions. This was confirmed by Jern et al. [2014], who prove a family of Bayesian network motifs that can lead to attitude polarisation. They go on to analyse the results of Lord et al. [1979] and offer two potential Bayesian network structures that could create attitude polarisation in this experiment. For instance, if an individual who has a strong pro-capital punishment prior also has a belief that the consensus is biased against capital punishment, then studies that are anti-capital punishment can be explained away as resulting from the biased consensus, while studies that are pro-capital punishment are strong evidence in support of capital punishment [Jern et al., 2014]. If individuals who are anti-capital punishment also believe that there is a bias in consensus, in this case a bias in favour of capital punishment, then it is rational for these individuals to also strengthen their beliefs when seeing the same data [Jern et al., 2014]. The pro- and anti- groups can rationally update their beliefs in opposite directions.

The rational basis for attitude polarisation was explored further by Henderson and Gebharter [2021] using a Bayesian network where evidence is influenced by the true values of the central hypothesis and source reliability, as in the BIASR model. They conclude that attitude polarisation can arise only if the individuals have different prior beliefs in both the central hypothesis and source reliability [Henderson and Gebharter, 2021]. This is a property of the Bayesian networks that generate rational attitude polarisation [Jern et al., 2014; Cook and Lewandowsky, 2016; Henderson and Gebharter, 2021] — they require different priors not only in the central hypothesis but also auxiliary beliefs.

The Bayesian network structures described by Jern et al. [2014] give a good explanation for attitude polarisation when central and auxiliary priors are different between polarising groups. However, the BIASR model generates attitude polarisation under the stricter condition that the pro- and anti- individuals differ only in their prior beliefs in the central hypothesis, and have the same auxiliary prior beliefs. In our model,

$$\begin{cases} P(H|D) > P(H), & \text{if} \quad P(H) > 0.5, P(R) = r \\ P(H|D) < P(H), & \text{if} \quad P(H) < 0.5, P(R) = r \end{cases}, \tag{4.12}$$

where $P(R) = r$ is the same prior belief in source reliability for both individuals.

As shown in Figure 4.5, the BIASR model leads to attitude polarisation when data is presented from two different sources, even when individuals only differ in their prior belief in the central hypothesis, $P(H = 0.8)$ and $P(H = 0.2)$. The simple and rational Bayesian models do not. As such the BIASR model meets both the general and stricter conditions we have defined for attitude polarisation. Strong prior beliefs either for or against the central hypothesis become more extreme overall.

Intuitively, let us consider the case where Alice starts with a strong prior belief in the dangers of vaccines. She is given two studies to read, one for and one against vaccine safety. The first study begins by stating that vaccines are safe — Alice starts to think that the study is not reliable, as she is confident that vaccines are dangerous. After reading on, the study makes another point about vaccine safety, Alice is now more easily able to dismiss this as she already has doubts over the study's reliability. As Alice reads on, she becomes convinced that the study is not credible and the later information has very little impact on her beliefs. The second study raises questions about vaccine safety. As Alice reads this study, her confidence in its credibility grows as it provides information that aligns with her existing beliefs in the dangers of vaccines, and she uses this evidence to bolster those same beliefs. Alice's beliefs about the dangers of vaccines and the reliability of the studies become correlated, and if she forgets about these correlations then attitude polarisation emerges.

4.5.3 Belief Perseverance

People can persevere in their beliefs with greater tenacity than the evidence would warrant [Klayman, 1995]. Belief perseverance is typically defined with a temporal aspect in the sense that once a belief is formed it will persist even once the evidence

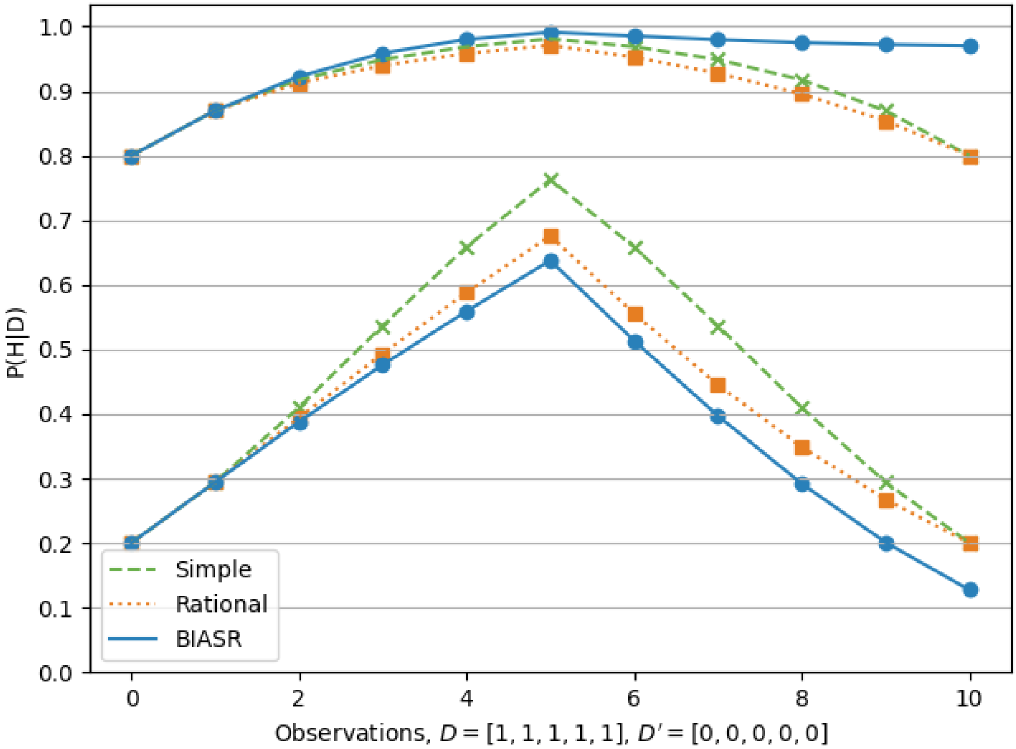


Figure 4.5: Attitude polarisation. Data from a source is followed by data from a new second source with the opposite view. With a strong initial prior (top), the BIASR model shows positive biased assimilation from the first data source followed by negative biased assimilation, overall increasing belief in the central hypothesis. With a low initial prior belief (bottom) we also see biased assimilation of both sources of data, overall decreasing belief in the hypothesis. Both positions become more extreme from seeing the same set of data under the BIASR model, showing attitude polarisation.

that formed its basis is discredited [Anderson et al., 1980; Ross et al., 1975]. In an early experimental study Ross et al. [1975] gave participants false feedback on a task (either good, average or bad). This (reasonably) influenced the participants' opinion of their task performance, but the participants held onto these opinions even after they were told that the feedback was fictitious. This effect was explored further by Anderson et al. [1980], who gave participants fictitious data suggesting that firefighters who were courageous were more likely to be successful in their jobs. This induced participant beliefs that persisted even once the data was revealed to be fictitious.

It has been noted that belief perseverance is connected to the primacy effect

[Nickerson, 1998], where data observed earlier has a larger impact on belief than data seen more recently. Bruner and Potter [1964] showed participants images, and found that they were slower to recognise those images when they came into focus slowly, as compared with participants who saw the same image without first seeing it out of focus. They attributed this effect to the perseverance of hypotheses generated while the image was out of focus. This was followed in the late 60s with a series of studies that tested participants' ability to form opinions through sampling, finding that early data could induce beliefs that were then held onto more strongly than would be Bayesian rational in light of later evidence against the belief [Geller and Pitz, 1968; Peterson and DuCharme, 1967; Jones et al., 1968].

In belief perseverance, the order that beliefs are formed is important. And if beliefs are formed from observed data, then the order that data is received is important. In contrast to the exchangeability principle of Bayesian rationality, i.e. that the order of the data received should not make a difference to the posterior beliefs, we define *belief perseverance* as the observation that data received earlier has a stronger influence on final beliefs than opposing data received later,

$$P(H|D) > P(H|D)_{rational} \quad , \quad D = [D_{for}, D_{against}] \,. \tag{4.13}$$

Figure 4.6 shows a simulation of belief perseverance. Starting from a neutral prior, $P(H) = 0.5$, both the simple and rational models end up with the same posterior belief as they began with, after seeing an equal amount of evidence for and against. In the BIASR model, the initial data drives belief in H beyond what is rational. Once the belief is ingrained, negative biased assimilation then slows down disconfirmation of belief. Here, we have simulated data as coming from separate sources. If we instead used a single source then we still observe belief perseverance but the effect is not as strong.

4.5.4 Confirmation Bias in Selection of Sources

Confirmation bias is usually defined not only in terms of assimilating information, but also in the selection of information in a way that supports existing beliefs. Taber and Lodge [2006] replicated and extended Lord et al. [1979]'s study on attitude polarisation. Participants were chosen who held strong beliefs about either gun control or affirmative action. They were then shown sources for and against those positions, but some participants also had the opportunity to choose the sources they wished to read. Those with strong prior beliefs selected the sources that were likely to agree with their position. Redlawsk [2002] found a similar effect in a behavioural

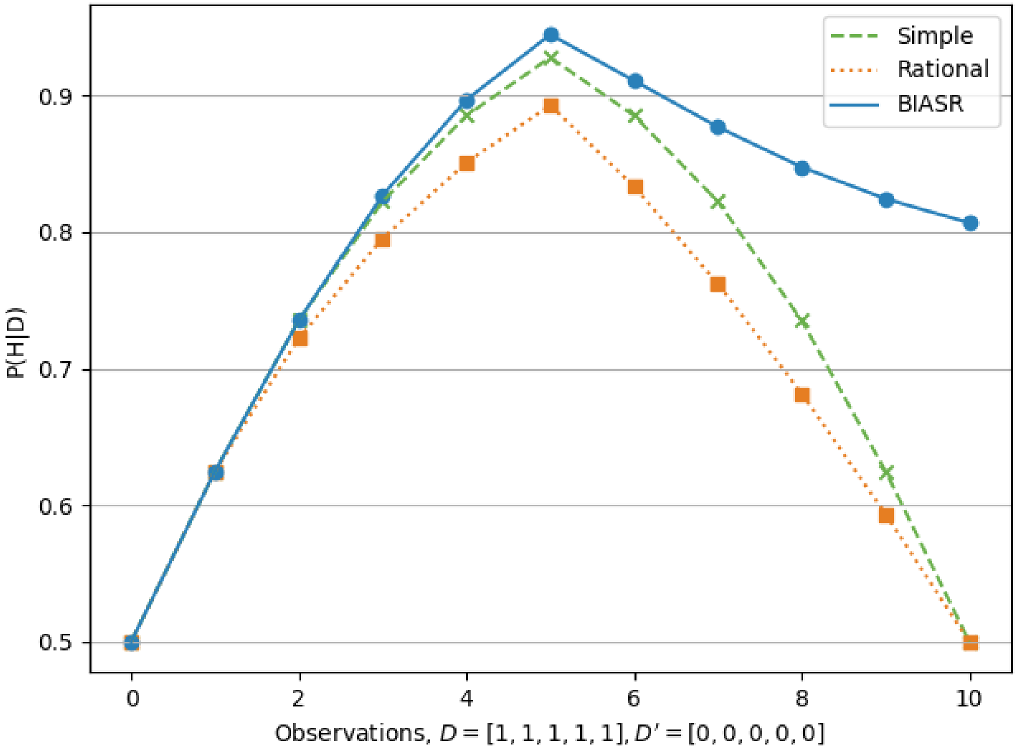


Figure 4.6: Belief perseverance. Data for, then data against, the central hypothesis are received from different sources given neutral initial priors in both the central hypothesis and source reliability. Under simple and rational models, the belief in the central hypothesis returns to the prior belief. With the BIASR model, biased assimilation dynamics mean that the data received earlier has a stronger effect on posterior beliefs than data received later.

experiment where they simulated a presidential primary election. Once participants had developed a preference for a candidate, they were more likely to search for information about that candidate. This form of confirmation bias may go beyond the selection of external sources, and Kunda [1990] also suggested a confirmation bias in the selection of memories and cognitive processes.

In order to extend our model to selection of sources we must add an extra assumption – agents are limited in that they cannot consume data from all sources and must be selective. An optimal selection would presumably be based on some kind of value function on the sources. This is difficult to model as value is subjective and would need to take into account complicated utility functions [Klayman and Ha, 1987]. Principled approaches [Klayman and Ha, 1987; Klayman, 1995] for quantifying the value of a source include a) quantifying the expected change in the probability of an agent guessing correctly about a hypothesis following information from the source, b) the diagnosticity of a question (or source) given by the expected log likelihood ratio. The likelihood ratio quantifies the degree of belief change given some data, equivalent to the ratio of posterior and prior odds. Here we use the expected log likelihood ratio as a measure of the value of a source, Q, defined as

$$Q = \left| P(D=1) log \left(\frac{P(D=1|H=1)}{P(D=1|H=0} \right) \right| + \left| P(D=0) log \left(\frac{P(D=0|H=1)}{P(D=0|H=0} \right) \right| . \tag{4.14}$$

This is adapted from Slowiaczek et al. [1992]. If agents select sources based on their value in terms of diagnosticity, then we can define confirmation bias in the selection of sources as valuing a confirmatory source's relative diagnosticity more so than would be rational,

$$\frac{Q|D_{for}}{Q|D_{against}} > \left(\frac{Q|D_{for}}{Q|D_{against}} \right)_{rational} . \tag{4.15}$$

We can calculate the diagnosticity of a source using equation 4.14, for the simple, rational and BIASR models. See the Supplementary Information for a derivation using the joint belief distribution.

Figure 4.7 shows the diagnosticity of sources for agents updating under the simple, rational and BIASR models. In the simple case the diagnosticity of sources is invariant. In the rational case there is little difference in the diagnosticity of sources, and the disconfirmatory sources are actually slightly preferred. With the BIASR model, there is a much greater difference in the diagnosticity of sources, with confirmatory sources much preferred. An individual that can choose only one source would much prefer the confirmatory source under the BIASR model, if that choice was made based on the source diagnosticity.

4.6 Empirical Evidence Aligned with the Independence Approximation

We have shown that the BIASR model can generate a range of confirmation bias type behaviours. If the model is capturing, in some sense, how people actually behave then we would expect to see a difference in behaviour depending on whether information is processed incrementally or all at once. Processing data all at once will give the same result as rational incremental processing, i.e. no bias. However, according to the BIASR model, sequential processing will show path dependence.

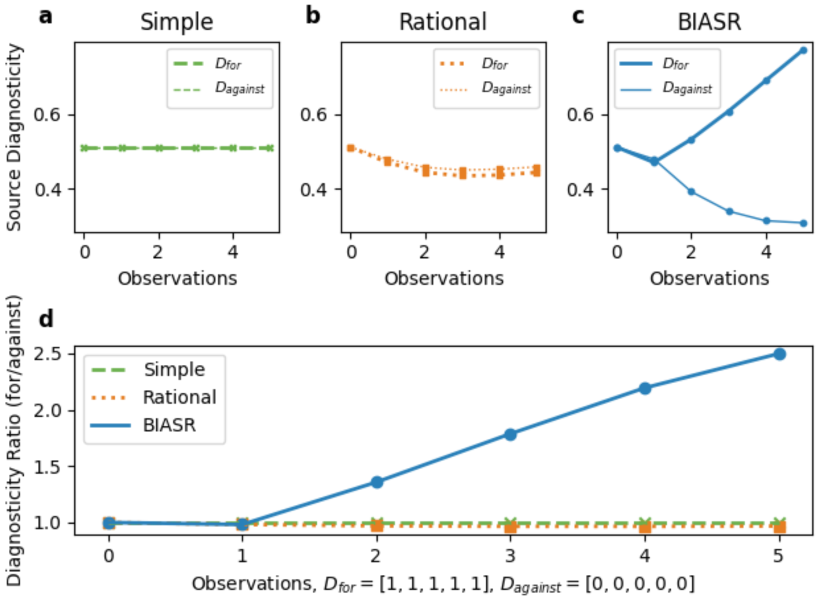


Figure 4.7: Confirmation bias in the selection of sources. a) The source diagnosticity is constant with the simple version of Bayes' theorem. b) In the rational model, the source diagnosticities are similar, and in fact the disconfirmatory sources have a slightly higher diagnosticity. c) With the BIASR model, the diagnosticity of confirmatory sources is much greater than disconfirmatory sources. d) The ratio of diagnosticity is much higher in the BIASR model, while in the simple and rational cases this ratio stays around 1.

An experimental manipulation would be to encourage participants to either a) process information incrementally or b) process information all at once. We expect to see more confirmation bias when the information is processed incrementally. We found two previous experimental studies where this distinction was made.

4.6.1 Redlawsk (2002)

Redlawsk [2002] describes the difference between *on-line processing*, where information is evaluated immediately and sequentially versus *memory processing* where information is remembered and then evaluated all at once when a decision is required. In an experiment, they simulated a presidential election and gave participants information about candidates. In the on-line condition, no further instructions were given as on-line processing is assumed to be the default behaviour. In the memory-based condition participants were encouraged to remember the information that they saw; they were told that they would be tested on it later, as well as being told that they would need to justify their choice to an experimenter. They investigated how participants reacted to incongruent (negative) information once they had developed a preference for a candidate. In the online condition this negative information actually increased the preference for the candidate, while in the memory condition the negative information reduced the candidate rating: In the on-line condition the incongruent information seems to be negatively evaluated to such an extent that it provides evidence in favour of the candidate.

Redlawsk [2002] attribute the difference to an additional accuracy motivation to the memory-based processors, within the framework of motivated reasoning [Kunda, 1990]. Within this framework, the memory-based processors are motivated for greater accuracy due to the instruction that they will need to justify their choices, and they achieve this by processing the information all at once (or remembering the dependencies between beliefs). The BIASR model suggests that the bias arises because of path dependence in the on-line condition. Figure 4.8 shows the data as presented by Redlawsk [2002] alongside a simulated replication of the effect with the BIASR model. Here, we used the same model as in the earlier simulations, with the change that unreliable sources are now anti-reliable, so are more likely to give false information, i.e. $P(D = 1|H = 1, R = 0) = 0.35$.

4.6.2 Carlson, Meloy and Russo (2006)

In this experiment [Carlson et al., 2006], participants were asked to make a choice between two restaurants after seeing each of the restaurants' attributes. The six attributes were typically neutral but included one for each restaurant that was much in its favour (for example, one restaurant has a professional dessert chef while the other has a small assortment of standard desserts). The order of attributes were manipulated so that the target restaurant had its very positive attribute revealed first, and the opposing restaurant had the attribute in fourth position. As a further treatment, in Study 1 the attributes were shown sequentially, while in Study 3 the attributes were shown together on a single page for each restaurant. They found a significant preference for the target restaurant in Study 1, but not in Study 3. Confirmation bias was not detected when information was presented in one block, but was detected when the same information was presented sequentially.

The authors of the study interpret the result within Russo's *predicisional distortion of information* framework. When incorporating information sequentially, a positive first attribute creates an initial preference for the target restaurant that then biases the interpretation of subsequent data so that overall the target restaurant is preferred. This framework is similar to our model and the findings here support both perspectives. Within the BIASR model, the preference for the target restaurant

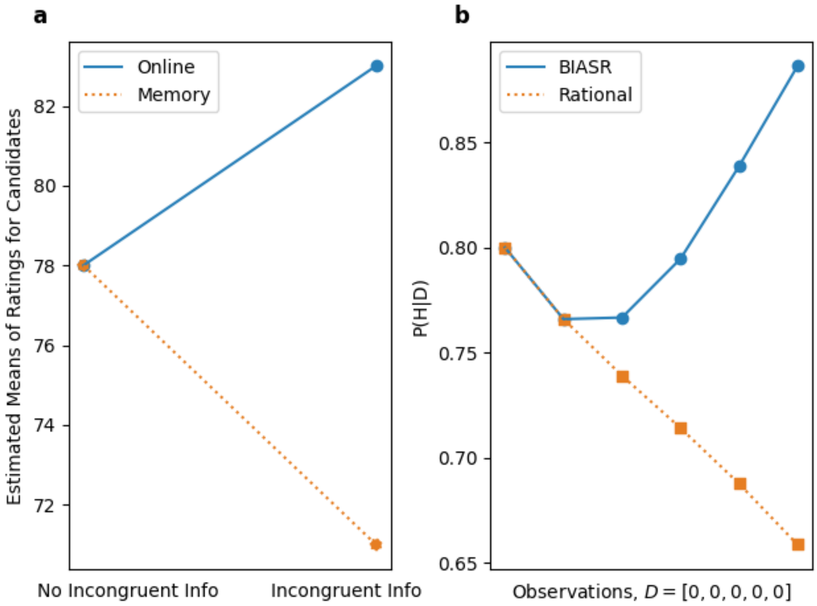
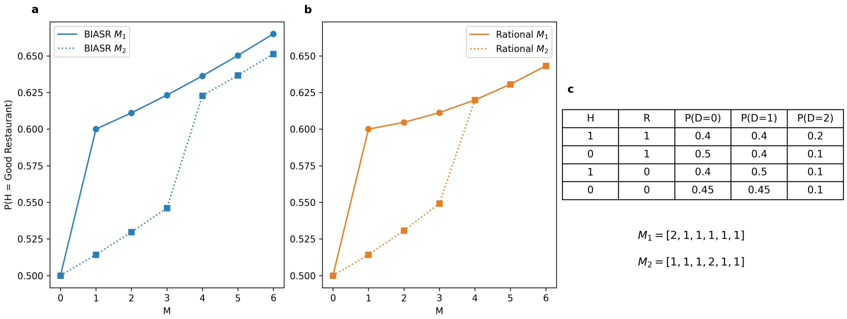


Figure 4.8: Replication of the Redlawsk result. a) The data presented by Redlawsk. Following negative information about a preferred candidate, the online processors increase their rating for the candidate, while the memory processors decrease their rating. b) A replication with the BIASR model. We start with a prior preference for the candidate and a neutral prior in source reliability (not shown). Unreliable sources are considered anti-reliable, i.e. negative information from an unreliable source actually acts as evidence for a candidate. We simulate receiving a series of negative pieces of information from the source. Similarly to Redlawsk: in the BIASR condition belief in the candidate increases, while it decreases in the rational condition.

in Study 1 is described by belief perseverance, i.e. the first attribute observed has a greater weight on the final choice than the fourth attribute. Alternatively, when data is shown all at once it is more likely to be processed together, which is equivalent to remembering the history of belief dependencies. We have simulated this result within the BIASR model (Figure 4.9). We used a similar setup as in the earlier simulations, but now messages can be negative, slightly positive or very positive, $D \in [0, 1, 2]$ respectively. We chose this setup because it replicates the result with a minimal change to the existing model.

4.7 Discussion

The traditional normative argument is that rational behaviour should enjoy higher evolutionary fitness [Daw et al., 2008]. As argued here and noted before, a normative



H	R	P(D=0)	P(D=1)	P(D=2)
1	1	0.4	0.4	0.2
0	1	0.5	0.4	0.1
1	0	0.4	0.5	0.1
0	0	0.45	0.45	0.1

$M_1 = [2, 1, 1, 1, 1, 1]$

$M_2 = [1, 1, 1, 2, 1, 1]$

Figure 4.9: Replication of the Carlson, Meloy and Russo result. Information is received about attributes of a restaurant. The attributes that are received are either "1": neutral (or slightly positive); or "2": strongly positive. a) In the BIASR model the strong initial positive message induces a bias towards restaurant 1 (M_1) which persists. b) In the rational model, the order that data is received does not influence the final beliefs and both restaurants are judged to be of equal expected quality. The beliefs are simulated using c) the conditional probability distribution.

account should also include cognitive limitations [Dasgupta et al., 2020; Klayman, 1995; Daw et al., 2008], such that when considering computationally intractable problems evolutionary pressures will favour organisms with efficient approximations to rationality [Daw et al., 2008]. We have argued that maintaining dependencies within large belief networks is computationally intractable given realistic memory constraints. We showed how human cognition can overcome this limitation through the BIASR model (Bayesian updating with an Independence Approximation and Source Reliability). And this approximation leads directly to many confirmation bias behaviours. Our results are general, and similar problems will be encountered by artificial agents with large world models.

Previous information processing models of confirmation bias either introduce irrationality without a complete explanation, or they explain the bias as rational given a certain belief updating structure. Irrationality can be included by, for example, adding a factor to reduce the weight of disconfirmatory evidence [Gerber and Green, 1999]. Our contribution offers a principled source of irrationality based on a boundedly rational approximation to Bayesian rationality. This approximation leads to a simplification of the rational model which is equivalent to the "belief-based" updating described in previous research [Merdes et al., 2020; Bovens et al., 2003; Hahn et al., 2018; Olsson, 2011]. Additionally, our single model is able to generate many forms of confirmation bias. We do not claim that the BIASR model is the full story, and for example Bayesian networks [Jern et al., 2014] can explain much of the empirical evidence for attitude polarisation. However, the BIASR model demonstrates a variety of other confirmation biases that the Bayesian rational model does not, suggesting that it is capturing an important aspect of boundedly rational cognition.

We have focused on a very simple Bayesian network to demonstrate that confirmation bias can arise from the BIASR model. We do not claim that this simple model is how people actually update their beliefs. However, the behaviour is robust and emerges under a wide range of conditional probability distributions. The assumptions also hold (and are even strengthened) with more complex belief structures. We present a model based on two *types* of sources (reliable or not). However, inference that includes beliefs about types of sources in general would be susceptible to confirmation bias in the same way (including for example biased or anti-reliable sources).

We have included an assumption that human information processing is described by the mathematics of Bayesian networks, and that human memory can be analysed in the same way as computer memory. Our feeling is that these principles are fundamental to information processing and so it is reasonable to assume that human cognition is at least partly bound by them.

4.7.1 Social and Individual Explanations

The BIASR model is at the information processing level. However, there have also been explanations of confirmation bias at the social and individual level. Our model is not in opposition to these explanations, but instead complements them.

There have been a range of **social explanations** for confirmation bias. Mercier and Sperber [2011] claim that confirmation bias can improve group cognition. If biased individuals argue to support their own belief then the result can be that overall there is a more efficient group search through hypothesis space, which is then reconciled through debate. This idea could describe the scientific process. Indeed, scientists are not immune to confirmation bias [Koehler, 1993; Dunbar, 1995; Mahoney, 1977] and history is littered with individual scientists who steadfastly held onto their beliefs despite disconfirmatory evidence [Nickerson, 1998]. Building on this idea, Norman [2016] argues that the purpose of human reasoning in general is to align group intentions and confirmation bias helps in this regard by strongly entrenching group mythology and beliefs that can persevere over time and so maintain group cohesion. Another perspective is that believing something strongly can influence others and help to bring it about [Peters, 2020], a form of self-fulfilling belief [Snyder, 1984].

At the **individual level**, confirmation bias may help to navigate asymmetric error utilities [Nickerson, 1998; Friedrich, 1993] (being wrong about believing there isn't a lion is more problematic than being wrong that there is a lion). From an adaptive perspective, a wider utility function is being optimised beyond truth seeking, and confirmation bias helps to drive behaviour towards a beneficial outcome in this wider game. This is almost certainly true if we assume that human behaviour is adaptive. While the threat of being eaten is obvious, the principle applies to other threats such as identity or self-perception. Kunda [1990] made the case that reasoning is motivated only sometimes by accuracy, and other times by a desire to arrive at certain conclusions. In this account, the individual's motivation will determine which cognitive processes are put to use. If accuracy is desired, then deeper processing is carried out. But if an individual has a motivation to e.g. preserve their self-image or identity, then they can introduce biases in their reasoning that lead to the preservation of those beliefs [Kunda, 1990]. People are not completely free to believe whatever they want, and are instead constrained by the available cognitive resources and by the need for coherence within beliefs, at least to the extent that they could justify themselves to someone else [Kunda, 1990]. The desire for coherence is an older idea that is also a part of the influential cognitive dissonance theory [Festinger, 1962]. Biased evaluation, biased assimilation and belief perseverance can be understood as the reconciliation of the dissonant beliefs "I believe that I am someone who holds correct beliefs" and "this evidence disconfirms my beliefs" [Kunda, 1990]. However, notably it has been argued that dissonance theory does not easily predict attitude polarisation [Lord, 1989]. Motivated reasoning and the avoidance of dissonance are a part of the puzzle, but it still leaves the question open of describing the cognitive processes involved.

As social behaviour emerges from individual behaviour, so individual behaviour emerges from cognitive processing. Kunda's perspective on motivated reasoning [Kunda, 1990] is enriched by our framework. Motivated reasoning relies on the assumption of different cognitive faculties that have differential levels of accuracy and effort. Our model provides a clear account of using extra cognitive resources to improve the accuracy of reasoning. An individual with an accuracy motivation could update their beliefs without applying the independence approximation, and instead use extra memory resources to consider dependencies between beliefs and avoid biases. We can also see the link to emotional states and hot vs cold cognition [Kunda, 1990] — one can imagine a hot-headed individual quickly jumping to false conclusions while a cooler head carefully thinking through the evidence and belief dependencies.

Given that confirmation bias exists, we can speculate that it would make sense for adaptive pressures to build other behaviours around this bias — nature is parsimonious. A purely rational agent would reason about the world and then decide on their actions based on these beliefs combined with an expected utility distribution. In certain situations it may be more cognitively efficient to shortcut this two-step process by leveraging confirmation bias to drive behaviour based on less than rational beliefs. Given that confirmation bias exists at the individual level, we can speculate that adaptive pressures built useful group dynamics upon it such as argumentation and debate [Mercier and Sperber, 2011], persistent group ideologies and mythologies [Norman, 2016] and even the will to force reality towards our beliefs [Peters, 2020].

4.8 Conclusion

The BIASR model is based on principled assumptions, generates many confirmation bias type behaviours, and aligns well with both empirical evidence and other explanations in the literature. The main principle of the BIASR model is that put forward by Daw et al. [2008], who contend that rationality is not the appropriate normative standard when studying human and animal behaviour. Instead, where rational computation is expensive we should expect to see efficient approximations to rationality. We demonstrate that an independence approximation is one way in which cognition can overcome intractable computational demands, providing a fuller normative explanation for the "belief-based" updating described in earlier work [Merdes et al., 2020; Olsson, 2011; Bovens et al., 2003; Hahn et al., 2018]. Given its general nature, the independence approximation deserves further investigation as a more general cognitive mechanism for boundedly rational reasoning with memory constraints.

4.9 Supplementary Information — The Diagnosticity of a Source

In the main text, we wrote down the diagnosticity of a question as the expected log likelihood ratio of the answers,

$$Q = \left| P(D=1) log \left(\frac{P(D=1|H=1)}{P(D=1|H=0} \right) \right| + \left| P(D=0) log \left(\frac{P(D=0|H=1)}{P(D=0|H=0} \right) \right| . \tag{4.16}$$

Here, we show how this can be applied to joint belief distributions in $P(H, R)$. The likelihood of data given a hypothesis can be written as,

$$P(D|H) = \frac{P(D,H)}{P(H)}\,. \tag{4.17}$$

In our case, we would prefer this in a form based on the joint belief distribution that includes the source reliabilities. By the law of total probability,

$$P(D|H) = \frac{P(D,H,R=1) + P(D,H,R=0)}{P(H,R=1) + P(H,R=0)}\,. \tag{4.18}$$

The numerator can be rewritten in terms of conditional probabilities on (H,R),

$$P(D|H) = \frac{P(D|H,R=1)P(H,R=1) + P(D|H,R=0)P(H,R=0)}{P(H,R=0) + P(H,R=1)}\,. \tag{4.19}$$

The terms of the right hand side are now all known given the simulated models in the main text. And this relation can be used to calculate the numerators and denominators of the likelihood ratios in Equation 4.16.

Additionally we can calculate

$$P(D) = \sum_{H,R} P(D|H,R)P(H,R)\,. \tag{4.20}$$

We can therefore straightforwardly calculate the expected diagnosticity of a source given a joint belief distribution. In the rational case, we can use the full joint distribution, and update that rationally as we receive more data. In the BIASR case, we can approximate the joint distribution by $P(H,R) = P(H)P(R)$, and use that approximation both in the diagnosticity equation and between subsequent datums from a source.

4.10 Supplementary Information — Simulations of the Full Joint Belief Distribution

We found it useful for building intuitions to examine the behaviour of the full belief distribution, $P(H,R)$, when receiving evidence. This is presented here in the case of confirmatory and disconfirmatory evidence. We follow the same simulation setup as described in the main paper.

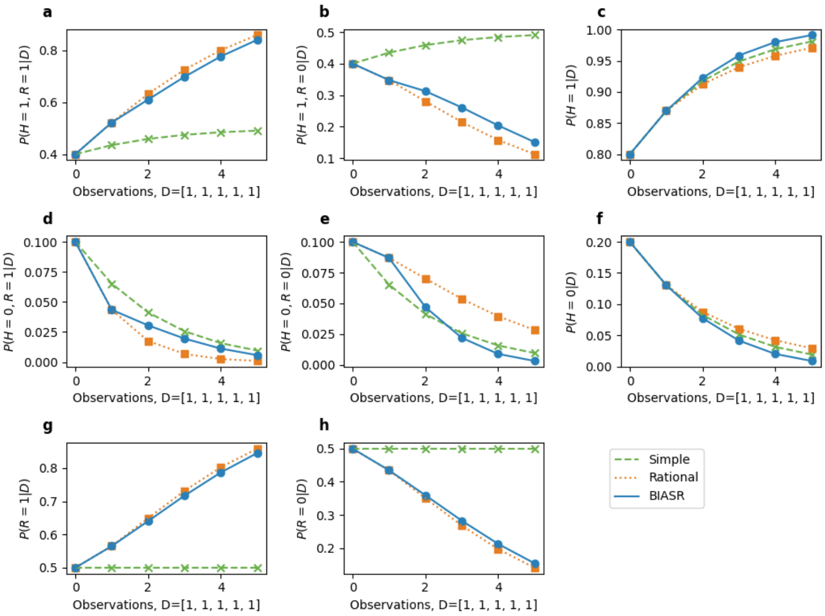


Figure 4.10: Joint belief distribution given confirmatory evidence. The rational model has more probability mass in the correlated beliefs a) $P(H = 1, R = 1)$ and e) $P(H = 0, R = 0)$, than the BIASR model. In the BIASR model, most of this correlated probability mass is transferred to the belief b) $P(H = 1, R = 0)$, and a small amount to d) $P(H = 0, R = 1)$. Overall this means that for the marginal beliefs the BIASR model has a c) stronger posterior belief in the central hypothesis, $P(H = 1)$, and g) a smaller posterior belief in the source reliability, $P(R = 1)$, compared to the rational case.

4.10.1 Confirmatory Evidence

In the confirmatory case (Figure 4.10), the belief dependencies act in a way to induce a positive correlation between H and R. The independence approximation, by definition, removes this correlation. For this reason the BIASR model (compared to the rational model) gives less probability weight to the beliefs $P(H = 1, R = 1)$ and $P(H = 0, R = 0)$, and so more to the beliefs $P(H = 1, R = 0)$ and $P(H = 0, R = 1)$. As $P(H = 0, R = 1)$ is very small compared to the other beliefs, most of the correlated probability mass is transferred to $P(H = 1, R = 0)$ during the independence approximation. In the BIASR model this strengthens the posterior belief in the central hypothesis beyond what would be rational. And it weakens the posterior belief in the source reliability compared to the rational case.

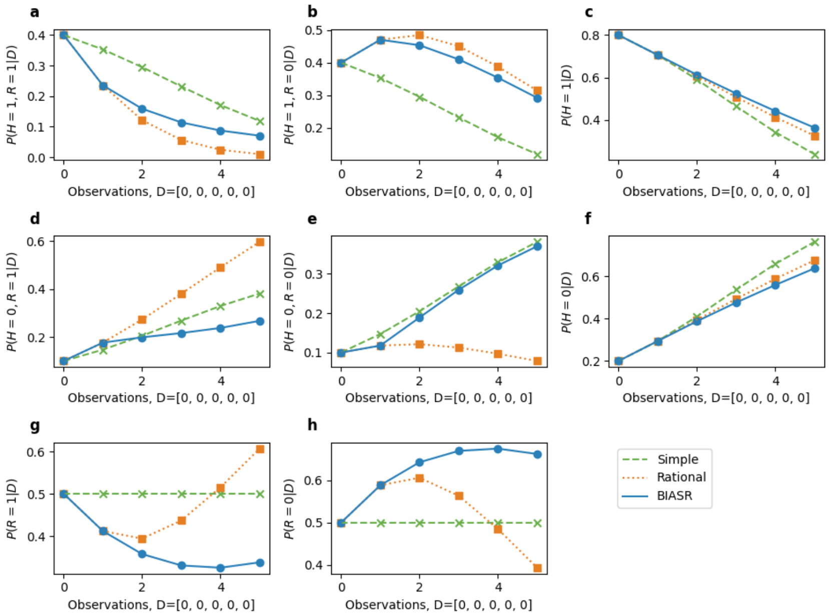


Figure 4.11: Joint belief distribution given disconfirmatory evidence. The rational model has more probability mass in the negatively correlated beliefs b) $P(H = 1, R = 0)$ and d) $P(H = 0, R = 1)$, than the BIASR model. In the BIASR model, most of this probability mass is transferred to the belief e) $P(H = 0, R = 0)$, and a small amount to a) $P(H = 1, R = 1)$. Overall this means that for the marginal beliefs the BIASR model has, compared to the rational model, a c) stronger posterior belief in the central hypothesis, $P(H = 1)$, and g) a much weaker posterior belief in the source reliability, $P(R = 1)$ compared to the rational case.

4.10.2 Disconfirmatory Evidence

We see a similar pattern with disconfirmatory evidence (Figure 4.11). Now there is a negative correlation and the beliefs $P(H = 1, R = 0)$ and $P(H = 0, R = 1)$ gain more probability mass in the rational case compared to the BIASR model. With the independence approximation, the correlated probability mass is transferred to the beliefs $P(H = 0, R = 0)$ and $P(H = 1, R = 1)$. When summing over the joint belief distribution, this results in the BIASR model having a higher marginal belief in the central hypothesis , $P(H = 1)$, and a weaker belief in the source reliability, $P(R = 1)$. The independence approximation allows the individual to diminish their belief that the source is reliable and the hypothesis is false.

Chapter 5

Discussion

5.1 Summary of Contributions

Chapter 2: The Rising Entropy of English in the Attention Economy began with empirical results that show that word entropy (and more broadly lexical diversity) has been rising steadily in American English since around 1900. Additionally, in the modern era, the word entropy of short-form media (news, magazines) is higher than in long-form media (fiction books, non-fiction); this was found in both American and British English. Even shorter-form media in the form of social media feeds from Twitter and Reddit were shown to have even higher word entropy than traditional media. A model of information foraging in the attention economy was developed that describes the rising word entropy as well as the media differences. This model also predicts that very short-form media (such as social media) is only competitive in a world with easy access to information (fast switching).

Chapter 3 - Bias in Zipf's Law Estimators concerns the methodological problem of fitting power laws models to rank-frequency data (i.e. Zipf's law). Systematic bias in prevailing maximum likelihood estimators was shown to be due to an inappropriate likelihood function. This bias is fundamentally due to correlated errors in frequency and frequency-rank and low sampling in empirical data from the tail of the underlying data generating process. The correct likelihood function is derived for words drawn from an underlying power law probability distribution. Unfortunately, computation of the maximum likelihood was found to be computationally intractable for real-world text samples. A method of Approximate Bayesian Computation was explored and shown to be effective at approximating the maximum likelihood with much less bias. However, applying this algorithm to real-world text samples introduces arbitrary biases depending on the selection of the summary statistic that is minimised in the algorithm. It was argued that these biases are fundamental to fitting Zipf's law, and that all current estimators for Zipf's law in language are biased as they do not represent the true data generating process.

Chapter 4 - Confirmation Bias Emerges from an Approximation to Bayesian Reasoning involves a comprehensive review of the confirmation bias literature and the presentation of the BIASR model (Bayesian Updating with an Independence Approximation). The 5 main forms of confirmation bias were clearly described and defined (biased evaluation, biased assimilation, attitude polarisation, belief perseverance, selection of sources). Existing explanations for confirmation bias were summarised and reviewed. It was argued that in a world with potentially unreliable sources, data received can provide information about central hypotheses and source reliability. Inference from data in such a world is represented as a collider type Bayesian network containing beliefs about central hypotheses and source reliability. Given this model, maintaining a joint belief distribution is a rational approach as it maintains information dependencies between beliefs that are introduced as data is received. Maintaining these joint belief distributions introduces unrealistic memory demands for agents with large world models. The cognitive burden can be greatly reduced by taking an independence approximation between beliefs. This represents the arguments underlying the BIASR model. Simulations show that each of the 5 listed forms of confirmation bias are generated under the BIASR model, as well as other empirical results in the literature.

5.2 Future directions

5.2.1 Chapter 2 - The Rising Entropy of English in the Attention Economy

A key assumption made in the model is that people are attracted to high entropy text. There are existing experimental results from eye-tracking experiments that show that people are attracted to complex information in advertisements [Radach et al., 2003] and towards surprising information when watching films [Itti and Baldi, 2009]. There are also evolutionary arguments as to why people would be attracted to high density information. A future direction would be to test this assumption experimentally when showing people high and low entropy text. This kind of experiment could also be extended to test the information foraging model more completely by observing how people choose between different items of textual information, how long they spend reading each item and the time spent switching and searching between items. Information prevalence and/or time required to switch between items could be manipulated in different experimental conditions.

The chapter considers English text. Future work could investigate other languages and types of text. A difficulty here will be finding suitable text corpora. The Corpus of Historical American English (used to investigate trends in the chapter) is a particularly high quality text corpora. One option is Google Ngrams [Michel et al., 2011], which has historical text samples for many languages. However, this is not a balanced corpora and the corpus composition changes over time. For example, the collections for US and British English have been shown to significantly change composition over historical time towards more scientific terms [Pechenick et al., 2015]. This could be due to Google collecting data by scanning books in University libraries, which introduces large amounts of textbooks in the second half of the 20th century. Any trends in word entropy found in Google Ngrams may reflect either changes in language use or changes in corpus composition.

Attention economy effects may be more pronounced in some other languages than in American English. In many countries there has been a more rapid and recent transition from a low competition media environment to a highly connected and competitive media environment, with the global phenomena of the internet, smartphones and social media. As such, we may see more pronounced attention economy effects on language use in these kinds of countries, which may be detectable without long-term balanced text corpora.

5.2.2 Chapter 3 - Bias in Zipf's Law Estimators

There do not seem to be clear future directions in terms of finding better Zipf's law estimators. Instead, the chapter argues that there is inherent bias in the estimators due to a disconnect between the complexity of the data generating process (human cognition) and the assumptions in the estimators. However, there is scope for future work that encourages a more consistent application of estimators. One approach would be to release an easy to use python or R package designed specifically for analysing Zipf's law in language. This could be released alongside an accompanying paper that clearly explains the various forms of Zipf's law and the estimators. Ideally this could be developed in collaboration with prominent Zipf's law researchers such as Ferrer-i-Cancho's group in Spain. This would provide consistency in the use of estimators as well as data cleaning processes. A challenge in relation to this is the development of hypothesis tests tailored specifically to Zipf's law, which would be expected in a statistical package. Existing hypothesis tests for power-law models use bootstrapping of data from models to generate p-values [Clauset et al., 2009] — it is unclear whether this is entirely suitable for rank-frequency data.

5.2.3 Chapter 4 - Confirmation Bias Emerges from an Approximation to Bayesian Reasoning

The BIASR model makes some clear predictions about human behaviour, which can be tested through behavioural experiments. Specifically, the model predicts that confirmation bias behaviour should appear when people update their beliefs sequentially when receiving data. This could be manipulated in alternative experimental conditions where people are either shown information sequentially or all at once, which should result in less confirmation bias. Alternatively, the alternative experimental condition could be to ask participants to remember the information that they have seen, which should encourage the participant to remember information dependencies. Similar previous experiments were described in the main chapter [Redlawsk, 2002; Carlson et al., 2006]. An experiment with pre-registered hypotheses based specifically on the BIASR model would carry more weight and represent a good test of the model. This experiment could also test when and how people use conditional dependencies when updating beliefs. For example, the participants could receive some negative information about a person from a particular news source, then later on find out that the news source is very unreliable. We can test if the beliefs about the person are updated or not after learning that the news source is poor quality.

Confirmation bias has been implicated as a factor in increasing political polarisation [Del Vicario et al., 2017]. The BIASR model lends itself well to simulating group behaviour. A group model could be based on a network of agents who maintain beliefs in a central hypothesis as well as beliefs about the reliability of each of the other agents. A similar group level model has been investigated before [Martins, 2013] and shown to generate polarisation in some conditions. The BIASR model on a network represents a Bayesian opinion dynamics model with simultaneously updating edge weights (beliefs in reliability), which is similar to a "trust matrix" as proposed by Degroot [1974]. As well as shedding light on political polarisation in the modern era, the updating of source reliability may also be an adaptation that overcomes fundamental biases in wisdom of crowd effects [Becker et al., 2017], as network weights (and influence) become correlated with agent accuracy. Simulation and/or behavioural experiments along these lines could be valuable contributions.

5.3 The PhD Experience

I wanted to add a personal note to describe the PhD experience, which is not captured by the thesis so far.

Six months into the PhD the COVID pandemic hit. This was very disruptive but fortunately I had spent the first 6 months having regular meetings with my main supervisor Thomas Hills, and had established a good working relationship. While I was only 6 months into the PhD, I was 18 months into the programme at the Mathematics of Real-World Systems Center for Doctoral Training (CDT), which began with a year-long Masters course. In this Masters year I had built good relationships with my coursemates and faculty which helped immensely during the COVID period in terms of support. More generally I would recommend a CDT to any PhD candidate — I found the training to be very valuable. And having a cohort of coursemates was even more valuable.

During COVID I moved away from the campus at Warwick and back home to Manchester for 18 months. As things began opening up again following the pandemic, I started a placement at the Alan Turing Institute in London. This meant that the majority of my PhD experience was remote. This raised lots of challenges including some degree of academic isolation. Fortunately I was able to find support from my supervisor Thomas Hills, other faculty staff, coursemates, other PhD students, and friends and family.

I was fortunate in that I was able to do a lot of in-person teaching before the COVID lockdown, as a teaching assistant in 4 courses. This helped me to gain confidence in public speaking, thoroughly learn the taught material, and generally gain some teaching experience. I continued teaching some courses remotely during the PhD, which I believe was much easier having had the in-person experience. I find teaching very valuable. I find it fulfilling to be able to support students in their own learning. And the practice of explaining things to people at a variety of levels really helps build my own understanding and intuitions.

Of course I discovered that in research there are a lot of dead-ends. The work presented in this thesis represents research projects that came to fruition, but there were others that did not. This included spending a long time trying to find a better estimator for calculating the entropy of text, using for example Lempel-Ziv compression [Schürmann and Grassberger, 1996; Ziv and Lempel, 1978] — I discovered that this is a very difficult problem. During the groundwork for Chapter 2 we also analysed Google Ngrams [Michel et al., 2011], but in the end did not use this dataset due to concerns of the unbalanced composition of the text samples [Pechenick et al., 2015]. I also spent some time investigating generative models for Zipf's law [Cancho and Solé, 2003] in connection with Chapters 2 and 3. In that case I ran out of time and might come back to the research question in the future. Overall these dead-ends were necessary parts of the research process and did give benefits in terms of deepening my domain knowledge in those areas and also improving the meta-skill of recognising good research projects.

As mentioned above, as a part of the Center for Doctoral Training programme I spent a year before the PhD competing a Masters. This included a research project under the supervision of Weisi Guo. I was fortunate enough to be given a very tractable problem by Weisi and we made some decent progress. I was able to publish the Masters project as an article in Nature Scientific Reports, "Organisational Social Influence in Heirarchical Graphs: From Anarchy to Tyranny" [Pilgrim et al., 2020]. This was published a few months into the PhD and it gave me a good "early win". Publishing something early also helped me to get over fear and trepidation in sending articles to journals. I think that an easy trap to fall into is to be overly cautious about sending things for publication and to delay too long (a trap which I also fell into on occasion).

During the PhD there were other projects beyond the work presented in this thesis. These were not included as they were either not research focused or my contribution was not easily extracted from the group collaboration. In the former case (not research focused) I developed and published a python package along with an article in the Journal of Open Source Software titled "piecewise-regression in python" [Pilgrim, 2021]. I also wrote an opinion article for PLOS Computational Biology "Ten Simple Rules for Working with Other Peopleś Code" [Pilgrim et al., 2023]. I was involved in other projects where I was not the lead author including a chapter in a book with Joe Austerweil and Kesong Cao, "Burstier Events: Analysing Human Memory over a Century of Events Using the New York Times" [Austerweil et al., 2022]; and an article in the Conference on Autonomous Agents and Multiagent Systems with Stas Zhydkov, Jacques Bara and Paolo Turrini, "The Grapevine Web: Analysing the Spread of False Information in Social Networks with Corrupted Sources", which has been accepted and due for publication in Summer 2022. And last but not least I was involved in several behavioural studies with Eugene Malthouse, Daniel Sgroi and Thomas Hills, the first of which, "When Fairness is Not Enough: The Disproportionate Contributions of the Poor in a Collective Action Problem", is in press and due for publication at the Journal of Experimental Psychology General.

Overall, I found the PhD experience challenging but ultimately worth it. I think it is important to remember to enjoy and celebrate the successes and to put the difficulties and challenges in perspective. I have changed a lot during the process, and I believe that I have become a better researcher and scientist. I intend to stay in academia and continue to research how people communicate and work together.

Bibliography

Project Gutenberg, Jun 2020. URL `https://www.gutenberg.org`. [Online; accessed 16. Jul. 2020].

Jon Agar. *Constant touch: A global history of the mobile phone.* Icon Books Ltd, 2013.

Manindra Agrawal. Determinant Versus Permanent. *Proceedings oh the International Congress of Mathematicians, Vol. 3, 2006-01-01, ISBN 978-3-03719-022-7, pags. 985-998*, 3, Jul 2008.

Craig A Anderson, Mark R Lepper, and Lee Ross. Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of personality and social psychology*, 39(6):1037, 1980.

Cecilie Schou Andreassen, Joël Billieux, Mark D Griffiths, Daria J Kuss, Zsolt Demetrovics, Elvis Mazzoni, and Ståle Pallesen. The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: A large-scale cross-sectional study. *Psychology of Addictive Behaviors*, 30(2):252, 2016.

Joseph L Austerweil, Charlie Pilgrim, and Kesong Cao. Burstier events: Analysing human memory over a century of events using the new york times. 2022.

Francis Bacon. *Novum organum.* 1620.

Jaume Baixeries, Brita Elvevåg, and Ramon Ferrer-i Cancho. The evolution of the exponent of zipf's law in language ontogeny. *PloS one*, 8(3):e53227, 2013.

C Daniel Batson. Rational processing or rationalization? the effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, 32(1):176, 1975.

H. Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. *Eur. Phys. J. B*, 58(2):167–173, Jul 2007. ISSN 1434-6036. doi: 10.1140/epjb/e2007-00219-y.

David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2): 180–191, November 2008. ISSN 0165-5515. doi: 10.1177/0165551508095781.

Mark A. Beaumont. Approximate Bayesian Computation in Evolution and Ecology. *Annu. Rev. Ecol. Evol. Syst.*, 41(1):379–406, Nov 2010. ISSN 1543-592X. doi: 10.1146/annurev-ecolsys-102209-144621.

Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035, Dec 2002. ISSN 0016-6731. URL `https://www.genetics.org/content/162/4/2025`.

Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, Nov 2009. ISSN 1464-3510. doi: 10.1093/biomet/asp052.

Joshua Becker, Devon Brackbill, and Damon Centola. Network dynamics of social influence in the wisdom of crowds. *Proc. Natl. Acad. Sci. U.S.A.*, 114(26):E5070–E5076, June 2017. doi: 10.1073/pnas.1615978114.

Richard E Bellman. *Adaptive control processes.* Princeton university press, 2015 (1957).

Christian Bentz, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery. Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLoS One*, 10(6):e0128254, Jun 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0128254.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275, 2017.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate bayesian computation with the wasserstein distance. *arXiv preprint arXiv:1905.03747*, 2019.

Robert L Bettinger and Mark N Grote. Marginal value theorem, patch choice, and human foraging response in varying environments. *Journal of Anthropological Archaeology*, 42:79–87, 2016.

Tanmay Bhowmik, Nan Niu, Wentao Wang, Jing-Ru C Cheng, Ling Li, and Xiongfei Cao. Optimal group size for software change tasks: A social information foraging perspective. *IEEE transactions on cybernetics*, 46(8):1784–1795, 2015.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

Paul Bloom and Frank C Keil. Thinking through language. *Mind & Language*, 16 (4):351–367, 2001.

Amos Bouskila and Daniel T. Blumstein. Rules of Thumb for Predation Hazard Assessment: Predictions from a Dynamic Model. *American Naturalist - AMER NATURALIST*, 139(1), January 1992. ISSN 0003-0147. doi: 10.1086/285318.

Luc Bovens, Stephan Hartmann, et al. *Bayesian epistemology.* Oxford University Press on Demand, 2003.

James H Brown, Vijay K Gupta, Bai-Lian Li, Bruce T Milne, Carla Restrepo, and Geoffrey B West. The fractal nature of nature: power laws, ecological complexity and biodiversity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1421):619–626, 2002.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Jerome S Bruner and Mary C Potter. Interference in visual recognition. *Science*, 144(3617):424–425, 1964.

Lou Burnard. Reference Guide for the British National Corpus (XML Edition), Jan 2007. URL `http://www.natcorp.ox.ac.uk/docs/URG`. [Online; accessed 18. Mar. 2021].

Ramon Ferrer i. Cancho and Ricard V. Solé. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. U.S.A.*, 100(3):788–791, Feb 2003. ISSN 0027-8424. doi: 10.1073/pnas.0335980100.

O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *J. Comput. Graph. Stat.*, 13(4):907–929, Dec 2004. ISSN 1061-8600. doi: 10.1198/106186004X12803.

Kurt A Carlson, Margaret G Meloy, and J Edward Russo. Leader-driven primacy: Using attribute order to affect consumer choice. *Journal of Consumer Research*, 32(4):513–518, 2006.

Eric L. Charnov. Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.*, 9(2):129–136, Apr 1976. ISSN 0040-5809. doi: 10.1016/0040-5809(76)90040-X.

Jianqing Chen and Jan Stallaert. An economic analysis of online advertising using behavioral targeting. *Mis Quarterly*, 38(2):429–A7, 2014.

Morten H Christiansen and Nick Chater. Language as shaped by the brain. *Behav Brain Sci*, 31(5):489–509, 2008.

Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. The production of information in the attention economy. *Scientific reports*, 5(1):1–6, 2015.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

Peter J Collins, Ulrike Hahn, Ylva Von Gerber, and Erik J Olsson. The bi-directional relationship between source characteristics and message content. *Frontiers in psychology*, 9:18, 2018.

John Cook and Stephan Lewandowsky. Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in cognitive science*, 8 (1):160–179, 2016.

Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artif. Intell.*, 42(2):393–405, March 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90060-D.

Alvaro Corral, Isabel Serra, and Ramon Ferrer-i Cancho. The distinct flavors of zipf's law in the rank-size and in the size-distribution representations, and its maximum-likelihood fitting. *arXiv preprint arXiv:1908.01398*, 2019.

Katalin Csilléry, Michael G. B. Blum, Oscar E. Gaggiotti, and Olivier François. Approximate Bayesian Computation (ABC) in Practice. *Trends Ecol. Evol.*, 25 (7):410–418, Jul 2010. ISSN 0169-5347. doi: 10.1016/j.tree.2010.04.001.

John M Darley and Paget H Gross. A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1):20, 1983.

Charles Darwin. *On the Various Contrivances by which Orchids are Fertilized by Insects (1862).* University of Chicago Press, 2011.

Ishita Dasgupta, Eric Schulz, Joshua B Tenenbaum, and Samuel J Gershman. A theory of learning to infer. *Psychological review*, 127(3):412, 2020.

Mark Davies. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190, 2009.

Mark Davies. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Edinburgh University Press 22 George Square, Edinburgh EH8 9LF UK*, Nov 2012. doi: 10.3366/cor.2012.0024.

Hank Davis and S Lyndsay McLeod. Why humans value sensational news: An evolutionary perspective. *Evolution and Human Behavior*, 24(3):208–216, 2003.

Nathaniel D Daw, Aaron C Courville, and Peter Dayan. Semi-rational models of conditioning: The case of trial order. *The probabilistic mind*, pages 431–452, 2008.

Morris H. Degroot. Reaching a Consensus. *J. Am. Stat. Assoc.*, 69(345):118–121, March 1974. ISSN 0162-1459. doi: 10.1080/01621459.1974.10480137.

Michela Del Vicario, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. Modeling confirmation bias and polarization. *Sci. Rep.*, 7(40391): 1–9, January 2017. ISSN 2045-2322. doi: 10.1038/srep40391.

Daniel C Dennett. *Kinds of minds: Toward an understanding of consciousness.* Basic Books, 2008.

Guy Deutscher et al. " overall complexity": a wild goose chase? 2009.

Kevin Dunbar. How scientists really reason: Scientific reasoning in real-world laboratories. *The nature of insight*, 18:365–395, 1995.

Robin IM Dunbar. Gossip in evolutionary perspective. *Review of general psychology*, 8(2):100–110, 2004.

W. Ebeling and T. Pöschel. Entropy and Long-Range Correlations in Literary English. *EPL*, 26(4):241–246, May 1994. ISSN 0295-5075. doi: 10.1209/0295-5075/26/4/001.

David S Evans. The economics of attention markets. *Available at SSRN 3044858*, 2020.

R. Ferrer i. Cancho. The variation of Zipf's law in human language. *Eur. Phys. J. B*, 44(2):249–257, March 2005. ISSN 1434-6036. doi: 10.1140/epjb/e2005-00121-8.

Leon Festinger. *A theory of cognitive dissonance*, volume 2. Stanford university press, 1962.

Baruch Fischhoff and Ruth Beyth-Marom. Hypothesis evaluation from a bayesian perspective. *Psychological review*, 90(3):239, 1983.

James Friedrich. Primary error detection and minimization (pedmin) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological review*, 100(2):298, 1993.

Wai-Tat Fu and Peter Pirolli. Snif-act: A cognitive model of user navigation on the world wide web. *Human–Computer Interaction*, 22(4):355–412, 2007.

Robert G Gallager. *Discrete stochastic processes*, volume 321. Springer Science & Business Media, 2012.

E Scott Geller and Gordon F Pitz. Confidence and decision speed in the revision of opinion. *Organizational Behavior and Human Performance*, 3(2):190–201, 1968.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Alan Gerber and Donald Green. Misperceptions about perceptual bias. *Annual review of political science*, 2(1):189–210, 1999.

Martin Gerlach and Francesc Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126, 2020.

Samuel J Gershman. How to never be wrong. *Psychonomic bulletin & review*, 26 (1):13–28, 2019.

Thomas Gilovich. Biased evaluation and persistence in gambling. *Journal of personality and social psychology*, 44(6):1110, 1983.

David G. Glynn. The permanent of a square matrix. *European Journal of Combinatorics*, 31(7):1887–1891, Oct 2010. ISSN 0195-6698. doi: 10.1016/j.ejc.2010.01.010.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

Michael H Goldhaber. The attention economy and the net. *First Monday*, 1997.

Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Problems with Fitting to the Power-Law Distribution. *European Physical Journal B*, 41(2), Feb 2004. ISSN 1434-6028. doi: 10.1140/epjb/e2004-00316-5.

Michael E Gorman. Error, falsification and scientific inference: An experimental investigation. *The Quarterly Journal of Experimental Psychology Section A*, 41 (2):385–412, 1989.

Ulrike Hahn and Adam JL Harris. What does it mean to be biased: Motivated reasoning and rationality. In *Psychology of learning and motivation*, volume 61, pages 41–102. Elsevier, 2014.

Ulrike Hahn, Christoph Merdes, and Momme von Sydow. How good is your evidence and how would you know? *Topics in Cognitive Science*, 10(4):660–678, 2018.

Rudolf Hanel, Bernat Corominas-Murtra, Bo Liu, and Stefan Thurner. Fitting power-laws in empirical data with estimators that work for all exponents. *PLoS One*, 12(2):e0170920, Feb 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0170920.

Andrew B. Hargadon and Beth A. Bechky. When Collections of Creatives Become Creative Collectives: A Field Study of Problem Solving at Work. *Organization Science*, August 2006. URL `https://pubsonline.informs.org/doi/abs/10.1287/orsc.1060.0200?casa`$_{t}oken = qj3t5_{o}F54IAAAAA : gLpsCCXM9PC_{p}iitg5hs - dJ0wYpNGgwwwWFipXZNVvrr95g5fqego2k3gVRlLz_{P}ETzVY_{r}rigrp$.

P Sol Hart and Erik C Nisbet. Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication research*, 39(6):701–723, 2012.

William Hart, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. Feeling Validated Versus Being Correct:A Meta-Analysis of Selective. *Psychol. Bull.*, 135(4):555, July 2009. doi: 10.1037/a0015701.

Martie G Haselton, Gregory A Bryant, Andreas Wilke, David A Frederick, Andrew Galperin, Willem E Frankenhuis, and Tyler Moore. Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, 27(5):733–763, 2009.

Harold Stanley Heaps. *Information retrieval, computational and theoretical aspects.* Academic Press, 1978.

Leah Henderson and Alexander Gebharter. The role of source reliability in belief polarisation. *Synthese*, pages 1–24, 2021.

Thomas T Hills. Animal foraging and the evolution of goal-directed cognition. *Cognitive science*, 30(1):3–41, 2006.

Thomas T Hills. The dark side of information proliferation. *Perspectives on Psychological Science*, 14(3):323–330, 2019.

Thomas T Hills and James S Adelman. Recent evolution of learnability in american english from 1800 to 2000. *Cognition*, 143:87–92, 2015.

Thomas T Hills, Michael N Jones, and Peter M Todd. Optimal foraging in semantic memory. *Psychological review*, 119(2):431, 2012.

Thomas T Hills, Peter M Todd, David Lazer, A David Redish, Iain D Couzin, Cognitive Search Research Group, et al. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, 19(1):46–54, 2015.

Crawford S Holling. Some characteristics of simple types of predation and parasitism. *Canadian entomologist*, 91(7):385–398, 1959.

Md Manjurul Hussain and Ishtiak Mahmud. pymannkendall: a python package for non parametric mann kendall family of trend tests. *Journal of Open Source Software*, 4(39):1556, 2019.

Ramon Ferrer i Cancho and Ricard V Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3): 788–791, 2003.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009.

Andreas Jarvstad and Ulrike Hahn. Source reliability and the conjunction fallacy. *Cognitive Science*, 35(4):682–711, 2011.

Alan Jern, Kai-Min K Chang, and Charles Kemp. Belief polarization is not always irrational. *Psychological review*, 121(2):206, 2014.

Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. Source monitoring. *Psychological bulletin*, 114(1):3, 1993.

Edward E Jones, Leslie Rock, Kelly G Shaver, George R Goethals, and Lawrence M Ward. Pattern of performance and ability attribution: An unexpected primacy effect. *Journal of Personality and Social Psychology*, 10(4):317, 1968.

Hillard Kaplan and Kim Hill. The evolutionary ecology of food acquisition. In *Evolutionary ecology and human behavior*, pages 167–202. Taylor and Francis Inc., 2017.

Kimmo Kettunen. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245, 2014.

Joshua Klayman. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418, 1995.

Joshua Klayman and Young-Won Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2):211, 1987.

Jonathan J Koehler. The influence of prior beliefs on scientific judgments of evidence quality. *Organizational behavior and human decision processes*, 56(1):28–55, 1993.

Ziva Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3):480, 1990.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

Hazel Lacohée, Nina Wakeford, and Ian Pearson. A social history of the mobile telephone with a view of its future. *BT Technology Journal*, 21(3):203–211, 2003.

Kevin N. Laland, Kim Sterelny, John Odling-Smee, William Hoppitt, and Tobias Uller. Cause and Effect in Biology Revisited: Is Mayr's Proximate-Ultimate Dichotomy Still Useful? *Science*, 334(6062):1512–1516, December 2011. ISSN 0036-8075. doi: 10.1126/science.1210879.

Joseph Lawrance, Christopher Bogart, Margaret Burnett, Rachel Bellamy, Kyle Rector, and Scott D Fleming. How programmers debug, revisited: An information foraging theory perspective. *IEEE Transactions on Software Engineering*, 39(2): 197–215, 2010a.

Joseph Lawrance, Margaret Burnett, Rachel Bellamy, Christopher Bogart, and Calvin Swart. Reactive information foraging for evolving goals. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 25–34, 2010b.

Christoph Leuenberger and Daniel Wegmann. Bayesian Computation and Model Selection Without Likelihoods. *Genetics*, 184(1):243–252, Jan 2010. ISSN 0016-6731. doi: 10.1534/genetics.109.109058.

Akiva Liberman and Shelly Chaiken. Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin*, 18(6):669–679, 1992.

Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A. Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163): 713–716, Oct 2007. ISSN 1476-4687. doi: 10.1038/nature06137.

Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.*, 43:e1, 2020. ISSN 0140-525X. doi: 10.1017/S0140525X1900061X.

Charles G Lord. The“ disappearance” of dissonance in an age of relativism. *Personality and Social Psychology Bulletin*, 15(4):513–518, 1989.

Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

Charles G Lord, Mark R Lepper, and Elizabeth Preston. Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, 47(6):1231, 1984.

Philipp Lorenz-Spreen, Bjarke Mørch Mønsted, Philipp Hövel, and Sune Lehmann. Accelerating dynamics of collective attention. *Nat. Commun.*, 10(1759):1–9, Apr 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09311-w.

Gary Lupyan and Rick Dale. Language Structure Is Partly Determined by Social Structure. *PLoS One*, 5(1):e8559, Jan 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0008559.

Robert H MacArthur and Eric R Pianka. On optimal use of a patchy environment. *The American Naturalist*, 100(916):603–609, 1966.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Michael J Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175, 1977.

Eugene Malthouse. Confirmation bias and vaccine-related beliefs in the time of COVID-19. *J. Public Health*, page fdac128, November 2022. ISSN 1741-3842. doi: 10.1093/pubmed/fdac128.

Benoit Mandelbrot. An informational theory of the statistical structure of language. *Communication theory*, 84:486–502, 1953.

Benoit B Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

André CR Martins. Trust in the coda model: Opinion dynamics and the reliability of other agents. *Physics Letters A*, 377(37):2333–2339, 2013.

Ernst Mayr. Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506, 1961.

John W McHoskey. Case closed? on the john f. kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic and Applied Social Psychology*, 17(3):395–409, 1995.

P. M. Meire and A. Ervynck. Are oystercatchers (Haematopus ostralegus) selecting the most profitable mussels (Mytilus edulis)? *Anim. Behav.*, 34(5):1427–1435, October 1986. ISSN 0003-3472. doi: 10.1016/S0003-3472(86)80213-5.

Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

Christoph Merdes, Momme Von Sydow, and Ulrike Hahn. Formal models of source reliability. *Synthese*, pages 1–29, 2020.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014): 176–182, Jan 2011. ISSN 0036-8075. doi: 10.1126/science.1199644.

Arthur G Miller, John W McHoskey, Cynthia M Bane, and Timothy G Dowd. The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology*, 64(4):561, 1993.

George A Miller. Some effects of intermittent silence. *The American journal of psychology*, 70(2):311–314, 1957.

Marcelo A Montemurro and Damián H Zanette. New perspectives on zipf's law in linguistics: from single texts to large corpora. *Glottometrics*, 4:87–99, 2002.

Isabel Moreno-Sánchez, Francesc Font-Clos, and Álvaro Corral. Large-Scale Analysis of Zipf's Law in English Texts. *PLoS One*, 11(1):e0147073, Jan 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0147073.

Geoffrey D Munro and Peter H Ditto. Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, 23(6):636–653, 1997.

Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Andy Norman. Why we reason: Intention-alignment and the genesis of human rationality. *Biology & Philosophy*, 31(5):685–704, 2016.

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

Erik J Olsson. A simulation approach to veritistic social epistemology. *Episteme*, 8 (2):127–143, 2011.

John T Ormerod and Matt P Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

Martin T Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, 17(11):776, 1962.

James F O'Connell and Kristen Hawkes. Alyawara plant use and optimal foraging theory. *Hunter-gatherer foraging strategies: Ethnographic and archaeological analyses*, pages 99–125, 1981.

Chang Sup Park and Barbara K Kaye. Smartphone and self-extension: Functionally, anthropomorphically, and ontologically extending self via the smartphone. *Mobile Media & Communication*, 7(2):215–231, 2019.

Judea Pearl. *Causality.* Cambridge university press, 2009.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041, 2015.

Uwe Peters. What is the function of confirmation bias? *Erkenntnis*, pages 1–26, 2020.

Alexander M Petersen, Joel N Tenenbaum, Shlomo Havlin, H Eugene Stanley, and Matjaž Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific reports*, 2(1):1–10, 2012.

Cameron R Peterson and Wesley M DuCharme. A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, 73(1):61, 1967.

Steven T. Piantadosi. Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions. *Psychon. Bull. Rev.*, 21(5):1112–1130, Oct 2014. ISSN 1531-5320. doi: 10.3758/s13423-014-0585-6.

John R Pierce. *An introduction to information theory: symbols, signals and noise.* Courier Corporation, 2012.

Charlie Pilgrim. Piecewise-regression (aka segmented regression) in python. *Journal of Open Source Software*, 6(68):3859, 2021.

Charlie Pilgrim and Thomas T Hills. Bias in zipf's law estimators. *arXiv preprint arXiv:2008.00903*, 2020.

Charlie Pilgrim, Weisi Guo, and Samuel Johnson. Organisational social influence on directed hierarchical graphs, from tyranny to anarchy. *Scientific Reports*, 10 (1):1–13, 2020.

Charlie Pilgrim, Paul Kent, Kasra Hosseini, and Ed Chalstrey. Ten simple rules for working with other people's code. *PLOS Computational Biology*, 19(4):e1011031, 2023.

Steven Pinker. Language as an adaptation to the cognitive niche. *Studies in the Evolution of Language*, 3:16–37, 2003.

David J Piorkowski, Scott D Fleming, Irwin Kwan, Margaret M Burnett, Christopher Scaffidi, Rachel KE Bellamy, and Joshua Jordahl. The whats and hows of programmers' foraging diets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3063–3072, 2013.

Peter Pirolli. An elementary social information foraging model. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 605–614, 2009a.

Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4): 643, 1999.

Peter LT Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2009b.

Scott Plous. Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, 21(13):1058–1082, 1991.

Horst Pöttker. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

Chaelin K Ra, Junhan Cho, Matthew D Stone, Julianne De La Cerda, Nicholas I Goldenson, Elizabeth Moroney, Irene Tung, Steve S Lee, and Adam M Leventhal. Association of digital media use with subsequent symptoms of attention-deficit/hyperactivity disorder among adolescents. *Jama*, 320(3):255–263, 2018.

Ralph Radach, Stefanie Lemmer, Christian Vorstius, Dieter Heller, and Karina Radach. Eye movements in the processing of print advertisements. In *The Mind's Eye*, pages 609–632. Elsevier, 2003.

David P Redlawsk. Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, 64(4): 1021–1044, 2002.

Victoria J Rideout, Ulla G Foehr, and Donald F Roberts. Generation m 2: Media in the lives of 8-to 18-year-olds. *Henry J. Kaiser Family Foundation*, 2010.

Lee Ross, Mark R Lepper, and Michael Hubbard. Perseverance in self-perception and social perception: biased attributional processes in the debriefing paradigm. *Journal of personality and social psychology*, 32(5):880, 1975.

Karolina Rudnicka. Variation of sentence length across time and genre. *Diachronic corpora, genre, and language change*, pages 220–240, 2018.

J Edward Russo. The predecisional distortion of information. In *Neuroeconomics, judgment, and decision making*, pages 109–128. Psychology Press, 2014.

J Edward Russo. Bayesian revision vs. information distortion. *Frontiers in psychology*, 9:1550, 2018.

J Edward Russo, Victoria Husted Medvec, and Margaret G Meloy. The distortion of information during decisions. *Organizational behavior and human decision processes*, 66(1):102–110, 1996.

Herbert John Ryser. *Combinatorial mathematics*, volume 14. American Mathematical Soc., 1963.

Geoffrey Sampson. A linguistic axiom challenged. *Language complexity as an evolving variable*, 2:18, 2009.

Adam N Sanborn. Types of approximation for probabilistic cognition: Sampling and variational. *Brain and cognition*, 112:98–101, 2017.

Adam N Sanborn and Ricardo Silva. Constraining bridges between levels of analysis: A computational justification for locally bayesian learning. *Journal of Mathematical Psychology*, 57(3-4):94–106, 2013.

Pamela Effrein Sandstrom. An optimal foraging approach to information seeking and use. *The library quarterly*, 64(4):414–449, 1994.

Thomas Schürmann and Peter Grassberger. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):414–427, 1996.

Aaronson Scott. A linear-optical proof that the permanent isP-hard. *Proc. R. Soc. A.*, 467(2136):3393–3405, Dec 2011. ISSN 1471-2946. doi: 10.1098/rspa.2011.0232.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX, 2010.

HL Seal. The maximum likelihood fitting of the discrete pareto law. *Journal of the Institute of Actuaries (1886-1994)*, 78(1):115–121, 1952.

Jaime E Settle. *Frenemies: How social media polarizes America.* Cambridge University Press, 2018.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *J. Comput. Soc. Sc.*, 3(2):279–317, November 2020. ISSN 2432-2725. doi: 10.1007/s42001-020-00086-5.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

Herbert A Simon. Designing organizations for an information-rich world. *Brookings Institute Lecture*, 1969.

Herbert A. Simon. Bounded Rationality. In *Utility and Probability*, pages 15–18. Palgrave Macmillan, London, England, UK, 1990. doi: 10.1007/978-1-349-20568-4_5.

Herbert A Simon et al. Designing organizations for an information-rich world. *Computers, communications, and the public interest*, 72:37, 1971.

S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U.S.A.*, 104(6):1760–1765, Feb 2007. ISSN 0027-8424. doi: 10.1073/pnas.0607208104.

Louisa M Slowiaczek, Joshua Klayman, Steven J Sherman, and Richard B Skov. Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, 20(4):392–405, 1992.

Eric Alden Smith, Robert L Bettinger, Charles A Bishop, Valda Blundell, Elizabeth Cashdan, Michael J Casimir, Andrew L Christenson, Bruce Cox, Rada Dyson-Hudson, Brian Hayden, et al. Anthropological applications of optimal foraging theory: a critical review [and comments and reply]. *Current Anthropology*, 24(5): 625–651, 1983.

Kenny Smith and Simon Kirby. Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3591–3603, 2008.

Mark Snyder. When belief creates reality. In *Advances in experimental social psychology*, volume 18, pages 247–305. Elsevier, 1984.

David W Stephens and John R Krebs. *Foraging theory*, volume 1. Princeton University Press, 1986.

David E Sumner. *The magazine century: American magazines since 1900*, volume 9. Peter Lang, 2010.

Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate Bayesian Computation. *PLoS Comput. Biol.*, 9(1):e1002803, Jan 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002803.

Charles S Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769, 2006.

Charles S Taber, Damon Cann, and Simona Kucsova. The motivated processing of political arguments. *Political Behavior*, 31(2):137–155, 2009.

Tiziana Terranova. Attention, economy and the brain. *Culture Machine*, 13, 2012.

Robin Thompson. Radicalization and the use of social media. *Journal of strategic security*, 4(4):167–190, 2011.

Peter M Todd and Thomas T Hills. Foraging in mind. *Current Directions in Psychological Science*, 29(3):309–315, 2020.

Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293, 1983.

Leslie G Valiant. The complexity of computing the permanent. *Theoretical computer science*, 8(2):189–201, 1979.

Robert P Vallone, Lee Ross, and Mark R Lepper. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of personality and social psychology*, 49(3):577, 1985.

Regina J. J. M. van den Eijnden, Jeroen S. Lemmens, and Patti M. Valkenburg. The Social Media Disorder Scale. *Computers in Human Behavior*, 61:478–487, August 2016. ISSN 0747-5632. doi: 10.1016/j.chb.2016.03.038.

Peter Vorderer. It's all entertainment—sure. But what exactly is entertainment? Communication research, media psychology, and the explanation of entertainment experiences. *Poetics*, 29(4):247–261, November 2001. ISSN 0304-422X. doi: 10.1016/S0304-422X(01)00037-7.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Soc. Sci. Med.*, 240:112552, November 2019. ISSN 0277-9536. doi: 10.1016/j.socscimed.2019.112552.

Peter C Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140, 1960.

Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

Earl E Werner and Gary G Mittelbach. Optimal foraging: field tests of diet choice and habitat switching. *American Zoologist*, 21(4):813–829, 1981.

Bruce Winterhalder. Diet choice, risk, and food sharing in a stochastic environment. *Journal of anthropological archaeology*, 5(4):369–392, 1986.

Heather Cleland Woods and Holly Scott. #Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem. *J. Adolesc.*, 51:41–49, August 2016. ISSN 0140-1971. doi: 10.1016/j.adolescence.2016.05.008.

Haoran Zhu and Lei Lei. Is modern english becoming less inflectionally diversified? evidence from entropy-based algorithm. *Lingua*, 216:10–27, 2018.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology.* Hafner, 1949.

Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, 24(5):530–536, 1978.