

VISUALIZING AND MODELING JOINT BEHAVIOR OF
CATEGORICAL VARIABLES WITH A LARGE NUMBER OF
LEVELS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Justin S. Dyer

November 2010

© Copyright by Justin S. Dyer 2010

All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Art B. Owen) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Guenther Walther)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Thomas Cover)

Approved for the University Committee on Graduate Studies

Abstract

Large social, internet, or biological networks frequently have a long tailed degree distribution. Often there are two long tail phenomena. For bipartite networks two different entity types may each have a power law degree distribution, and for directed networks both in-degree and out-degree may have power law distributions. These can all be viewed as datasets with a large number of categorical levels along multiple dimensions. We frequently observe a head-to-tail affinity between two types of entities. For example with movie ratings data, viewers making the fewest ratings are overrepresented in movies receiving the most ratings and conversely.

We present a graphical display to visualize the affinities. The display also exposes some interpretable anomalies in network data. Our graphic is based on ordering the entities by size. We show in a Zipf–Poisson model that the largest entities accounting for the majority of data are correctly ordered with probability near one. A saturation model produces head to tail affinities similar to those seen in the data, though a bipartite preferential attachment model gives a better fit to the marginal distributions. Sharp bounds are obtained on the behavior of the margins in both of these cases. An extended model incorporating temporal aspects of the evolution of networks reproduces some observed characteristics of real data using only a few parameters.

As our graphical display has a close connection to copulas, we explore the utility of parametric and nonparametric copula models for our data. In the nonparametric case, we present two new ways of generating smooth valid copula-density estimates. In the first, we modify an algorithm in the existing literature to get valid copula densities. In the second, we formulate the smoothing problem in a convex-programming framework, thus folding the copula constraints directly into the optimization problem, and then provide insight into the character of the smoothing parameter.

We also develop a new class of discrete-choice models in which the choice set grows with time. By exploiting the structure that arises when the covariate vector takes a special form, we are able to fit a rich model via maximum-likelihood to a data set with over fifty million observations. Interesting issues arise when attempting to simulate from the fitted model, which leads to a proposal for how to sample from discrete distributions with a very large number of possible outcomes and with probabilities that evolve over time.

Acknowledgments

It seems that standard mythology regarding Ph.D. research tells the story of a student toiling alone by him or herself in an attic (or a basement) over the course of several years in order to produce enough original work to put together a thesis and graduate. I suspect, however, that it is the very rare student who produces a sufficient body of original work with no guidance or support whatsoever from anyone else. I certainly was not such a student.

There are many people without whose support, guidance, encouragement, and patience I would not have been able to complete this degree. While some might view that as a weakness, I have slowly learned to view it as a strength.

First and foremost is my advisor, Art Owen. From the very beginning, when I met with him to discuss the possibility of working with him, he has been a constant source of encouragement and enthusiasm. I have certainly not been the easiest doctoral student to advise—for many reasons—and yet, he stuck with me and supported me even in ways that I did not learn about until later, through other people. I have done my best to study him in an effort to better understand and emulate his intuition, statistical and mathematical acumen, and approach to communicating with other researchers, both inside and outside the field of statistics. His respect for others and evident *joie de vivre* make him a true pleasure to be around. In the end, I could not

have asked for a better advisor.

Second, my wife, Hazel, has made the imminent completion of this thesis possible. Without her support, patience, level head, and quick wit, I would likely not have survived my first year in the program. She forfeited the better part of that year of our marriage so that I could pursue what, at the time, seemed like a frustrating and quixotic quest. Since then, she has done a tremendous job to help keep my head on straight, and has (rightfully) adjusted it, as necessary.

My peers in the program have been a constant source of intellectual stimulation and wonderfully pleasant company. Ya Xu and Murat Ahmed, in particular, have challenged my thinking, improved my knowledge and intuition, and always been around for a chat when I needed to unload. Camilo Rivera was instrumental in helping me submit my thesis after I had already moved away from California. Donal McMahon, Sarah Emerson, Genevera Allen, and Nick Johnson are also due many thanks for their friendship and statistical insights.

Hewlett–Packard Laboratories, and especially, Hsiu-Khuern Tang, Fereydoon Safai, and Jaap Suermondt, provided an additional avenue for me to continue to grow as a statistician. The National Science Foundation provided much appreciated financial support through a Graduate Research Fellowship during the majority of my time at Stanford. My M.S. thesis advisor, Bala Natarajan, continues to be a good friend and advisor whom I respect immensely. Without his support in helping me get an early taste of academic research, I doubt I would have had the opportunity to come to Stanford.

Thanks are also due to the other thesis readers and members of my orals committee: Guenther Walther, Tom Cover, Rob Tibshirani, and Simon Jackman. The job of

attending an oral exam and reading a thesis seems to be one of the many activities of a professor that goes largely unnoticed and underappreciated. Thank you all. When I tore my ACL on the first day of Spring Break of my first year while playing basketball with Rob, he accompanied me to the hospital and stayed with me there for an entire Saturday afternoon, treating me as he would a son, despite the fact he could have been elsewhere with his own family. I will never forget that expression of generosity and humanity.

In addition, Prof. Cover has been extremely generous with his time, and the opportunity to be a TA for his Mathematics of Sports course was one of the highlights of my stay at Stanford. During the course of that quarter, and subsequent ones, conversations with him have stimulated many hours of extra thought and study on my part; I consider him one of the most interesting people I have ever had the chance to be around.

Finally, my parents and brother have always been there for me, and for that I cannot thank them enough. I am fortunate to count each one of them not only as family, but also as among my very best and closest friends. Being the product of two parents with doctorates in engineering, it is easy to assume that there was pressure on me to take a certain direction in life. Quite the contrary, the freedom my parents gave me from a very young age is an indication of the respect they had, and continue to have, for my development as an individual. The unwavering support that they've given is a sign of their immense love. Thank you.

I am surely leaving out someone else that deserves mention. For that, my sincerest apologies.


I am proud to say that I was unable to complete my Ph.D. all by myself.

Contents

Abstract	iv
Acknowledgments	vi
1 Introduction	1
1.1 Background on power laws	4
1.1.1 Pareto’s observations on income distribution	5
1.1.2 Yule’s continuous-time process for evolution	6
1.1.3 Zipf’s law, distribution, and plot	8
1.1.4 Simon’s model for text	11
1.1.5 Scaling laws in internet web graphs	12
1.1.6 Bivariate power laws	15
1.2 Background on copulas	18
1.2.1 Definition and first properties	19
1.2.2 Sklar’s theorem and its consequences	21
1.2.3 Additional properties	25
2 The data display	29
2.1 Introduction	29

2.2	Forming the display	30
2.2.1	Variations on the theme	33
2.3	Examples	34
2.3.1	Netflix movie ratings	35
2.3.2	Yahoo! music ratings	37
2.3.3	IMDB movie-actor database	39
2.3.4	Epinions	41
2.3.5	Wikipedia	44
2.3.6	Snapfish	46
2.3.7	arXiv hep-th citations	49
2.3.8	Enron email network	50
2.3.9	California roads	51
2.4	Numerical summaries	53
2.5	How accurate is the display?	55
2.6	Summary	57
3	Parametric copula models	58
3.1	Classical parametric copulas	59
3.1.1	Sklar’s theorem and the inversion method	59
3.1.2	Archimedean copulas	62
3.1.3	Remarks on symmetry	66
3.2	Generating new copulas from old	67
3.2.1	Reflections and mixtures	68
3.2.2	Liebscher’s method	69
3.3	Remarks on simulating from parametric copula models	71

3.4	Summary	73
4	Nonparametric copula models	74
4.1	Genest et al.'s nonparametric wavelet copulas	74
4.1.1	Sinkhorn–Knopp for valid copula-density estimates	77
4.1.2	Netflix example	79
4.2	A convex-programming approach to nonparametric copula estimation	80
4.2.1	Problem formulation	80
4.2.2	Properties	82
4.2.3	Selecting the smoothing parameter	84
4.3	Summary	85
5	Simple generative models	86
5.1	A saturation model	87
5.1.1	Numerical examples	89
5.1.2	Parameter estimation	91
5.2	Bipartite preferential attachment	92
5.2.1	Numerical example	94
5.2.2	Extensions	96
5.2.3	Parameter estimation	98
5.3	Bipartite models incorporating temporal preferential attachment . . .	98
5.3.1	Model specification	99
5.3.2	Numerical Examples	100
5.4	Summary	102

6	Discrete-choice models	103
6.1	The model	103
6.1.1	The likelihood	106
6.2	Updatable features	107
6.3	Fitting via maximum likelihood	109
6.4	Sampling from the fitted model	112
6.4.1	Method 1: Binary search with cumulative updates . Indeed;	113
6.4.2	Method 2: Modular storage binning	114
6.5	Netflix example	116
6.5.1	Data preprocessing	116
6.5.2	Modeling arrival rates	116
6.5.3	Modeling entity selection conditional on edge class	118
6.5.4	Simulating from the model	121
6.6	Summary	125
A	Proofs	126
A.1	Proof of Theorem 2.1	127
A.1.1	 Correct relative ordering , equation (2.1)	127
A.1.2	Correct absolute ordering, equation (2.2)	129
A.1.3	Limit to correct ordering, equation (2.3)	132
A.1.4	Sharper results on the limits to the correct ordering	134
A.1.5	Probability of tail exceedance is small even in the deep tail	137
A.2	Proof of Theorem 5.1	140
A.3	Proof of Theorem 5.2	142

List of Tables

2.1	Summary statistics for example datasets	35
2.2	Numerical summary of corner affinities of example datasets.	54
6.1	Fitted coefficients of rate model for Netflix data.	118
6.2	Fitted coefficients of conditional entity selection model for Netflix data.	121

List of Figures

1.1	Degree distribution (<i>left</i>) and Zipf (<i>right</i>) plots for an i.i.d. sample from a Zipf distribution.	10
2.1	Toy example of data visualization. (Permission to use this figure was kindly provided by Art Owen.)	32
2.2	Zipf plot of Netflix margins.	36
2.3	Data display for Netflix example.	37
2.4	Zipf plot of Yahoo! music ratings margins.	38
2.5	Data display for Yahoo! music ratings example.	39
2.6	Zipf plot of IMDB margins.	40
2.7	Data display for IMDB example.	40
2.8	Zipf plot of Epinions margins.	42
2.9	Degree distribution plot of Epinions margins.	43
2.10	Data display for Epinions example.	43
2.11	Zipf plot of Wikipedia margins.	45
2.12	Data display for Wikipedia example.	45
2.13	Zipf plot of Snapfish margins.	47
2.14	Data display for Snapfish example.	47

2.15	Zipf plot of arXiv hep-th margins.	49
2.16	Data display for arXiv hep-th example.	50
2.17	Zipf plot of Enron margins.	51
2.18	Data display for Enron example.	52
2.19	Data display for CA roads example.	53
3.1	Example Gaussian copulas	61
3.2	Example Clayton copulas	65
3.3	Example Frank copulas	66
3.4	Example asymmetric copulas from “generating” Frank copula with $\theta = 5$. 71	
4.1	Original and smoothed versions of histogram-normalized Netflix display. 79	
5.1	Degree distribution plot of bipartite preferential attachment ($a = 1.5$, $b = 1.8$, $t = 10^7$) margins.	89
5.2	Zipf plot of saturation model margins with $a = 1.5$, $b = 1.8$ and $N = 10^7$. 90	
5.3	Data display for two saturation model examples.	91
5.4	Degree distribution plot of bipartite preferential attachment ($p = 4/9$, $t = 10^6$) margins.	94
5.5	Zipf plot of bipartite preferential attachment ($p = 4/9$, $t = 10^6$) margins. 95	
5.6	Data display for bipartite preferential attachment ($p = 4/9$, $t = 10^6$) example.	96
5.7	Examples of real datasets and their extended-model counterparts. . . 101	
6.1	Empirical and fitted rates of arrival for each of the four classes for the Netflix example. Plots are on a log–log scale.	119
6.2	Zipf plots from simulated and real Netflix data.	123

6.3	Copula plots from simulated and real Netflix data.	124
-----	--	-----

Chapter 1

Introduction

Network data typically exhibit a long tail phenomenon in which some nodes or entities account for much more of the data than others. For bipartite networks with links between two distinct types of entities, there are usually two power laws, one for each type of entity. Similarly, directed networks may have two power laws, one for in-degree and one for out-degree.

The phenomenon we are interested in here is the nature of the association between the two power law quantities. Very commonly we see that the entities from the head of one distribution have an affinity for entities at the tail of the other. In this thesis, we look at graphical displays of the entire joint distribution. We see several different patterns that can then be interpreted in terms of the original data. A given correlation coefficient could be consistent with many different data patterns. The patterns we see are often concentrated at the extreme ranges, head or tail, of the data.

A famous example of ratings data comes from the Netflix prize with just over 100 million movie ratings made by 480,189 customers on 17,770 movies. Inspecting the Netflix data it becomes clear that the busy raters tend to rate the less popular

movies and that the popular movies tend to attract the less active raters. A similar phenomenon happens in other data sets. But the strength and nature of these affinities differ from data set to data set. Furthermore the affinities don't have to be symmetric: popular movies are less strongly associated with rare raters than busy raters are with unpopular movies.

In the remainder of this chapter, we cover some of the background surrounding this problem. We mention some classical works that study univariate power-law behavior from the perspective of stochastic processes and data analysis. We begin with Pareto's work on income distribution, following up with Yule's continuous-time stochastic-process for evolution. Zipf and Simon's independent investigations of vocabulary usage patterns in the English language are perhaps some of the most well-known early examples of examination of power-law behavior where the response variable is categorical. More recently, Faloutsos and many others have studied the degree distributions of web and social networks. The Barabási-Albert preferential-attachment model is perhaps the most widely known work in this area and we review the formulation and rigorous results obtained by Bollobás, et al. Our intent will not be to do a full literature survey, but rather a representative sampling to give an overall picture of the field.

Copulas also play a central role in this thesis, albeit from a slightly atypical perspective. We review some of the basic properties in the last part of this chapter in preparation for the remaining chapters.

Chapter 2 proposes and discusses a gray scale display to show the joint distribution of two long tailed quantities. The displays show the shape, size and sometimes surprising location of the affinities in real data sets. We examine numerous examples

taken from web, email, and social networks, large-scale ratings databases and transportation networks. In making these displays we have implicitly assumed that sorting entities by their observed size in the data set puts them into the correct order that we would see in an infinite sample. For large data sets, we introduce a Zipf–Poisson ensemble model that gives us reason to expect that the largest entities, accounting for most of the data will in fact be properly ordered. The model has an infinite number of entities, of which only finitely many appear in a given sample. With a Zipf parameter $\alpha > 0$ and expected number $N\zeta(\alpha)$ of observations, the top $(3N/(32\log(N)))^{1/(\alpha+2)}$ entities are correctly ordered with probability tending to 1 as $N \rightarrow \infty$.

Chapters 3 and 4 discuss parametric and nonparametric methods, respectively, for explicitly fitting a copula density to the observed data display. In the parametric case, it proves difficult to generate the asymmetries seen in real data and the resulting mathematical forms are unwieldy and unintuitive at the same time. This motivates a nonparametric approach and we examine some of the work of Genest in this area. Unfortunately, nonparametric estimates seldom yield valid copulas. We propose a convex-programming formulation for smoothing the empirical copula estimate corresponding to the data display. The smoothing is done by fixing a linear basis and introducing a shrinkage parameter to control the degree of smoothness. This shrinking parameter turns out to have a very natural interpretation when the “correct” basis is used.

Head-to-tail affinities for raters and rated items can be explained in terms of experienced raters having more varied and sophisticated tastes than beginners. In Chapter 5, we propose three generative baseline models that provide alternative explanations. One is a saturation model in which raters and items are independently

sampled but subject to a threshold in which no pair is counted more than once. For the bipartite case, saturation creates a head-to-tail affinity but we show that it generates unrealistic marginal distributions. A second model invokes bipartite preferential attachment. This model provides reasonable marginal distributions and we find head-to-tail affinities. We derive some properties of these two models and demonstrate their ability to generate head-to-tail affinities. The third model extends the bipartite preferential-attachment by degree framework to include a temporal preferential-attachment component as well. To decide which combination of degree and temporal preferential-attachment to use, a multinomial trial is introduced, adding a few more parameters to the model. We compare the simulations under these models using various parameter settings to those observed in real datasets and show qualitatively similar fits.

Chapter 6 considers a unique discrete-choice model for bipartite ratings or graph data. The model departs from conventional ones in that the choice-set increases along with the number of observations. For certain types of features, the likelihood equations can be updated surprisingly easily, greatly reducing the computational complexity while at the same time allowing for a rich dependence structure between the bivariate response variables.

An Appendix contains the proofs of various propositions and theorems introduced in the text.

1.1 Background on power laws

Power-law like behavior has been observed empirically for over one hundred years. In this subsection, we start with Pareto's observations on income distribution and work

our way up through the Barabási-Albert preferential-attachment model of internet graphs and the rigorous analysis of its properties. In the process, we will touch on example works in evolution, linguistics, social dynamics, and internet topology to name a few. A recent and quite complete literature survey on this topic can be found in [51], and so we do not seek to replicate that here.

1.1.1 Pareto’s observations on income distribution

Vilfredo Pareto was perhaps the first to make explicit reference to a power-law distribution in an observed process. The Italian economist was interested in political economy and in a set of lecture notes on the subject [59], he proposes the following “law” for income distribution:

$$\log N(x) = -\alpha \log x + b ,$$

where $N(x)$ is the number of people with an income exceeding x . Looking at the ratio of $N(x)/N(h)$ for $h < x$, the proportion of people with income exceeding x to that exceeding h is

$$U(x) = \left(\frac{h}{x}\right)^\alpha .$$

This has come to be known as Pareto’s Law.

Now, if h is the minimum possible income (assumed non-zero), $U(x)$ can be interpreted as the upper-tail distribution of a random variable X since $U(h) = 1$ and $\lim_{x \rightarrow \infty} U(x) = 0$, provided of course that $\alpha > 0$.

This provides our first example of a *power law* distribution. For the purposes of this thesis, a power-law distribution is a distribution function F such that $S = 1 - F$

satisfies

$$\lim_{x \rightarrow \infty} \frac{S(x)}{x^{-\alpha}} = c, \quad (1.1)$$

for some $\alpha > 0$ and $c > 0$. Heuristically, what this means is that if we plot the values of $S(x)$ against x on a log-log plot, we see something that looks like a line, particularly as x grows large.

Pareto wasn't quite content with his first "law" and so he proposed a modification by introducing another parameter a . Under his revised model he had

$$\log N(x) = -\alpha \log(x + a) + b,$$

yielding upper-tail distribution function

$$U(x) = \left(\frac{h + a}{x + a} \right)^\alpha. \quad (1.2)$$

Note that $U(x)$ in (1.2) satisfies the condition in (1.1). Despite being proposed by Pareto, the discrete analog of the distribution corresponding to (1.2) is often referred to as the Zipf–Mandelbrot law.

More perspective and history of a statistical nature on Pareto's laws can be found in [10].

1.1.2 Yule's continuous-time process for evolution

Yule [68] provides perhaps the first process related to the preferential-attachment graph models for which Barabási and Albert have received much recent fame. Yule was interested in evolution and specifically in modeling the evolution of the number

of overall species over time. In modern parlance, his model can be described as a birth-death process in which no death actually occurs (i.e., a *pure birth* process).

The initial description of the process [68, pg. 33] starts off considering discrete time steps from which, via a limiting argument, a continuous time process is constructed. Suppose that there are N initially existing genera, with each genus consisting of a single species. At each time step, with probability p , each species within each genus bifurcates into two new species, and with probability q each species fails to bifurcate. So, after one step there are about Nq genera of a single species and Np of two species. After the second time instant, the mean number of single-species genera is Nq^2 , of two-species genera is $N(pq + (pq^2)) = Npq(1 + q)$, of three species is $N(2p^2q)$ and of four species is Np^3 .

A continuous limit is then achieved by taking the time between intervals to be small and the number of intervals large, i.e., $t = n\Delta t$ and setting $p = s\Delta t$ for some constant s . The term $q^n = (1 - p)^n$ arises naturally in the analysis as the number of generations grows large and so, from this we get $(1 - p)^n \rightarrow e^{-st}$ as t remains fixed and n goes to infinity. The proportion of genera of size k is $e^{-st}(1 - e^{-st})^{k-1}$.

Now, if we suppose that not only can species bifurcate at each time interval, but genera may as well, albeit at a different rate, we may ask ourselves how many of the total genera over all time have a certain size. Let the probabilistic rate of genus bifurcation be g (where, for species bifurcation it was s) and define $\rho = s/g$, which should naturally be greater than unity.

The total number of genera at time t is then Ne^{gt} and the rate of new genera is Nge^{gt} . So at time T , there are about $Nge^{T-x} dx$ of age x to $x + dx$; hence, the proportion at time T of age x is $ge^{g-x} dx$. From this, we can calculate the proportion

of genera containing a single species as

$$g \int_0^\infty e^{-sx} e^{-gx} dx = g(g+s)^{-1} = (1+\rho)^{-1},$$

since there are e^{-sx} single-species genera created at time x .

For genera of two species we get a proportion $\rho(1+\rho)^{-1}(1+2\rho)^{-2}$, and in general, the proportion of genera containing k species will be

$$f_k = \frac{(k-1)!\rho^{k-1}}{(1+\rho)(1+2\rho)\cdots(1+k\rho)}.$$

Then, it can be shown via Stirling's approximation or similar arguments that $f_k \sim k^{-(1+1/\rho)}$ as $k \rightarrow \infty$, i.e., the proportion of total genera containing k species has a power-law distribution.

Considering a much different application and a different generative process, in this thesis, we arrive at an essentially identical power-law for the degree-distribution of a single partition of a bipartite graph generated via a simple bipartite preferential-attachment scheme in a discrete-time setting. See Chapter 5 and the appendix for more details. Whereas Yule's analysis is certainly elegant, it is not entirely rigorous. We provide rigorous statements and proofs of our results and augment those of Yule by also providing upper and lower bounds on the degree distributions of the processes we are interested in.

1.1.3 Zipf's law, distribution, and plot

Zipf was interested in, among other things, the usage patterns of vocabulary. He observed [69] that across multiple languages, including Latin, Chinese, and English,

there was a consistent relationship between the number of unique words and the number of occurrences of each of these words. He also provides his first suggestion for a visualization of these relationships.

Zipf suggests plotting on a double-logarithmic plot, the number of unique words with a given number of occurrences along the abscissa and the corresponding number of occurrences along the ordinate. A modern statistician would likely be strongly tempted to swap the ordinate with the abscissa and normalize the count of unique words for each data point by the total number of unique words, thus obtaining an estimated frequency distribution on the number of occurrences for unique words in the language.

Upon examining the data from multiple languages and plotting them, Zipf proposes the relationship

$$ab^2 = k ,$$

where a denotes the number of unique words that occur b times in the text and k is a constant depending on the length of the text under consideration. In other words, he proposes a power-law distribution for the number of occurrences with scaling coefficient of 2. His estimates of four corpora of English text are 2.15, 2.05, 2.1 and 2.15, respectively.

Another visualization is also suggested in [69] and discussed further in [70]. Zipf attributes the visualization to an unnamed friend [69, pg. 44]. He suggests considering the unique words in English as independent entities and plotting their relative frequencies in ranked order on a double-logarithmic plot. Hence, the most popular word would be have a coordinate of one on the abscissa and its relative frequency in the text on the ordinate. This is the visualization most commonly known as a *Zipf*

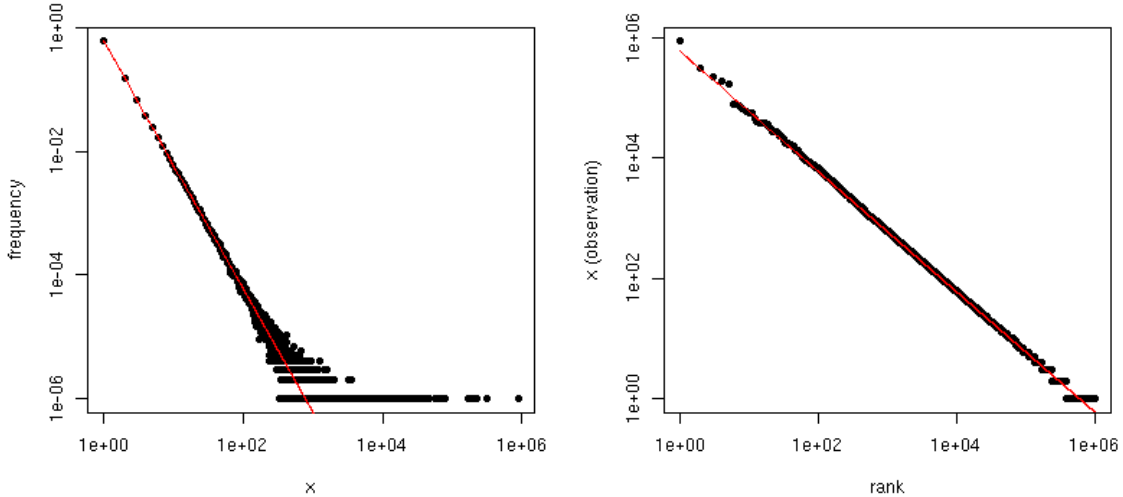


Figure 1.1: Degree distribution (*left*) and Zipf (*right*) plots for an i.i.d. sample from a Zipf distribution.

plot. We will make heavy use of it in the rest of this thesis; see Chapters 2 and 5, in particular.

This also motivates the Zipf distribution, which is the discrete counterpart of the Pareto distribution. Let $x \in \mathbb{N}$ and $\alpha > 1$. Then a random variable X follows the Zipf distribution if

$$\mathbb{P}(X = x) \propto x^{-\alpha}.$$

It is easy to see that the proportionality constant is the Riemann zeta function evaluated at α , i.e.,

$$\zeta(\alpha) = \sum_{x=1}^{\infty} x^{-\alpha}.$$

Figure 1.1 shows the empirical distribution (left pane) and Zipf (right pane) plots of an independent and identically (i.i.d.) sample of size 1 000 000 from a Zipf distribution with parameter $\alpha = 2$. The plots are both double-logarithmic; the slope of the

line in the left pane is 2 and in the right pane is 1. The left-pane slope corresponds to the scaling exponent of the distribution. To explain the slope in the right pane, let $\hat{r} = S_n/n$ where $S_n = \sum_{i=1}^n \mathbf{1}_{(X>x)}$, i.e., \hat{r} is the empirical proportion of random variables in the sample that exceed x . Now,

$$\mathbb{P}(X = x) \sim c_1 x^{-\alpha} \Rightarrow \mathbb{P}(X > x) \sim c_2 x^{-(\alpha-1)},$$

and so $\hat{r} \xrightarrow{P} c_2 x^{-(\alpha-1)}$. Inverting this relationship, we have $(c_2/\hat{r})^{1/(\alpha-1)} \xrightarrow{P} x$ by the continuous mapping theorem, and so we conclude that magnitude of the $(\hat{r}n)$ th order statistic is proportional to $\hat{r}^{-1/(\alpha-1)}$. This is exactly the quantity we need to determine the slope of the Zipf plot. So, at least for i.i.d. samples, if we sample from a power law distribution with scaling exponent α , we expect the Zipf plot to have slope $1/(\alpha - 1)$. Several different accounts of this property are available in the literature, mostly from physics. See [9, 3, 2, 56]. All of the aforementioned references implicitly assume an i.i.d. sample, as we have explicitly. Indeed, our analysis does not strictly require an i.i.d. sample, but rather just that \hat{r} is consistent for $\mathbb{P}(X > x)$.

1.1.4 Simon's model for text

In 1955, Simon [61] proposed a model for how words arise in a text document. His model is simple. Let $f(i, k)$ be the number of *unique* words that have appeared i times out of the first k words. Then, the probability that the $(k + 1)$ st word is one that has already appeared $f(i, k)$ times is proportional to $if(i, k)$ and there is a constant probability α that the $(k + 1)$ st word is an as-yet-unobserved one.

This model leads to a recurrence very similar to the preferential-attachment model

of Barabási and Albert, [6], which we will soon discuss in detail. Specifically,

$$\mathbb{E}(f(i, k + 1) \mid f(i', k), \forall i' \leq k) - f(i, k) = \frac{1 - \alpha}{k} ((i - 1)f(i - 1, k) - if(i, k)) ,$$

and taking expectations on both sides we arrive at a recurrence for $\mathbb{E}f(i, k)$. Starting at $i = 1$, we can solve explicitly and then by induction obtain an expression for each i and k .

The main difference between this model and the Barabási-Albert model is that in Simon's model, the addition of a new element at each step is stochastic in nature and only one of the values of $f(i, k)$ gets updated, whereas in the latter model, the addition of a new element is deterministic (i.e., a previously unobserved entity is *always* added at each step) and two values of $f(i, k)$ must be updated.

We will consider a model in Chapter 5 which combines features of both the Simon and Barabási-Albert models. From one perspective, our model will be composed of two simultaneously running versions of the Simon model which are dependent on one another. Another interpretation is that of a Barabási-Albert model adapted to a bipartite graph, as opposed to the unipartite case that was originally considered.

Simon's model generated a good deal of interest after it was introduced and aspects of it were debated vigorously in the literature. For more details, see [61, 45, 62, 46].

1.1.5 Scaling laws in internet web graphs

Starting in the 1990's, it was observed empirically that the degrees of internet web graphs [27] and certain aspects of ethernet traffic [41] tended to follow a distribution that could be well approximated by a power law. Extensive study of both the empirical properties of such data [57, 23, 66] and their modeling and analysis from a

stochastic-process point of view was undertaken and continues to this day.

Probably the most well-known model for describing the evolution of web graphs was introduced by Barabási and Albert [6] and is popularly known as the *preferential attachment* model. It has roots going back to Yule, Simon, and others. Barabási and Albert used mean-field heuristics to conclude that their model generated a power-law degree distribution with exponent of 3. This scaling exponent matched closely some observed graphs taken from the internet. Further extensions of this model are possible which include mixing adding new edges to old nodes, rewiring of old links and stochastic addition of new nodes [4]. Depending on the choice of the parameters, different behavior of the degree distributions can be obtained.

We now describe the basic model and generative process. We begin with an initial graph G_0 and will construct an undirected graph. In the simplest case, G_0 can be a graph of two nodes with one link connecting them. At each time step t , a new vertex is added and creates m new edges which connect “randomly” to the previous vertices such that each vertex is chosen with probability proportional to its current degree. From this, we see that there are $(t + 2)$ nodes at time t and $mt + 1$ edges.

It remains to define precisely what is meant by connecting “randomly” to the previous nodes. Unfortunately Barabási and Albert were not explicit in this regard. We adopt the convention introduced in [8] in which we start with the case $m = 1$, which is unambiguous and then “collapse” blocks of m vertices into one, preserving the edges as we go. This will result, of course, in some self-loops and multiple edges, giving a multi-graph in general. However, a rigorous analysis is substantially easier once such a convention has been adopted.

Barabási and Albert argue that

$$P(n_i(t)/t > x) \propto x^{-3},$$

where $n_i(t)$ is the number of vertices at time t that are of degree i (i.e., have i undirected edges).

Bollobás et al. (2001) [8] undertake a rigorous analysis of the preferential-attachment model and show that, at least for small degrees,

$$\frac{n_i(t)}{t} \xrightarrow{p} \frac{2m(m+1)}{(i+m)(i+m+1)(i+m+2)}.$$

This result is proved by first showing that the right-hand side is $\mathbb{E} \frac{n_i(t)}{t}$ and then using a martingale concentration inequality to obtain the convergence in probability. We use a similar approach in the appendix for the proof of our Theorem 5.2 and so we omit the details here.

Many other authors have considered similar and more general models and have used martingales for their analysis. Good examples of such work can be found in [39, 12, 20]. Additional connections to stochastic processes arising in ecological and evolutionary networks can be found in [21].

As previously mentioned, the preferential attachment model was originally proposed as a means for explaining how web networks arise “organically”. Recent interesting and fascinating work has suggested that some of the empirical observations may be due as much to sampling bias as actual dynamics. Lakhina, et al. [40] observe that traceroute sampling of a random network with Poisson-distributed edges—like those of a classical Erdős-Rényi graph [25, 26]—results in an interesting artifact: power-law

degree distributions. This, and more, is proved in [1].

1.1.6 Bivariate power laws

Compared to the vast literature on univariate power laws, and in particular, univariate statistics of various types of networks, the literature on bivariate power laws is quite sparse. Indeed, at the beginning of the project that led to this thesis, a Google search for “bivariate Zipf” turned up exactly zero links. (My advisor, Art Owen, has rectified this situation by surreptitiously slipping the phrase into one of his webpages.)

Within the last ten years, some work has begun at trying to model and understand graphs that have a more bipartite or directed structure. Interesting questions arise regarding the dependence between the entities when they take on more than one type or when the edges that connect them have directionality.

Most of the current work up to this point has been focused on simple statistical summaries that attempt to capture some aspect of the dependence between entities. These have mostly taken the form of some estimate of correlation, albeit many times rechristened. Various models have also been proposed to explain the observed behavior; many of them are adaptations of the basic preferential-attachment framework.

Earlier work on assortative and disassortative mixing by degree [55] focused on a correlation coefficient. Social networks were predominantly positively assortative by degree while diverse kinds of biological network were negatively assortative. Positive association among highly connected elements is also called the rich-club ordering [11].

Zlatic et al. [71] examine different language versions of Wikipedia and show empirically that they have similar degree distributions and reciprocity. Since Wikipedia

links are directed, reciprocity seeks to capture an estimate of the proportion of “bidirectional” links, i.e., the number of times that pages mutually point to one another. They use the statistic

$$\hat{r} = \frac{L_{\text{bd}}/L - \bar{L}}{1 - \bar{L}},$$

where L is the total number of links, $\bar{L} = L/N(N-1)$ is the proportion of links to the total possible number of links and L_{bd} is the number of bidirectional links, where each bidirectional link is double-counted. This statistic originates in [28] where it is shown that it is simply Pearson’s product-moment correlation where the “observations” are taken to be indicator random variables for all of the potential edges. Similar work in this vein can be found in [52], though the application is different. Liu et al. [43] consider a collaborative-filtering algorithm for recommender networks which incorporates such correlation measures.

Krapivsky, et al. [38] consider a model for directed graphs which has some similarities to the bipartite preferential attachment scheme that we present in Chapter 5. In that work, at each time step, with probability p , a new node is created and a link is generated to an existing node according to preferential attachment by in-degree. With probability q , two old nodes are connected, where the originating node is chosen by preferential attachment on the out-degree and the destination node is chosen by preferential attachment on the in-degree. The preferential attachment scheme is generalized by allowing for the addition of a constant to the degree. The constants are different for in-degree and out-degree preferential attachments and these constants allow for more flexibility in the possible resulting power laws.

The model differs from the one we present in Chapter 5 in a couple of subtle ways. First, our focus is on bipartite networks in which we have two distinguishable *types*

of entities. The aforementioned model is for directed graphs over one type of entity. Second, although our most basic model does not allow for new links between two existing nodes, augmenting the model with this capability is quite straightforward and we discuss it briefly in Chapter 5. We have elected not to do this simply to keep the analysis clean and clear. The Krapivsky model can be viewed as a slight modification of ours if we instead add a node to *both* sides of our bipartite graph at the same time and then only create one new link, always originating from the same partition of the graph, say the “left” partition. All links between old nodes also originate only from the “left” partition and terminate in the “right” partition.

Part of the focus of this thesis is on visualization techniques for examining the dependence between two types of categorical entities with a huge number of levels. Beyond two-dimensional scatterplots that seek to illuminate various correlations, it seems minimal work in this vein has been done. Newman [55] gives an example two-dimensional heatmap-like plot of an observed adjacency matrix. From this, the skew of the degree distributions is apparent, but, perhaps, not too much else.

Maslov and Sneppen [47] examine protein and gene interaction networks for one particular species of yeast. Compared to the examples in this thesis, the networks are quite small. However, they provide a display that is similar in nature to the one we propose here. Specifically they show “correlations” between in-degree and out-degree in the networks as a two-dimensional heatmap. Their “correlation” can also be interpreted as a certain likelihood ratio for the observed graph relative to a “null model” graph obtained by rewiring the edges. This “null model” is estimated by rewiring the graph, which is computationally much more expensive than our approach and introduces unnecessary additional sampling variation. Furthermore, each of the

margins of the display is logarithmically binned, according to the authors, “to improve the statistics”. If each of the margins were exactly power laws, this binning would have the same effect as the nonlinear transformation we use, but is different when the margins are not exact power laws.

We depart from [47] by assuming a different “null model” against which to compare that does not require large Monte Carlo simulations to estimate. This leads to a connection with copulas, which we discuss in Chapter 2, and an interpretation of the display as a likelihood ratio with respect to independence. Furthermore, we apply a nonlinear transformation on the margins so as to make them uniform, regardless of their original distribution. As previously mentioned, if the marginal distributions are both exact power laws, this would be essentially equivalent to logarithmic binning. However, in none of the examples we consider do we observe *perfect* power laws and in some cases we see fairly drastic deviations. We also prefer gray-level plots as one can interpret the darkness of any part of the display as being proportional to the density of the number of counts in the corresponding cell. While color heatmaps may be more eye-catching, many times it is difficult to properly interpret the meaning of the colors and the strong nonlinearity of the human eye in interpreting color can cause problems. Furthermore, because our display has uniform margins, each cell of the graph has an obvious interpretation in terms of the joint probabilities of observing a pair of entities within given quantiles of their corresponding degree distributions.

1.2 Background on copulas

The display presented in Section 2.2 can be interpreted as a copula density with respect to a discrete measure on $[0, 1]^2$. In order to better understand the interpretation

of the display, some facts about copulas are useful, which we review here. Different classes of copulas are discussed in Chapters 3 and 4.

The study of copulas goes back at least to the 1950's. A good introductory article is [30]. Useful introductory texts are [34, 53]. We will follow [53, Ch. 2] closely in the rest of this section. Deheuvels [15, 16] discusses empirical copulas and their relation to sample statistics for various dependence measures. In [16], a nonparametric test of independence based on the empirical copula is also proposed.

Copulas have found their greatest success (and some would say, failure) in applications in the world of financial modeling. Durrleman, et al. [22] is one of the first to take up the question of finding the “right” copula to use in those applications. Several parametric and nonparametric alternatives are considered there. The merits of copula modeling have recently become a topic of debate, from both philosophical and mathematical points of view. A mostly positive view is provided in [24], while a rather negative one is available in [50]. The latter includes several discussion responses (most taking a much more positive view than the original author) and a rejoinder.

1.2.1 Definition and first properties

We focus on the bivariate case for simplicity; the extensions to higher-dimensional cases are apparent.

Definition 1. A bivariate copula is a function $C : \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfying

1. $C(u, v) \geq 0$ for all $u, v \in [0, 1]$,
2. $C(0, t) = C(t, 0) = 0$ for all $t \in [0, 1]$,

3. $C(1, t) = C(t, 1) = t$ for all $t \in [0, 1]$, and
4. $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$ for all $0 \leq u_1 \leq u_2 \leq 1$ and $0 \leq v_1 \leq v_2 \leq 1$.

In words, C is a bivariate distribution function on $[0, 1]^2$ with uniform marginal distributions.

To understand the behavior of a class of functions, it is often useful to obtain bounds and simple properties. First note that $C(1, t) = C(t, 1) = t$ implies that both margins are uniform on $[0, 1]$.

Theorem 1.1 (Fréchet–Hoeffding bounds). *Let C be a copula. Then, for each $(u, v) \in [0, 1]^2$,*

$$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v) .$$

Proof. By Definition 1, we have that

$$C(u, 1) - C(u, v) - C(0, 1) + C(0, v) \geq 0 ,$$

and since $C(0, 1) = C(0, v) = 0$ and $C(u, 1) = u$, we conclude $C(u, v) \leq u$. Likewise, $C(u, v) \leq v$, and so the upper bound is established.

Similarly,

$$C(1, 1) - C(1, v) - C(u, 1) + C(u, v) \geq 0 ,$$

whence, $C(u, v) \geq u + v - 1$. Since, also, $C(u, v) \geq 0$, the lower-bound is established. □

Corollary 1.1. *The Fréchet–Hoeffding bounds are copulas.*

Proof. Check the conditions of Definition 1. For the upper bound, clearly, $\min(u, v) \geq 0$, $\min(u, 1) = u$, $\min(1, v) = v$ and

$$\min(u_2, v_2) - \min(u_1, v_2) - \min(u_2, v_1) + \min(u_1, v_1) \geq 0,$$

for all $0 \leq u_1 \leq u_2 \leq 1$ and $0 \leq v_1 \leq v_2 \leq 1$. The proof for the lower bound is similar. \square

1.2.2 Sklar's theorem and its consequences

The next result will be useful in establishing that for every bivariate distribution function, at least one corresponding copula exists. This result is known as *Sklar's theorem* and provides a rather direct means for constructing valid copulas.

Theorem 1.2 (Uniform continuity). *For every $(u_1, v_1) \in [0, 1]^2$ and $(u_2, v_2) \in [0, 1]^2$,*

$$|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1|.$$

Proof. If $u_1 \leq u_2$ then an argument similar to that of Theorem 1.1 shows that

$$C(u_2, v) - C(u_1, v) \leq u_2 - u_1.$$

for each $v \in [0, 1]^2$. Likewise, if $u_1 \geq u_2$, then

$$C(u_1, v) - C(u_2, v) \leq u_1 - u_2,$$

and these two together imply

$$|C(u_2, v) - C(u_1, v)| \leq |u_2 - u_1|.$$

Similarly, $|C(u, v_2) - C(u, v_1)| \leq |v_2 - v_1|$ for all $u, v_1, v_2 \in [0, 1]$. So, by the triangle inequality, we have

$$\begin{aligned} |C(u_2, v_2) - C(u_1, v_1)| &= |C(u_2, v_2) - C(u_1, v_2) + C(u_1, v_2) - C(u_1, v_1)| \\ &\leq |C(u_2, v_2) - C(u_1, v_2)| + |C(u_1, v_2) - C(u_1, v_1)| \\ &\leq |u_2 - u_1| + |v_2 - v_1|. \end{aligned}$$

□

Remark 1.1. It should be clear that a version of Theorem 1.2 exists for general bivariate distribution functions. If H is a bivariate distribution function defined on $\overline{\mathbb{R}} \times \overline{\mathbb{R}}$ —where $\overline{\mathbb{R}}$ denotes the extended real line—and has margins F and G , then

$$|H(x_2, y_2) - H(x_1, y_1)| \leq |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)|.$$

The following definition will come in handy for establishing some of the subsequent results.

Definition 2. Let C' satisfy all of the conditions of a copula, but restricted to a subset $S = S_1 \times S_2 \subset [0, 1]^2$. Then, C' is a *subcopula*.

Lemma 1.1. *Let H be a joint distribution on $\overline{\mathbb{R}}^2$, with margins F and G . There exists a unique subcopula C' satisfying*

$$1. \text{ dom } C' = \text{range } F \times \text{range } G.$$

$$2. \forall x, y \in \overline{\mathbb{R}}, H(x, y) = C'(F(x), G(y)).$$

Proof. Let (x_1, y_1) and (x_2, y_2) be points in $\overline{\mathbb{R}}^2$. Then,

$$|H(x_2, y_2) - H(x_1, y_1)| \leq |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)|,$$

from which we conclude that $F(x_1) = F(x_2)$ and $G(y_1) = G(y_2)$ implies that $H(x_1, y_1) = H(x_2, y_2)$. Hence, $C' : \text{range } F \times \text{range } G \rightarrow [0, 1]$ defined by the ordered pairs

$$\{((F(x), G(y)), H(x, y)) : x, y \in \overline{\mathbb{R}}\}$$

is a well-defined function. That C' is a subcopula follows from the properties of H . \square

Lemma 1.2. *If C' is a subcopula, there exists a copula C such that $C(u, v) = C'(u, v)$ on $\text{dom } C'$. The copula C is not, in general, unique.*

Proof. One way to construct a copula that extends a subcopula onto $[0, 1]^2$ is via bilinear interpolation. The intuition is obvious, but the proof is tedious. See [53, pg. 19, Lemma 2.3.5] for details. \square

The following theorem leads to the canonical method for constructing copulas and will be important in Chapter 3.

Theorem 1.3 (Sklar's theorem). *Let H be a joint distribution function with margins F and G . Then, there exists a copula C such that for all $(x, y) \in \overline{\mathbb{R}}^2$, $H(x, y) = C(F(x), G(y))$. Conversely, if C is a copula and F and G are marginal distributions, then H as defined is a valid joint distribution function with margins F and G .*

Proof. Lemmas 1.1 and 1.2 shows that we can construct a copula C such that $H(x, y) = C(F(x), G(y))$. Given C , F , and G , it is straightforward to verify that H is a distribution function. Indeed

1. $\lim_{y \rightarrow -\infty} H(x, y) = \lim_{y \rightarrow -\infty} C(F(x), G(y)) = C(F(x), 0) = 0$. Likewise, for the first argument.
2. $\lim_{y \rightarrow \infty} H(x, y) = \lim_{y \rightarrow \infty} C(F(x), G(y)) = C(F(x), 1) = F(x)$.
3. For $(x_1, y_1), (x_2, y_2) \in \overline{\mathbb{R}}^2$ such that $x_1 \leq x_2$ and $y_2 \leq y_1$, there exist $u_1 = F(x_1) \leq F(x_2) = u_2$ and $v_1 = G(y_1) \leq G(y_2) = v_2$. Then, we have

$$\begin{aligned} & H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1) \\ &= C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \\ &\geq 0, \end{aligned}$$

where the last inequality follows from the fact that C is a copula. □

Corollary 1.2. *If F and G are both continuous, then the copula C in Theorem 1.3 is unique.*

Proof. If C' is the corresponding subcopula of Lemma 1.1, then $\mathbf{dom} C' = \mathbf{range} F \times \mathbf{range} G = [0, 1]^2$ and since the domain is all of the unit square, then C' must also be a copula. Hence $C = C'$ and the uniqueness of C follows. □

Definition 3 (Quantile function). If F is a (cumulative) distribution function, then

the *quantile* or *inverse cumulative distribution function* is defined as

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}.$$

Corollary 1.3. *Let H be a distribution function with margins F and G and let C' be the corresponding unique subcopula. Then, for all $(u, v) \in \mathbf{dom} C'$,*

$$C'(u, v) = H(F^{-1}(u), G^{-1}(v)).$$

Proof. Check the definitions. □

Corollary 1.3 shows how to construct (sub)copulas from bivariate distribution functions. If F and G are continuous, then the subcopula is unique and is actually a copula. In the event that F and G are not continuous, then we only obtain a subcopula, but we can use Lemma 1.2 to construct a full copula. In this thesis, we are primarily interested in the latter case.

1.2.3 Additional properties

Copulas have long been interesting because of their relation to concepts of dependence. Indeed, most of the focus on copulas have been in relation to distributions with continuous margins. Under this special case, several strong statements can be made connecting copulas and dependence.

Our first result shows that in many cases, ranks are all that matter for determining a copula.

Proposition 1.1. *Let X and Y be continuous random variables with copula C and*

suppose h_1 and h_2 are strictly increasing functions on **range** X and **range** Y , respectively. Then $(h_1(X), h_2(Y))$ also has copula C .

Proof. Suppose F and G are the distribution functions of X and Y , respectively. Let H_1 and H_2 be the distribution functions of $h_1(X)$ and $h_2(Y)$ and $C'(u, v)$ be the copula of $(h_1(X), h_2(Y))$. Then,

$$H_1(x) = \mathbb{P}(h_1(X) \leq x) = \mathbb{P}(X \leq h_1^{-1}(x)) = F(h_1^{-1}(x)),$$

and $H_2(y) = G(h_2^{-1}(y))$. Furthermore,

$$\begin{aligned} C'(H_1(x), H_2(y)) &= \mathbb{P}(h_1(X) \leq x, h_2(Y) \leq y) \\ &= \mathbb{P}(X \leq h_1^{-1}(x), Y \leq h_2^{-1}(y)) \\ &= C(F(h_1^{-1}(x)), G(h_2^{-1}(y))) \\ &= C(H_1(x), H_2(y)), \end{aligned}$$

where the first and last lines are applications of Sklar's theorem. Hence $C' = C$, as desired. \square

By Sklar's theorem, we know that if X and Y are continuous random variables, then the Fréchet–Hoeffding bounds (see Theorem 1.1) are achieved when $X = f(Y)$ a.s. for some strictly monotone function f . Conversely, even without the assumption of continuity, if the copula of X and Y attains one of the Fréchet–Hoeffding bounds, then there exists a function f such that $X = f(Y)$ a.s. and f is strictly monotone on **range** Y .

A similar statement can be made with regard to independence.

Definition 4. The *independence copula* is a function $\Pi : [0, 1]^2 \rightarrow [0, 1]$ defined by $\Pi(u, v) = uv$.

Lemma 1.3 (Independence copula). *Suppose that X and Y have copula $\Pi(u, v)$ for almost all $(u, v) \in [0, 1]^2$. Then, X and Y are independent.*

Proof. A straightforward application of Sklar's theorem asserts that

$$\mathbb{P}(X \leq x, Y \leq y) = \Pi(F(x), G(y)) = F(x)G(y),$$

and so X and Y are independent, as claimed. \square

Corollary 1.4. *If X and Y are continuous, independent random variables, then their corresponding copula is the independence copula.*

The connection between copulas and dependence concepts is very nice in the case of continuous random variables. Many of the various concordance measures can be written explicitly in terms of the copula of a pair of random variables. When dealing with inherently discrete data, which is the subject of this thesis, many of these properties break down if care is not taken.

First of all, *the* copula for a pair of discrete random variables X and Y no longer exists, i.e., many copulas satisfy Sklar's theorem. Furthermore, as shown in [32], this class of copulas can be quite large. However, a particular choice leads to recovery of most of the nice properties present in the continuous case. In particular, extending the unique *subcopula* corresponding to a discrete X and Y via bilinear interpolation yields the following properties:

1. The bilinearly interpolated copula C_{bl} is absolutely continuous.

2. X and Y are independent if and only if $C_{\text{bl}} = \Pi$.
3. If ρ is a concordance measure expressible in terms of a function of the copula, then $\rho(H) = \rho(C_{\text{bl}})$ where H is the joint distribution of (X, Y) .

We can interpret the copula C_{bl} as arising from a continuous approximation of the discrete distribution, in which we add independent and identically distributed uniform random variables to X and Y . In terms of the copula density (since C_{bl} is absolutely continuous), this amounts to spreading the mass uniformly in every square formed by the lattice points of the bivariate discrete distribution. This is (essentially) the construction we choose in the derivation of our copula estimates in Chapter 2.

The extension procedure is not perfect, though. For example, the Fréchet–Hoeffding bounds are never achieved and some dependence measures cannot attain their extremal values via this extension [54]. These limitations are of little concern to us for the purposes of this thesis, while the benefits of the construction are useful.

Chapter 2

The data display

2.1 Introduction

Large bivariate transposable datasets with marginal entities taking on a large number of categorical levels may have dependencies between the row entities and column entities. This chapter presents a simple display for these data that allows for visualizing the dependency in a meaningful and interpretable manner. The primary intent of our visualization is essentially exploratory in nature: We want to be able to see interpretable structure in the data with the hope that it will guide any further analyses that one may wish to conduct.

We present several examples to demonstrate the usefulness of the display and provide some potential interpretations of the dependencies we see. Some of the dependencies are, at least qualitatively, somewhat unsurprising to those with domain expertise while others would not be readily guessed without the aid of this visualization. In both cases, we have found that the patterns brought out by these displays tend to generate considerable discussion and directions for future exploration of the

underlying data.

The display can also be viewed as particular and inherent to a certain dataset or as an estimate of an underlying quantity taken from a sample represented by the dataset in question. In this latter viewpoint, it is natural to ask how accurate a depiction the data display provide of the underlying quantity of interest. In trying to answer this question, we focus on the case where the marginal distributions are assumed to follow a power law and the marginal sums give noisy estimates of these distributions. Under assumptions made precise in what follows, we demonstrate that the vast majority of the data display is accurate.

2.2 Forming the display

We suppose that our data are given as a matrix X_{ij} for $1 \leq i \leq n$ and $1 \leq j \leq m$. Most of the data sets we're interested in have $X_{ij} \geq 0$. As examples, X_{ij} could represent whether user i rated item j , the number of edges from node i to node j , the dollar value of transactions from purchaser i to vendor j and so on. We are interested in an estimate of the empirical copula “density”. In our case, our estimate can be viewed as an approximation of a function $c(u, v) = \frac{\partial C}{\partial Q}$ where C is an empirical copula and Q is a certain discrete measure chosen for the purposes of visualization.

Given the data matrix, we form the marginal sums $X_{i\bullet} = \sum_{j=1}^m X_{ij}$ and $X_{\bullet j} = \sum_{i=1}^n X_{ij}$. We assume that the row entities have been sorted so that $X_{1\bullet} \geq X_{2\bullet} \geq \dots \geq X_{n\bullet}$ and similarly $X_{\bullet j} \geq X_{\bullet j+1}$. It is convenient to refer to large and small entities, where the size of an entity is simply its marginal sum.

Our display is a gray level plot in the unit square $[0, 1]^2$. The row entities are on the horizontal axis arranged from smallest at 0 to largest at 1. The amount of

horizontal space given to row entity i is proportional to $X_{i\bullet}$. The column entities are similarly arranged on the vertical axis, again with smallest at 0, largest at 1 and space proportional to $X_{\bullet j}$. The unit square is split into rectangles shown in gray. The gray level of a rectangle is proportional to the sum of the observed X_{ij} values in it divided by the area of that rectangle.

With this convention a dark upper right corner means that the large row entities favor the large column entities more than they would under independence, while a light upper corner means that the heads of the two distributions avoid each other. Similar interpretations apply to the other three corners.

A simple toy example often helps clarify matters. Suppose we have bipartite data where there are three entities A , B , and C that purchase items from the set $\{I, II, III, IV\}$. Assume that A purchases $\{IV\}$, B purchases $\{III, IV, IV\}$ (i.e., item IV is purchased twice) and C purchases $\{I, II, II, III, III, IV\}$. This is a more general example than the majority of the datasets that we examine later in that we allow for multiple copies of the same dyad, e.g., (B, IV) is represented twice. Figure 2.1 shows the resulting visualization. The left pane shows the raw counts, with the entities along the margin receiving a width proportional to their marginal counts. The right pane shows the resulting two-dimensional visualization. To explain the values and colors, consider the bottom right patch corresponding to (C, I) . There is one count out of a total of ten in that patch. Because C accounts for six out of the ten marginal counts, it gets a width of 0.6. Because I has one count out of ten, it gets a height of 0.1. If there was no association between the two types of entities, we would expect a proportion $0.1 \cdot 0.6 = 0.06$ of the counts to appear in this patch. Instead we observe 0.1 and so the associated ratio is $0.1/0.06 \approx 1.67$. The greyscale

Discrepancy plot for miniature example

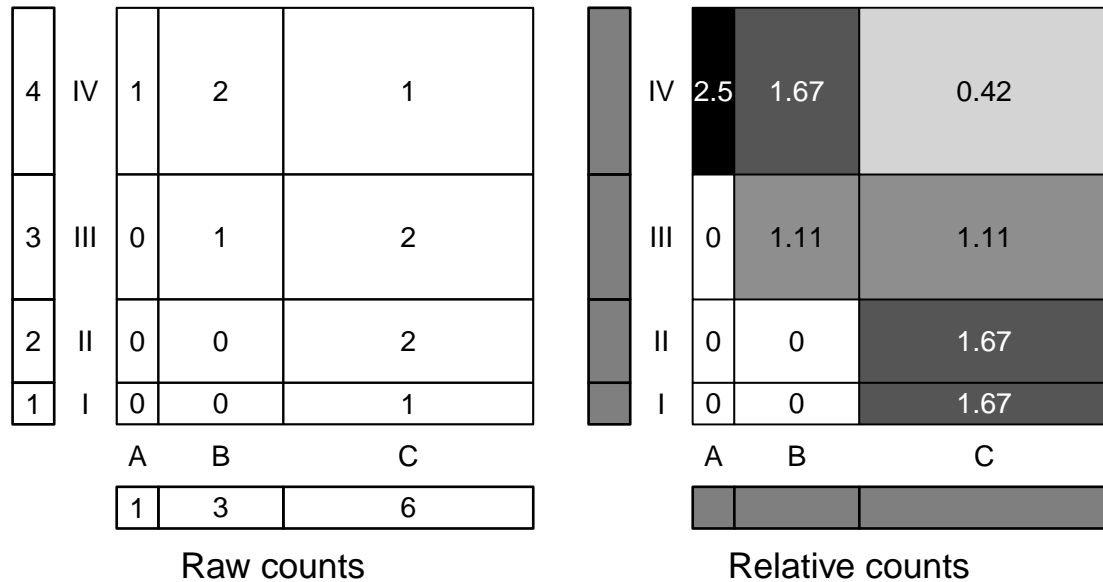


Figure 2.1: Toy example of data visualization. (Permission to use this figure was kindly provided by Art Owen.)

level for this patch is chosen linearly with respect to the range $[0, 2.5]$. Because a total of zero counts is the minimum possible, we always choose this as the lower bound of the range. We use the maximum observed ratio for the upper bound so as to make sure we use the entire color range.

In practical datasets there may be hundreds of thousands or even millions of distinct entities along each dimension. Thus, for the purposes of visualization, it is useful to aggregate many levels into bins so that the display can be reduced to a manageable number of cells. In our examples, we have aggregated so that there are 100 cells along each dimension. The discrete measure Q mentioned in the previous section is precisely that derived from choice of the number and placement of bins along each dimension.

In many cases, the data corresponding to a given entity will “straddle” the boundary of two (or more) bins. In this event, a proportional number of counts are allocated to each bin according to the relative percentage of the data on each side of the boundary. Additionally, sometimes there are many small entities with the same marginal total, such as two total links, that in aggregate comprise a large proportion of the image. We adopt the convention of assuming no additional ordering preference among the entities. Hence, in such cases, our plot aggregates all such data and distributes the counts evenly across the affected bins, with bins along the boundary of the affected area receiving partial counts. In effect, this makes the smallest number of assumptions about counts corresponding to entities with the same marginal sums and is similar to adopting a “uniform prior” or “maximum entropy” assumption.

Note that the conventions we have adopted always yield a copula estimate that satisfies the copula constraints of the previous section. Many estimation methods of both copulas and copula densities available in the literature do not necessarily satisfy the copula constraints.

2.2.1 Variations on the theme

Many possible variations of this basic display can be used. We will show one such variation in the examples that follow. The basic display assigns a pixel value linearly with respect to the number of counts corresponding to a particular bin. This is useful for seeing the general structure of the dependence and has a simple interpretation.

In many instances, particularly when there are very heavy extreme-value dependencies, only a few pixels will be extremely dark (nearly black) while the rest of the image will be very light, and so it can be difficult to ascertain the more minute

characteristics of the dependence. In this instance, one could consider nonlinear transformations of the counts in each bin (e.g., logged counts) to accentuate the patterns in the copula estimate.

Another approach which we've found very useful and which we show repeatedly in the examples that follow is to normalize the histogram of the display. Thus, each grey level is used the same number of times in the image. This amounts to a data-dependent nonlinear transformation of the data and has an easy interpretation: Each gray level corresponds to a uniformly-spaced estimated level set of the copula density. This particular transformation tends to make the dependence structures much more visible, providing additional insights into the data.

2.3 Examples

Here we consider several large example data sets, most of which are taken from internet resources. All but one of the data sets we examine are publicly available. We provide only a selection of datasets to illustrate certain features of the display and the interpretations we can draw from them.

The examples roughly fall into three classes: (a) bivariate data where the row and column entities are of different types, for example movies and customers, (b) network data in which the row and column entities are the same and directed links exist between entities and (c) network data with the same entities along the rows and columns, but undirected links between entities. It is easy to see that, in general, the display we get for the first two cases will be asymmetric and for the last case, is symmetric, by construction.

Table 2.1 provides a summary of each of the datasets. Given are the total number

Dataset	Edges	Row entities			Column entities		
		# entities	max. cnt	# uniq.	# entities	max. cnt	# uniq.
Netflix (users, movies)	100480507	480189	17653	2782	17770	232944	6275
Yahoo! (users, songs)	699640226	1823179	131523	9658	136736	323512	18763
IMDB (actors, movies)	1470417	383640	646	261	127822	294	147
Epinions (truster, trusted)	508836	60341	1801	326	51957	3035	421
Wikipedia (out, in)-degree	45030389	3465604	7061	1185	2488225	187342	3099
Snapfish (sharer, sharee)	24861827	2736598	8470	588	18126196	4738	133
arXiv hep-th (citer, cited)	352807	25059	562	155	23180	2414	280
Enron email addr. (sym)	367662	36692	1383	334	36692	1383	334
CA intersections (sym)	5533214	1965206	12	11	1965206	12	11

Table 2.1: Summary statistics for example datasets

of “edges” (i.e., total observations in the dataset), and for each margin, the number of observed entities, the maximum marginal count, and the number of unique marginal counts observed. All minimum counts were one except for the Netflix data (smallest number of ratings for a movie was three) and the Yahoo! data (smallest number of ratings by a user was 20, smallest number of ratings for a song was 929).

2.3.1 Netflix movie ratings

Our first example is of the Netflix movie ratings data. These data were made available by Netflix for the purposes of the Netflix Prize, a \$1,000,000 cash award to any team that could improve the root-mean-squared error performance of their Cinematch prediction system by at least 10% [7, 60]. The prize was won by combining over 700 predictions from different machine-learning models.

The data consist of (user, movie, rating) triplets in which user i rates movie j , giving the movie an integer rating from 1 (bad) to 5 (excellent). There are approximately 18,000 distinct movies and 480,000 distinct users in the dataset and a total of over 100 million ratings. We are interested in the data matrix composed of elements X_{ij} where $X_{ij} = 1$ if user i rates movie j . This matrix is extremely sparse

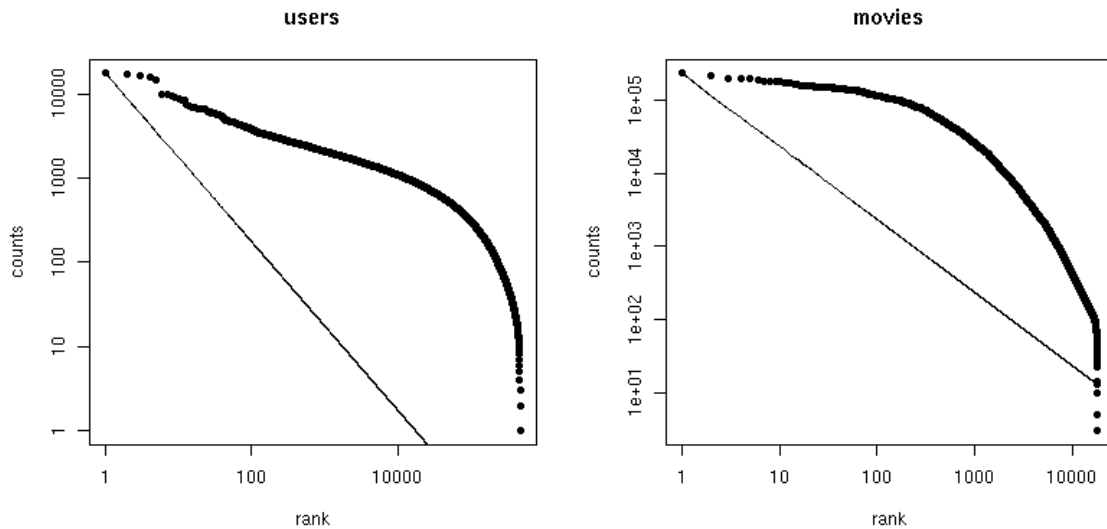


Figure 2.2: Zipf plot of Netflix margins.

$(10^8 / (1.8 \times 10^4 \cdot 4.8 \times 10^5) < 0.012)$.

Figure 2.2 shows a Zipf plot of the marginal number of ratings by user and movie, respectively. In this example, the margins are not very power-law like. This is apparent from the large degree of curvature of each of the plots. The reference line has a slope of negative one. Assuming an infinite number of levels and a power-law distribution for the margins, the Zipf plot would need to fall strictly under the reference line in order to be summable. In practice, however, we will observe only a finite sample and the margins may not be perfectly (or even weakly) power-law and so the Zipf plot will often lie above the reference line.

Figure 2.3 depicts two versions of our data display. The left pane shows the standard display, with a linear scale for the ratings counts. The right pane shows the histogram-equalized version. There are strong head-to-tail affinities concentrated among the most extreme entities. The one connecting large users to small movies is stronger than the converse. The asymmetry of the copula estimate is particularly

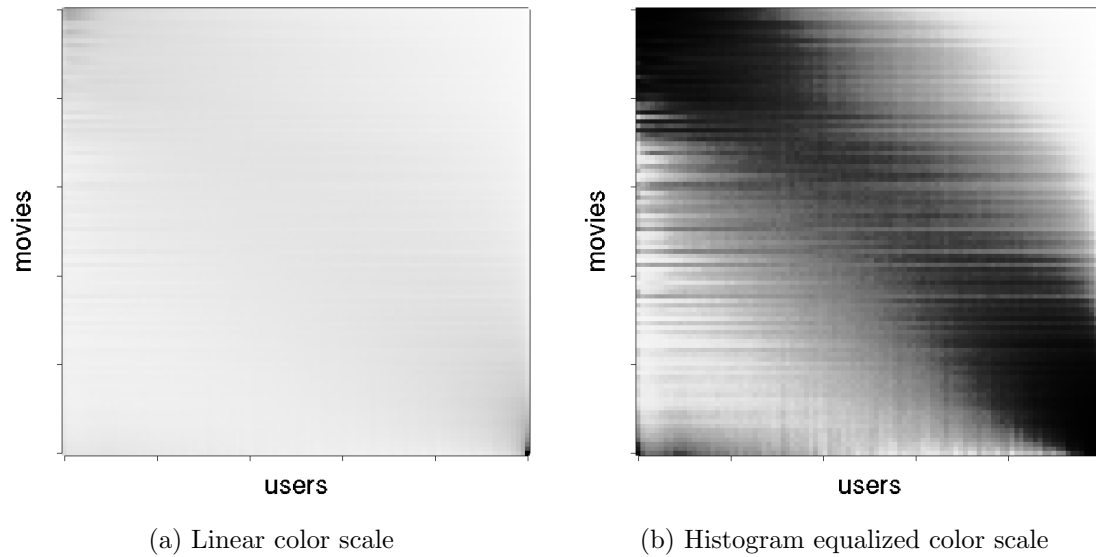


Figure 2.3: Data display for Netflix example.

evident from the right pane of the figure.

It would be reasonable to weight the user ratings. We investigated two extreme weightings, one that just counts five-star ratings and one that only counts one-star ratings. Those plots (data not shown) both had strong affinities between inactive users and frequently rated movies, but much less affinity between busy users and rarely rated movies.

2.3.2 Yahoo! music ratings

Yahoo! recently released a large set of music ratings from their online music rating service via their WebScope project [67]. The data is very similar to the Netflix movie ratings, except at an even larger scale. In this case, approximately 1.8 million users rated over 136 thousand unique songs. The total number of ratings is about 700 million. Again, we (conceptually) form a data matrix where $X_{ij} = 1$ if user i rates

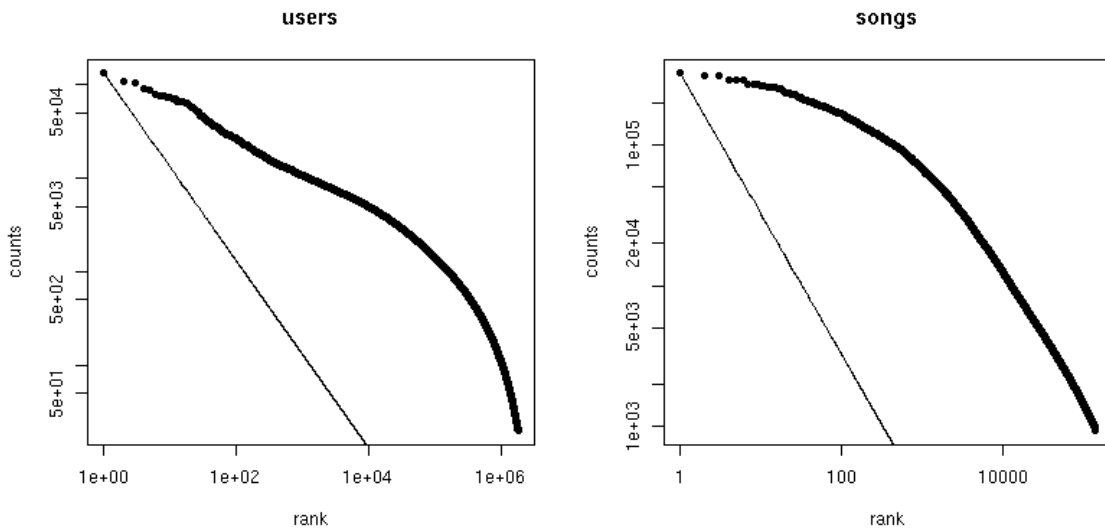


Figure 2.4: Zipf plot of Yahoo! music ratings margins.

song j and, otherwise, $X_{ij} = 0$.

Figure 2.4 depicts the Zipf plots of the margins. In this case we see an approximate power-law in the head of the distribution for the users, with a thin tail. For the songs, we observe a flatter-than-power-law head, transitioning into a power-law tail with a scaling coefficient of approximately one.

Figure 2.5 shows quite similar patterns to those observed in the Netflix case. There are strong affinities between inactive users and popular songs as well as very active users and less popular songs. The overall shape of the copula is qualitatively similar. By comparing the linear-color versions of the Yahoo! and Netflix displays side-by-side, we can see that the affinities in the corners are not as strong in the Yahoo! case as in the Netflix data, indicating a greater proportion of raters and items are involved in the former dataset. This is apparent from the overall darker average greyscale value of the Yahoo! display. In both data sets, the large users avoid the large items

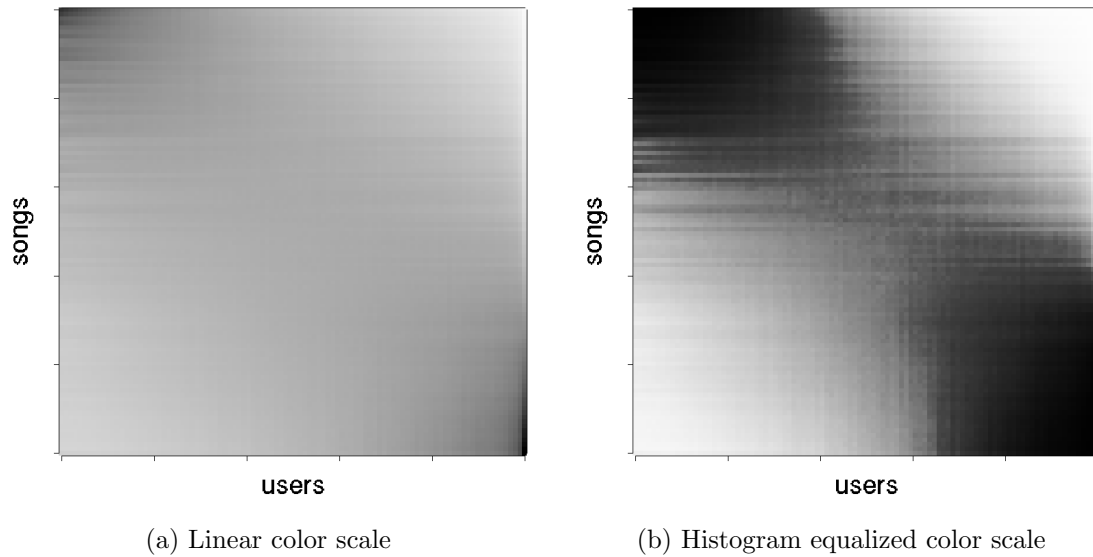


Figure 2.5: Data display for Yahoo! music ratings example.

more strongly than the small users avoid the small items.

The similarities in these first two examples perhaps lead one to hope for the discovery of some underlying explanation for the patterns in the data. It may also give one (somewhat false) hope of describing the data via parametric copula models. As we will see in later examples, the structure of some data sets is so peculiar that no “reasonable” parametric model is likely to capture the observed features.

2.3.3 IMDB movie-actor database

The Internet Movie Database (IMDB) [13] contains casting information for a very large collection of movies. A bivariate transposable dataset can be formed from this database by considering actors as separate entities and the movies that they are cast in as another set of entities. Then $X_{ij} = 1$ if actor i is cast in movie j and $X_{ij} = 0$ otherwise.

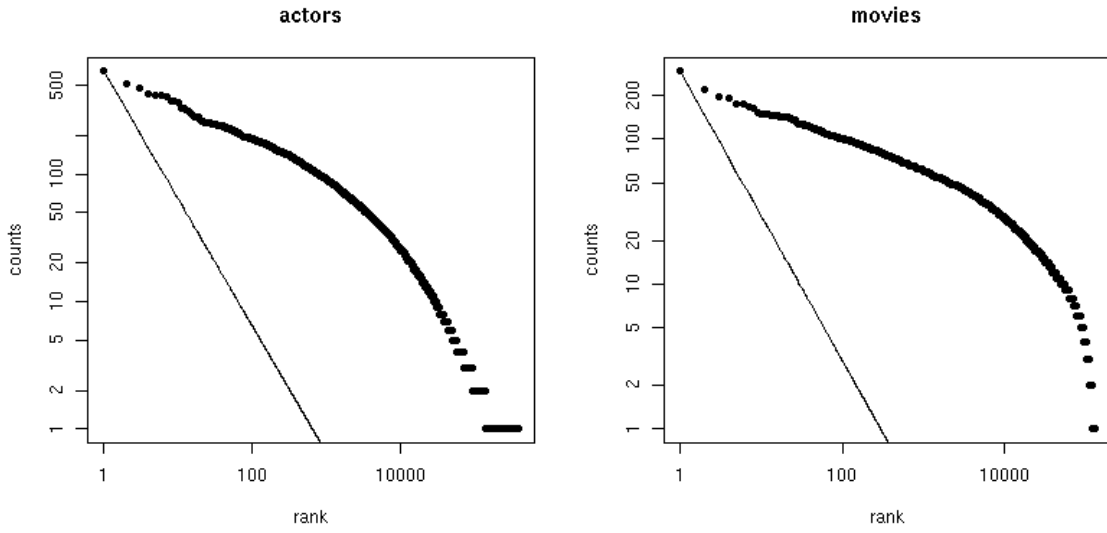
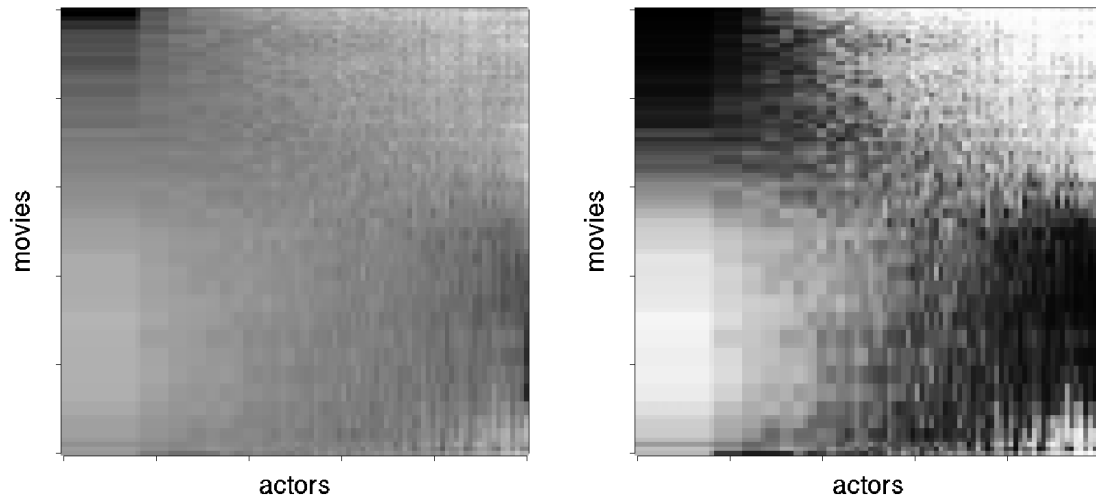


Figure 2.6: Zipf plot of IMDB margins.



(a) Linear color scale

(b) Histogram equalized color scale

Figure 2.7: Data display for IMDB example.

Figure 2.6 shows power-law like behavior for the head of both the actor and movie margins. The scaling coefficient is about $1/2$, and the tail appears thinner than a power law in the case of movies. For actors, it appears as though something approximating a second power-law with a larger coefficient takes over for actors who've been in fewer than 30 movies or so.

Approximately 16% of actors in the database have been in only one movie. Thus, in the data display in Figure 2.7, the first sixteen bins along each row are given the same number of counts according to the uniformity convention adopted in Section 2.2. However, each movie has a different distribution of actors in terms of the number of movies they've been cast in. Hence, though along each row we distribute counts evenly across the first 16 bins, a color gradient exists down the column.

We further note that actors who worked rarely (e.g. only once) are overrepresented in movies with the largest casts. On the other hand, the busiest actors were overrepresented in movies with small casts, but not the very smallest casts.

2.3.4 Epinions

The Epinions data provides the first example where the entities along the two margins are nominally the same. Epinions allows users to read and write reviews on a large array of products. In addition, users can choose to “trust” other users based on the perceived quality of their reviews. These “trust” relationships are also viewable. For the dataset we consider, $X_{ij} = 1$ if user i trusts user j . We can treat X as the adjacency matrix of a directed graph of the users. (Note that the trust relationships are not necessarily reciprocal.) There are about 60,000 users who've elected to trust someone and about 51,000 users that are trusted. About half a million such

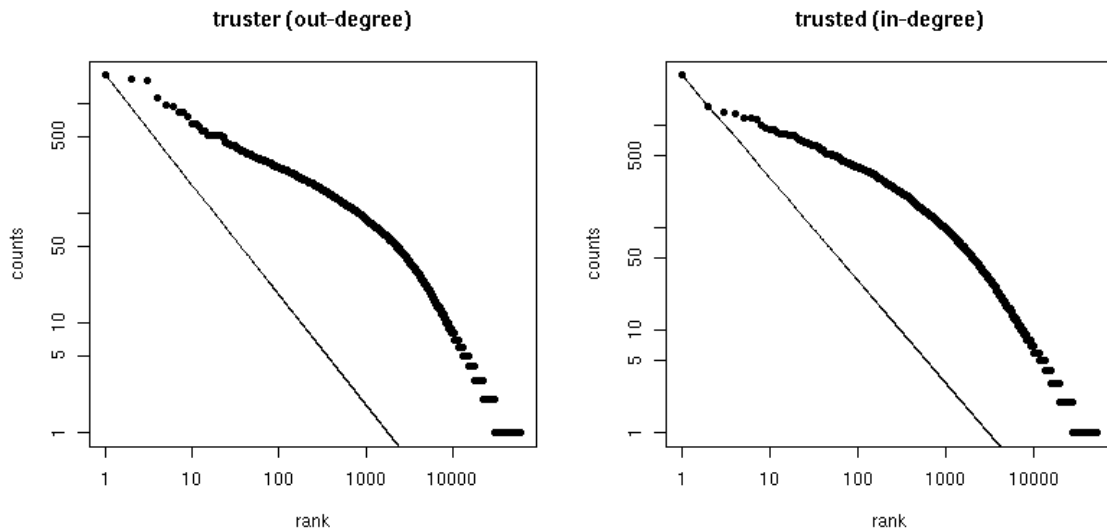


Figure 2.8: Zipf plot of Epinions margins.

relationships exist.

From Figure 2.8, we see a broken-line power-law relationship for both truster and trusted users, with a slope of less than unity at the head and greater than unity in the tail. If one looks at a degree-distribution plot (Figure 2.9) instead of a Zipf plot, the power-law distribution for small degrees (i.e., for those that trust only a few people or that are trusted by only a few) is even more apparent.

Figure 2.10 shows the copula estimates for the Epinions data. We see two dark clusters at the lower left and right. The least trusted users are trusted by two distinct types of user. The first type trusts a great many other users. The second type trusts other seldom trusted users.

So, some users appear to be very indiscriminate about who they trust. It would be interesting to investigate what the second type of user corresponds to. These could be users who are introduced to the service, rate a few products, trust a couple of

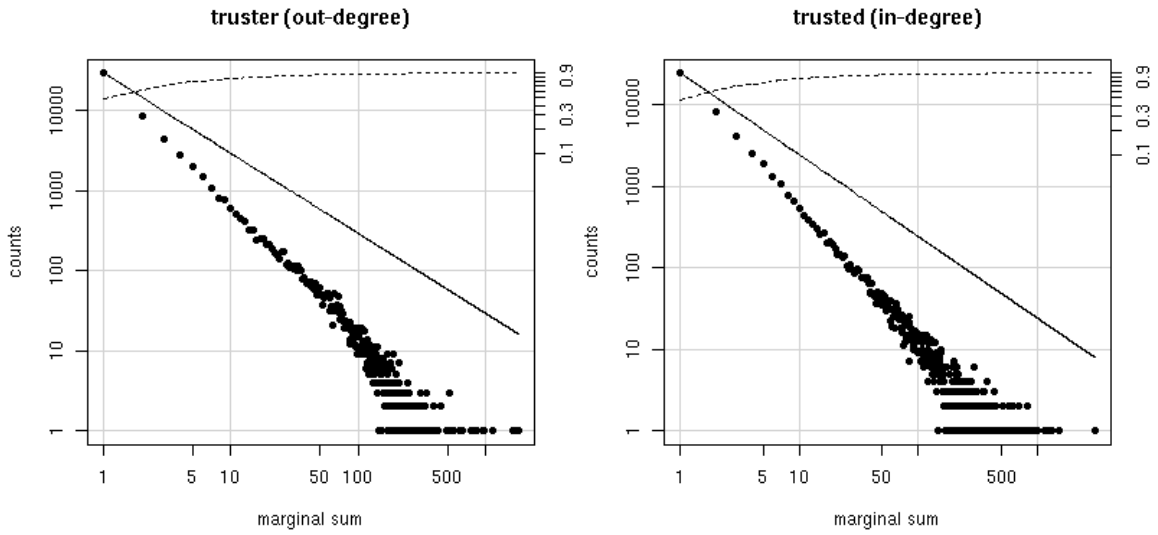


Figure 2.9: Degree distribution plot of Epinions margins.

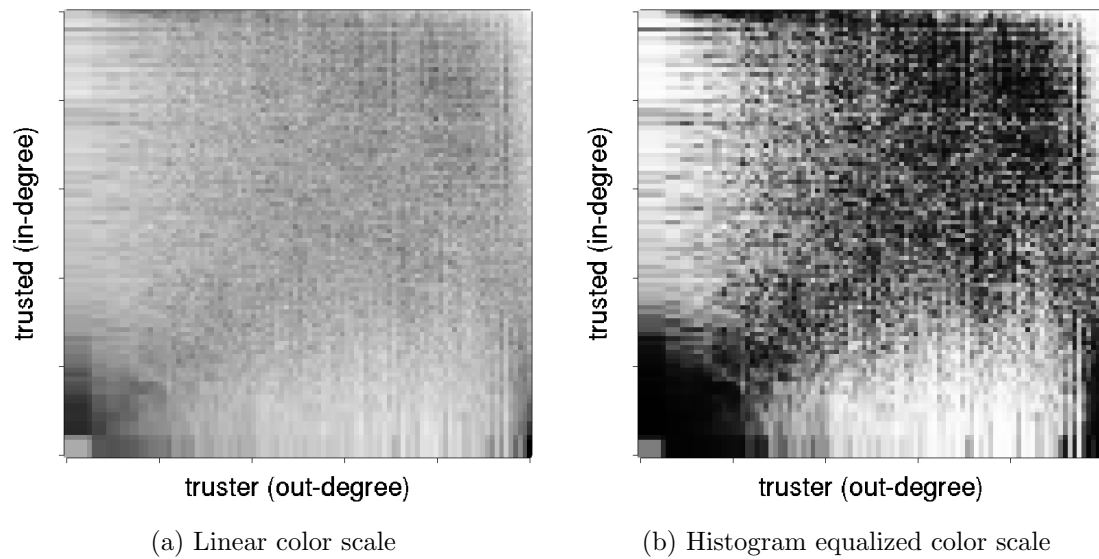


Figure 2.10: Data display for Epinions example.

people and then disappear. Or they might be people that know each other personally and decide to trust them even though few other people on the site do.

Thus, it would be interesting to delve further by looking at the number of people a person trusts in relation to the number of items they've rated and the number of items rated by the people they trust. One might also want to look at reciprocity: If someone decides to trust me, how likely am I to follow suit and trust that person back? We leave these questions for further study.

2.3.5 Wikipedia

The Wikipedia dataset is a snapshot of the directed graph of Wikipedia pages taken in 2007 by David Gleich. It contains over 3.4 million pages and 45 million links between pages. Some pages have as few as one out-link and others have as many as 7000. Some pages are only linked to from one other page and others receive links from 100s of thousands of different pages.

Interesting questions include: Do pages with lots of links tend to link to other popular pages or more obscure ones? Do pages with a single link point to the most popular pages? Do the degree distributions look anything like what one would expect from a preferential-attachment or other generative graph model?

Figure 2.11 suggests that the marginal distributions are approximately power law. The data display of Figure 2.12 is quite unique. We might expect to see hubs and authorities [36] in this graph. There is a cluster of topics with large out-degree and small in-degree. It included many lists, as we might expect for hubs. One striking mode represents pages with a medium-small number of out links and a high, but not maximal, number of in links. Upon inspection, this hotspot included many topics

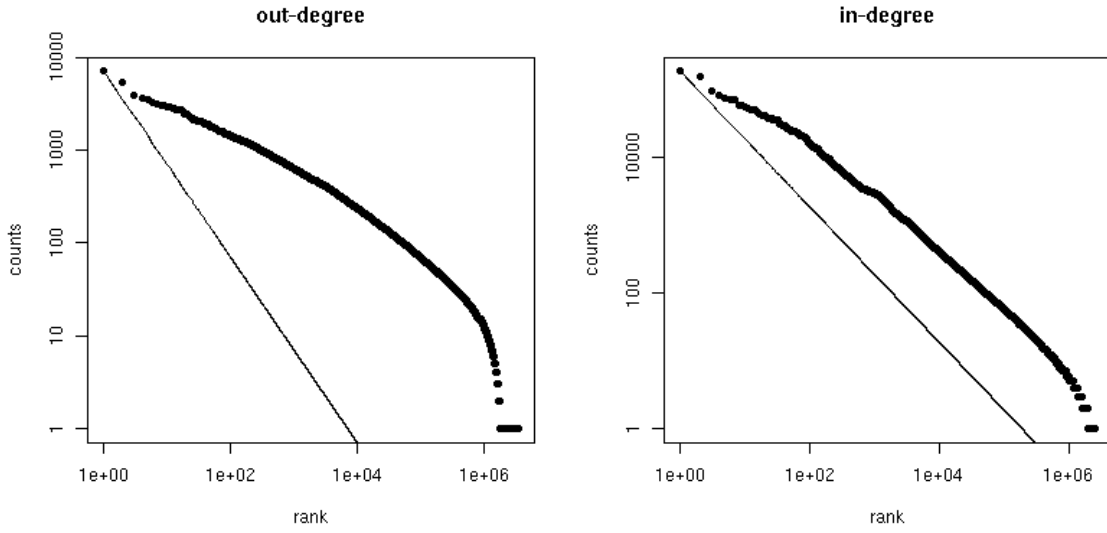
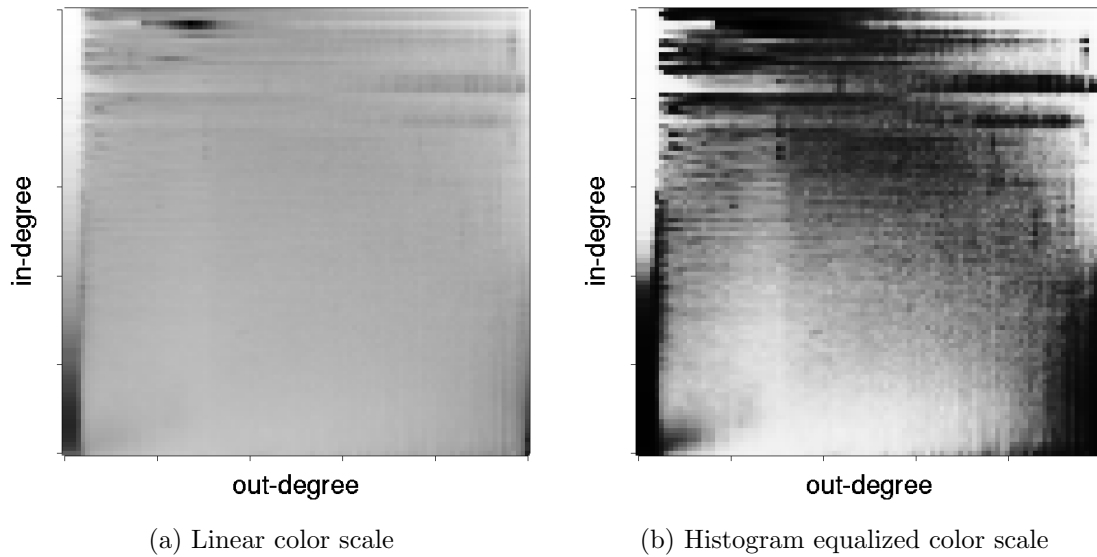


Figure 2.11: Zipf plot of Wikipedia margins.



(a) Linear color scale

(b) Histogram equalized color scale

Figure 2.12: Data display for Wikipedia example.

that are either years or locations—in particular, countries. This is roughly what we might expect for authorities but they did not quite land in the upper left hand corner where one might have expected. Some pages have both small in-degree and small out-degree. Many of those are stubs. It is exceedingly rare for a topic to have both low out-degree and high in-degree.

Though the data display does not specifically seek to illuminate properties such as “clusters” in the graph, it does provide insight into the data that we would not get from other methods. This example also hints at the inadequacy of purely parametric models for fitting copulas to this particular type of data. The hotspot near the upper-left corner of Figure 2.12 would simply not be fit by any of the common parametric copulas.

2.3.6 Snapfish

The Snapfish photo web service allows users to upload digital photos and purchase various objects (e.g., prints, coffee mugs, etc.) imprinted with the photos. It also allows users to organize their photos into albums and share them with other users. In order to view a shared album, one must be a Snapfish user.

The Snapfish sharing graph is a directed graph in which $X_{ij} = 1$ if user i has shared at least one album with user j . This example is different from the previous directed-graph examples in that both user i and user j must take some positive action in order for a link to be created. In this case, j must explicitly accept the sharing request in order to generate an edge in the graph. There are over two million sharers in the graph and over 18 million sharees (i.e., people who’ve accepted a sharing request). Almost 25 million albums had been shared when the data was collected.

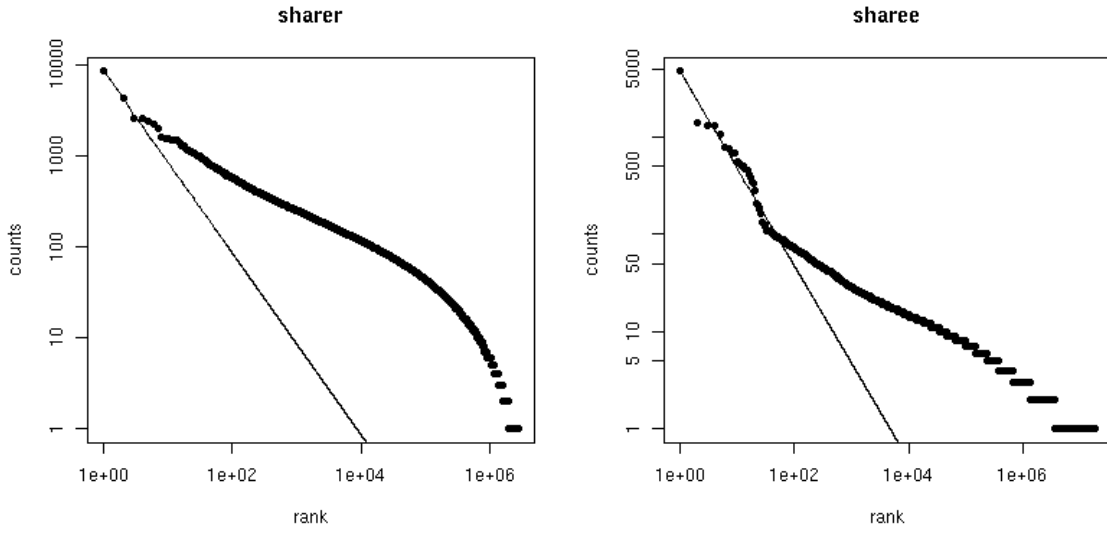
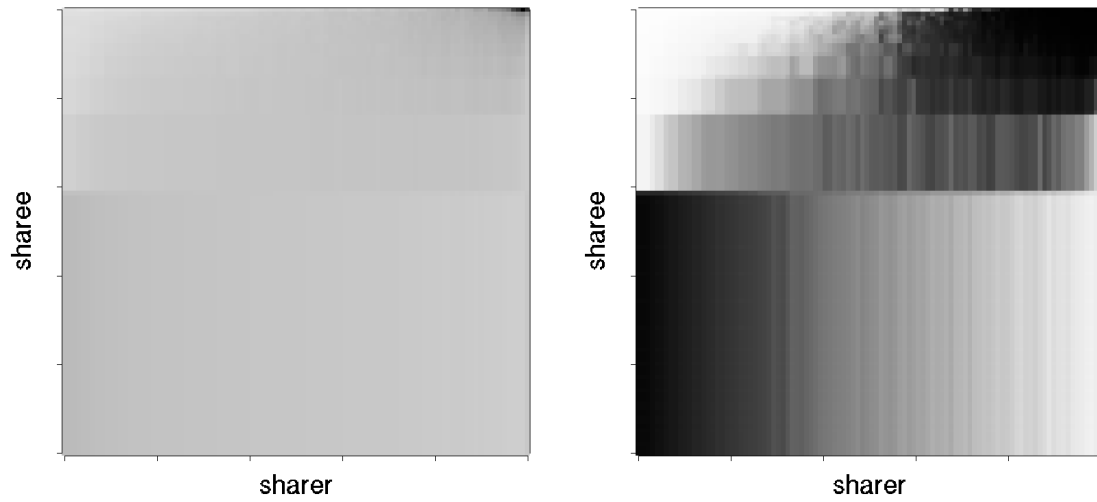


Figure 2.13: Zipf plot of Snapfish margins.



(a) Linear color scale

(b) Histogram equalized color scale

Figure 2.14: Data display for Snapfish example.

Many aspects of this dataset are unique. The Zipf plots of Figure 2.13, while showing some power-law behavior, are surprisingly convex. In virtually every other example, deviations from a power-law in the Zipf plot normally manifest themselves in a concave form. Figure 2.14 shows several interesting features. First, about 60% of sharees have accepted only a single sharing request. Indeed, this is one of the primary ways that Snapfish attracts new users: Snapfish users can send email to share albums with non-Snapfish users. But, in order to actually view the album, the latter must register with Snapfish, at which time they automatically accept the sharing request. It is possible that many of these users quickly become inactive. As indicated by the somewhat darker lower left corner of the plot, inactive sharers tend to share with those that also accept very few sharing requests.

Also, there is a very strong, asymmetric, head-to-head affinity in the graph, which has not been seen in other directed-graph examples. Upon examination, it was noted that there are a few incredibly active sharers and sharees. One sharer has shared successfully with over 8000 different sharees and there are a few sharees that have accepted sharing requests from thousands of other users. In one case, a sharer managed to share close to 100 thousand unique albums. These very few active sharers likely account for the strong head-to-head affinity and may also account for the convexity of the marginal distributions. Some evidence suggests that these very active users may actually be automated. Once these anomalous users are excluded from the data, we see a shift in the slope of the marginal distributions.

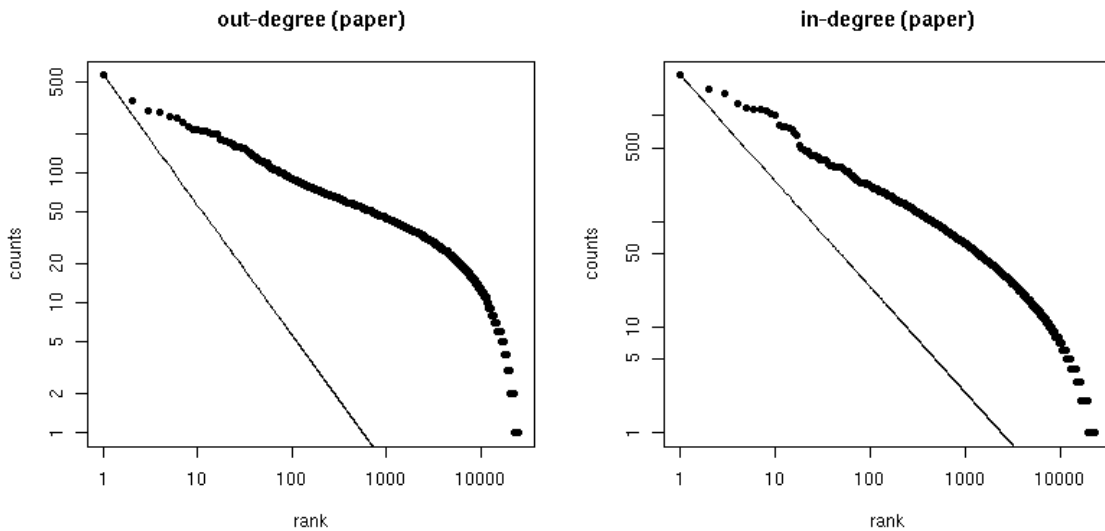


Figure 2.15: Zipf plot of arXiv hep-th margins.

2.3.7 arXiv hep-th citations

The arXiv theoretical high-energy physics (hep-th) citation network consists of hep-th papers as entities with directed links from i to j if paper i cites paper j . Both papers must be submitted to the arXiv in order to be counted. There are over 25 thousand papers in the dataset and over 350 thousand citations.

The degree distributions of citation networks have received considerable interest over the last few years. Such work has also included efforts to cluster papers and/or authors together based on characteristics of the directed graph.

Figure 2.15 shows some approximate power-law behavior in the head of the marginal distributions. One might guess that there would be an affinity between papers with few arXiv links and those with many arXiv links, e.g., survey papers or extremely popular ones that are relevant to a large part of the field. Figure 2.16 indicates that this appears to not be the case. Indeed, the strongest affinity is between papers with

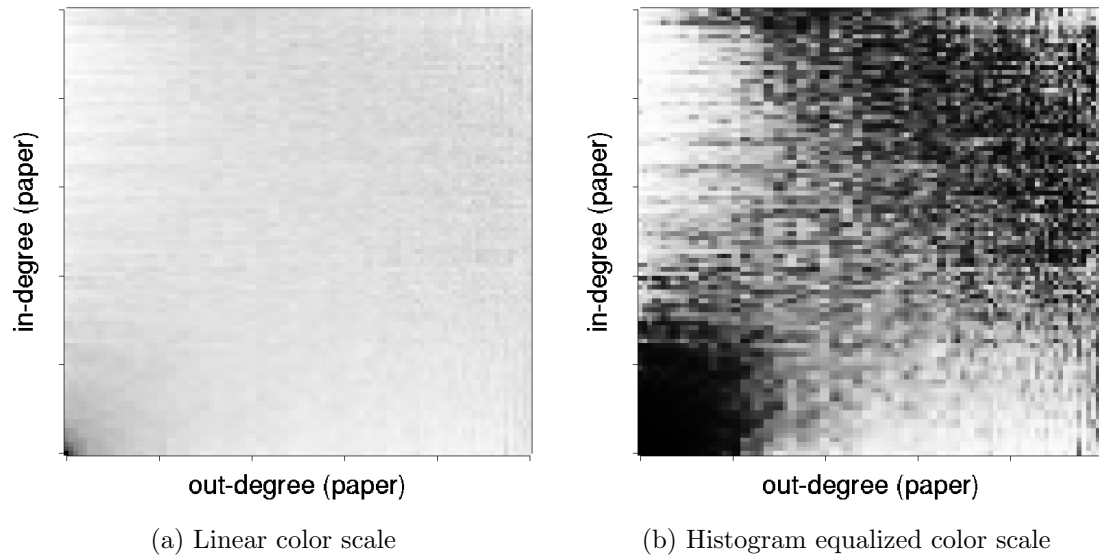


Figure 2.16: Data display for arXiv hep-th example.

a very small number of citations. Only citations within hep-th are counted so the affinity at the low end is not just among papers with few total citations or references. Instead of a strong head-to-tail affinity, we instead see a broad head-to-head affinity, where relatively popular papers cite other well-cited ones. The copula is also quite symmetric, though not perfectly so.

2.3.8 Enron email network

The Enron network consists of emails sent to and from Enron employees in the events surrounding the lead-up to the bankruptcy of the company. Among emails between Enron employees, we observe the full set of emails, but the dataset does not include the full set of emails of people outside of Enron that corresponded with Enron employees. Because of how the dataset was processed, the graph is undirected. So, $X_{ij} = X_{ji} = 1$ if and only if user i sent user j an email *or* vice versa.

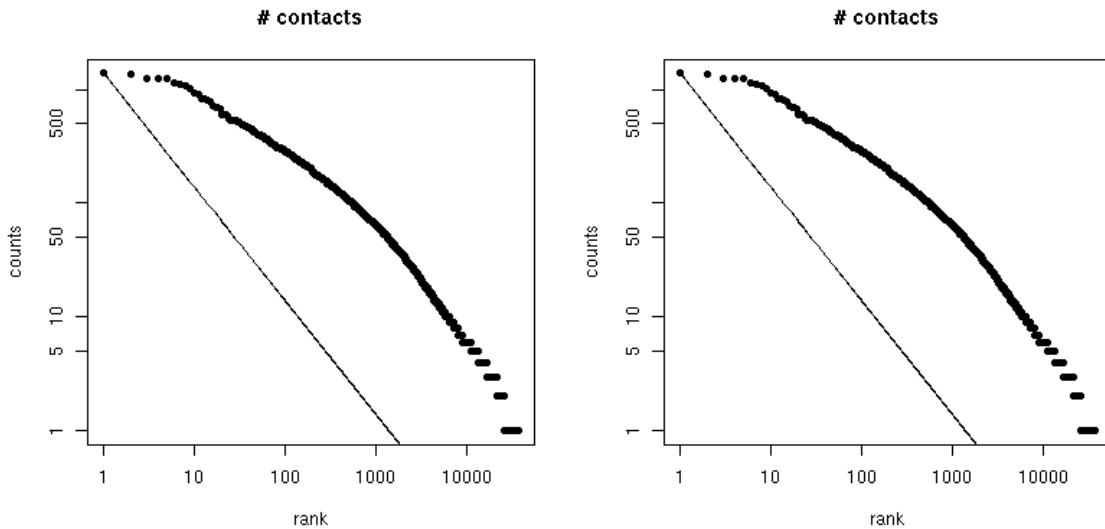


Figure 2.17: Zipf plot of Enron margins.

The marginal distributions in Figure 2.17 have a slope somewhat close to one and also some curvature. The copula estimates in Figure 2.18 are symmetric by construction. There are some head-to-tail affinities that we would expect because some email is broadcast to all or most accounts. There are also some affinities among smaller participants of equal size.

2.3.9 California roads

The final example we consider is of a network of California roads. This dataset is quite different from the others in that it is not derived from human-to-human interaction. This is another example of an undirected graph. The entities are intersections and $X_{ij} = 1$ if intersections i and j share a road segment. Because of the nature of the data, the graph is almost planar. The maximum degree is 12, indicating that there is an intersection where 12 separate road segments meet. The most common

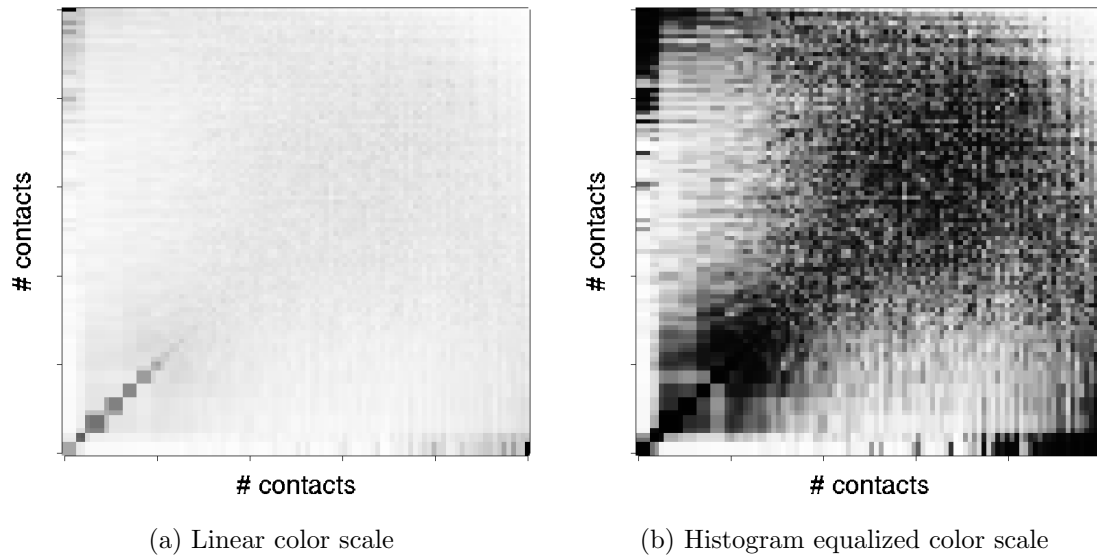


Figure 2.18: Data display for Enron example.

degree is 3. At first, this seems surprising since one might instead expect 4 to be the most common in an urban environment. However, many of the roads in the data set apparently correspond to forest highways and we hypothesize that this accounts for a degree of 3 being most popular, as the roads may fork multiple times. *T*-junctions are also fairly popular in suburban environments.

Because of the very limited number of unique marginal sums, the Zipf plot and histogram-normalized copulas are not very instructive. Figure 2.19 shows the copula estimate for these data. There is a very strong head-to-head affinity, as major intersections connect to each other. Road segments with few connections tend to connect to other such roads, with one important exception. Intersections composed of only one connection do not connect to each other.

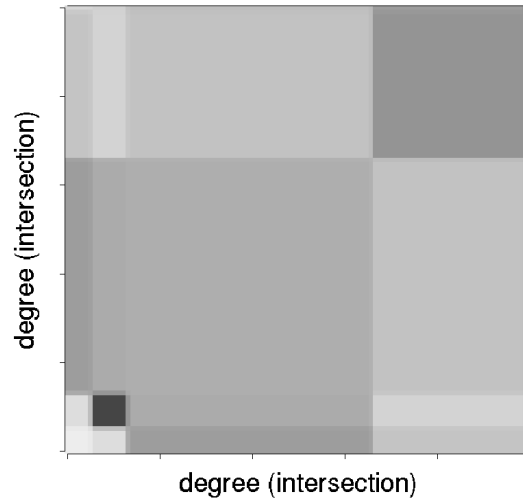


Figure 2.19: Data display for CA roads example.

2.4 Numerical summaries

In addition to data visualization, simple numerical summaries can also be useful, especially when comparing two different datasets. For example, noting the qualitative similarities between the copula estimates for the Netflix and Yahoo! ratings data, we may want a quantitative measure for assessing the affinities in each of the four corners.

There are many ways that one could go about doing this. The simplest is simply to sum the counts in rectangles of a given size in each corner and divide by $X_{\bullet\bullet}$ times the area of the rectangle. It is natural to consider squares. We have found that a 0.05×0.05 square gives good results over a wide range of datasets. Of course, for any particular one, this might not be the best choice. This ratio gives us a lift statistic with a value of 1 corresponding to neutral affinity.

Table 2.2 shows these affinities in each of the four corners for the nine example datasets. From the table, we see that the Netflix data does appear to have stronger

Data	(lo,lo)	(lo,hi)	(hi,lo)	(hi,hi)
Netflix (users, movies)	0.981	2.776	3.192	0.225
Yahoo! (users, songs)	0.551	2.127	2.163	0.202
IMDB (actors, movies)	0.871	2.000	0.787	0.528
Epinions (truster, trusted)	1.084	0.608	1.864	0.358
Wikipedia (out, in)-degree	2.213	0.100	1.722	0.251
Snapfish (sharer, sharee)	1.187	0.575	0.881	1.979
arXiv hep-th (citer, cited)	3.928	0.377	0.631	0.733
Enron email addr. (sym)	3.225	3.972	3.972	0.202
CA intersections (sym)	0.240	0.717	0.717	1.507

Table 2.2: Numerical summary of corner affinities of example datasets.

head-to-tail affinities than the Yahoo! data. However, the strongest corner affinities (at least, by the simple measure we've adopted) among the examples is found in the arXiv paper citation and Enron networks.

Another simple variant on this estimate is to instead calculate the quantity

$$\hat{a}_{\ell\ell}^{(\alpha)} = \inf\{t : \hat{c}(t, t) = \alpha\}$$

which is the smallest square that accounts for a proportion α of the data (i.e., total counts). So, this gives a measure of the tail-to-tail affinity. The analogous affinities for each of the other three corners are similarly defined. For the examples we have considered, this numerical summary yields similar results to the one we've already considered.

Many other possible summaries of the corner affinities exist, including, e.g., non-parametric estimates of tail dependence.

2.5 How accurate is the display?

Having proposed a visualization for the types of data we are interested in and examined several examples from which we attempt to draw conclusions, it is natural to wonder how accurate the display is. Without fixing a specific data-generation model, this question is a little difficult to answer, and even upon imposing such a model, we must be sensitive to how lack-of-fit of the data to that model might change our conclusions.

In this section, we consider a model for the marginal sums of our data which, while fairly flexible, allows us to make some quantitative statements regarding the accuracy of the display. We are mostly interested in data that have margins that follow an approximate power-law and so we focus our model around this idea.

To this end, we introduce the Zipf–Poisson ensemble. For integers $i \geq 1$ let $X_i \sim \text{Poi}(\lambda_i)$ be independent where $\lambda_i = N\theta_i$ for $N > 0$ and $\theta_i = i^{-\alpha}$ for a parameter $\alpha > 1$.

Here X_i represents the marginal total number of observations for the i 'th largest entity. Given these X_i we will sort them getting $X_{[1]} \geq X_{[2]} \geq X_{[3]} \geq \dots$. Entity $[i]$ is the i 'th largest in the sample, while entity i is the true i 'th largest entity. We will see that $i = [i]$ for small i in the limit as $N \rightarrow \infty$.

We have chosen intensities θ_i proportional to the probability mass function of a Zipf law. A plot of $\log(\theta_i)$ versus $\log(i)$ is linear with slope $-\alpha$. When real data are plotted this way, we seldom see exactly a linear plot. It is more usual to see some curvature, with different steepness at large and small i . We will however use the Zipf distribution for its simplicity; other distributions with nearly power law behavior will have similar results.

Theorem 2.1. *Let X_i be sampled from the Zipf–Poisson ensemble with parameter $\alpha > 1$. If $n = n(N) \leq (AN/\log(N))^{1/(\alpha+2)}$ for $A = \alpha^2(\alpha + 2)/4$, then*

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 1. \quad (2.1)$$

If $n = n(N) \leq (BN \log(N))^{1/(\alpha+2)}$ for $B < A$, then

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n > \max_{i>n} X_i) = 1. \quad (2.2)$$

If instead $n = n(N) \geq CN^{1/(\alpha+2)}$ for any fixed $C > 0$, then

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 0. \quad (2.3)$$

Proof. The first and third claims follow from Corollaries A.1 and A.2, respectively. The second follows from Lemma A.4. These are all proved in the Appendix. \square

Theorem 2.1 shows that the first $n = (AN/\log(N))^{1/(\alpha+2)}$ entries are correctly ordered relative to each other with probability tending to one. There is also the possibility that some of the smaller entities might jump ahead of the n 'th one. But if we replace A in Theorem 2.1 by any $B < A$, then such jumping is ruled out. As a result, for any $\alpha > 1$ we get at least the first $(3N/(4 \log(N)))^{1/(\alpha+2)}$ entities in the correct order, with probability tending to one.

For large N , the top $n_\epsilon = N^{1/(\alpha+2)-\epsilon}$ entities get properly ordered with very high probability for $0 < \epsilon < 1/(\alpha+2)$. The tail beyond n_ϵ accounts for a proportion of data close to $\zeta(\alpha)^{-1} \int_{n_\epsilon}^{\infty} x^{-\alpha} dx = O(n_\epsilon^{-\alpha+1}) = O(N^{(1-\alpha)/(\alpha+2)+\epsilon'})$ for $\epsilon' = \epsilon(\alpha - 1)$. Taking small ϵ and recalling that $\alpha > 1$ we find that the fraction of data from improperly

ordered entities is asymptotically negligible in the Zipf–Poisson ensemble.

2.6 Summary

In this chapter, we have presented a new data visualization for bivariate transposable data with a large number of categorical levels along each dimension. Furthermore, we have examined multiple examples in which we have strived to demonstrate the usefulness of the visualization in drawing new insight and conclusions about the data. Our aim has been to provide an exploratory tool to guide more in-depth analyses. Through a Zipf–Poisson model for the marginal sums, we have shown that the vast majority of the data display is accurate with high probability as the sample size goes to infinity.

Chapter 3

Parametric copula models

This chapter deals with some of the simplest, and most common, parametric copula models. Parametric models have found various applications over the last decade, perhaps most (in)famously in quantitative and computational finance. There are two main strategies for generating parametric copulas: the inversion method and Archimedean families. These copulas tend to have a high degree of symmetry. On the other hand, the empirical examples from Chapter 2 are highly asymmetric, making standard parametric copulas a poor choice for modeling.

We follow this discussion with some fairly simple methods of constructing new copulas from old ones while introducing various forms of asymmetry. The methods for generating asymmetry discussed in this chapter are not new, but do not seem to be widely known or taken advantage of in applications. Still, we find that the resulting copulas have no straightforward interpretation and are cumbersome mathematically both as a representation of the dependence as well as from the standpoint of trying to fit the models to observed data.

3.1 Classical parametric copulas

In this section, we discuss the two most popular ways of constructing parametric copulas. We first discuss the inversion method which is implied by Sklar's theorem and give the most famous example. We then present Archimedean copulas. These result in somewhat flexible shapes for the bivariate dependence, but their inherent symmetry is problematic.

3.1.1 Sklar's theorem and the inversion method

Sklar's theorem (Theorem 1.3) provides a method for generating a valid parametric copula. Suppose that $\{H_\theta\}$ is a family of bivariate distribution functions indexed by a parameter θ and with margins F_θ and G_θ . Then Corollary 1.3 tells us that we can construct a copula via

$$C_\theta(u, v) = H_\theta(F_\theta^{-1}(u), G_\theta^{-1}(v)) .$$

In general, the above construction yields only a subcopula, but the results of Chapter 1 show that we can easily extend the subcopula to a full copula.

The canonical example of a copula constructed in this way is the *Gaussian copula*.

Definition 5. The *Gaussian copula* of correlation ρ is a function $C_\rho : [0, 1]^2 \rightarrow [0, 1]$ defined by

$$C_\rho(u, v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v)) ,$$

where $\Phi_\rho(\cdot, \cdot)$ indicates the standard bivariate normal distribution with correlation ρ and $\Phi^{-1}(\cdot)$ is the inverse standard-normal distribution function.

The Gaussian copula has a density with respect to Lebesgue measure and is easily obtained by double differentiation of C_ρ .

Definition 6. The *Gaussian copula density* of correlation ρ is a function $c_\rho : [0, 1]^2 \rightarrow \mathbb{R}$ satisfying

$$c_\rho(u, v) = \frac{\varphi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))}{\varphi(\Phi^{-1}(u))\varphi(\Phi^{-1}(v))},$$

where

$$\varphi_\rho(x, y) = (2\pi\sqrt{1-\rho^2})^{-1} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right)$$

is the bivariate standard normal density with correlation ρ and $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ is the standard normal density.

Note that from the definition of the Gaussian copula density, it is easy to see that the density is symmetric in its two arguments. Actually, the Gaussian copula satisfies the stronger condition of radial symmetry since it is generated from an elliptical distribution. A copula is radially symmetric if and only if, for all $(u, v) \in [0, 1]^2$,

$$C(u, v) = C(1 - u, 1 - v) + u + v - 1.$$

In terms of a copula density, if it exists, this means that

$$\int_0^u \int_0^v c(u', v') \, dv' \, du' = \int_{1-u}^1 \int_{1-v}^1 c(u', v') \, dv' \, du',$$

from which the geometric meaning is more apparent, i.e., the amount of mass in any “lower-tail rectangle” is always the same as the corresponding mass in the “upper-tail rectangle” of the same dimension.

Figure 3.1 shows three example Gaussian copulas, of correlation $\rho = 0.2, 0.4, 0.8$.

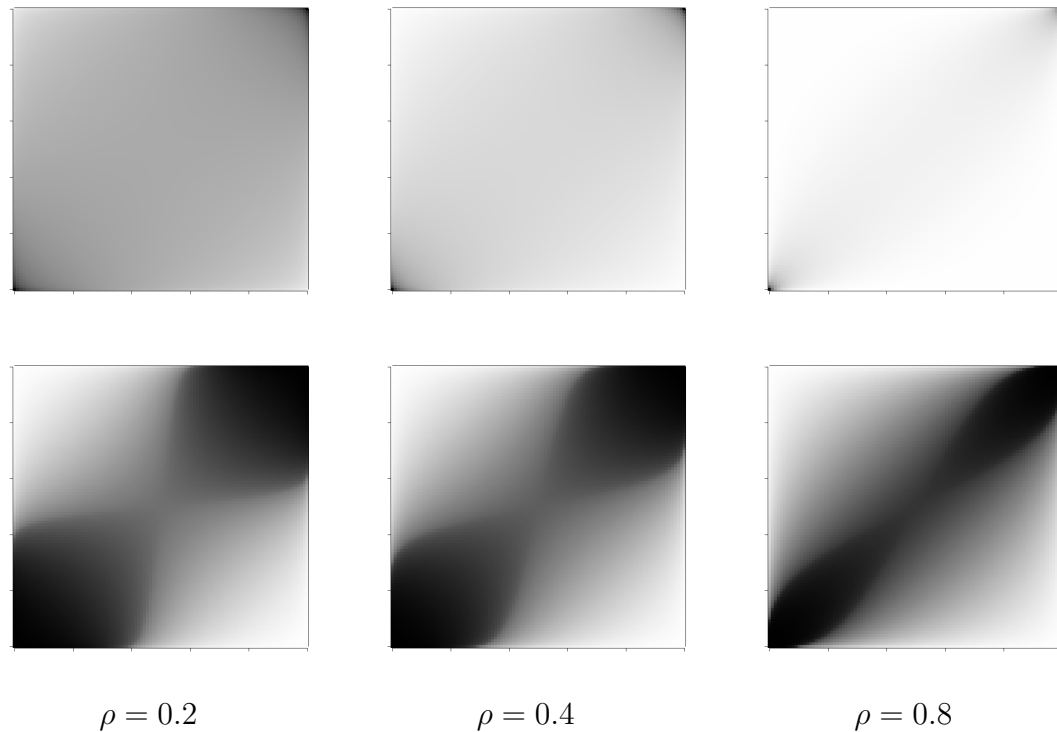


Figure 3.1: Example Gaussian copulas

The top panels show the copula densities under a linear scaling and the bottom panels show their histogram-equalized counterparts. While all of these examples show positive dependence, the Gaussian copula can, of course, generate negative dependence by selecting $\rho < 0$. Furthermore, the Gaussian copula attains the Fréchet–Hoeffding bounds by choosing $\rho = \pm 1$. The choice $\rho = 0$ corresponds to Π , the independence copula.

These properties and, perhaps, the ubiquity of the Gaussian distribution in other areas of statistics have led to the widespread use of the Gaussian copula in applications. However, from the examples, it is apparent that it cannot be used to model our examples very well and its use has also recently attracted much criticism in other

application areas.

3.1.2 Archimedean copulas

Archimedean copulas are a class of copulas that arise out of a generator function, which is usually parametrized by at least one parameter. The ease with which an Archimedean copula can be constructed is, in fact, one of its principle attractions. We primarily follow [53, 30].

Definition 7. An *Archimedean generator* is a function $\varphi : [0, 1] \rightarrow [0, \infty]$ that is convex, strictly decreasing, and satisfies $\varphi(1) = 0$.

The pseudo-inverse of φ , denoted φ^{-1} , is defined as the usual inverse on $[0, \varphi(0)]$ and is zero otherwise. Then, $\varphi^{-1}(\varphi(t)) = t$ on $[0, 1]$ and

$$\varphi(\varphi^{-1}(t)) = \min(t, \varphi(0)).$$

If the generator satisfies $\lim_{t \rightarrow 0^+} \varphi(t) = \infty$, then φ^{-1} is the usual inverse.

Lemma 3.1. For $(u, v) \in [0, 1]^2$ and an Archimedean generator φ , let $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$. Then, for $t \geq 0$,

$$\varphi^{-1}(\varphi(u) + \varphi(v) + \varphi(t)) = C(C(u, v), t).$$

Proof. If $\varphi(u) + \varphi(v) < \varphi(0)$ then

$$\varphi(u) + \varphi(v) = \varphi(\varphi^{-1}(\varphi(u) + \varphi(v))),$$

and so $\varphi^{-1}(\varphi(u) + \varphi(v) + \varphi(t)) = C(C(u, v), t)$ by the definition.

If $\varphi(u) + \varphi(v) \geq \varphi(0)$, then $\varphi(u) + \varphi(v) + \varphi(t) \geq \varphi(0)$ and so

$$0 = \varphi^{-1}(\varphi(u) + \varphi(v) + \varphi(t)) = \varphi^{-1}(\varphi(u) + \varphi(v)) = \varphi^{-1}(\varphi(0) + \varphi(t)),$$

and so $\varphi^{-1}(\varphi(u) + \varphi(v) + \varphi(t)) = C(C(u, v), t)$, as desired. \square

Proposition 3.1. *Given an Archimedean generator φ , let $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$, defined on $[0, 1]^2$. Then, C is a copula.*

Proof. $C(u, 0) = \varphi^{-1}(\varphi(u) + \varphi(0)) = 0$ since $\varphi(u) + \varphi(0) > \varphi(0)$. Similarly, $C(u, 1) = \varphi^{-1}(\varphi(u) + \varphi(1)) = \varphi^{-1}(\varphi(u)) = u$. Symmetry implies $C(0, v) = 0$ and $C(1, v) = v$.

Now, since φ is convex, so is φ^{-1} . Fix $0 \leq u_1 \leq u_2 \leq 1$ and $0 \leq v_1 \leq v_2 \leq 1$. Then, $\varphi(u_1) \geq \varphi(u_2) \geq 0$ and $\varphi(v_1) \geq \varphi(v_2) \geq 0$. Set

$$\gamma = \frac{\varphi(u_1) - \varphi(u_2)}{\varphi(u_1) - \varphi(u_2) + \varphi(v_1)}.$$

By convexity of φ^{-1} , we have

$$\varphi^{-1}(\varphi(u_1)) \leq (1 - \gamma)\varphi^{-1}(\varphi(u_2)) + \gamma\varphi^{-1}(\varphi(u_1) + \varphi(v_1)),$$

and

$$\varphi^{-1}(\varphi(u_2) + \varphi(v_1)) \leq \gamma\varphi^{-1}(\varphi(u_2)) + (1 - \gamma)\varphi^{-1}(\varphi(u_1) + \varphi(v_1)),$$

and adding these two yields

$$u_1 + \varphi^{-1}(\varphi(u_2) + \varphi(v_1)) \leq u_2 + \varphi^{-1}(\varphi(u_1) + \varphi(v_1)),$$

where we have used the fact that $\varphi^{-1}(\varphi(u)) = u$. By the definition of C , this last

inequality can be rewritten as

$$C(u_2, v_1) - C(u_1, v_1) \leq u_2 - u_1 .$$

Now, for $0 \leq v_1 \leq v_2 \leq 1$, by continuity of φ and φ^{-1} , there exists $t \geq 0$ such that $\varphi(v_1) = \varphi(v_2) + \varphi(t)$. Hence,

$$\begin{aligned} C(u_2, v_1) - C(u_1, v_1) &= \varphi^{-1}(\varphi(u_2) + \varphi(v_1)) - \varphi^{-1}(\varphi(u_1) + \varphi(v_1)) \\ &= C(C(u_2, v_2), t) - C(C(u_1, v_2), t) \\ &\leq C(u_2, v_2) - C(u_1, v_2) , \end{aligned}$$

where the last line follows from the hypothesis.

Therefore, C satisfies all the conditions of a copula. \square

Example 3.1. *The family of Clayton copulas is defined via the generator $\varphi(t) = t^{-\theta} - 1$ for $\theta \geq -1$. The associated copula is $C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$. The copula density is*

$$c(u, v) = \frac{\partial C(u, v)}{\partial u \partial v} = (\theta + 1) u^{-\theta-1} v^{-\theta-1} (u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}-2} .$$

The Clayton copula family is parametrized by a single nonpositive parameter θ . It is easy to check that φ satisfies the conditions of a valid Archimedean generator.

Figure 3.2 provide some examples of Clayton copulas.

Example 3.2. *The Frank copula is indexed by the parameter $\theta \in \mathbb{R}$ and has generator*

$$\varphi(t) = -\log \left(\frac{e^{-\theta x} - 1}{e^{-\theta} - 1} \right) ,$$

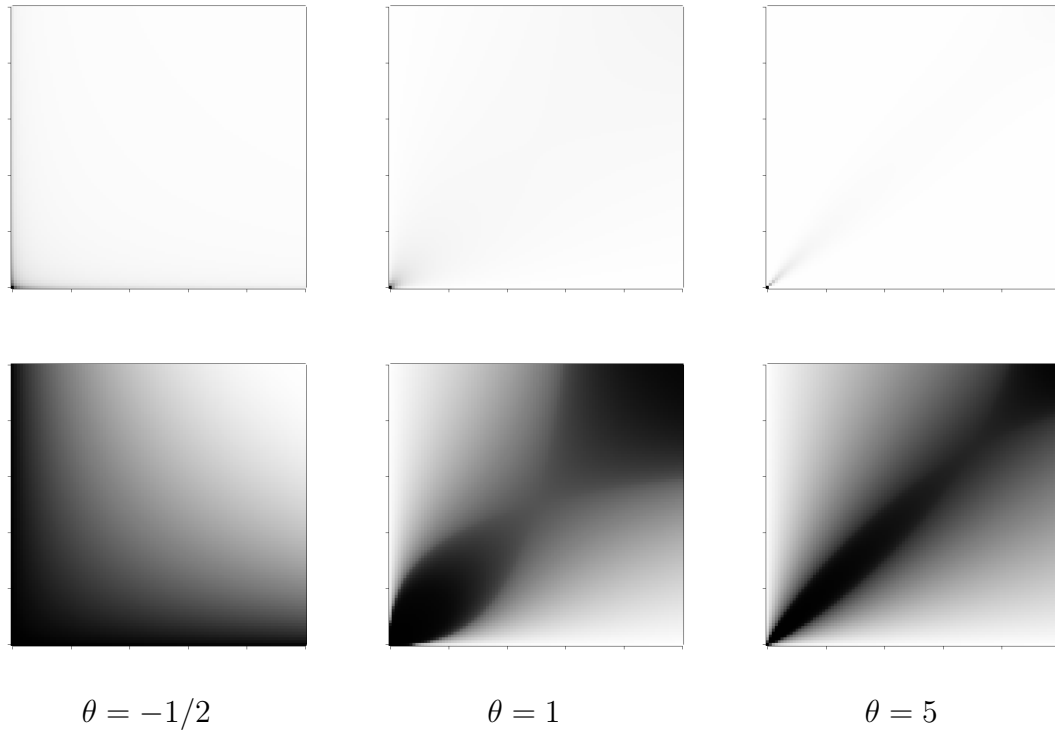


Figure 3.2: Example Clayton copulas

copula

$$C(u, v) = -\frac{1}{\theta} \log \left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right),$$

and copula density

$$c(u, v) = \frac{-\theta e^{-\theta(u+v)}(e^{-\theta} - 1)}{((e^{-\theta} - 1) + (e^{-\theta u} - 1)(e^{-\theta v} - 1))^2},$$

respectively.

Figure 3.3 show some examples of copulas from the Frank family.

Both of these copulas are interesting in that special cases yield the Fréchet–Hoeffding bounds and the independence copula. In the case of the Clayton copula,

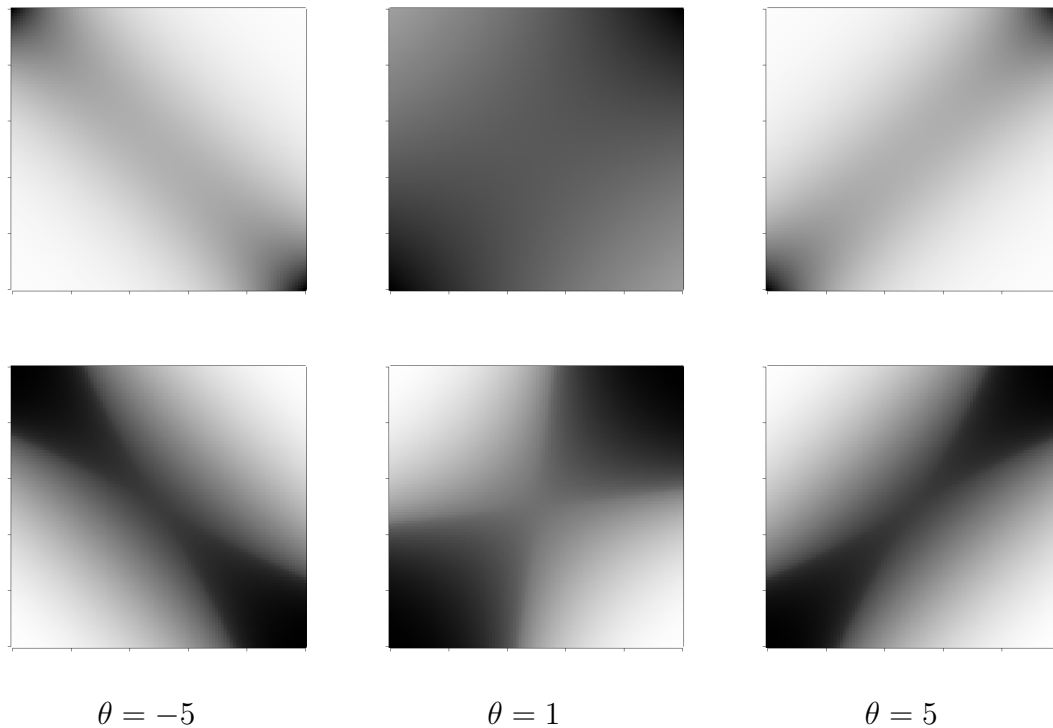


Figure 3.3: Example Frank copulas

the parameters are $\theta = -1$ for the lower bound, $\theta = 0$ for Π and $\theta = \infty$ for the upper bound. For the Frank copula, the parameters are $\theta = -\infty$, $\theta = 0$, and $\theta = 1$, respectively. However, attaining all three of these “special” copulas via specific choices of parameter values is the exception, rather than the rule, for Archimedean copulas.

3.1.3 Remarks on symmetry

From the figures, it is apparent that the Gaussian copula and Archimedean copulas exhibit a high degree of symmetry. As previously mentioned, the Gaussian copula is radially symmetric, which is a fairly strong notion of symmetry.

All Archimedean copulas are examples of distributions of *exchangeable* random

variables. In other words, if (X, Y) have Archimedean copula C and common marginal distribution F , then (X, Y) and (Y, X) have the same distribution. This is most easily seen from the construction of an Archimedean copula through its generator, i.e.,

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)) = C(v, u).$$

Thus, $H(x, y) = C(F(x), F(y)) = C(F(y), F(x)) = H(y, x)$ via Sklar's theorem. Note that the Gaussian copula also leads to exchangeable random variables if the marginal distributions of X and Y are the same.

Unfortunately, essentially all of the examples considered in Chapter 2 exhibit asymmetry (excepting the symmetric-graph examples). Thus, Gaussian and Archimedean copulas are of limited value in these instances. In the next section, we discuss how to generate new parametric copulas from old ones. One method provides a mechanism for generating some forms of asymmetry.

3.2 Generating new copulas from old

In this section, we begin with a few simple methods of constructing new copulas from old ones. Unfortunately, even these methods will not resolve the problems of symmetry discussed in the previous section. The second method, introduced by Liebscher [42], does produce some desirable asymmetries, but comes with its own drawbacks.

3.2.1 Reflections and mixtures

There are a few very simple methods of generating new copulas from old. None of them break the exchangeable nature of Archimedean copulas or the Gaussian copula, but they can easily modify other dependence properties. For example, many of the Archimedean copulas exhibit only positive dependence, i.e., the probability of getting an extremely large or extremely small value of one variable conditional on the other is greater than the associated marginal probability. This is what we have referred to in Chapter 2 as head-to-head and tail-to-tail affinities. In many of the examples, however, there are head-to-tail affinities.

Suppose that $c(u, v)$ is an empirical copula density. Then, translating the properties of a copula distribution to that of a density, we have that

1. $\int_0^1 c(u, v) dv = 1$ for all $u \in [0, 1]$,
2. $\int_0^1 c(u, v) du = 1$ for all $v \in [0, 1]$, and,
3. $c(u, v) \geq 0$ for all $(u, v) \in [0, 1]^2$.

From this it is apparent that, given a copula density c , we can generate new ones by reflecting and rotating them. That is, all of the following are also valid copulas

1. $c_1(u, v) = c(u, 1 - v)$,
2. $c_2(u, v) = c(1 - u, v)$, and,
3. $c_3(u, v) = c(1 - u, 1 - v)$.

Even if a copula C does not have a corresponding density c , such constructions are easily obtained. We leave it to the reader as a simple exercise to verify the associated expressions in terms of C .

These reflections can produce copulas with head-to-tail affinities from ones with head-to-head and/or tail-to-tail affinities.

A second simple method for constructing new copulas is via mixtures. It is trivial to verify, for example, that finite convex combinations of copulas also result in a valid copula. That is, if C_1, \dots, C_n are copulas and $\lambda_i \geq 0 \forall 1 \leq i \leq n$ with $\sum_{i=1}^n \lambda_i = 1$, then

$$C(u, v) = \sum_{i=1}^n \lambda_i C_i(u, v)$$

is also a copula.

The continuous analog to this is also true. Suppose $\{C_\theta\}$ is a family of copulas indexed by a continuous parameter $\theta \in \Omega$. Let Λ be a probability distribution over Ω . Then

$$C(u, v) = \int_{\Omega} C_\theta(u, v) d\Lambda(\theta)$$

is a valid copula. This can also be interpreted from a Bayesian perspective as putting a prior on θ .

3.2.2 Liebscher's method

Liebscher [42] presents two methods for constructing new copulas from old ones, where the resulting copula is, in general, asymmetric, even when the “generating copulas” are symmetric. We will focus on the first, more general method; the second one is applicable specifically to Archimedean copulas.

Suppose we are given functions g_{ij} for $1 \leq i \leq n$ and $1 \leq j \leq d$ such that g_{ij} is either strictly increasing on $[0, 1]$ or is identically equal to 1. Suppose further that,

for every j , $\prod_{i=1}^n g_{ij}(v) = v$ and $\lim_{v \rightarrow 0^+} g_{ij}(v) = g_{ij}(0)$. Then,

$$C(u_1, \dots, u_d) = \prod_{i=1}^n C_i(g_{i1}(u_1), \dots, g_{id}(u_d))$$

is a valid copula. For the proof, see [42].

One example of a valid set of functions g_{ij} is $g_{ij}(v) = v^{\theta_{ij}}$ where $0 \leq \theta_{ij} \leq 1$ and $\sum_{i=1}^n \theta_{ij} = 1$. The paper provides further examples.

Example 3.3. Let C_1 and C_2 be copulas and $\alpha, \beta \in [0, 1]$. Then,

$$C(u, v) = C_1(u^\alpha, v^\beta)C_2(u^{1-\alpha}, v^{1-\beta})$$

is a copula.

A special case of interest is obtained when $C_2 = \Pi$. Then, given copula C ,

$$\bar{C}(u, v) = C(u^\alpha, v^\beta)u^{1-\alpha}v^{1-\beta},$$

is a copula, and if C has a density c and partial derivatives $c_u = \partial C / \partial u$ and $c_v = \partial C / \partial v$, then

$$\begin{aligned} \bar{c}(u, v) = & \alpha\beta c(u^\alpha, v^\beta) + \alpha(1-\beta)c_u(u^\alpha, v^\beta)v^{-\beta} + \\ & (1-\alpha)\beta c_v(u^\alpha, v^\beta)u^{-\alpha} + (1-\alpha)(1-\beta)C(u^\alpha, v^\beta)u^{-\alpha}v^{-\beta} \end{aligned}$$

is the copula density of \bar{C} .

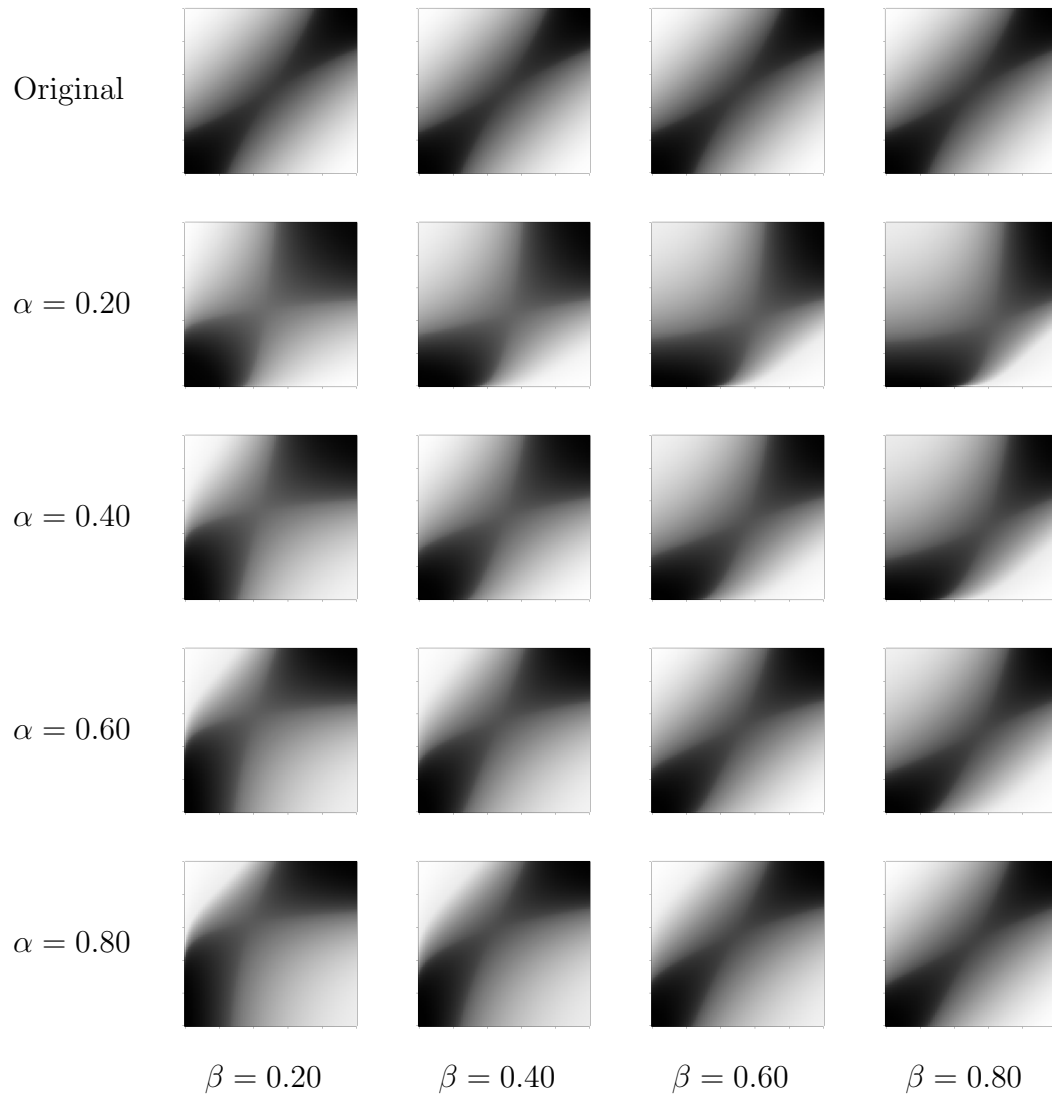


Figure 3.4: Example asymmetric copulas from “generating” Frank copula with $\theta = 5$.

3.3 Remarks on simulating from parametric copula models

Simulating from copula models that use Sklar’s theorem as a method of construction is quite easy. All one needs to do is generate (X, Y) according to the underlying

joint distribution H and then return $(U, V) = (F(X), G(Y))$ where F and G are the corresponding marginals.

The case of Archimedean copulas is not much harder. For simplicity, we will assume that $\varphi(0^+) = \infty$ and so C admits a density. We can then use a “conditional distribution” approach to generating the bivariate copula distribution. Specifically, note that

$$C(u, v) = \mathbb{P}(U \leq u, V \leq v) = \int_0^v \mathbb{P}(U \leq u \mid V = v') \mathbb{P}(V = v') \, dv',$$

where $P(V = v')$ denotes the density of V . But, since $\mathbb{P}(V = v) = 1$ for $0 \leq v \leq 1$, then we have

$$\mathbb{P}(U \leq u \mid V = v) = \frac{\partial C(u, v)}{\partial v} = \frac{\varphi'(v)}{\varphi'(C(u, v))}.$$

We can now use the standard inversion method on the conditional distribution. Let $w = \mathbb{P}(U \leq u \mid V = v)$. Then,

$$t \equiv C(u, v) = \varphi'^{(-1)}(\varphi'(v)/w),$$

so that

$$u = \varphi^{-1}(\varphi(t) - \varphi(v)).$$

Thus, the following algorithm will generate a random variate from the Archimedean copula with generator φ .

1. Generate independent, uniform random variables V and W .
2. Set $T = \varphi'^{(-1)}(\varphi'(V)/W)$.

3. Set $U = \varphi^{-1}(\varphi(T) - \varphi(V))$.
4. Return (U, V) .

3.4 Summary

In this chapter, we have reviewed the properties of the main classes of parametric copulas. These classes tend to have a high degree of symmetry and so we have discussed ways to use them as “generators” in order to get asymmetric copulas. Some examples demonstrate the kinds of asymmetry that we might expect.

Unfortunately, the mathematical forms of the resulting copulas are not very amenable to analysis and the parameters are tied into the models in a way that makes them inconvenient, at best, to estimate. Thus, we have conspicuously avoided the problem of fitting parametric copula models in this chapter. Indeed, the main fitting approaches rely on the assumption that the observations from the copula model are i.i.d., which the data we are interested in do not follow. Parametric models arising from Sklar’s theorem or via an Archimedean generator function *are*, however, easy to simulate from.

Another disadvantage of parametric models in our situation is that they do not tend to have shapes that are very similar to the idiosyncratic patterns in our observed examples. This is true, even when we allow for asymmetry via Liebscher’s method.

Furthermore, there does not appear to be a clear, sensible generative model incorporating a parametric copula from which our data would arise.

Chapter 4

Nonparametric copula models

The previous chapter dealt with parametric models for copulas. The present one explores the nonparametric case. We begin by examining a recent paper [31] which gives a general methodology for smoothing copula-density estimates via wavelets and which provides some nice examples. This paper leaves a slight gap in that the copula density estimates that it describes are not actually valid copula densities. We propose a “patch” for this problem and then spend the second part of the chapter generalizing somewhat the basic ideas of [31]. Our approach leads to a convex-programming problem with a smoothing parameter that is quite interpretable provided the basis functions used satisfy a minor condition.

4.1 Genest et al.’s nonparametric wavelet copulas

In [31], Genest et al., take up the problem of estimating copula densities from observed data, where they assume that (X, Y) pairs of random variables are sampled in i.i.d. fashion. This section gives a brief overview of the ideas; details can be found in the

paper.

We begin by considering the empirical-distribution function of each margin. Let

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(X_i \leq x)}$$

and

$$\hat{G}_n(y) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(Y_i \leq y)} .$$

Then on the lattice consisting of points $(i/n, j/n)$ for $0 \leq i, j \leq n$, we can get empirical-copula estimate

$$\begin{aligned} \hat{C}_n(i/n, j/n) &= \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{(\hat{F}_n(X_k) \leq i/n)} \mathbf{1}_{(\hat{G}_n(Y_k) \leq j/n)} \\ &= \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{(R_k \leq i, S_k \leq j)} , \end{aligned}$$

where R_k is the rank of X_k in the sample and S_k is the rank of Y_k . If we knew the marginal distributions F and G , then we could use the maximum-likelihood estimate

$$\tilde{C}_n(i/n, j/n) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{(F(X_k) \leq i/n)} \mathbf{1}_{(G(Y_k) \leq j/n)} .$$

Since we do not (usually) know F or G , we instead use their MLEs as plug-in estimates to obtain \hat{C}_n . As alluded to in [31], this extra layer of estimation does not present much of a problem in terms of asymptotics.

Unfortunately, while rapid convergence of empirical distribution functions is the norm, estimates of densities are not often as nicely behaved. In particular, empirical density estimates are very noisy. To smooth out such estimates of the copula

density $c(u, v)$, Genest et al. employ wavelets. Authoritative references on wavelets are [14, 44]. Though wavelets have many interesting properties and advantages—computationally, mathematically, and intuitively—for our purposes, we will mostly view them as a specific case of an orthonormal set of basis functions.

Instead of using the full lattice $\{(i/n, j/n) : 1 \leq i \leq n, 1 \leq j \leq n\}$, it is useful to bin over a regular sublattice. Genest et al. uses $N = 2^m$ points along each dimension such that m is an integer satisfying $2^m \leq \sqrt{n} < 2^{m+1}$, as suggested by wavelet theory.

The basic algorithm then goes as follows.

1. For each bin, calculate the proportion of the data within the bin, i.e.,

$$\hat{b}_{i,j} = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\left(\frac{i-1}{N} < \frac{R_k}{n} \leq \frac{i}{N}, \frac{j-1}{N} < \frac{S_k}{n} \leq \frac{j}{N}\right)}.$$

2. Apply the Fast Wavelet Transform to the matrix $B = (\hat{b}_{ij})$ and construct wavelet estimates of the copula density such that the estimate at wavelet-resolution level ℓ is

$$\tilde{c}_\ell(u, v) = \sum_{(i,j) \in \mathbb{Z}^2} \tilde{b}_{i,j}^{(\ell)} \psi_{i,j}^{(\ell)}(u, v),$$

where $\tilde{b}_{i,j}^{(\ell)}$ are the wavelet coefficients at resolution level ℓ and $\{\psi\}$ is the specific wavelet basis.

3. Choose $c^* = \tilde{c}_{m-1}$ as the estimate of the copula density c .

Many variations on this theme are possible. Instead of simply choosing the final estimate as the one at resolution level one less than the total resolution, any one of the popular wavelet-thresholding schemes could be substituted.

However, a slightly more subtle problem remains.

4.1.1 Sinkhorn–Knopp for valid copula-density estimates

Even in the case of the empirical-copula estimate, the resulting density is not a proper copula density unless particular care is taken. In the case of the Genest et al. algorithm, we essentially start out with this estimate and then do additional smoothing. So, we can expect the problem to persist. Indeed, the wavelet-based estimate may not even be positive over the whole unit square, or even at the lattice points.

In particular, the copula-density estimate c^* is not guaranteed to have uniform marginal distributions, and, usually will not.

The issue of nonnegative is relatively easy to address. We can simply truncate the copula estimate wherever it goes negative. So, we get an (intermediate) estimate

$$c_t^*(u, v) = c^*(u, v) \mathbf{1}_{(c^*(u, v) \geq 0)}.$$

Such a modification will, of course, affect the normalization of the copula density as well. But, this is not really an additional problem, since we already don't expect the marginal constraints to be satisfied.

To fix the marginal constraints, we propose the use of the *Sinkhorn–Knopp* algorithm [63, 37] to follow the truncation step. This is also known as *iterative proportional fitting* in the categorical-data-analysis literature. The intuition is simple. Starting with a nonnegative matrix $A^{(k)}$, we compute

$$r^{(k)} = A^{(k)} \mathbf{1}_n$$

and

$$c^{(k)} = (A^{(k)})^T \mathbf{1}_n$$

where $\mathbf{1}_n$ is an n -length vector of ones. Let $R^{(k)} = \text{diag}(1/r_1^{(k)}, \dots, 1/r_n^{(k)})$ and define $C^{(k+1)} = \text{diag}(1/c_1^{(k)}, \dots, 1/c_n^{(k)})$. Take

$$A^{(k+1)} = C^{(k)} A^{(k)} R^{(k)} .$$

Then, subject to a mild condition on the original nonnegative matrix $A^{(0)}$, the sequence of matrices $A^{(k)}$ is guaranteed to converge to a doubly stochastic matrix.

More formally,

Theorem 4.1 (Sinkhorn–Knopp). *Let A be a nonnegative square matrix. Then the Sinkhorn–Knopp iteration will converge to a unique matrix $A_\infty = \lim_{k \rightarrow \infty} A^{(k)}$ with a positive diagonal if and only if for every strictly positive entry of A , there is a permutation of the columns such that the entry in question lies on the main diagonal and this main diagonal has strictly positive entries.*

To make the connection with the copula-density estimate, for each truncated and smoothed copula-density estimate $c_t^*(u, v)$ evaluated on the lattice $\{(i/N, j/N) : 1 \leq i \leq N, 1 \leq j \leq N\}$ there is a corresponding nonnegative matrix C_t^* . To achieve a proper copula density, we need some modification of this matrix to be doubly-stochastic. The Sinkhorn–Knopp algorithm guarantees a unique modification of C_t^* which is doubly stochastic, provided the minor technical condition of the theorem is satisfied. Call this new matrix C_{sk}^* . Then using the bilinear-interpolation scheme briefly described in Chapter 3, it is easy to construct a valid copula-density estimate out of C_{sk}^* .

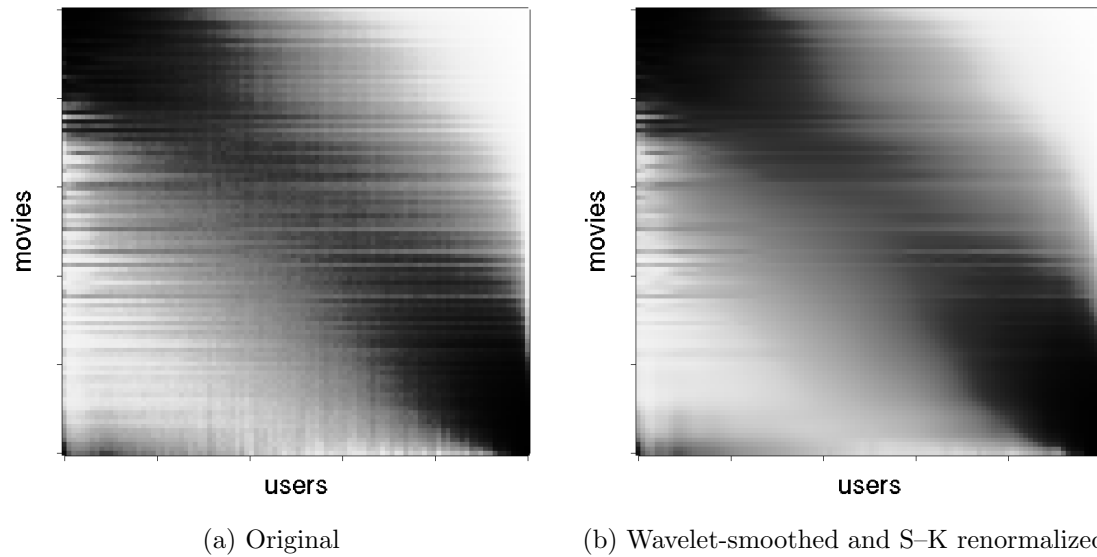


Figure 4.1: Original and smoothed versions of histogram-normalized Netflix display.

4.1.2 Netflix example

To illustrate the characteristics of this method, we apply it to the Netflix example. We choose $N = 128$ bins (to get close to 100 and still be a power of 2) and employ the “symlet6” tensor-product wavelet basis. However, as in most applications of wavelets, the resulting estimate is largely invariant with respect to the choice of basis. We fit the wavelet coefficients to the raw Netflix 128×128 copula-density estimate, keeping the detail coefficients only up to level $m - 1 = \log_2 N - 1 = 6$. After reconstructing the estimate of the copula density, we then truncate any negative values and renormalize via Sinkhorn–Knopp.

Figure 4.1 shows the result. In the left-hand pane is the original histogram-equalized version of the data display. In the right-hand pane is the wavelet-smoothed copula. The clearest differences are apparent around the transition between high and low density along the diagonal starting at the top left and going to the bottom right.

4.2 A convex-programming approach to nonparametric copula estimation

This section is dedicated to a slightly different approach than the previous section. It seems somewhat unsatisfactory to begin with a copula-density estimate, smooth it out to remove some noise, and then end up with an estimate that no longer satisfies the necessary constraints to be a copula density. In order to get back to an estimate with the necessary properties, we had to “fix up” the original smoothed estimate. From the previous chapter, we also saw that it is quite challenging to construct realistic copula densities analytically.

The goal of this section is to try to strike a compromise between the two. We seek a nonparametric estimate of the copula density that explicitly satisfies the necessary constraints. At the same time, we want to be able to modulate the smoothness of this estimate.

4.2.1 Problem formulation

The previous section employed a linear wavelet basis $\Psi = \{\psi\}$ onto which the data were projected in order to then obtain a smoothed estimate. While an interesting and useful choice, for our purposes, there is nothing particularly special about wavelets. Instead, we can generalize to the case where Ψ is simply an orthonormal set of basis functions.

Suppose we have a copula-density estimate \hat{C} represented as an $n \times n$ matrix. Let $\hat{c} = \mathbf{vec}(\hat{C})$ be the vectorized form of this matrix, i.e., an n^2 -length vector where

the columns are stacked one on top of the next. Then, the coefficients of the copula-density estimate in the orthonormal basis are simply $\theta = \Psi^T \hat{c}$. Also, because \hat{C} is a proper copula-density estimate, then $\hat{C}_{ij} \geq 0$ for all i, j and $\hat{C}1_n = \hat{C}^T 1_n = 1_n$; that is, \hat{C} is a nonnegative, doubly stochastic matrix.

In the Genest et al. algorithm, the optimization is (implicitly) done with respect to squared-error loss. While this makes the analysis convenient—and is the classical loss function, even for density estimation—it is not the only, nor necessarily best, choice. For example, Devroye and Györfi [18] strongly advocate the use of L_1 loss for nonparametric density estimation due to its invariance under arbitrary one-to-one transformations and its relationship to total-variation distance, among other properties. We will focus on L_1 loss as well, with a view also towards simplifying the computations. However, our approach is readily adaptable to squared-error loss, as we briefly discuss below.

To get a smoothed estimate, we seek a vector $w \in \mathbb{R}^{n^2}$ that minimizes $\|\hat{c} - \Psi w\|_1$ subject to the constraints that $\tilde{c} = \Psi w$ is a proper copula density. With no constraints other than the copula-density ones, this problem leads to the obvious solution that $w = \Psi^T \hat{c}$ and no smoothing is obtained. However, if we introduce a parameter $\lambda \in \mathbb{R}$ and force $\|w\|_1 \leq \lambda$, then λ acts as a regularization parameter and the resulting estimate \tilde{c} will be smoother than the original \hat{c} .

Let \tilde{C} be the matrix satisfying $\tilde{c} = \mathbf{vec}(\tilde{C})$. Then, there are matrices $R \in \mathbb{R}^{n \times n^2}$ and $L \in \mathbb{R}^{n \times n^2}$ such that $\tilde{C}1_n = 1_n$ if and only if $R\tilde{c} = 1_n$ and, likewise, $\tilde{C}^T 1_n = 1_n$

if and only if $L\tilde{c} = 1_n$. Thus, we seek a solution to the optimization problem

$$\begin{aligned}
& \text{minimize} && \|\hat{c} - \tilde{c}\|_1 \\
& \text{subject to} && \tilde{c} = \Psi w, \tilde{c} \geq 0 \\
& && L\tilde{c} = R\tilde{c} = 1_n \\
& && \|w\|_1 \leq \lambda,
\end{aligned} \tag{4.1}$$

where “ \geq ” with respect to at least one vector is interpreted elementwise. The optimization is done with respect to the parameters w .

4.2.2 Properties

A careful examination of (4.1) reveals that this problem is not only convex, but corresponds to a simple linear program. Thus, from a computational standpoint, for moderate size n , the problem is quite solvable. (If we replace L_1 with L_2 loss as the optimization function, the problem is still convex.)

We are left with the issue of the regularization parameter λ . Though “natural” in the (modern) regularization framework, its interpretation appears tricky. A minor—and natural—constraint on the choice of the basis Ψ yields a straightforward interpretation and is the subject of the following proposition.

Proposition 4.1. *Assume that the orthonormal basis Ψ contains the constant vector $n^{-1}1_{n^2}$. Let $\underline{\lambda} = \inf\{\lambda : (4.1) \text{ is feasible}\}$ and $\bar{\lambda} = \sup\{\lambda : \|\hat{c} - \tilde{c}(\lambda)\|_1 > 0\}$. Then, $\underline{\lambda} = 1$ and the corresponding solution to (4.1) is $\tilde{c}(\underline{\lambda}) = n^{-1}1_{n^2}$. Furthermore, $\bar{\lambda} = \|\Psi^T \hat{c}\|_1$ with corresponding solution $\tilde{c}(\bar{\lambda}) = \hat{c}$.*

Proof. We begin by proving the infeasibility result. Suppose that $\sum_i |w_i| < 1$. Then, $\sum_i |w_i|^2 < 1$ as well. Since Ψ is orthonormal, Parseval’s relation asserts $\|w\|_2^2 = \|\tilde{c}\|_2^2$

and the copula constraints on \tilde{c} imply that $\tilde{c}^T \mathbf{1}_{n^2} = n$. An application of the Cauchy–Schwarz inequality yields a contradiction since $\|\tilde{c}\|_2^2 \geq n^{-2}(\tilde{c}^T \mathbf{1}_{n^2})^2 = 1$.

Now, suppose that $\sum_i |w_i| = 1$. Then either $\sum_i |w_i|^2 < 1$ or $\sum_i |w_i|^2 = 1$, the latter being true if and only if $w = e_k$ for some k . Because Ψ is orthonormal and contains the constant vector $n^{-1}\mathbf{1}_{n^2}$ —which we can assume to be the first vector in the basis without loss of generality—then $w = e_1$ is the only vector that simultaneously satisfies the copula constraints and the constraint that $\|w\|_1 = 1$. Hence $\tilde{c} = \Psi w = n^{-1}\mathbf{1}_{n^2}$ yielding that \tilde{C} is the independence copula density.

For the last part of the result, suppose that $\lambda \geq \|\Psi^T \hat{c}\|_1$. Then, $w = \Psi^T \hat{c}$ is feasible and $\tilde{c} = \Psi w = \hat{c}$, so $\|\hat{c} - \tilde{c}\|_1 = 0$. For any $\lambda_0 < \lambda$, $\Psi^T \hat{c}$ is not a feasible choice for w and so $\|\hat{c} - \tilde{c}(\lambda_0)\|_1 > 0$. \square

The interpretation of Proposition 4.1 is that (4.1) yields a smoothed copula density that “interpolates” between the independence copula density and the original empirical copula-density estimate, where the interpolation is controlled by the regularization parameter λ . In order to get a desirable interpretation of the smoothing parameter, we have constrained Ψ to include the constant vector. This is natural insofar as any choice of basis should allow for a maximally sparse representation of the constant functions.

A careful examination of the proof shows that if $\|\cdot\|_1$ is replaced by $\|\cdot\|_2$ everywhere that it appears in (4.1) and in Proposition 4.1, then the proof goes through just the same and λ will have the same interpretation. One can even use L_1 as the optimization criterion and L_2 as the regularization criterion or vice versa, again with the same interpretation on λ . However, the resulting solution $\tilde{c}(\lambda)$ will, of course, be different in each case. The version in which L_2 is used for both the optimization function and

the regularization constraint corresponds to a linearly-constrained nonnegative ridge regression.

4.2.3 Selecting the smoothing parameter

We are left with the problem of how to select the regularization parameter λ . Cross-validation is currently the most popular method for doing so, but the obvious approaches yield some difficulties in our problem due to the equality constraints.

Specifically, leave-one-out cross-validation, in which a single element of \hat{c} is left out is not helpful, since whichever element is left out remains completely determined by the equality constraints on the copula density. Leaving out single whole rows or columns has the same problem.

Removing multiple rows and/or multiple columns is one option. But, it seems clear that any solution obtained is invariant to permutations of the rows and columns of the solution vector with respect to the left out rows and columns, since the latter would not enter into the optimization function.

Leaving out random elements, as long as there are at least two in each row and column with left-out elements, doesn't suffer from these problems and so is likely to work better.

An altogether different approach is to leave out entire subsets of entities for the rows and columns, fit a new empirical copula density estimate, use this to fit the smoothed estimate for a given λ and then evaluate performance on the original empirical copula-density estimate formed by considering all the entities along the rows and columns. This has the advantage that a full copula-density estimate is always being fit to for a prespecified λ . On the other hand, such an approach would be much

more computationally intensive. Also, for some data sets, a single entity corresponds to a large proportion of the total data and so if it were left out, the copula-density estimate would be drastically changed.

We leave the further exploration of this matter to future work.

4.3 Summary

In this chapter, we have considered two new, but separate, methods for obtaining valid nonparametric estimates of the copula density. The first involves “fixing up” the wavelet estimate provided by Genest et al.’s algorithm via truncation and an application of Sinkhorn–Knopp. Except in cases in which Genest et al.’s algorithm yields a truly pathological estimate, our algorithm should always have a valid, unique solution.

The second method attempts to avoid the somewhat ad-hoc nature of the first approach. Instead of constructing an estimate and then trying to fix it so that it has the right properties, we start off by enforcing the necessary properties *a priori*. We focus on optimizing an L_1 criterion. This yields a linear program which can output a set of estimates indexed by a continuous parameter λ . Furthermore, the interpretation of the effect of the choice of λ is natural, provided the basis used for the fit satisfies a minor, logical condition. A completely analogous approach with similar interpretations exists if we replace the L_1 criterion by an L_2 one or, even, mix the two.

Chapter 5

Simple generative models

In Chapters 3 and 4 we explored methods for explicitly estimating the copula densities corresponding to the data display of Chapter 2. In some sense, these approaches were somewhat unsatisfactory on both mathematical and conceptual grounds.

In this chapter, we take a different tack. We explore simple generative models for these large datasets that result in certain copula structures, instead of trying to explicitly estimate the copulas. The reasons for this are two-fold: (a) modeling the data at a higher level than through the copula itself provides intuition and (b) more quantitative mathematical statements about the underlying models and the resultant marginal distributions and copulas can be made.

The copula plots we show give the density of the variables with respect to an underlying model of independence. The point is not that independence is plausible, just that ratios of the observed density to one under independence are easy to interpret. Even so, it would be interesting to model dependence.

We present some simple generative models for networks. Some caution is in order because very distinct generative models can produce similar output, while subtly

different generative models can mean the difference between lognormal and power law distributions. See [51]. Also, while simple models with a handful of parameters can describe a few salient properties of real data, they do not necessarily produce fully realistic graphs. See for example [33]. Despite these caveats, generative models provide useful insights and hypotheses to test.

Here we consider some simple probability models that describe dependence between row and column entities. The patterns we see in each corner of a copula plot might require three to six parameters to describe, the latter for a full quadratic surface. Four independently varying corners could require 12 to 24 parameters. We will look at models with only one or two parameters and accept that only qualitative matches to some observed features are possible.

5.1 A saturation model

In ratings data we often see strong head-to-tail affinities between raters and items. This may be explained through the idea that sophisticated raters have branched out to the less well known items, while the neophytes mostly stay with the items of massive popularity.

Before adopting such a taste-based explanation we should at least consider a much simpler one. A rater is very unlikely to rate the same item twice. Even if this happens, the system may well retain only the last rating that was made. There are just not enough popular items for a busy rater to rate. These saturation effects alone would induce negative dependence. A large saturation effect has already been noted by [48] for symmetric networks, such as the internet, where there are greatly diminished connections among the most connected nodes.

We introduce a model that has independence apart from the limitation of one rating per rater-item pair. We begin with a bivariate Zipf–Poisson ensemble with latent variables $Y_{ij} \sim \text{Poi}(Nci^{-a}j^{-b})$, where $c = c_a c_b = 1/(\zeta(a)\zeta(b))$ and $\zeta(x)$ denotes the Riemann-zeta function. Then we apply a thresholding step yielding observed values $X_{ij} = 1$ if $Y_{ij} \geq 1$ and $X_{ij} = 0$ otherwise. In the latent model, row and column entities are generated independently. The thresholding that turns Y_{ij} into X_{ij} is more likely to deplete the head-to-head combinations than any others.

The following theorem provides bounds on the expected observed marginal distributions after saturation.

Theorem 5.1. *Let X_{ij} be sampled as the thresholded bivariate Zipf–Poisson ensemble described above. Let $X_{i\bullet}$ and $X_{\bullet j}$ be the marginal sums. Then*

$$\begin{aligned}\mathbb{E}(X_{i\bullet}) &\leq \min \left\{ \Gamma(1 - 1/b)(Nc)^{1/b}i^{-a/b}, Nc_a i^{-a} \right\}, \\ \mathbb{E}(X_{i\bullet}) &\geq (1 + Nci^{-a})^{1/b} - 1,\end{aligned}$$

and, as $i \rightarrow \infty$,

$$\mathbb{E}(X_{i\bullet}) \sim Nc_a i^{-a}.$$

By symmetry, analogous results hold for $X_{\bullet j}$ where we swap the roles of i and j , and a and b , respectively.

From Theorem 5.1, the marginal distribution of the row entity behaves as a power law that starts at slope a/b for the largest entities and transitions to slope a for the small ones. Conversely the column entities have slope starting at b/a and transitioning to b . As a consequence, the large entities of one type follow a power law with rate ≤ 1 while large entities of the other type have a rate ≥ 1 . We have not seen that pattern

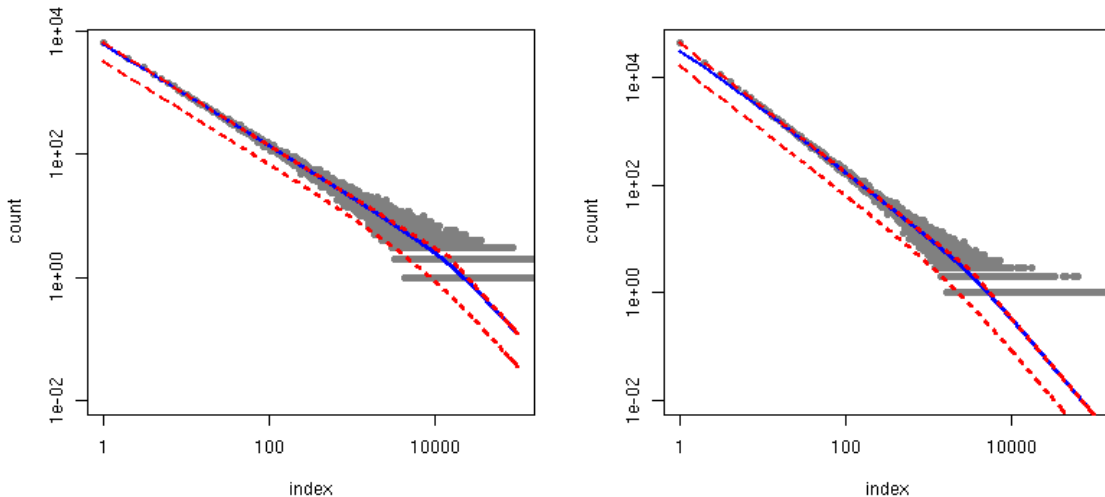


Figure 5.1: Degree distribution plot of bipartite preferential attachment ($a = 1.5$, $b = 1.8$, $t = 10^7$) margins.

in any of the data sets we've investigated. As a result we believe that the head-to-tail affinity often seen in ratings data is not simply due to saturation. Models invoking taste therefore seem more plausible.

5.1.1 Numerical examples

We consider two examples in this section, though we focus mainly on one. For the first example, we simulate from the model with $a = 1.5$, $b = 1.8$ and $N = 10^7$. For the second model, we take $b = 2.5$ instead.

Figure 5.1 shows the observed Zipf plot and bounds from Theorem 5.1 for the example simulated with $a = 1.5$ and $b = 1.8$. Instead of sorting the entities by observed counts, we have kept the original ordering. We see that these expected counts have curvature and the top entities appear well-ordered. Also the slope near

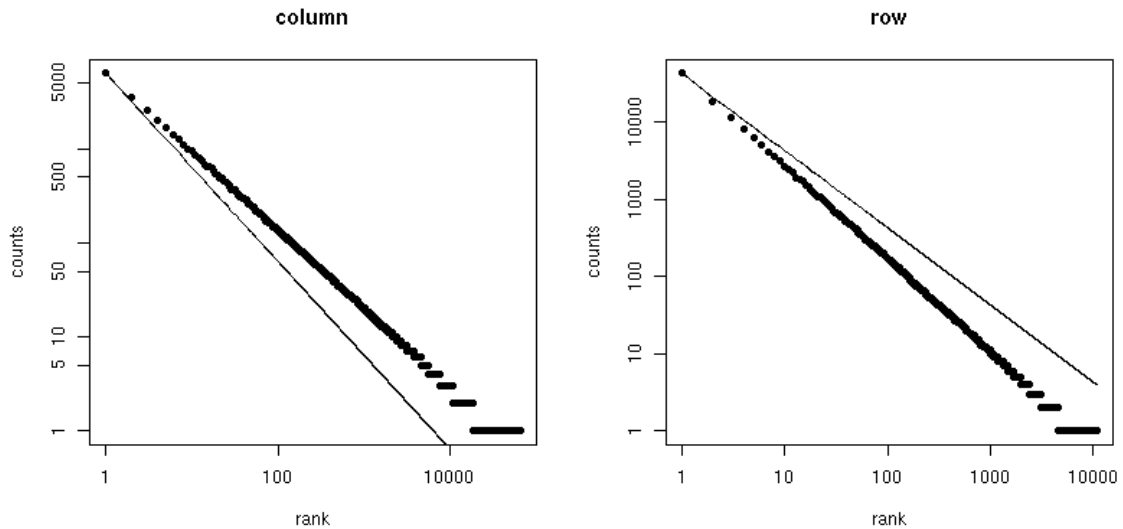


Figure 5.2: Zipf plot of saturation model margins with $a = 1.5$, $b = 1.8$ and $N = 10^7$.

the origin is not a , but has instead been altered by the sampling process. The expected number of counts was estimated by calculating a sum over millions of values of y for each fixed x (and vice versa).

Figure 5.2 shows the Zipf plots for the $b = 1.8$ example. The power-law nature of the data is very evident and there is minimal curvature. We also see the aforementioned features that the slopes are reflected versions of one another around 1.

Figure 5.3 shows the copula estimates for both $b = 1.8$ and $b = 2.5$ examples. We do indeed see an asymmetric negative dependence in the corners, though the affinity extends farther towards the center of the square than we have seen in real data. The parameter clearly alters the strength of the head-to-tail affinities that are present.

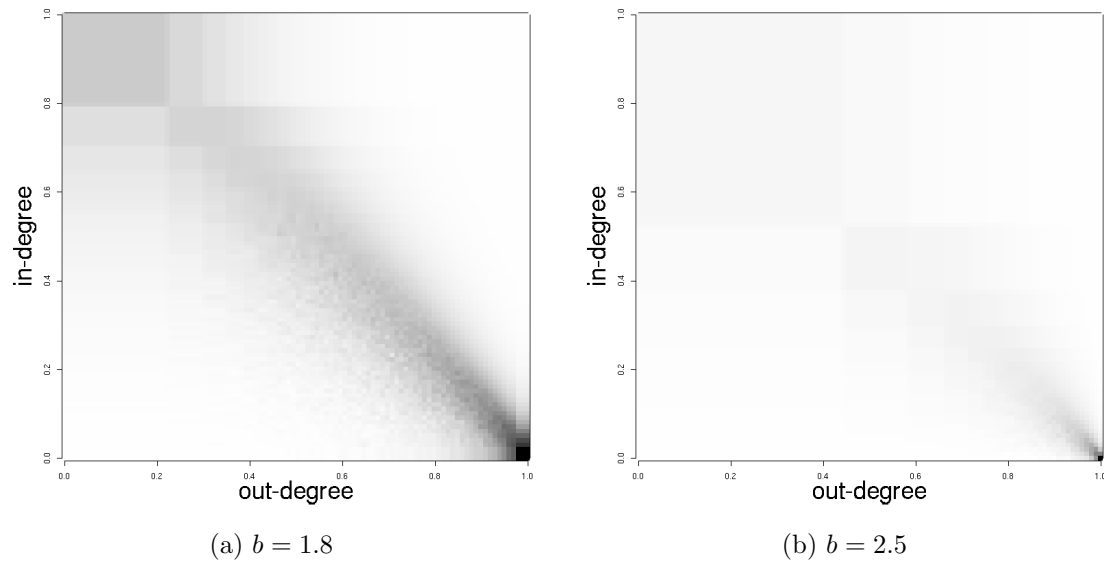


Figure 5.3: Data display for two saturation model examples.

5.1.2 Parameter estimation

Our view in this chapter so far has been to present simple generative models that are interpretable and can reproduce some qualitative features of the data. At the same time, we recognize that with only a couple of parameters we cannot hope to fit such a simple model to the rich datasets that motivate this work. Hence, our focus is not on actual parameter estimation for these models.

However, if one really found it desirable to try to do parameter estimation for this simple model, the most straightforward way would be to fit linear regressions to the head of Zipf plot. Since we do not know the true ordering of the categorical levels, we cannot fit to the head of the unordered version depicted in Figure 5.1. Ideally, one would fit a linear regression to the head of both Zipf plots. This would provide an estimate for the ratio a/b as well as a lack-of-fit estimate since we expect the slopes

to be reciprocals of one another.

5.2 Bipartite preferential attachment

A second model for these data is a bipartite preferential attachment model. The model constructs a bipartite graph via a simple extension of the Barabási-Albert model [6].

There are several generative models for bipartite graphs. Bipartite graphs in which each node type have the same degree distribution can have edges randomly assigned as in [58]. Graphs in which one kind of node has a prescribed degree distribution and the other kind are sample by preferential attachment have also been considered [33].

We investigate a preferential attachment model that generates the degree distributions along with the edge connectivity. Bipartite preferential attachment describes a random graph with two parameters, an integer valued time $t \geq 1$ and a probability $p \in (0, 1)$. There are two node sets, \mathcal{M} and \mathcal{N} , corresponding to row and column entities respectively. At time t the graph has nodes $i = 1, \dots, m(t)$ from \mathcal{M} and nodes $j = 1, \dots, n(t)$ from \mathcal{N} . We will assume that the entity sets are distinct. The graph is represented by an $\infty \times \infty$ matrix with elements $X_{ij} = X_{ij}(t) \in \{0, 1\}$, of which only t elements are nonzero.

The process starts at $t = 1$ with $m(1) = n(1) = 1$ and a single edge connecting node 1 of \mathcal{M} with node 1 of \mathcal{N} . That is $X_{11} = 1$ and $X_{ij} = 0$ if $i > 1$ or $j > 1$. At each time $t \geq 2$, we sample $U_t \sim U(0, 1)$. If $U_t \leq p$, then we add a new node $i = n(t) = n(t-1) + 1$ to \mathcal{M} and connect it at random to one of the nodes $j \in \{1, \dots, n(t-1)\}$ in \mathcal{N} , thereby setting $X_{ij} = 1$. If $U_t > p$, then we add a new node to \mathcal{N} and connect it at random to one of the nodes $1, \dots, m(t-1)$ of \mathcal{M} . The random

connections are always made by preferential attachment. A new node of one type is connected to a particular old node of the other type with probability equal to the degree of that old node at time $t - 1$, divided by the total number $t - 1$ of edges.

In [6] it is argued that the degree distribution of the vertices in (unipartite) preferential attachment graphs decays as a power law with exponent 3. That is, if p_k is the proportion of vertices of degree k , then $p_k = \Theta(k^{-3})$ as the number of vertices goes to infinity. This was further formalized in [8]. Degree distributions in bipartite preferential attachment are fundamentally different from those in the unipartite case.

Theorem 5.2. *Let $X_{ij}(t)$ be sampled from the bivariate preferential attachment model with $p \in (0, 1)$ and $q = 1 - p$. Let $X_{i\bullet}(t)$ and $X_{\bullet j}(t)$ be the marginal sums and let $M(k, t) = \sum_i \mathbf{1}_{(X_{i\bullet}(t)=k)}$ and $N(k, t) = \sum_j \mathbf{1}_{(X_{\bullet j}(t)=k)}$ be the number of vertices of degree k in \mathcal{M} and \mathcal{N} , respectively at time t . Then*

$$\begin{aligned} \frac{M(k, t)}{t} &\rightarrow \frac{p(k-1)!}{q \prod_{i=1}^k (i+1/q)} \sim \frac{p}{q} \Gamma(1+1/q) k^{-1-1/q} \\ \frac{N(k, t)}{t} &\rightarrow \frac{q(k-1)!}{p \prod_{i=1}^k (i+1/p)} \sim \frac{q}{p} \Gamma(1+1/p) k^{-1-1/p} \end{aligned}$$

where the convergence is both in mean and in probability as $t \rightarrow \infty$, and the asymptotic equivalence holds as $k \rightarrow \infty$.

Theorem 5.2 shows that both marginal distributions follow power laws. Unlike the Barabási-Albert model, the degree distributions for the bipartite model do not always have a scaling coefficient of 3. One margin has a coefficient in the range $(2, 3]$ and the other coefficient is in the range $[3, \infty)$. In real-life networks, it has been often observed that scaling coefficients tend to fall between 2 and 4. Some extensions to the basic Barabási-Albert model do generate scaling laws with coefficients in $(2, \infty)$ [20, Ch.

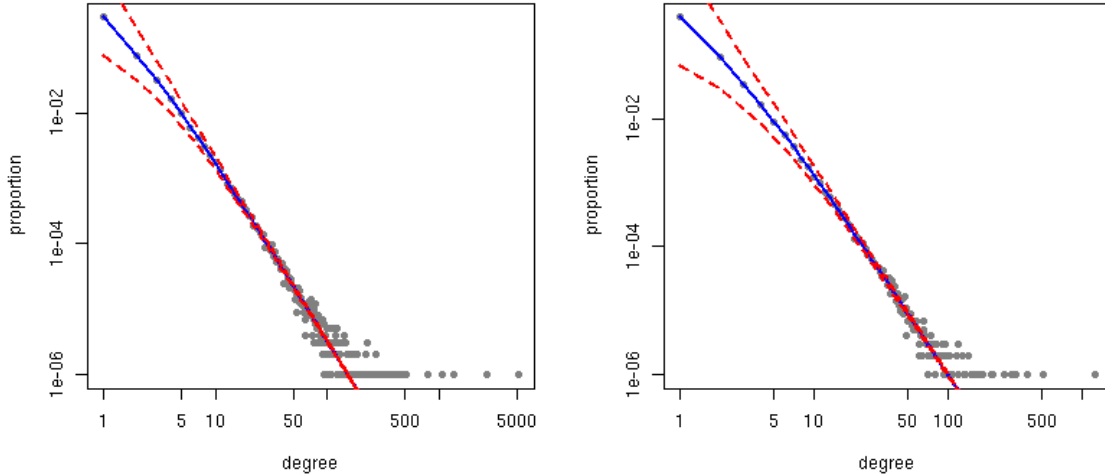


Figure 5.4: Degree distribution plot of bipartite preferential attachment ($p = 4/9$, $t = 10^6$) margins.

4].

Like the saturation model, the bipartite model generates head-to-tail affinities, but they do not concentrate in the corners. In the appendix, we provide explicit upper and lower bounds in addition to the asymptotic statements in Theorem 5.2.

5.2.1 Numerical example

Here we consider an example with parameter $p = 4/9$. We run the simulation until there are a total of $t = 1,000,000$ nodes. An immediate consequence of Theorem 5.2 is that we expect the total number of nodes of degree one in \mathcal{M} to be $tp/(2-p) = 285714.6$ and in the \mathcal{N} to be $tq/(2-q) = 384615.8$. The observed values were 285029 and 385030, respectively.

Figure 5.4 shows the degree distributions of \mathcal{M} and \mathcal{N} , respectively, along with

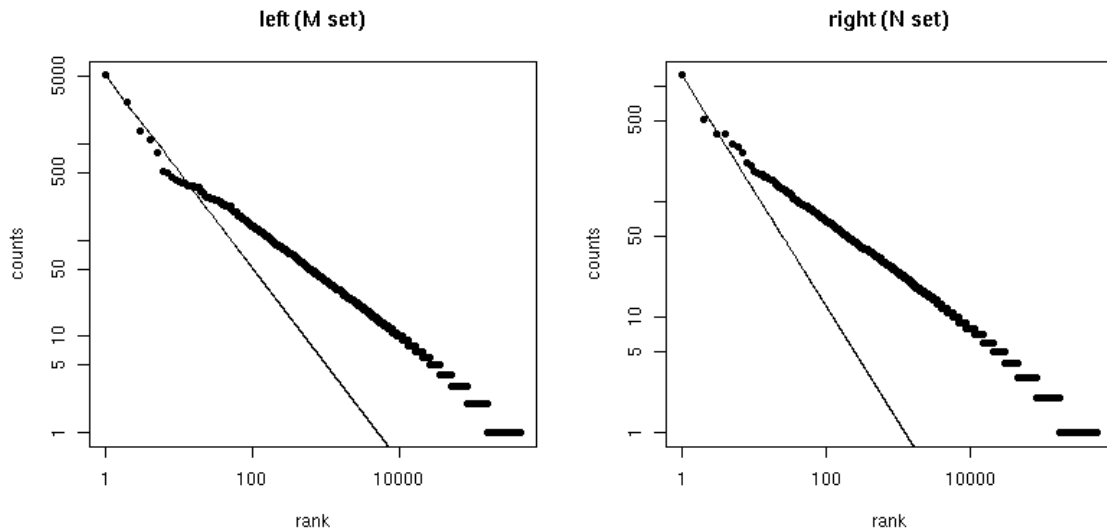


Figure 5.5: Zipf plot of bipartite preferential attachment ($p = 4/9$, $t = 10^6$) margins.

the upper and lower bounds that arise out of the development of the asymptotic statement of Theorem 5.2. See Corollaries A.4 and A.5 of the appendix. We see that these asymptotic bounds become quite tight even for quite small degrees. The theoretical finite- t expectation is also given and we see that it is a near perfect match for small degrees. Note that the expected degree distribution predicts some concavity with respect to a perfect power-law. This is most obvious for small degrees.

Figure 5.5 shows the corresponding Zipf plots. Here we see some degree of concavity at the head of the distribution, somewhat similar to the Snapfish example. Overall, there is an evident power-law behavior.

Figure 5.6 shows the linear-color data display. From this, the large number of single-degree nodes are apparent. There is also a presence of head-to-tail affinities.

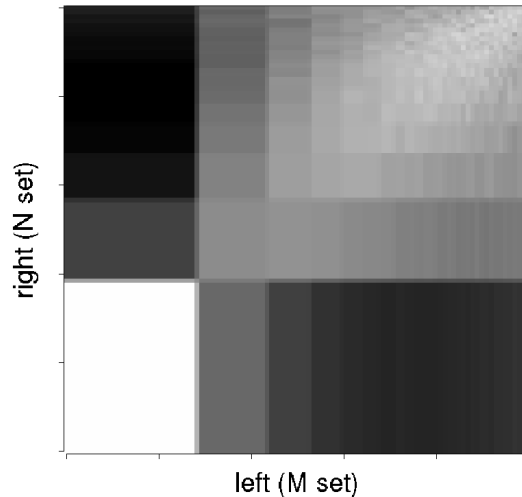


Figure 5.6: Data display for bipartite preferential attachment ($p = 4/9$, $t = 10^6$) example.

5.2.2 Extensions

The model we have considered only has one true parameter (besides the parameter that specifies the number of edges). It is easy to come up with many possible extensions of this model, many of which are also amenable to analysis.

One obvious extension is to introduce two additional parameters, $r_m \in [0, 1]$ and $r_n \in [0, 1]$. These denote the probabilities of choosing between “uniform” and preferential attachments conditional on the selection of the \mathcal{M} or the \mathcal{N} for adding a new node. Thus, the data generation scheme is modified as follows: Draw a $\text{Ber}(p)$ random variable to decide whether to add to \mathcal{M} or \mathcal{N} . Conditional on adding to the \mathcal{M} , choose a $\text{Ber}(r_m)$ to decide whether to attach to \mathcal{N} via uniform or preferential attachment. Uniform attachment here means that we select an existing node in \mathcal{N} to attach to uniformly at random.

It is easy to see that preferential attachment corresponds to selecting an existing edge uniformly at random and then connecting to the corresponding vertex that it is attached to in the appropriate set. Thus, the extension above trades off between uniform attachment of vertices and edges. A recursion equation very similar in nature to that from the simple bipartite preferential attachment model arises and can be solved using essentially the same approach.

One possibility that is excluded from our model is the connection between two new vertices, one from each side of the partition. Simple variations of the basic model can address this as well. Indeed, we could introduce three parameters p_1, p_2 , and p_3 corresponding to a four-outcome multinomial and draw from this multinomial to determine whether to choose an old node or a new node from each side of the partition at every time t . Then, conditional on deciding to pick an old node, we can draw a Bernoulli random variable, as before, to trade off uniform and preferential attachment. This results in a total of five parameters, which introduces some additional flexibility into the model.

The preferential attachment scheme can also be altered. Popular choices for the probability of attachment to a given vertex are to choose proportional to d_i^γ for some power γ or proportional to $d_i + \alpha$ for some $\alpha > -1$. See [20] for more details.

One, perhaps, disturbing aspect of the basic model we have considered is that a large proportion of nodes of degree one are created and it is not possible to make the proportion on both sides of the partition arbitrarily small. One would hope that via some of these simple modifications, such flexibility would be obtained. However, it can be easily shown by solving the necessary recurrence equations that even the five-parameter model is not enough to generate an arbitrary proportion of nodes of

degree one on either side of the partition.

5.2.3 Parameter estimation

Parameter estimation in the bipartite preferential attachment model is exceedingly simple and hardly worth mentioning. Since the number of nodes in \mathcal{M} is a sum of Bernoulli random variables, the uniformly minimum variance unbiased estimator is $\hat{p} = |\mathcal{M}|/t$.

5.3 Bipartite models incorporating temporal preferential attachment

In the previous two sections, we discuss two simple models for generating asymmetric head-to-tail affinities in bivariate copulas arising from entities with a very large number of categorical levels. We focused on saturation and preferential-attachment as the mechanisms for generating these affinities.

One other common feature in data of this type is some form of temporal preferential attachment. For example, a new movie or new song may be more likely to be rated than an old movie or song. We would like a simple extension to our models that incorporates this behavior. Another feature we would like to incorporate is a time-dependent rate of entry of newly observed levels of each entity. In many cases, one observes a large rate of growth initially followed by a successively smaller rate of growth.

We use as a template, one of the proposed extensions of the previous section, but replace uniform attachment with a temporal aspect. Since we are also aiming for a

limited number of parameters, we will combine the temporal preferential attachment and the generation of new nodes of the graph under a single framework. While our model will be conceptually simple to describe, it appears to be surprisingly difficult to analyze. Thus, we will limit ourselves mostly to empirical evaluations, with one of our main objectives being to demonstrate similar affinities to those seen in real data.

5.3.1 Model specification

At each time point t , a multinomial random variable will be drawn to determine whether to select via temporal or degree preferential attachment. That is, we draw C_t according to

	Temporal	Degree
Temporal	π_{tt}	π_{td}
Degree	π_{dt}	π_{dd}

where (Temporal, Degree) indicates that the left entity is chosen via temporal preferential attachment and the right entity is chosen independently from the left via degree preferential attachment. Hence, the entities of each partition are conditionally independent given C_t . If they were chosen completely independently, we would observe a (nearly) uniform copula. However, stratifying by the outcome C_t is enough to generate interesting affinities in the copula estimates.

Preferential attachment by degree is performed as before. Temporal preferential attachment is done as follows: A parametrized distribution is chosen for each partition of the graph. Let $T_m \equiv T_m(t)$ be a positive integer-valued random variable with distribution F_m , drawn at time t . Analogously, we have T_n and F_n for the right partition. Let $S_m(t)$ be the cardinality of \mathcal{M} at time t and let $S_n(t)$ be the corresponding

cardinality of \mathcal{N} . Both quantities are, of course, random. Then if $T_m \leq S_m(t)$, we connect to node T_m on the left. If $T_m > S_m(t)$, then we add a *new* node to \mathcal{M} and connect to it. Temporal preferential attachment to \mathcal{N} is entirely analogous.

Thus, the rate of new nodes is specified, in part, by the tail of the distributions F_m and F_n . In particular,

Proposition 5.1. *Assuming $\pi_{tt} + \pi_{td} > 0$ and $\mathbb{E}T_m = \infty$, we have $S_m(t) \uparrow \infty$ a.s.*

Proof. Let X_1, X_2, \dots be a sequence of i.i.d. r.v's distributed according to F_m . Let $S_0 = 0$ and $S_n = S_{n-1} + \mathbf{1}_{(X_n > S_{n-1})}$. Then, clearly, $S_n \leq n$. Let $A_n = \{X_n > S_{n-1}\} \supset \{X_n > n - 1\} = B_n$. Hence, $\mathbb{P}(B_n \text{ i.o.}) \leq \mathbb{P}(A_n \text{ i.o.})$. But, the B_n are independent and

$$\mathbb{E}X_1 = \sum_n \mathbb{P}(B_n) = \infty,$$

and so $\mathbb{P}(B_n \text{ i.o.}) = \mathbb{P}(A_n \text{ i.o.}) = 1$ by the second Borel–Cantelli lemma. This clearly implies that $S_n \uparrow \infty$ a.s..

Now, to get the conclusion, note that if $\pi_{tt} + \pi_{td} > 0$, then there will (almost surely) be an infinite subsequence of times t_1, t_2, \dots for which a draw from F_m will be made. Let $T_m(t_i) = X_i$, so that $S_m(t_i) = S_i$, and the result follows. \square

5.3.2 Numerical Examples

We consider three numerical examples. Our intent is to demonstrate that we can generate similar features to several observed copulas. The three examples we are motivated by are the Netflix, Snapfish, and arXiv data sets. We choose these because in the first we see head-to-tail affinities, in the second we see primarily a head-to-head affinity and in the last there is a strong tail-to-tail affinity and a broad head-to-head

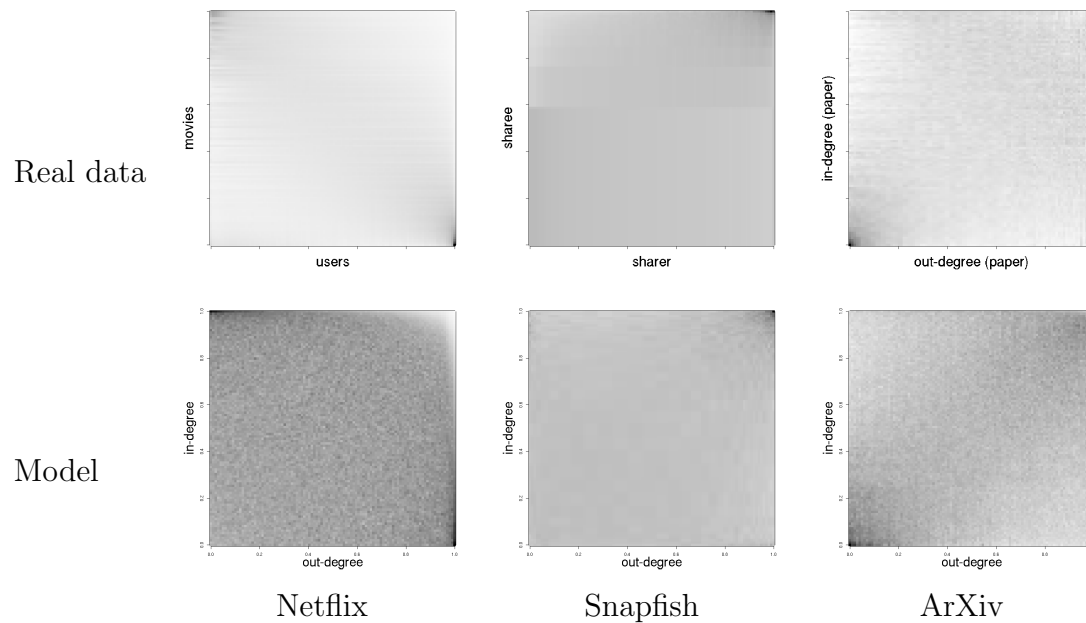


Figure 5.7: Examples of real datasets and their extended-model counterparts.

one.

Figure 5.7 shows the three example copulas from Chapter 2 and three corresponding copula estimates from simulations of the extended model. For all three examples, the temporal distribution is $\text{Zipf}(\alpha)$. For the Netflix example, $\alpha = 1.475$, $\pi_{tt} = 0$ and $\pi_{td} = \pi_{dt} = 0.3$. For the Snapfish example, $\alpha = 1.225$, $\pi_{tt} = 0.4$ and $\pi_{td} = \pi_{dt} = 0.3$. For arXiv, $\alpha = 1.45$, $\pi_{tt} = 0.3$ and $\pi_{td} = \pi_{dt} = 0.15$.

We reiterate that our aim is only to demonstrate the flexibility of generating the larger qualitative features of each copula. For the Netflix example, we see that we can generate head-to-tail affinities with severe depletion in the head-to-head corner of the distribution. To match to the Snapfish example, we have generated a strong head-to-head affinity. For the arXiv dataset, we have produced a strong tail-to-tail affinity along with a broad head-to-head affinity. The parameters were “fit” by trial-and-error

based on the intuition of how each should affect the copula. Because of the complex relationship between the parameters and the features of the copula, sometimes our intuition failed us.

5.4 Summary

We have presented three different generative models for describing the patterns seen in the marginal distributions and the data display. Only a rough resemblance exists between our models and the patterns in real data sets. Nor do the data resemble classical parametric copula densities that we have seen. As noted above a close match would require many more parameters than we have used.

The saturation and bipartite preferential attachment models we have used are conceptually simple and provide some intuition regarding how the various affinities may arise. In addition, they are amenable to mathematical analysis and we have tried to pin down their properties as much as practicable through fairly straightforward techniques.

The intuition gained from these simple models allows for extending them in various ways to generate new affinities. In Section 5.3 we have pursued one of these directions.

Chapter 6

Discrete-choice models

In the previous chapters we have looked at various approaches to estimating, visualizing, and modeling the observed dependence between two types of entities with a large number of categorical levels. For the most part, we have eschewed more classical statistical approaches, in part, because we sought to *understand* as much as *model* the observed dependencies and also because lack of fit of any model reasonably easy to fit should be large since the data collected are so large and complex.

This chapter deals with a somewhat more classical approach via discrete-choice models—with a new twist.

6.1 The model

We are interested in modeling the evolution of large bipartite graphs with links between two subsets of nodes, which we call \mathcal{A} and \mathcal{B} . At time t , we observe a graph $\mathcal{G}_t \subset \mathcal{A} \times \mathcal{B}$.

At times $t = 1, 2, \dots, T$, the t 'th edge (A_t, B_t) arrives and so $\mathcal{G}_t = \mathcal{G}_{t-1} \cup \{(A_t, B_t)\}$.

For each “edge time” t , there is also a corresponding “calendar time” C_t so that edge t arrives at calendar time $C_t \geq C_{t-1}$.

At time t , we have seen only subsets $\mathcal{A}_t \subset \mathcal{A}$ and $\mathcal{B}_t \subset \mathcal{B}$ of the total sets of entities. So, $\mathcal{G}_t \subset \mathcal{A}_t \times \mathcal{B}_t$ and the set on the right-hand side is the smallest product set that contains \mathcal{G}_t .

We seek a model to predict \mathcal{G}_t given \mathcal{G}_{t-1} and some features on the members \mathcal{A}_{t-1} and \mathcal{B}_{t-1} . The new edge at time t could come from any element of $\mathcal{A}_{t-1} \times \mathcal{B}_{t-1}$ or possibly from $(\mathcal{A} \times \mathcal{B}) - (\mathcal{A}_{t-1} \times \mathcal{B}_{t-1})$. However, we assume nothing (including any features) about as-of-yet unobserved members of \mathcal{A} and \mathcal{B} .

The model allows $A_t \in \tilde{\mathcal{A}}_t \equiv \mathcal{A}_{t-1} \cup \{A^*\}$ where the symbol A^* denotes the event that a new value from \mathcal{A} is observed at time t . The analogous quantities and notation for \mathcal{B} are also defined.

We aim to capture phenomena similar to preferential attachment by degree and preferential attachment by age, as well as the arrival rates of new entities. For example, preferential attachment by age appears to be a strong feature in the Netflix data. New raters tend to start with a burst of activity. Bursts can occur at other times as well. New entities are also an important aspect. New customers join the service at some rate that changes with time. New movies also come out and are added to the data base, though this rate appears to be somewhat more constant than the arrival rate of customers.

The model we employ for (A_t, B_t) given the history up to time t is

$$\mathbb{P}(A_t = a_t, B_t = b_t) = \frac{e^{Z_t(a_t, b_t)^T \gamma}}{\sum_{(a, b)} e^{Z_t(a, b)^T \gamma}}, \quad (6.1)$$

where $Z_t(a, b)$ are the vector of features corresponding to the pair (a, b) , γ is a parameter vector and the sum in the denominator is taken over $\tilde{\mathcal{A}}_t \times \tilde{\mathcal{B}}_t$.

The model considered in (6.1) is a form of discrete-choice, multinomial-regression model commonly found in various fields of application, most notably economics and marketing. However, our model is nonstandard in that we think of a as a consumer and b as a product, in which case both the consumer and the product are being “generated” by the process. Notice that the set of candidates changes over time as new entities are observed.

Because the candidate sets for our applications are extremely large, we will focus almost exclusively on parameter vectors $Z_t(a, b)$ that depend elementwise on either a or b , but not both or neither. (The latter is ruled out because the likelihood is easily shown to be invariant to such a parameter.) It is easy to see by examining (6.1) that if $Z_{t,i}(a, b)$ does not depend on either a or b , then the likelihood is invariant to this feature. Thus, we consider features of the form $Z_t(a, b) = (X_t(a), Y_t(b))$ and this yields a separable model

$$\mathbb{P}(A_t = a_t, B_t = b_t) = \frac{e^{X_t(a_t)^\top \alpha}}{\sum_a e^{X_t(a)^\top \alpha}} \frac{e^{Y_t(b_t)^\top \beta}}{\sum_b e^{X_t(b)^\top \beta}}, \quad (6.2)$$

where $\gamma = (\alpha, \beta)$.

The interpretation of such a model is that A_t and B_t are conditionally independent given the past history and the feature vector. This does not mean that the two entity types are unrelated. For example, when we pick a random edge $(a, b) \in \mathcal{G}_t$, the degrees of a and b are not necessarily independent.

Employing such separability lead to significant savings in computation.

6.1.1 The likelihood

The likelihood for the more general model (6.1) is

$$\mathcal{L}(\gamma) = \prod_{t=1}^T \mathbb{P}(A_t = a_t, B_t = b_t) = \prod_{t=1}^T \frac{e^{Z_t(a_t, b_t)^\top \gamma}}{S_t(\gamma)}, \quad (6.3)$$

and the log-likelihood is

$$\ell(\gamma) = \sum_{t=1}^T (Z_t(a_t, b_t)^\top \gamma - \log S_t(\gamma)), \quad (6.4)$$

where

$$S_t(\gamma) = \sum_{(a,b) \in \tilde{\mathcal{A}}_t \times \tilde{\mathcal{B}}_t} e^{Z_t(a,b)^\top \gamma}.$$

To get simple expressions for the gradient and Hessian, define

$$\pi_t(a, b; \gamma) = \mathbb{P}(A_t = a, B_t = b) = \frac{e^{Z_t(a,b)^\top \gamma}}{S_t(\gamma)}$$

and the weighted average

$$\tilde{Z}_t(\gamma) = \sum_{(a,b) \in \tilde{\mathcal{A}}_t \times \tilde{\mathcal{B}}_t} \pi_t(a, b; \gamma) Z_t(a, b; \gamma).$$

Then, the gradient of the log-likelihood is

$$g(\gamma) = \partial \ell(\gamma) / \partial \gamma = \sum_{t=1}^T Z_t(a_t, b_t) - \tilde{Z}_t(\gamma), \quad (6.5)$$

and the Hessian is

$$H(\gamma) = \frac{\partial^2 \ell}{\partial \gamma \partial \gamma^\top} = \sum_{t=1}^T \tilde{Z}_t(\gamma) \tilde{Z}_t(\gamma)^\top - \sum_{(a,b)} \pi_t(a, b; \gamma) Z_t(a, b) Z_t(a, b)^\top. \quad (6.6)$$

For the general model (6.1), the log-likelihood, gradient and Hessian involve nested sums over t , then over (a, b) given t . In general, evaluating such sums requires cost $O(T|\mathcal{A}_T||\mathcal{B}_T|)$. For the full Netflix data, we have about $10^6 \times 5 \times 10^5 \times 1.5 \times 10^4 = 7.5 \times 10^{17}$ operations.

For the separable model (6.2), the total likelihood factors into a product of two terms, one depending on \mathcal{A} and the other on \mathcal{B} , and each having the same general form as (6.3), but where the sums are over one type of entity only. Hence, evaluating the log-likelihood, gradient and Hessian costs $O(T(|\mathcal{A}_T| + |\mathcal{B}_T|))$, a huge savings.

If we are careful about our use of features, we can reduce this further still. In particular, if the totality of $\pi_t(a, b; \gamma)$ can be updated in $O(1)$ work, then the cost is reduced to $O(T)$. We explore this further in the next section.

6.2 Updatable features

We seek classes of features that make updating the likelihood equations “easy”. We will use references to entities in \mathcal{A} through, understanding that everything holds just as well for those of \mathcal{B} .

Clearly, the simplest such class consists of the “constant” features, i.e., features that depend only on the particular $a \in \mathcal{A}$ and not on t . Then, these features are fixed once a entered into \mathcal{A}_t and no further adjustment to the likelihood, gradient, or Hessian is needed.

The second class of features corresponds to those for which only $O(1)$ levels change at time t . Suppose feature i is such a feature. Then

$$S_t(\alpha) = \sum_{a \in \tilde{\mathcal{A}}_t} e^{X_{t,i}(a)\alpha_i} e^{X_{t,(-i)}(a)^\top \alpha_{(-i)}} = S_{t-1}(\alpha) + \sum_{a \in \Delta_{t,i}} e^{X_{t-1,(-i)}^\top \alpha_{(-i)}} (e^{X_{t,i}\alpha_i} - e^{X_{t-1,i}\alpha_i}),$$

where $\Delta_{t,i}$ is the set of entities for which the i th feature changed at time t and $X_{t,(-i)}$ and $\alpha_{(-i)}$ are the vectors of features and coefficients with the i th coordinate removed. If multiple such features exist, then the sum can be performed over $\Delta_t = \cup_i \Delta_{t,i}$ where for each distinct a we update only the necessary coordinates.

The next class of features that are easily updatable are ones that are linear in t . Suppose $X_{t,i}(a) = X_{t-1,i}(a) + \delta$ for all $a \in \mathcal{A}_{t-1}$. Then,

$$S_t(\alpha) = e^{\alpha_i \delta} S_{t-1}(\alpha) + \sum_{a \in \mathcal{A}_t - \mathcal{A}_{t-1}} e^{X_{t,i}\alpha_i},$$

where the second sum is vacuous if $\mathcal{A}_t \subseteq \mathcal{A}_{t-1}$. However, if $X_{t,i} = \delta_i$ for $a \in \mathcal{A}_t - \mathcal{A}_{t-1}$ then no update is needed as the likelihood remains invariant. In either event, because of this invariance, this class essentially reduces to the previous one. We will encounter a similar such a situation in the example considered below. Calendar age and edge age are both of this type. However, due to the invariance, edge age as a feature is simply equivalent to using a feature corresponding to the first observed time for each entity.

6.3 Fitting via maximum likelihood

The main challenge in fitting by maximum likelihood is simply the sheer volume of data that must be processed. Thus, we seek efficient schemes to update the likelihood equations. Once we have these, then a straightforward, though careful, application of Newton's method can be used.

To this end, we introduce the auxiliary sums

$$S^\ell(t) = \sum_{a \in \tilde{\mathcal{A}}_t} e^{X_t(a)^\top \alpha} \quad (6.7)$$

$$S^g(t) = \sum_{a \in \tilde{\mathcal{A}}_t} X_t(a) e^{X_t(a)^\top \alpha} \quad (6.8)$$

$$S^H(t) = \sum_{a \in \tilde{\mathcal{A}}_t} X_t(a) X_t(a)^\top e^{X_t(a)^\top \alpha}, \quad (6.9)$$

where we have suppressed the dependence on α in our notation.

The analysis below combines all three classes of easily updatable features as described in the previous section. That is, we simultaneously allow for

1. Nearly all feature vectors $\{X_t(a) : a \in \mathcal{A}_t\}$ to be updated by a common feature vector δ_t , that does not depend on a .
2. A “small” set of elements of \mathcal{A}_t to have their own “independent” feature-vector updates, i.e., each new feature vector depends on the particular a of interest. (By “small” we mean uniformly bounded in t .)
3. New entities to arise with their own new feature vectors.

Partition $\tilde{\mathcal{A}}_t = \mathcal{Z}_t \cup \Delta_t \cup \mathcal{A}_t^*$, where \mathcal{Z}_t is the set of $a \in \mathcal{A}_t$ such that the feature vector corresponding to a changes only by some fixed amount δ_t , independently of a ;

Δ_t is the set of $a \in \mathcal{A}$ where the change in the features depends on the particular a ; and, \mathcal{A}_t^* is either empty if a_t is not a new entity or is a_t in the case of a newly observed entity.

Then

$$\begin{aligned}
S^\ell(t) &= \sum_{\bar{z}_t} e^{X_t(a)^\top \alpha} + \sum_{\Delta_t} e^{X_t(a)^\top \alpha} + \sum_{\mathcal{A}_t^*} e^{X_t(a)^\top \alpha} \\
&= \sum_{\bar{z}_t} e^{X_{t-1}(a)^\top \alpha} e^{\delta_t^\top \alpha} + \sum_{\Delta_t} e^{X_t(a)^\top \alpha} + \sum_{\mathcal{A}_t^*} e^{X_t(a)^\top \alpha} \\
&= \sum_{\bar{\mathcal{A}}_{t-1}} e^{X_{t-1}(a)^\top \alpha} e^{\delta_t^\top \alpha} - \sum_{\Delta_t} e^{X_{t-1}(a)^\top \alpha} e^{\delta_t^\top \alpha} + \sum_{\Delta_t} e^{X_t(a)^\top \alpha} + \sum_{\mathcal{A}_t^*} e^{X_t(a)^\top \alpha} \\
&= e^{\delta_t^\top \alpha} \left(S^\ell(t-1) - \sum_{\Delta_t} e^{X_{t-1}(a)^\top \alpha} \right) + \sum_{\Delta_t \cup \mathcal{A}_t^*} e^{X_t(a)^\top \alpha},
\end{aligned}$$

hence, if $|\Delta_t| = O(1)$, then the calculation of $S^\ell(T)$ is $O(T)$.

Similar sequential updating of $S^g(t)$ and $S^H(t)$ are also available. Tedious algebra shows that

$$\begin{aligned}
S^g(t) &= e^{\delta_t^\top \alpha} \left(S^g(t-1) - \sum_{\Delta_t} X_{t-1}(a) e^{X_{t-1}(a)^\top \alpha} \right) \\
&\quad + \delta_t e^{\delta_t^\top \alpha} \left(S^\ell(t-1) - \sum_{\Delta_t} e^{X_{t-1}(a)^\top \alpha} \right) \\
&\quad + \sum_{\Delta_t \cup \mathcal{A}_t^*} X_t(a) e^{X_t(a)^\top \alpha}
\end{aligned}$$

and

$$\begin{aligned}
S^H(t) &= e^{\delta_t^\top \alpha} \left(S^H(t-1) - \sum_{\Delta_t} X_{t-1}(a) X_{t-1}(a)^\top e^{X_{t-1}(a)^\top \alpha} \right) \\
&\quad + e^{\delta_t^\top \alpha} \left(S^g(t-1) - \sum_{\Delta_t} X_{t-1}(a) e^{X_{t-1}(a)^\top \alpha} \right) \delta_t^\top \\
&\quad + e^{\delta_t^\top \alpha} \delta_t \left(S^g(t-1) - \sum_{\Delta_t} X_{t-1}(a) e^{X_{t-1}(a)^\top \alpha} \right)^\top \\
&\quad + e^{\delta_t^\top \alpha} \delta_t \delta_t^\top \left(S^\ell(t-1) - \sum_{\Delta_t} e^{X_{t-1}(a)^\top \alpha} \right) \\
&\quad + \sum_{\Delta_t \cup \mathcal{A}_t^*} X_t(a) X_t(a)^\top e^{X_t(a)^\top \alpha}.
\end{aligned}$$

We now turn to the likelihood equations. Adapting (6.4), (6.5), and (6.6) to one component of the separable model, we see that they can be written compactly as

$$\begin{aligned}
\ell(T; \alpha) &= \sum_{t=1}^T (X_t(a_t)^\top \alpha - \log S^\ell(t)) \\
g(T; \alpha) &= \sum_{t=1}^T X_t(a_t) - \sum_{t=1}^T \frac{S^g(t)}{S^\ell(t)} \\
H(T; \alpha) &= \sum_{t=1}^T \frac{S^g(t) S^g(t)^\top}{(S^\ell(t))^2} - \sum_{t=1}^T \frac{S^H(t)}{S^\ell(t)}.
\end{aligned}$$

Since each of the sums has cost $O(T)$ under the assumption that $|\Delta_t| = O(1)$, then we see that each Newton's step also has cost $O(T)$. This yields a very efficient means for solving the maximum likelihood equations.

6.4 Sampling from the fitted model

Having fit the parameters for the model, it is often desirable to be able to sample from it. Since our model essentially consists of a vector of probabilities, this is nominally an issue of sampling from a discrete probability distribution.

However, our problem is unique in at least a couple of aspects. First of all, the number of outcomes with strictly positive probability of occurrence is growing as we sample. The number of such outcomes is $|\tilde{\mathcal{A}}_t|$. Furthermore, at each time point, not only are the number of outcomes potentially growing, but the probability of each is changing as the covariates change.

Note that $\pi_t(a; \hat{\alpha}) = e^{X_t(a)^T \hat{\alpha}} / S^\ell(t; \hat{\alpha})$ is the probability of each outcome in $\tilde{\mathcal{A}}_t$. From the previous section, we know that, if we keep track of $S^\ell(t; \hat{\alpha})$, then we can update all of the $\pi_t(a; \hat{\alpha})$ in $O(1)$ time.

Hence, we generalize a bit and focus on the problem of sampling from a discrete distribution with $p_i > 0$ for $1 \leq i \leq n$ and $\sum_i p_i = 1$, where we assume n to be “large”.

A “naïve” sampler would work roughly as follows.

1. Form the cumulative sums $c_i = \sum_{j=1}^i p_j$.
2. Draw a standard uniform random number U .
3. Return $X = \max\{i : c_i \leq U\}$.

If the p_i 's do not change, then the c_i 's can be precomputed and stored. To make the sampling more efficient, the p_i 's can also be sorted in decreasing order before the cumulative sums are calculated. One can also construct a Huffman tree and perform

a binary search over it. More efficient schemes exist for the case where the p_i 's do not change. See [17, Ch. 3].

The complication that arises in our model is that the individual probabilities p_i are changing as a function of time and we do not know a priori which ones will change nor at what time. The “naïve” method is still a candidate, but we have to update $O(n)$ cumulative probabilities on average every time any of the p_i 's change. Schemes which involve considerable setup time, such as maintaining sorted lists or binary trees seem inefficient as well.

Here, we propose two sampling methods that, when combined, turn out to be surprisingly useful.

6.4.1 Method 1: Binary search with cumulative updates

This first method attempts to handle the issue of changing probabilities, while at the same time utilizing binary search where possible, to reduce the number of necessary comparisons.

Suppose we have n strictly positive probabilities p_i that sum to unity. We also have an accompanying cumulative probability array of c_i 's and a pointer x into the cumulative probability array. We begin with $x = 0$.

Each time we change one of the p_i , we check to see if $x \geq i$. If so, then we set $x = i$ and $c_x = c_{x-1} + p_i$, where p_i is the new probability value at index i . Otherwise, we leave x unchanged.

The sampling then proceeds as follows.

1. Draw a uniform random variable U .
2. If $U \leq c_x$, perform binary search on the subvector (c_1, \dots, c_x) to identify X .

Return this value.

3. If $U > c_x$, then, starting at c_x , update the cumulative sums until finding the maximum x' such that $c_{x'} \leq U$. Set $X = x = x'$ and return X .

Notice that the sampling scheme essentially performs a lazy update of the cumulative probability vector, i.e., it updates the index x every time it is necessary. Thus, if the probabilities p_i never change, the algorithm “converges” to the binary-search algorithm. If the probabilities change, then we keep track of the largest possible index x for which the cumulative probabilities (c_1, \dots, c_x) are still valid. For our particular example, replacing a naive sampler with binary search with cumulative updates, resulted in the simulation running about 8 times faster.

6.4.2 Method 2: Modular storage binning

The previous method reduces the average number of comparisons considerably with respect to the naïve sampler. A simple modification provides another considerable speedup as the number of potential outcomes n grows large. It is inspired by the method of guide tables [17, Ch. 3], as we now see.

Fix a number of bins B . For each bin, we maintain a bin total $b_i \geq 0$ such that $\sum_{i=1}^B b_i = 1$. Also associated with each bin i is a probability vector $(p_{i,1}, \dots, p_{i,n_i})$, where the number of elements in the probability vector of each bin will change (i.e., grow) with time.

Now, suppose that at time t , a new entity is observed and it is the j 'th new entity. We associate it with the bin $i = j \pmod{B}$ and update the bin total b_i by adding in the new probability p_j . We also add p_j to the probability vector associated with bin i .

Sampling then proceeds as follows.

1. Draw a uniform random variate U .
2. Using sampling Method 1 of the previous section, find the bin i to which U corresponds.
3. Within the probability vector associated with bin i , use Method 1 to locate the element j to which U corresponds.
4. Set $X = (j - 1)B + i$ and return X .

To keep the bin totals current, whenever p_i is changed, then we update the corresponding bin total for the bin $i \pmod{B}$ to which it belongs and also update the corresponding probability vector within the bin.

When sampling repeatedly from our discrete-choice model, we do not know the precise number of observed entities at any point in time. However, we may be able to get a rough estimate of the average by inspecting the fitted parameter values. This can aid us in selecting an “optimal” number of bins B . In our particular application, we also have the fortune that smaller values of i will tend to have higher p_i and so within each bin, the probabilities will tend to be somewhat ordered from largest to smallest.

More sophisticated binning strategies could also be used where we try to maintain the distribution across bins as close to uniform as possible. However, this makes determining the value of the discrete random variate a little more difficult as we must also store the index of each probability alongside its actual value.

6.5 Netflix example

This section is devoted to examining a simple model for the Netflix data. We aim to show the usefulness of the model and the attendant fitting procedures, while at the same time cautioning the reader that the model selected does not do a great job of capturing many of the features of the Netflix data, in particular, the shapes of dependence observed in the empirical copula. We discuss possible reasons for this at the end of the section. Before that, though, we focus on the “successes” of the approach, i.e., the ability to even fit such a model to such large data and the reasonable performance in estimating the rates of arrival of customers and movies.

6.5.1 Data preprocessing

We are interested in modeling rates of arrival and attempting to model dependence of the observed Netflix graph \mathcal{G}_t . To reduce bias in the modeling of arrival rates, we begin by removing any movie or user that appears within the first year of the data. This should eliminate artifacts due to, for example, the movie *Casablanca* appearing in the data from the outset (due to it being more than a half-century old).

This simple step reduces the number of observations from about 10^8 to about 5.5×10^7 .

6.5.2 Modeling arrival rates

Our basic strategy will be to model hierarchically and then fit each piece of the likelihood separately. This will allow for a “parallelization” of the parameter fitting, which

we exploit computationally by fitting each piece on a different core of a multicore processor.

The highest level of the hierarchy that we consider involves stratifying the observations into classes according to the “type” of arrival. Specifically, each observation (A_t, B_t) can be classified according to whether A_t and/or B_t have previously been observed. Hence, we obtain four classes: (old customer, old movie), (old customer, new movie), (new customer, old movie) and (new customer, new movie).

Let $\mathcal{C} = \{(\text{old}, \text{old}), (\text{old}, \text{new}), (\text{new}, \text{old}), (\text{new}, \text{new})\}$ be the set of classes and C_t be the class of the t 'th observation. The total likelihood is then partitioned according to class, so that

$$\mathcal{L} = \prod_{c \in \mathcal{C}} \prod_{t \in \{i: C_i = c\}} \mathbb{P}(A_t = a_t, B_t = b_t \mid C_t = c) \mathbb{P}(C_t = c).$$

Let t be the edge time and $T_c \equiv T_c(t)$ be the calendar time of each observation. Then, we model the probability of occurrence of class c at time t by

$$\begin{aligned} \log \mathbb{P}(C_t = (\text{new}, \text{new})) &\propto \mu_{00} + \alpha_{00} \log t + \beta_{00} \log T_c \\ \log \mathbb{P}(C_t = (\text{new}, \text{old})) &\propto \mu_{01} + \alpha_{01} \log t + \beta_{01} \log T_c \\ \log \mathbb{P}(C_t = (\text{old}, \text{new})) &\propto \mu_{10} + \alpha_{10} \log t + \beta_{10} \log T_c \\ \log \mathbb{P}(C_t = (\text{old}, \text{old})) &\propto \mu_{11} + \alpha_{11} \log t + \beta_{11} \log T_c. \end{aligned}$$

In order for the model to be identifiable, we take $\mu_{00} = \alpha_{00} = \beta_{00} = 0$. Since we expect different rates of arrival at different times, it makes sense to model with respect to both $\log t$ and $\log T_c$, though one would expect them to be highly correlated, especially if the number of elements added per calendar “day” is roughly constant.

	μ	α	β
(new,new)	0.0	0.0	0.0
(new,old)	-1.8268	1.0506	-0.8707
(old,new)	-1.3123	0.5380	-0.3085
(old,old)	-8.0465	0.9745	0.8324

Table 6.1: Fitted coefficients of rate model for Netflix data.

The fitted coefficients for the Netflix data can be found in Table 6.1.

Figure 6.1 shows the approximate rates of arrival of each class as a function of time along with the fit from the model just described. The estimates of the rates of arrival were obtained by averaging contiguous blocks of 500 edges at a time to give localized point estimates for each class. The fitted probabilities from the model were then superimposed. From the plots, we can see that, while there is a statistically significant lack of fit, the model captures the main trends in the data, including the curvature.

6.5.3 Modeling entity selection conditional on edge class

We now turn to modeling the conditional portion of the likelihood $\mathbb{P}(A_t = a_t, B_t = b_t \mid C_t = c)$. We choose a simple model to incorporate preferential attachment by degree and by time. Because we assume separability, we have

$$\mathbb{P}(A_t = a_t, B_t = b_t \mid C_t = c) = \mathbb{P}(A_t = a_t \mid C_t = c)\mathbb{P}(B_t = b_t \mid C_t = c),$$

and, so, we can fit each piece individually. Depending on the class, either one or both of the observed entities may be new at time t . For these pieces, no further modeling is necessary since that portion of the likelihood is captured by the rate

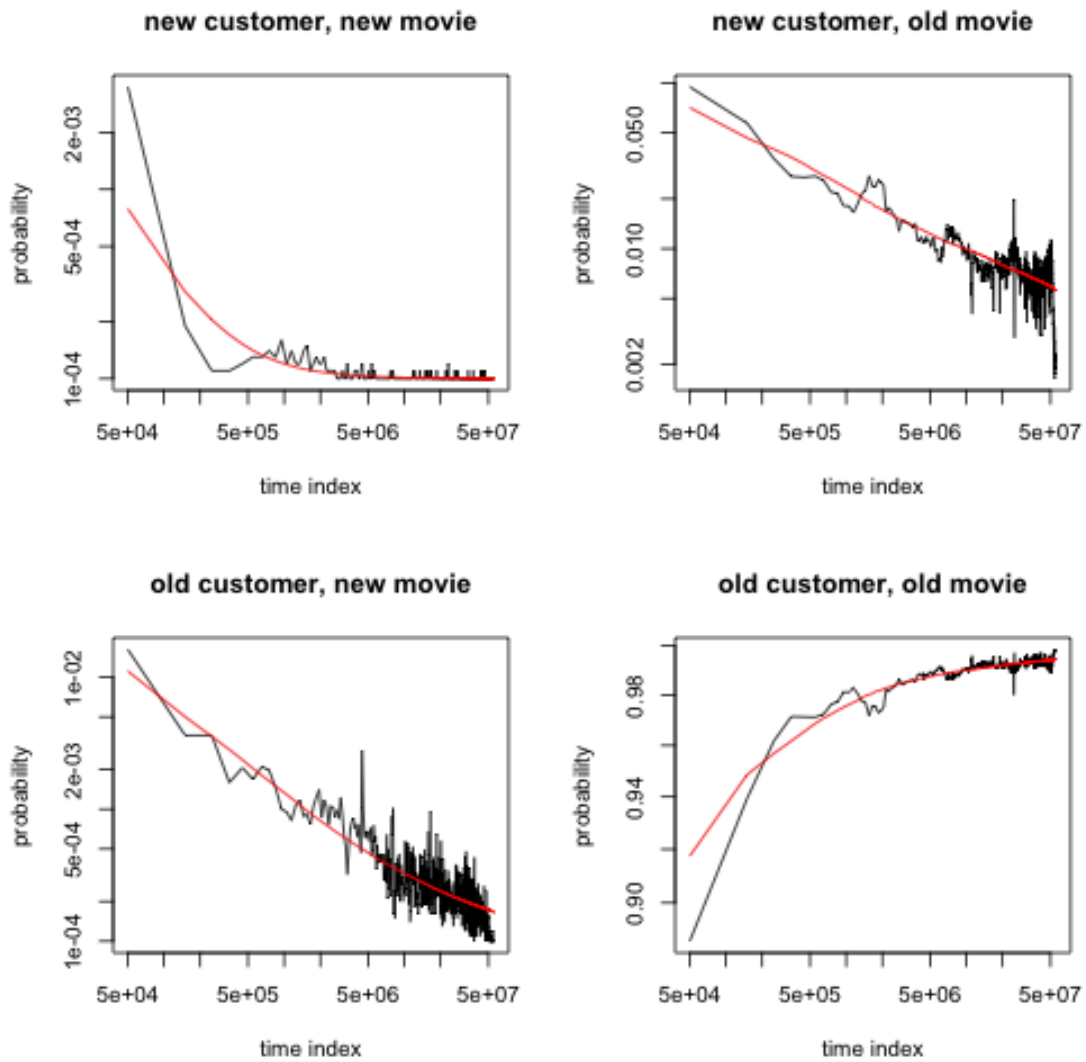


Figure 6.1: Empirical and fitted rates of arrival for each of the four classes for the Netflix example. Plots are on a log–log scale.

modeling. For example, for $C_t = c = \{(\text{new customer, old movie})\}$, there is nothing left to model about A_t since we know $A_t = A_t^*$ and so we are left only with the piece $\mathbb{P}(B_t = b_t \mid C_t = c)$.

Thus, the remaining portions to be modeled are

1. Probability of selection for each customer given that the observed edge class is (old customer, new movie),
2. Probability of selection for each movie given that the observed edge class is (new customer, old movie),
3. Probability of selection for each customer given that the observed edge class is (old customer, old movie), and
4. Probability of selection for each movie given that the observed edge class is (old customer, old movie).

Each portion of the likelihood is modeled as

$$\log \mathbb{P}(A_t = a \mid C_t = c) \propto \gamma_c(t - t_0(a)) + \delta_c \log d_t(a).$$

From the model, it should be clear that there is a component corresponding to preferential attachment by degree with exponent δ_c and a temporal preferential attachment component proportional to $e^{\gamma_c(t-t_0(a))}$. Here $t_0(a)$ denotes the time at which the a 'th entity was first observed, and so the feature is the “age” of a in edge time. As mentioned previously, due to invariance of the likelihood with respect to linear features, this is equivalent to choosing based on the feature $t_0(a)$, but the interpretation in terms of age is more direct.

	γ	δ
(new,old) movie	-1.0158×10^{-7}	1.1434
(old,new) customer	-5.8210×10^{-8}	2.2785
(old,old) customer	-8.6458×10^{-8}	0.6861
(old,old) movie	-7.7993×10^{-8}	0.9687

Table 6.2: Fitted coefficients of conditional entity selection model for Netflix data.

The fitted coefficients are found in Table 6.2. From the estimated coefficients, it appears that the temporal effect is strongest for movies when considering a new customer who shows up and rates a previously existing movie. The preferential attachment by degree component is strongest for existing customers who rate new movies. That is, more active raters appear much more likely to rate a new movie than less active raters. Classical preferential attachment would have a coefficient $\delta = 1$, so we see that choosing proportional to $d_t(a)^{2.2785}$ is much more strongly weighted towards the active raters. In what follows, we will see that this strong preferential attachment effect induces an artifact in the Zipf plots.

6.5.4 Simulating from the model

With the coefficient estimates given in the previous sections, we then simulated from the model until we obtained the same number of observations as in the original Netflix data filtered as described above. This resulted in about 55 million edges simulated from the model.

Figure 6.2 shows the corresponding Zipf plots from the simulated and actual data. The “in-degree” (i.e., movies) plots have similar shape and somewhat similar scale. The “most popular movie” from the simulated data appears to only have about a quarter of the raters that the actual Netflix data does, and the slope at the head

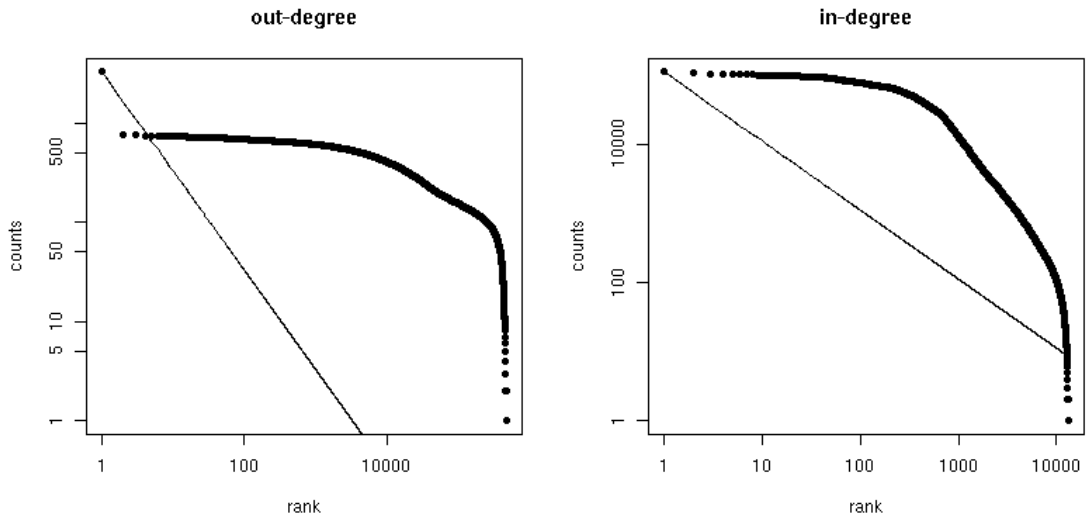
appears more shallow.

The “out-degree” (i.e., customers) plots are somewhat more disparate. In particular, a curious anomaly is that the most active “customer” in the simulated data is much more active than the second-most, yet substantially less active than the corresponding raters in the actual Netflix data. The Netflix data has five raters that are substantially more active than the rest.

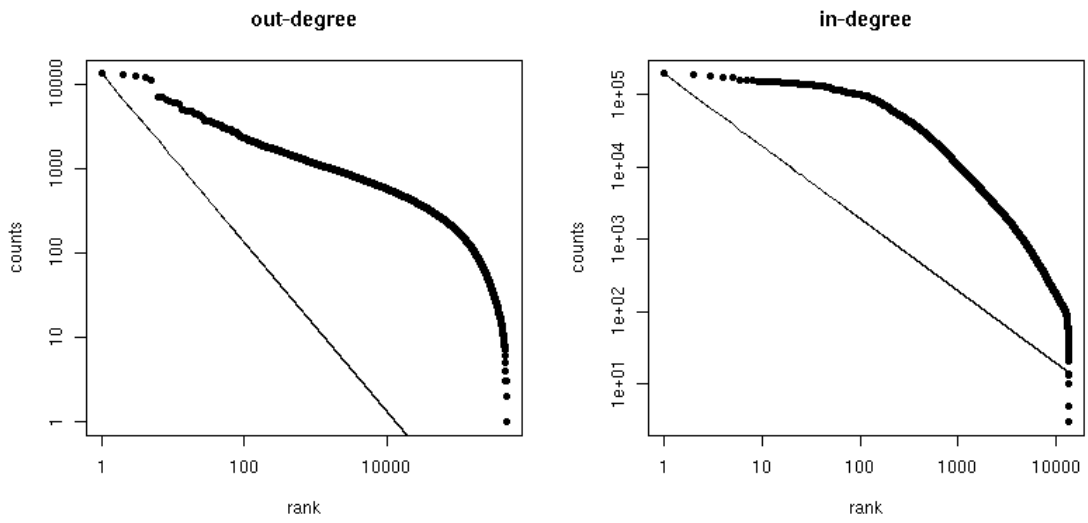
A close examination of the simulation reveals that the vast majority of the “excess” ratings for the most active rater come from the (old customer, new movie) conditional portion. This makes sense, since the coefficient corresponding to preferential attachment by degree is so large. Further testing revealed that this coefficient was actually fit correctly, but there appears to be a lack of fit to the actual data.

Figure 6.3 give the copula plots for the simulated data along with their counterparts from the actual Netflix data for comparison. As we can see the conditionally independent sampling model does not give the independence copula. In addition, the fitted model yields head-to-head and tail-to-tail affinities, instead of the head-to-tail affinities observed in the Netflix data. By observing the copulas at different time points across the simulation, it was seen that the Netflix copulas were quite stable, while the copulas for the simulated data did not stabilize until after about 30 million edges were observed. We hypothesize that because (old,old) entities dominate the total number of observations, then this is the majority of the structure we see in the simulated data.

Because these entities are chosen conditionally independent of one another, we might expect to see the independence copula, or some close analog of it. This would be true if the probability distributions were not changing as a function of time. However,

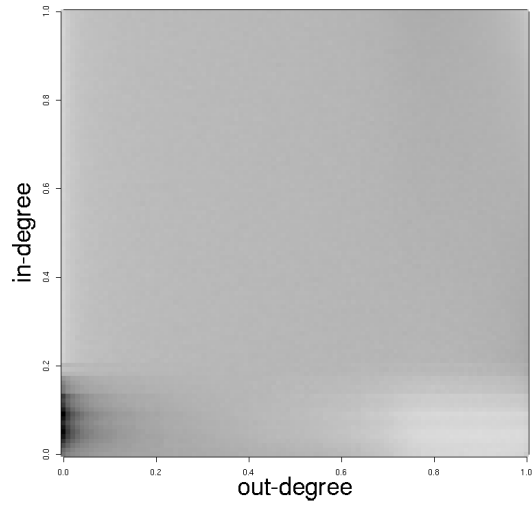


(a) Zipf plots from simulated model

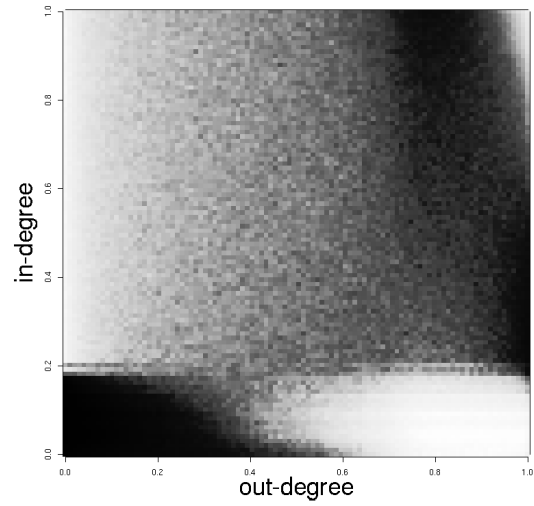


(b) Zipf plots from actual Netflix data

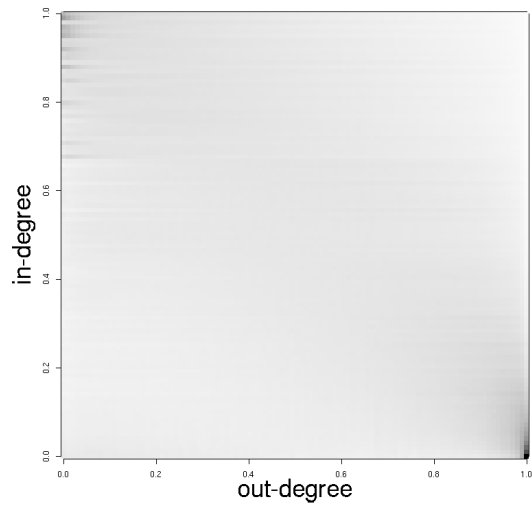
Figure 6.2: Zipf plots from simulated and real Netflix data.



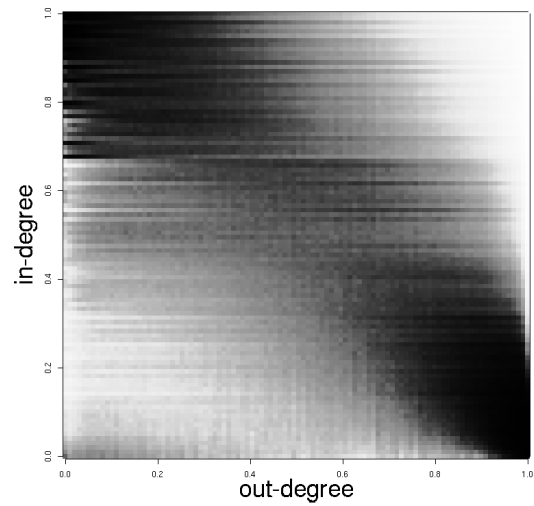
(a) Linearly scaled copula from simulation



(b) Renormalized copula from simulation



(c) Linearly scaled copula from actual data



(d) Renormalized copula from actual data

Figure 6.3: Copula plots from simulated and real Netflix data.

since “customers” and “movies” have associated probabilities that change each time a new edge is made to one or the other, we should not expect to see the independence copula. Still, the head-to-head and tail-to-tail affinities observed do not appear to have a simple, straightforward explanation.

We suspect that the (old customer, old movie) conditional likelihood requires a richer model, perhaps one that is not fully separable. One idea would be to pick a rater conditionally independent of the movie and then select the movie they rate based on some nonseparable feature which links the movie and the rater. If chosen appropriately, this might give rise to the proper head-to-tail affinities seen in the actual data.

6.6 Summary

In this chapter, we developed a discrete-choice model where both the consumer and product are jointly modeled. We also depart from traditional such models by letting the choice set increase as a function of time and by attempting to model the rate of arrival of new observations according to a predefined class. We develop an approach to maximum-likelihood estimation of the model parameters even when the number of observations are in the tens of millions and the total choice set involves over a billion possible (rater, movie) pairs. This is achieved by exploiting the structure inherent in models with separable and easily updatable features. To help sample from the model efficiently, we have described two sampling methods for discrete distributions with probability masses that evolve in time. The model was successfully fit to the Netflix data; though, it did not appear to adequately capture the main dependence structure in the observed copula.

Appendix A

Proofs

The proofs make use of some bounds on Poisson probabilities from the literature, collected here. We also use the difference of two independent Poisson random variables, which has a Skellam [64] distribution.

If $Y \sim \text{Poi}(\lambda)$ and $t \geq \lambda - 1$ then Klar [35] shows that

$$\mathbb{P}(Y \geq t) \leq \left(1 - \frac{\lambda}{t+1}\right)^{-1} \frac{e^{-\lambda} \lambda^t}{t!}. \quad (\text{A.1})$$

A classic result of Teicher [65] is that

$$\mathbb{P}(Y \leq \lambda) \geq \exp(-1) \quad (\text{A.2})$$

when $Y \sim \text{Poi}(\lambda)$. Also, if $Y \sim \text{Poi}(\lambda)$, then

$$\sup_{-\infty < t < \infty} \left| \mathbb{P}\left(\frac{Y - \lambda}{\sqrt{\lambda}} \leq t\right) - \Phi(t) \right| \leq \frac{0.8}{\sqrt{\lambda}}, \quad (\text{A.3})$$

where Φ is the standard normal CDF. Equation (A.3) follows by specializing a Berry-Esseen result for compound Poisson distributions [49, Theorem 1] to the case of a Poisson distribution.

We will also use Gautschi's inequality [29] on the Gamma function,

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s} \quad (\text{A.4})$$

which holds for $x > 0$ and $0 < s < 1$.

A.1 Proof of Theorem 2.1

Theorem 2.1 has three claims. The first follows from Corollary A.1 below. Adding in Lemma A.4 gives the second. The third claim follows from Corollary A.2. For clarity of presentation, we prove each piece of Theorem 2.1 in a separate subsection.

A.1.1 Correct relative ordering, equation (2.1)

The difference of two independent Poisson random variables has a [64] distribution. We begin with a Chernoff bound for the Skellam distribution.

Lemma A.1. *Let $Z = X - Y$ where $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\nu)$ are independent and $\lambda \geq \nu$. Then*

$$\mathbb{P}(Z \leq 0) \leq \exp(-(\sqrt{\lambda} - \sqrt{\nu})^2). \quad (\text{A.5})$$

Proof. Let $\varphi(t) = \lambda e^{-t} + \nu e^t$. Then φ is a convex function attaining its minimum at $t^* = \log(\sqrt{\lambda/\nu}) \geq 0$, with $\varphi(t^*) = 2\sqrt{\lambda\nu}$. Using the Laplace transform of the Poisson

distribution

$$m(t) \equiv \mathbb{E}(e^{-tZ}) = e^{\lambda(e^{-t}-1)} e^{\nu(e^t-1)} = e^{-(\lambda+\nu)} e^{\varphi(t)}.$$

For $t \geq 0$, Markov's inequality gives $\mathbb{P}(Z \leq 0) = \mathbb{P}(e^{-tZ} \geq 1) \leq \mathbb{E}(e^{-tZ})$. Taking $t = t^*$ yields (A.5). \square

Lemma A.2. *Let X_i be sampled from the Zipf–Poisson ensemble. Then for $n \geq 2$,*

$$\mathbb{P}(X_1 > X_2 > \cdots > X_n) \geq 1 - n \exp\left(-\frac{N\alpha^2}{4} n^{-\alpha-2}\right). \quad (\text{A.6})$$

Proof. By Lemma A.1 and the Bonferroni inequality, the probability that $X_{i+1} \geq X_i$ holds for any $i < n$ is no more than

$$\sum_{i=1}^{n-1} \exp(-(\sqrt{\lambda_i} - \sqrt{\lambda_{i+1}})^2) = \sum_{i=1}^{n-1} \exp(-N(\sqrt{\theta_i} - \sqrt{\theta_{i+1}})^2). \quad (\text{A.7})$$

For $x \geq 1$, let $f(x) = x^{-\alpha/2}$. Then $|\sqrt{\theta_i} - \sqrt{\theta_{i+1}}| = |f(i) - f(i+1)| = |f'(z)|$ for some $z \in (i, i+1)$. Because $|f'|$ is decreasing, (A.7) is at most $n \exp(-Nf'(n)^2)$, establishing (A.6). \square

Now we can establish the first claim in Theorem 2.1.

Corollary A.1. *Let X_i be sampled from the Zipf–Poisson ensemble. Choose $n = n(N) \geq 2$ so that $n \leq (AN/\log(N))^{1/(\alpha+2)}$ holds for all large enough N where $A = \alpha^2(\alpha+2)/4$. Then*

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 1.$$

Proof. For large enough N we let $n = (A_N N / \log(N))^{1/(\alpha+2)}$ for $A_N \leq A$. Then

$$\begin{aligned} n \exp\left(-\frac{N\alpha^2}{4} n^{-\alpha-2}\right) &= \left(\frac{A_N N}{\log(N)}\right)^{1/(\alpha+2)} N^{-\alpha^2/(4A_N)} \\ &= \left(\frac{A_N N}{\log(N)}\right)^{1/(\alpha+2)} N^{-\alpha^2/(4\alpha^2(\alpha+2)/4)} \\ &\leq \left(\frac{A}{\log(N)}\right)^{1/(\alpha+2)} \\ &\rightarrow 0. \end{aligned}$$

The proof then follows from Lemma A.2. \square

A.1.2 Correct absolute ordering, equation (2.2)

For the second claim in Theorem 2.1 we need to control the probability that one of the entities X_i from the tail given by $i > n$, can jump over one of the first n entities. Lemma A.3 bounds the probability that an entity from the tail of the Zipf–Poisson ensemble can jump over a high level τ .

Lemma A.3. *Let X_i for $i \geq 1$ be from the Zipf–Poisson ensemble with parameter $\alpha > 1$. If $\tau \geq \lambda_n$ then*

$$\mathbb{P}\left(\max_{i>n} X_i > \tau\right) \leq \frac{N^{1/\alpha}}{\alpha} \frac{\tau+1}{\tau+1-\lambda_n} \frac{\tau^{-1/\alpha}}{\tau-1/\alpha}. \quad (\text{A.8})$$

Proof. First, $\mathbb{P}(\max_{i>n} X_i > \tau) \leq \sum_{i=n+1}^{\infty} \mathbb{P}(X_i > \tau)$ and then from (A.1)

$$\mathbb{P}\left(\max_{i>n} X_i > \tau\right) \leq \left(1 - \frac{\lambda_n}{\tau+1}\right)^{-1} \sum_{i=n+1}^{\infty} \frac{e^{-\lambda_i} \lambda_i^\tau}{\Gamma(\tau+1)}.$$

Now $\lambda_i = Ni^{-\alpha}$. For $i > n$ we have $\tau > \lambda_i = Ni^{-\alpha}$. Over this range, $e^{-\lambda_i} \lambda_i^\tau$ is an

increasing function of λ . Therefore,

$$\begin{aligned} \sum_{i=n+1}^{\infty} e^{-\lambda_i} \lambda_i^\tau &\leq \int_n^{\infty} e^{-Nx^{-\alpha}} (Nx^{-\alpha})^\tau dx \\ &\leq \frac{N^{1/\alpha}}{\alpha} \int_0^{Nn^{-\alpha}} e^{-y} y^{\tau-1/\alpha-1} dy \\ &\leq \frac{N^{1/\alpha}}{\alpha} \Gamma(\tau - 1/\alpha). \end{aligned}$$

As a result

$$\mathbb{P}\left(\max_{i>n} X_i > \tau\right) \leq \frac{N^{1/\alpha}}{\alpha} \frac{\tau + 1}{\tau + 1 - \lambda_n} \frac{\Gamma(\tau - 1/\alpha)}{\Gamma(\tau + 1)}.$$

Now

$$\frac{\Gamma(\tau - 1/\alpha)}{\Gamma(\tau + 1)} = \frac{\Gamma(\tau + 1 - 1/\alpha)}{\Gamma(\tau + 1)} \frac{1}{\tau - 1/\alpha} < \frac{\tau^{-1/\alpha}}{\tau - 1/\alpha}$$

by Gautschi's inequality (A.4), with $s = 1 - 1/\alpha$, establishing (A.8). \square

For an incorrect ordering to arise, either an entity from the tail exceeds a high level, or an entity from among the first n is unusually low. Lemma A.4 uses a threshold for which both such events are unlikely, establishing the second claim (2.2) of Theorem 2.1.

Lemma A.4. *Let X_i for $i \geq 1$ be from the Zipf–Poisson ensemble with parameter $\alpha > 1$. Let $n(N)$ satisfy $n \geq (AN/\log(N))^{1/(\alpha+2)}$ for $0 < A < A(\alpha) = \alpha^2(\alpha + 2)/4$. Let $m \leq (BN/\log(N))^{1/(\alpha+2)}$ for $0 < B < A$. Then*

$$\lim_{N \rightarrow \infty} \mathbb{P}\left(\max_{i>n} X_i \geq X_m\right) = 0.$$

Proof. For any threshold τ ,

$$\mathbb{P}\left(\max_{i>n} X_i \geq X_m\right) \leq \mathbb{P}\left(\max_{i>n} X_i > \tau\right) + \mathbb{P}(X_m \leq \tau). \quad (\text{A.9})$$

The threshold we choose is $\tau = \sqrt{\lambda_m \lambda_n}$ where $\lambda_i = \mathbb{E}(X_i) = Ni^{-\alpha}$.

Write $n = (A_N N / \log(N))^{1/(\alpha+2)}$ and $m = (B_N N / \log(N))^{1/(\alpha+2)}$ for $0 < B_N < B < A_N < A < A(\alpha)$. Then $\tau = \sqrt{\lambda_m \lambda_n} = N(C_N N / \log(N))^{-\alpha/(\alpha+2)}$ where $C_N = \sqrt{A_N B_N}$. Therefore

$$\tau = O\left(N^{2/(\alpha+2)} (\log(N))^{\alpha/(\alpha+2)}\right).$$

By construction, $\tau > \lambda_n$ and so by Lemma A.3

$$\mathbb{P}\left(\max_{i>n} X_i > \tau\right) \leq \frac{N^{1/\alpha}}{\alpha} \frac{\tau + 1}{\tau + 1 - \lambda_n} \frac{\tau^{-1/\alpha}}{\tau - 1/\alpha}.$$

Because $\lambda_n/\tau = (B_N/A_N)^{\alpha/(2\alpha+4)}$, we have $(\tau + 1)/(\tau + 1 - \lambda_n) = O(1)$. Therefore

$$\mathbb{P}\left(\max_{i>n} X_i > \tau\right) = O(N^{1/\alpha} \tau^{-1/\alpha-1}) = O(N^{-1/(\alpha+2)} (\log(N))^{(\alpha+1)/(\alpha+2)})$$

and so the first term in (A.9) tends to 0 as $N \rightarrow \infty$.

For the second term in (A.9), notice that X_m has mean λ_m and standard deviation

$\sqrt{\lambda_m}$. Letting $\rho = \alpha/(\alpha + 2)$ and applying Chebychev's inequality, we find that

$$\begin{aligned} \mathbb{P}(X_m \leq \tau) &\leq \frac{\lambda_m}{(\tau - \lambda_m)^2} \\ &= \frac{B_N^\rho}{(B_N^\rho - C_N^\rho)^2} N^{-2/(\alpha+2)} (\log(N))^{-\rho} \\ &\leq \frac{1}{(A^{\rho/2} - B^{\rho/2})^2} N^{-2/(\alpha+2)} (\log(N))^{-\rho} \\ &\rightarrow 0 \end{aligned}$$

as $N \rightarrow \infty$. □

A.1.3 Limit to correct ordering, equation (2.3)

While we can get $(AN/\log(N))^{1/(\alpha+2)}$ entities properly ordered, there is a limit to the number of correctly ordered entities. We cannot get above $CN^{1/(\alpha+2)}$ correctly ordered entities, asymptotically. That is, the logarithm cannot be removed. We begin with a lower bound on the probability of a wrong ordering for two consecutive entities.

Lemma A.5. *Let X_i be from the Zipf–Poisson ensemble with $\alpha > 1$. Suppose that $AN^{1/(\alpha+2)} \leq i < i + 1 \leq BN^{1/(\alpha+2)}$ where $0 < A < B < \infty$. Then for large enough N ,*

$$\mathbb{P}(X_{i+1} \geq X_i) \geq \frac{1}{3} \Phi\left(-\alpha \frac{A^{\alpha/2}}{B^{\alpha+1}}\right).$$

Proof. First $\mathbb{P}(X_{i+1} \geq X_i) \geq \mathbb{P}(X_{i+1} > \lambda_i) \mathbb{P}(X_i \leq \lambda_i) \geq \mathbb{P}(X_{i+1} > \lambda_i)/e$ using Teicher's inequality (A.2). Next

$$\mathbb{P}(X_{i+1} > \lambda_i) = 1 - \mathbb{P}(X_{i+1} \leq \lambda_i) \geq \Phi\left(\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}}\right) - \frac{0.8}{\sqrt{\lambda_{i+1}}}.$$

Now,

$$\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}} = \sqrt{N} \frac{(i+1)^{-\alpha} - i^{-\alpha}}{\sqrt{(i+1)^{-\alpha}}} = -\alpha\sqrt{N} \frac{(i+\eta)^{-\alpha-1}}{\sqrt{(i+1)^{-\alpha}}}$$

for some $\eta \in (0, 1)$. Applying the bounds on i ,

$$\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}} \geq -\alpha\sqrt{N} \frac{(N^{1/(\alpha+2)}A)^{\alpha/2}}{(N^{1/(\alpha+2)}B)^{\alpha+1}} = -\alpha \frac{A^{\alpha/2}}{B^{\alpha+1}}.$$

Finally, letting $N \rightarrow \infty$ we have $\lambda_{i+1} \rightarrow \infty$ and so $0.8/\sqrt{\lambda_{i+1}}$ is eventually smaller than $(1-e/3)\Phi(-\alpha A^{\alpha/2}B^{-\alpha-1})$. Letting $\theta = -\alpha A^{\alpha/2}B^{-\alpha-1}$ we have, for large enough N ,

$$\mathbb{P}(X_{i+1} \geq X_i) \geq \left(\Phi(\theta) - \left(1 - \frac{e}{3}\right)\Phi(\theta) \right) \frac{1}{e} = \frac{1}{3}\Phi(\theta). \quad \square$$

To complete the proof of Theorem 2.1 we establish equation (2.3). For n beyond a multiple of $N^{1/(\alpha+2)}$, the reverse orderings predicted by Lemma A.5 cannot be avoided.

Corollary A.2. *Let X_i be sampled from the Zipf–Poisson ensemble. Suppose that $n = n(N)$ satisfies $n \geq CN^{1/(\alpha+2)}$ for $0 < C < \infty$. Then*

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 0.$$

Proof. Let $p \in (0, 1)$ be a constant such that $\mathbb{P}(X_{i+1} \geq X_i) \geq p$ holds for all large enough N and $(C/2)N^{1/(2+\alpha)} \leq i < i+1 \leq CN^{1/(2+\alpha)}$. For instance Lemma A.5

shows that $p = \Phi(-\alpha(C/2)^{\alpha/2}/C^\alpha)/3 = \Phi(-\alpha(2C)^{-\alpha/2})/3$ is such a constant. Then

$$\mathbb{P}(X_1 > X_2 > \cdots > X_n) \leq \prod_i^* \mathbb{P}(X_i > X_{i+1}) \quad (\text{A.10})$$

holds where \prod^* is over all odd integers $i \in [(C/2)N^{1/(\alpha+2)}, CN^{1/(\alpha+2)}]$. There are roughly $CN^{1/(\alpha+2)}/4$ odd integers in the product. For large enough N , the right side of (A.12) is below $(1-p)^{CN^{1/(\alpha+2)}/5} \rightarrow 0$. \square

A.1.4 Sharper results on the limits to the correct ordering

We can sharpen the results of the previous subsection to show that we really can only achieve $(BN/\log(N))^{1/(\alpha+2)}$ correctly ordered entities, for some $B > A$. The idea is to loosen the restriction in Lemma A.5 that the individual probabilities be bounded below by a constant. Instead, we bound below by some power of N . This rate is slow enough that if we consider enough entities in a modified version of Corollary A.2, the probability will still converge to zero.

In this subsection, we will employ a classical bound on the upper tail of the normal distribution as a crucial step in sharpening the results of the previous subsections. Durrett [19, pg. 7] shows that, for $x > 0$,

$$1 - \Phi(x) \geq (x^{-1} - x^{-3})\varphi(x) \quad (\text{A.11})$$

where $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$.

Lemma A.6 (Modified Lemma A.5). *Let X_i be from the Zipf–Poisson ensemble with $\alpha > 1$. Suppose that $(BN/\log N)^{1/(\alpha+2)} \leq i < i+1 \leq (CN/\log N)^{1/(\alpha+2)}$ where*

$\alpha^2(\alpha + 2)2^{\alpha-1} < B < C < \infty$. Then, for large enough N ,

$$\mathbb{P}(X_{i+1} \geq X_i) \geq N^{-1/(\alpha+2)+\varepsilon/2}$$

where $\varepsilon = 1/(\alpha + 2) - \alpha^2 2^{\alpha-1} B^{-1} > 0$.

Proof. First $\mathbb{P}(X_{i+1} \geq X_i) \geq \mathbb{P}(X_{i+1} > \lambda_i)\mathbb{P}(X_i \leq \lambda_i) \geq \mathbb{P}(X_{i+1} > \lambda_i)/e$ using Teicher's inequality (A.2). Next

$$\mathbb{P}(X_{i+1} > \lambda_i) = 1 - \mathbb{P}(X_{i+1} \leq \lambda_i) \geq \Phi\left(\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}}\right) - \frac{0.8}{\sqrt{\lambda_{i+1}}}.$$

Now,

$$\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}} = \sqrt{N} \frac{(i+1)^{-\alpha} - i^{-\alpha}}{\sqrt{(i+1)^{-\alpha}}} = -\alpha\sqrt{N} \frac{(i+\eta)^{-\alpha-1}}{\sqrt{(i+1)^{-\alpha}}} = -\alpha \left(\frac{i+\eta}{i+1}\right)^{-\alpha/2} (i+\eta)^{-(\alpha+2)/2}$$

for some $\eta \in (0, 1)$. Applying the bounds on i ,

$$\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}} \geq -\alpha 2^{\alpha/2} B^{-1/2} (\log N)^{1/2}.$$

Let $\beta = \alpha^2 2^{\alpha-1} B^{-1}$. Plugging this in above yields

$$\Phi(-\sqrt{2\beta \log N}) = 1 - \Phi(\sqrt{2\beta \log N}) \geq ((2\beta \log N)^{-1/2} - (2\beta \log N)^{-3/2}) \frac{e^{-\beta \log N}}{\sqrt{2\pi}}.$$

Now

$$\lambda_{i+1}^{-1/2} \leq N^{-1/2} (CN/\log N)^{\alpha/2(\alpha+2)} = (C/\log N)^{\alpha/2(\alpha+2)} N^{-1/(\alpha+2)}.$$

Then,

$$\mathbb{P}(X_{i+1} > \lambda_i) \geq ((\log N)^{-1/2} - (\log N)^{-3/2})N^{-\beta} - 0.8(C/\log N)^{a/2(\alpha+2)}N^{-1/(\alpha+2)}$$

. Finally, note that $\beta = 1/(\alpha + 2) - \varepsilon$, and so, for large enough N ,

$$\mathbb{P}(X_{i+1} > \lambda_i) \geq eN^{-1/(\alpha+2)+\varepsilon/2},$$

which completes the lemma. \square

To complete the proof of a modified version of Theorem 2.1 we establish a modified form of equation (2.3). For n beyond $(BN/\log N)^{1/(\alpha+2)}$, where $B > \alpha^2(\alpha + 2)2^{\alpha-1}$, the reverse orderings predicted by Lemma A.6 cannot be avoided.

Corollary A.3 (Modified Corollary A.2). *Let X_i be sampled from the Zipf–Poisson ensemble. Suppose that $n = n(N)$ satisfies $n \geq (BN/\log N)^{1/(\alpha+2)}$ for $B > \alpha^2(\alpha + 2)2^{\alpha-1}$. Then*

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 0.$$

Proof. By the previous lemma, $\mathbb{P}(X_{i+1} \geq X_i) \geq N^{-1/(\alpha+2)+\varepsilon/2}$ holds for all large enough N and $(BN/\log N)^{1/(2+\alpha)} \leq i < i + 1 \leq 2(BN/\log N)^{1/(2+\alpha)}$. Then

$$\mathbb{P}(X_1 > X_2 > \cdots > X_n) \leq \prod_i^* \mathbb{P}(X_i > X_{i+1}) \quad (\text{A.12})$$

holds where \prod^* is over all odd integers $i \in [(BN/\log N)^{1/(\alpha+2)}, 2(BN/\log N)^{1/(\alpha+2)}]$. There are roughly $(BN/\log N)^{1/(\alpha+2)}/2$ odd integers in the product. For large enough N , the right side of (A.12) is below $(1 - N^{-1/(\alpha+2)+\varepsilon/2})^{(BN/\log N)^{1/(\alpha+2)}/3} \rightarrow 0$. \square

Remark A.1. We have not done the most careful job of bounding the constants. It appears that a slight modification of Lemma A.6 will allow for $B > \alpha^2(\alpha + 2)2^{-1}$, so we can drop a factor of 2^α . In particular, the term $2^{\alpha/2}$ in Lemma A.6 arose from the bound

$$\left(\frac{i+1}{i+\eta}\right)^{\alpha/2} \leq 2^{\alpha/2}$$

where $\eta \in [0, 1]$. When N gets large enough, we should be able to bound instead by $1 + \varepsilon_1$ for any $\varepsilon_1 > 0$.

Remark A.2. Recall that in order to get the probability to converge to one in the limit then we had to choose $A < \frac{\alpha^2(\alpha+2)}{4}$. So, now we know that the number of correctly ordered elements normalized by $(\alpha^2(\alpha+2)N/\log N)^{1/(\alpha+2)}$ has to live within the interval $[2^{-2/(\alpha+2)}, 2^{(\alpha-1)/(\alpha+2)}]$ with probability tending to one. A slightly more delicate analysis in Lemma A.6 should shrink the interval to (essentially) $[2^{-2/(\alpha+2)}, 2^{-1/(\alpha+2)}]$.

A.1.5 Probability of tail exceedance is small even in the deep tail

This subsection is dedicated to an auxiliary result that shows that, as long as we allow for a small gap, no elements from the tail will jump over entity n for $n(N) = (AN/\log N)^{1/\beta}$ for every $A > 0$ and every $\beta > \alpha$. This is a sharper version of Lemma A.4 and relies on a different approach. We will need the following upper bound on the upper tail of a normal distribution. For $x \geq 0$,

$$1 - \Phi(x) \leq \sqrt{\frac{\pi}{2}}\varphi(x).$$

This result is essentially a modified version of the Chernoff bound for the normal distribution, but obtained in a different way. Note that

$$\int_x^\infty e^{-u^2/2} du = e^{-x^2/2} \int_x^\infty e^{-(u-x)^2/2-x(u-x)} du \quad (\text{A.13})$$

$$\leq e^{-x^2/2} \int_0^\infty e^{-z^2/2} dz, \quad (\text{A.14})$$

where we have begun by adding and subtracting x in the exponent, then used the change of variable $z = u - x$ and the trivial bound $e^{-xz} \leq 1$ to get the result. The sharpness of the constant is not important for the analysis that follows, but it is the sharpest possible, as can be seen by considering $x = 0$.

We now proceed with our sharpened version of Lemma A.4.

Proposition A.1 (Sharper version of Lemma A.4). *Fix $0 < B < A$. Let $\{X_i\}$ be a Zipf–Poisson ensemble with parameter α . Then, for any $\beta > \alpha$, $m \leq (BN)^{1/\beta}$ and $n \geq (AN)^{1/\beta}$.*

$$\lim_{N \rightarrow \infty} \mathbb{P}(\max_{i > n} X_i \geq X_m) = 0.$$

Proof. First fix $u = u(N)$ such that $m < n < u$. Then,

$$\begin{aligned} \mathbb{P}(\max_{i > n} X_i \geq X_m) &\leq \mathbb{P}(\max_{n < i \leq u} X_i \geq X_m) + \mathbb{P}(\max_{i > u} X_i \geq X_m) \\ &\leq \mathbb{P}(\max_{n < i \leq u} X_i \geq X_m) + \mathbb{P}(\sum_{i > u} X_i \geq X_m). \end{aligned}$$

We begin by estimating the first term. By Bonferroni and the Chernoff bound on

the Skellam distribution, we have

$$\begin{aligned} \mathbb{P}(\max_{n < i \leq u} X_i \geq X_m) &\leq \sum_{i=n+1}^u \mathbb{P}(X_i \geq X_m) \\ &\leq \sum_{i=n+1}^u e^{-N(m^{-\alpha/2} - i^{-\alpha/2})^2}. \end{aligned}$$

Now, the right-hand side is decreasing in i , so,

$$\sum_{i=n+1}^u e^{-N(m^{-\alpha/2} - i^{-\alpha/2})^2} \leq \int_n^u e^{-N(m^{-\alpha/2} - x^{-\alpha/2})^2} dx \quad (\text{A.15})$$

$$= \int_{n^{-\alpha/2}}^{u^{-\alpha/2}} e^{-N(m^{-\alpha/2} - y)^2} \left(-\frac{2}{\alpha} y^{-(1+2/\alpha)} \right) dy \quad (y = x^{-\alpha/2}) \quad (\text{A.16})$$

$$= \frac{2}{\alpha} \int_{u^{-\alpha/2}}^{n^{-\alpha/2}} y^{-(1+2/\alpha)} e^{-N(m^{-\alpha/2} - y)^2} dy \quad (\text{A.17})$$

$$\leq \frac{2}{\alpha} u^{1+\alpha/2} \int_{u^{-\alpha/2}}^{n^{-\alpha/2}} e^{-N(m^{-\alpha/2} - y)^2} dy \quad (\text{A.18})$$

$$\leq \frac{2}{\alpha} u^{1+\alpha/2} \int_{-\infty}^{n^{-\alpha/2}} e^{-N(m^{-\alpha/2} - y)^2} dy \quad (\text{A.19})$$

$$= \frac{2}{\alpha} \frac{u^{1+\alpha/2}}{\sqrt{2N}} \int_{-\infty}^{\sqrt{2N}(n^{-\alpha/2} - m^{-\alpha/2})} e^{-\frac{1}{2}z^2} dz \quad (\text{A.20})$$

$$= \frac{2}{\alpha} \frac{u^{1+\alpha/2}}{\sqrt{2N}} \int_{\sqrt{2N}(m^{-\alpha/2} - n^{-\alpha/2})}^{\infty} e^{-\frac{1}{2}z^2} dz \quad (\text{A.21})$$

$$\leq \frac{2}{\alpha} \frac{u^{1+\alpha/2}}{\sqrt{2N}} \left(\frac{\sqrt{\pi}}{\sqrt{2}} e^{-N(m^{-\alpha/2} - n^{-\alpha/2})^2} \right) \quad (\text{A.22})$$

$$= \frac{\sqrt{\pi}}{\alpha} u^{1+\alpha/2} N^{-1/2} e^{-N(m^{-\alpha/2} - n^{-\alpha/2})^2} \quad (\text{A.23})$$

$$= \frac{\sqrt{\pi}}{\alpha} \exp \left(-N(m^{-\alpha/2} - n^{-\alpha/2})^2 - \frac{1}{2} \log N + (1 + \alpha/2) \log u \right). \quad (\text{A.24})$$

Now, taking $m = (BN)^{1/\beta}$, $n = (AN)^{1/\beta}$ and $u = (CN)^{1/\gamma}$ where $\beta > \alpha$ and any $\beta > \gamma$ such that $m < n < u$, we see that the right-hand side vanishes in the limit as $N \rightarrow \infty$.

Consider the second term. $\sum_{i>u} X_i$ is Poisson with mean $N \sum_{i>u} i^{-\alpha} \leq Nu^{-(\alpha-1)}$. Letting $u = (CN)^{1/\gamma}$ for $\beta > \gamma/(1 - 1/\alpha)$ and large enough C guarantees that $\mathbb{P}(\sum_{i>u} X_i \geq X_m) \rightarrow 0$ by the Chernoff bound on the Skellam distribution. \square

Remark A.3. Note that for $n(N) = (AN)^{1/\alpha}$, the mean value is a constant A^{-1} . In particular, if $A \leq 1$, we could never hope to exceed the entire tail. Indeed, the probability that $X_n = 0$ is bounded away from zero as N goes to infinity. So, in that sense, the result of this subsection is essentially the best possible.

A.2 Proof of Theorem 5.1

We will make use of the following integral.

Lemma A.7. *Let $\alpha > 0$, $\beta > 1$. Then*

$$\int_0^\infty 1 - \exp(-\alpha t^{-\beta}) dt = \Gamma(1 - 1/\beta) \alpha^{1/\beta}.$$

Proof. Introduce the change of variable $u(t) = \alpha t^{-\beta}$. Then, the term $1 - e^{-u}$ will appear in the integrand. Write this as $\int_0^u e^{-w} dw$, apply Fubini's theorem and simplify. \square

The upper bound in Theorem 5.1 is now easy to obtain. Note that $\mathbb{E}(X_{i\bullet}) = \sum_{j=1}^\infty 1 - \exp(-Nc_i^{-a}j^{-b})$. By monotonicity, $\sum_j \mathbb{E}(X_{ij}) \leq \sum_j \mathbb{E}(Y_{ij}) = Nc_a i^{-a}$. For

the other part of the bound, $1 - \exp(-Nci^{-a}j^{-b})$ is decreasing in j , so

$$\mathbb{E}(X_{i\bullet}) \leq \int_0^\infty 1 - \exp(-Nci^{-a}y^{-b}) dy,$$

and an application of Lemma A.7 with $\alpha = Nci^{-a}$ and $\beta = b$ gives the result.

For the lower bound we will use the following elementary inequality:

$$1 - e^{-x} \geq \frac{x}{1+x}, \quad \text{for } 0 \leq x < \infty. \quad (\text{A.25})$$

An application of (A.25) to $\mathbb{E}(X_{i\bullet})$ yields

$$\mathbb{E}(X_{i\bullet}) \geq \sum_{j=1}^{\infty} \frac{Nci^{-a}j^{-b}}{1 + Nci^{-a}j^{-b}}.$$

Each term on the right-hand side decreases as j increases, and so

$$\begin{aligned} \mathbb{E}(X_{i\bullet}) &\geq \int_1^\infty \frac{Nci^{-a}y^{-b}}{1 + Nci^{-a}y^{-b}} dy \\ &= \int_1^\infty \frac{Nci^{-a}}{u + Nci^{-a}} b^{-1} u^{-1+1/b} du, \quad (u = y^b) \\ &= b^{-1} \int_1^\infty u^{1/b} (u^{-1} - (u + Nci^{-a})^{-1}) du \\ &\geq b^{-1} \int_1^\infty (u^{-1+1/b} - (u + Nci^{-a})^{-1+1/b}) du \\ &= (1 + Nci^{-a})^{1/b} - 1, \end{aligned}$$

as desired.

To get the asymptotic result, note that

$$1 \geq \frac{\mathbb{E}(X_{i\bullet})}{\mathbb{E}(Y_{i\bullet})} \geq \sum_{j=1}^{\infty} \frac{Nci^{-a}j^{-b}}{Nc_a i^{-a}(1 + Nci^{-a}j^{-b})} \geq c_b \sum_{j=1}^{\infty} (j^b + Nci^{-a})^{-1},$$

and, for all $j \geq 1$, $\varepsilon \geq 0$, we have $(j^b + \varepsilon)^{-1} \geq (1 + \varepsilon)^{-1}j^{-b}$. Hence, the right-hand side converges to one as $i \rightarrow \infty$.

A.3 Proof of Theorem 5.2

The results will mostly be established via application of two lemmas in [20] along with arguments adapted from [8].

Lemma A.8. *Let c and b be constants. Define the recurrence relation $x_{n+1} = c_n + (1 - b/n)x_n$. Then if $c_n \rightarrow c$, $x_n/n \rightarrow c/(1 + b)$.*

Proof. See [20, Lemma 4.1.1] and [20, Lemma 4.1.2]. □

Lemma A.9 (Azuma–Hoeffding inequality). *Let X_t be a martingale with uniformly bounded increments. Then*

$$\mathbb{P}(|X_n - X_0| > x) \leq e^{-x^2/(2c^2n)},$$

where c is the bound on the martingale increments.

We begin by observing that at any time t , there are exactly t edges in the bipartite graph. We will focus on the analysis of $M(k, t)$, keeping in mind that the results for $N(k, t)$ are entirely analogous, except that we replace the sampling probability p with $1 - p$.

First consider $M(1, t)$, i.e., the number of vertices in \mathcal{M} at time t with a single edge. At time $t + 1$, we either add a new vertex of unit degree with probability p , or preferential attachment is performed on \mathcal{M} with probability $q = 1 - p$. Hence

$$\mathbb{E}(M(1, t + 1) - M(1, t)) = p - \frac{q}{t}\mathbb{E}(M(1, t)).$$

Applying Lemma A.8 to $\mathbb{E}(M(1, t))$ with $c_t = c = p$ and $b = q$, we conclude that $\mathbb{E}(M(1, t))/t \rightarrow p/(2 - p)$.

Similarly for each $M(k, t)$, $k \geq 2$, we have the recurrence

$$\mathbb{E}M(k, t + 1) = \frac{(k - 1)q}{t}\mathbb{E}M(k - 1, t) + \left(1 - \frac{kq}{t}\right)\mathbb{E}M(k, t),$$

and a second application of Lemma A.8 yields

$$\frac{\mathbb{E}M(k, t)}{t} \rightarrow \frac{(k - 1)q}{1 + kq} \lim_{t \rightarrow \infty} \frac{\mathbb{E}M(k - 1, t)}{t},$$

where the limit on the right-hand side exists by induction. Solving the recursion, we get

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}M(k, t)}{t} = \frac{p(k - 1)!}{q \prod_{i=1}^k (i + 1/q)} \sim \frac{p}{q} \Gamma(1 + 1/q) k^{-1-1/q}. \quad (\text{A.26})$$

This establishes convergence in mean. Below we provide tight bounds in addition to the asymptotic statement above. To obtain convergence in probability, let $X(k, s) = \mathbb{E}(M(k, t) \mid \mathcal{F}_s)$ for $s \leq t$. Then $X(k, s)$ is a martingale, and by an elegant result from [8], $|X(k, s) - X(k, s - 1)| \leq 2$ for all s . Noting that $X(k, 0) = \mathbb{E}M(k, t)$, an application of Lemma A.9 with $x = \sqrt{t \log t}$ gives the desired convergence in

probability.

We can obtain explicit bounds in (A.26) as a simple consequence of two famous theorems.

Lemma A.10 (Bohr–Mollerup). *Let $\Gamma(x)$ denote the gamma function. Then, for all $x \in [0, 1]$*

$$\frac{n!n^x}{\prod_{k=0}^n (x+k)} \leq \Gamma(x) \leq \frac{n!n^x}{\prod_{k=0}^n (x+k)} \frac{x+n}{n}.$$

Proof. See Artin [5, page 14]. □

Corollary A.4. *Let $\Gamma_n(x) = n!n^x / (\prod_{k=0}^n (x+k))$. Then for all $x > 0$ and all n ,*

$$\Gamma_n(x) \geq \Gamma(x) \frac{n^{x+1}}{(n+x)^{x+1}}.$$

Proof. By Lemma A.10, for all $x \in (0, 1]$,

$$\Gamma_n(x) \geq \Gamma(x) \frac{n}{n+x} \geq \Gamma(x) \frac{n^{x+1}}{(x+n)^{x+1}},$$

and induction on x gives

$$\Gamma_n(x+1) = x\Gamma_n(x) \frac{n}{x+1+n} \geq \Gamma(x+1) \frac{n^{x+2}}{(x+1+n)^{x+2}}.$$

This holds for each n . □

Lemma A.11 (Gauss). *For all $x > 0$,*

$$\Gamma(x) = \lim_{n \rightarrow \infty} \frac{n!n^x}{\prod_{k=0}^n (x+k)}.$$

Proof. This follows from Lemma A.10 by induction on x . See [5, page 15]. □

Corollary A.5. For all $x \geq 1$ and every n , $\Gamma_n(x) \leq \Gamma(x)$.

Proof. For $x \geq 1$,

$$\frac{\Gamma_{n+1}(x)}{\Gamma_n(x)} = \frac{\left(1 + \frac{1}{n}\right)^x}{1 + \frac{x}{n+1}} \geq \frac{1 + \frac{x}{n}}{1 + \frac{x}{n+1}} \geq 1.$$

So, $\Gamma_n(x) \uparrow \Gamma(x)$ for all $x \geq 1$. □

Applying Corollary A.4 to $\lim_{t \rightarrow \infty} \mathbb{E}M(k(t))/t$, we obtain

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}M(k, t)}{t} = \frac{p(k-1)!}{q \prod_{i=1}^k (i + 1/q)} \geq \frac{p}{q} \Gamma(1 + 1/q) \left(\frac{k}{k + 1/q}\right)^{1+1/q} k^{-1-1/q}.$$

To get the upper bound, since $1/q \geq 1$, we can use Corollary A.5, yielding

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}M(k, t)}{t} = \frac{p(k-1)!}{q \prod_{i=1}^k (i + 1/q)} \leq \frac{p}{q} \Gamma(1 + 1/q) k^{-1-1/q}.$$

Bibliography

- [1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling. *Journal of the ACM*, 56(4), 2009. Article no. 21. [15]
- [2] L. A. Adamic. Zipf, power-laws, and Pareto—a ranking tutorial. <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>. [11]
- [3] L. A. Adamic and B. A. Huberman. Zipf’s law and the internet. *Glottometrics*, 3:143–150, 2002. [11]
- [4] R. Albert and A.-L. Barabási. Topology of evolving networks: Local events and universality. *Phys. Rev. Lett.*, 85(24):5234–5237, 2000. [13]
- [5] E. Artin. *The Gamma Function*. Holt, Rinehart and Winston, New York, 1964. [144]
- [6] A.-L. Barabási and R. Albert. The emergence of scaling in random networks. *Science*, 286:509–512, 1999. [12, 13, 92, 93]
- [7] J. Bennett and S. Lanning. The Netflix prize. In *Proceedings of KDD Cup and Workshop 2007*, 2007. [35]

- [8] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Struct. Algorithms*, 18(3):279–290, 2001. [13, 14, 93, 142, 143]
- [9] A. Bookstein. Informetric distributions, part I: Unified overview. *J. Amer. Soc. Info. Sci.*, 41(5):368–375, 1990. [11]
- [10] C. Bresciani-Turroni. On Pareto’s law. *Journal of the Royal Statistical Society*, 100(3):421–432, 1937. [6]
- [11] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature physics*, 2:110–115, 2006. [15]
- [12] C. Cooper and A. Frieze. A general model of web graphs. *Random Struct. Alg.*, 22:311–335, 2003. [14]
- [13] The Internet Movie Database. <http://www.imdb.com>. [39]
- [14] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992. [76]
- [15] P. Deheuvels. Caractérisation complète des lois extrêmes multivariées et de la convergence des types extremes. *Publ. Inst. Statist. Univ. Paris*, 23:1–37, 1978. [19]
- [16] P. Deheuvels. La fonction de dépendance empirique et ses propriétés. un test non paramétrique d’indépendance. *Acad. Roy. Belg. Bul. Cl. Sci. 5*, 65:274–292, 1979. [19]
- [17] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986. [113, 114]

- [18] L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons, New York, 1985. [81]
- [19] R. Durrett. *Probability: Theory and Examples*. Duxbury, Belmont, CA, 2005. [134]
- [20] R. Durrett. *Random Graph Dynamics*. Cambridge University Press, New York, 2006. [14, 93, 97, 142]
- [21] R. Durrett and J. Schweinsberg. Power laws for family sizes in a duplication model. *Ann. Prob.*, 33(6):2094–2126, 2005. [14]
- [22] V. Durrleman, A. Nikeghbali, and T. Roncalli. Which copula is the right one? Technical report, Groupe de Recherche Opérationnelle, Crédit Lyonnais, 2000. [19]
- [23] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: reasoning about a highly connected world*. Cambridge University Press, Cambridge, 2010. [12]
- [24] P. Embrechts. Copulas: A personal view. *J. Risk. and Insur.*, 76(3):639–650, 2009. [19]
- [25] P. Erdős and A Rényi. On random graphs, I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959. [14]
- [26] P. Erdős and A Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960. [14]

- [27] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM Comput. Commun. Rev.*, 29(4):251–262, 1999. [12]
- [28] D. Garlaschelli and M. I. Loffredo. Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.*, 93(26), Dec 2004. 268701. [16]
- [29] W. Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. Phys.*, 38:77–81, 1959. [127]
- [30] C. Genest and J. MacKay. The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283, 1986. [19, 62]
- [31] C. Genest, E. Masiello, and K. Tribouley. Estimating copula densities through wavelets. *Insurance: Mathematics and Economics*, 44:170–181, 2009. [74, 75]
- [32] C. Genest and J. Nešlehová. A primer on copulas for count data. *ASTIN Bulletin*, 37:475–515, 2007. [27]
- [33] J.-L. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. *Physica A*, 371:795–813, 2006. [87, 92]
- [34] H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, Lausanne, 1997. [19]
- [35] B. Klar. Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 14:161–171, 2000. [126]
- [36] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. [44]

- [37] P. A. Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM J. Matrix Anal. and Appl.*, 30(1):261–275, 2008. [77]
- [38] P. L. Krapivsky, G. J. Rodgers, and S. Redner. Degree distributions of growing networks. *Phys. Rev. Lett.*, 86(23):5401–5404, 2001. [16]
- [39] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57–65. IEEE Computer Society, 2000. [14]
- [40] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie. Sampling biases in IP topology measurements. In *IEEE INFOCOM*, pages 332–341, 2003. [14]
- [41] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *Networking, IEEE/ACM Transactions on*, 2(1):1–15, 1994. [12]
- [42] E. Liebscher. Construction of asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 99:2234–2250, 2008. [67, 69, 70]
- [43] J.-G. Liu, T. Zhou, B.-H. Wang, and Y.-C. Zhang. Degree correlation of bipartite network on personalized recommendation. *Intl. J. Mod. Phys. C*, 21(1):137–147, 2010. [16]
- [44] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, Burlington, MA, 3rd edition, 2008. [76]
- [45] B. Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by H. A. Simon. *Information and Control*, 2:90–99, 1959. [12]

- [46] B. Mandelbrot. Final note on a class of skew distribution functions: Analysis and critique of a model due to H. A. Simon. *Information and Control*, 4:198–216, 1961. [12]
- [47] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002. [17, 18]
- [48] S. Maslov, K. Sneppen, and A. Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physics A*, 333:529–540, 2004. [87]
- [49] R. Michel. On Berry–Esseen results for the compound Poisson distribution. *Insurance: Mathematics and Economics*, 13:35–37, 1993. [127]
- [50] T. Mikosch. Copulas: Tales and facts, with discussions and rejoinder. *Extremes*, 9:3–56, 2006. Discussions are provided by multiple other authors. [19]
- [51] M. Mitzenmacher. A brief history of generative models for power law and log-normal distributions. *Internet Mathematics*, 1(2):226–251, 2004. [5, 87]
- [52] C. R. Myers. Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs. *Phys. Rev. E*, 68, 2003. 046116. [16]
- [53] R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2nd edition, 2006. [19, 23, 62]
- [54] J. Nešlehová. On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis*, 98(3):544–567, 2007. [28]

- [55] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67:026126 1–13, 2003. [15, 17]
- [56] M. E. J. Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46(5):323–351, 2005. [11]
- [57] M. E. J. Newman, A.-L. Barabási, and D. J. Watts. *The structure and dynamics of networks*. Princeton University Press, Princeton NJ, 2005. [12]
- [58] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Finding and evaluating community structure in networks. *Proceedings of the National Academy of Science*, 99:2566–2572, 2002. [92]
- [59] V. Pareto. *Cours d’Economie Politique*, volume II. F. Rouge, Lausanne, 1897. [5]
- [60] Netflix Prize. <http://www.netflixprize.com>. [35]
- [61] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955. [11, 12]
- [62] H. A. Simon. Some further notes on a class of skew distribution functions. *Information and Control*, 3(1):80–88, 1960. [12]
- [63] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348, 1962. [77]
- [64] J. G. Skellam. The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society: Series A*, 109(3):296, 1946. [126, 127]

- [65] H. Teicher. An inequality on Poisson probabilities. *The Annals of Mathematical Statistics*, 26:147–149, 1955. [126]
- [66] D. J. Watts. The “new” science of networks. *Annu. Rev. Sociol.*, 30:243–270, 2004. [12]
- [67] Yahoo! Webscope. http://research.yahoo.com/Academic_Relations. [37]
- [68] G. U. Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Phil. Trans. R. Soc. Lond. B*, 213(3):21–87, 1925. [6, 7]
- [69] G. K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, 1935. [8, 9]
- [70] G. K. Zipf. *Human behavior and the principle of least effort; an introduction to human ecology*. Hafner, New York, 1949. [9]
- [71] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Phys. Rev. E*, 74, 2006. 016115. [15]