Visualizing bivariate long-tailed data

Justin S. Dyer
Stanford University

Art B. Owen
Stanford University

June 2011

Abstract

Variables in large data sets in biology or e-commerce often have a head, made up of very frequent values and a long tail of ever rarer values. Models such as the Zipf or Zipf–Mandelbrot provide a good description. The problem we address here is the visualization of two such long-tailed variables, as one might see in a bivariate Zipf context. We introduce a copula plot to display the joint behavior of such variables. The plot uses an empirical ordering of the data; we prove that this ordering is asymptotically accurate in a Zipf–Mandelbrot–Poisson model. We often see an association between entities at the head of one variable with those from the tail of the other. We present two generative models (saturation and bipartite preferential attachment) that show such qualitative behavior and we characterize the power law behavior of the marginal distributions in these models.

1 Introduction

It is increasingly common to see data sets in which two or more categorical variables each have a long-tailed distribution, of which the Zipf distribution is the best known example. In the Netflix data, some movies were much more popular than others, and some users were much more active than others. In large social or biological networks, both the in-degree of nodes and the out-degree of nodes may be long-tailed. For electronic commerce there can be many more than two such variables. For example, an online bookstore may keep track of customers' IP addresses, books' ISBNs, credit card numbers, query strings and book review identifiers.

In commercial settings, there is considerable interest in prediction, as exemplified by the Netflix prize (Bennett and Lanning, 2007), which centered on predicting the ratings that a user would give a movie. The prize was eventually won by a method that combined hundreds of more basic prediction rules.

Ensemble methods and other black box predictors have performance that is hard to beat, but they do not easily supply qualitative insights about the phenomena being modeled. Qualitative insights are useful for understanding the mechanisms underlying the data, especially when one contemplates changing the mechanism.

Our goal in this paper is to develop some methods for exploring multiple long-tailed variables. With ordinary pairs of variables one can easily form a scatterplot to explore their joint behavior. For a single long-tailed variable, a Lorenz curve shows, for example the fraction of wealth held by the poorest $\alpha\%$ of the population, as a function of α. Our plot shows a bivariate display of this type.

Earlier work on assortative and disassortative mixing by degree (Newman, 2003) focused on a correlation coefficient. Social networks were predominantly positively assortative by degree while diverse kinds of biological network were negatively assortative. Positive association between highly connected elements is also called the rich-club ordering (Colizza et al., 2006).

In this paper, 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. We also see some clusters at unexpected locations in the middle of the data ranges.

A famous example of ratings data comes from the Netflix prize (Bennett and Lanning, 2007) 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.

An outline of the paper is as follows. Section 2 presents a gray scale copula display to show the joint distribution of two long-tailed quantities.

The displays show the shape, size and sometimes surprising location of the affinities. Section 3 shows examples of the copula display on some large data sets typical of e-commerce applications.

In making the copula 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. Section 4 gives results showing that most of the data are placed near where they should be, for large sample sizes.

The ratings data we looked at typically showed head-to-tail affinities. 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. Section 5 proposes two mechanical baseline models that provide alternative explanations. One is a saturation model in which raters and items are independently sampled but subject to a limit such that no pair is counted more than once. For the bipartite case, saturation creates a head-to-tail affinity but we find that it generates unusual marginal distributions. A second model invokes bipartite preferential attachment. This model provides reasonable marginal distributions and we find head-to-tail affinities. Our conclusions are in Section 6 along with a discussion of additional similar plots that one could make. Theorem proofs are in an Appendix.

2 The data display

2.1 Construction

We suppose that our data are given as a matrix X_{ij} for $1 \le i \le n$ and $1 \le j \le m$. Most of the data sets we're interested in have $X_{ij} \ge 0$. 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.

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 $X_{1\bullet} \ge X_{2\bullet} \ge \cdots \ge X_{n\bullet}$ and similarly $X_{\bullet j} \ge X_{\bullet j+1}$. 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 variable	Row variable	X_{ij}
A	IV	1
B	III	1
B	IV	2
C	I	1
C	II	2
C	III	2
C	IV	1

Table 1: This table shows a small example with 7 nonzero counts summing to 10. One variable takes values A, B, C, $\cdots$ while the other takes values I, II, III, $\cdots$. The copula plot for this data is shown in Figure 1.

column entities are similarly arranged on the vertical axis, again with smallest at 0, largest at 1 and with length proportional to $X_{\bullet j}$. If one then selects a data point X_{ij} uniformly from those in the sample, it corresponds to a point in $[0,1]^2$ with uniformly distributed horizontal and vertical coordinates.

For our plot we split the unit square into rectangles and shade them 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. We call this a copula plot because the gray level is proportional to a bivariate density with uniform margins.

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

Table 1 shows a small example. It has seven nonzero numbers totalling 10. The data include four different row entities I, II, III and IV. There are three different column entities A, B and C. The first row indicates that the combination A-IV was observed one time.

The row entities, sorted from least to most frequent are I, II, III and IV with relative frequencies 0.1, 0.2, 0.3 and 0.4. The column entities in increasing order are A, B and C with relative frequencies 0.1, 0.3 and 0.6.

The raw counts and marginal totals are shown in the left panel of Figure 1. Each entity is given horizontal or vertical space proportional to its frequency. As a result the upper right corner, C-IV, has the largest rectangle. That

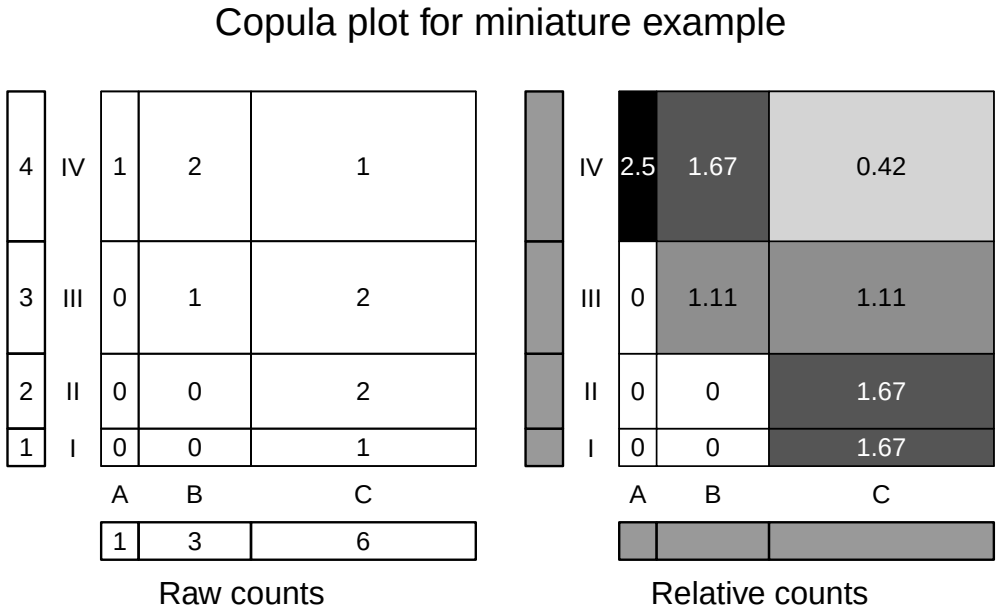


Figure 1: This figure shows the construction of the copula plot for the data from Table 1. The left panel shows joint and marginal counts. The right panel has gray levels showing joint counts relative to area and marginal counts relative to length. The latter are uniform by construction.

rectangle has area 0.6×0.4 or 24% of the unit square. But it only has 1 of 10 counts or 10% of the data. Therefore we plot it with a gray level proportional to $10/24 \doteq 0.42$. The combination A-IV was also observed 1 time in 10 but it has smaller area $0.1 \times 0.4 = .04$ and so it is plotted with a gray level proportional to $0.1/0.04 = 2.5$. The image was scaled so that the gray scale from 0 to 1 corresponds to a ratio from 0 to the maximum observed, in this case 2.5.

For large data sets it is not effective to plot one row of rectangles for each column entity and vice versa. Instead we aggregate the data to reduce the display to a manageable number of cells. In our large examples, we have aggregated so that there are 100 cells along each dimension. Sometimes there are many small entities with the same marginal total, such as 2 total links, that in aggregate comprise a large proportion of the image. In such cases, our plot does not split the rectangle corresponding to such a popular level.

2.2 Copulas and discrepancies

By construction, the data being plotted are nonnegative, and have line integrals equal to 1 if one component of $[0,1]^2$ is fixed and the other is the variable of integration. Therefore the gray level plot depicts a bivariate copula density (Nelsen, 2006).

The gray level plot allows us to compare the observed copula to a uniform one. A uniform copula would be uniformly gray. Darker and lighter gray indicate regions with higher (respectively lower) joint density than they would have under uniformity. Metrics comparing distributions to uniformity on a hypercube are called discrepancies. See Niederreiter (1992) for a discussion of discrepancies in quasi-Monte Carlo sampling. Our data display reveals a local discrepancy, with near black pixels showing positive discrepancy, and near white ones showing a negative discrepancy.

The random variables whose copula we plot require a few remarks. They are ordered categorical variables derived from the original categories by sorting according to relative abundance. For example, movies are an unordered categorical variable. They can be sorted into an ordered categorical variable based on the number of ratings that they got. In practice, we sort them on their sample popularity, which may deviate from their popularity in the process from which they are sampled.

2.3 Maslov and Sneppen's display

Our display superficially resembles one used by Maslov and Sneppen (2002) to show patterns in protein interaction networks. They plot a normalized empirical probability that two proteins of given degree are connected. For a bipartite graph joining prey proteins to bait proteins, the probability is the fraction of all edges that connect a prey protein with degree K_0 to a bait protein of degree K_1. The normalization is the same probability, for a network constructed with random edges where all nodes retain their original degrees.

Whereas their normalization is based on randomly rewiring the edges in the graph, ours is based on a deterministic respacing of the axes. The Netflix data set has about 100,000,000 edges and the Yahoo! song ratings data set is even larger. For such large data sets, it is computationally burdensome to randomly resample the graph, but quite simple to rescale the coordinate axes.

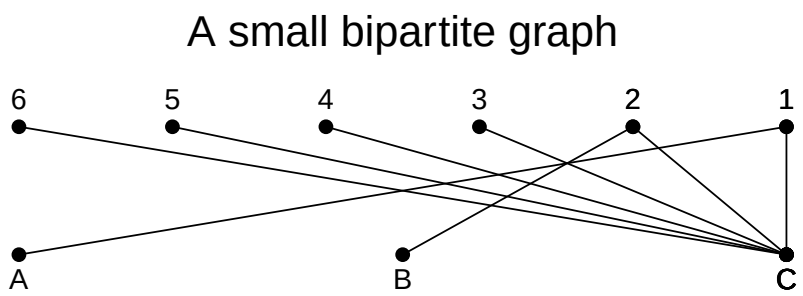


Figure 2: A small bipartite graph illustrating the difference between the Maslov and Sneppen display and the one discussed herein. Any random rewiring of the graph leaves all degrees unchanged, so the Maslov and Sneppen display takes a value of one everyone. Our copula plot would be non-uniform for this case, because there are no connections from $\{A, B\}$ to $\{3, 4, 5, 6\}$.

We find it interesting that Maslov and Sneppen (2002) procedure is *not* equivalent to ours. To show that they are different, it is enough to use a very tiny example. Consider data of the form:

A1, B2, C1, C2, C3, C4, C5, C6,

as illustrated by the graph in Figure 2.

Each step in their random graph algorithm selects two edges, say uv and xy. Then if uy and xv are not edges in the graph, the selected edges are removed and replaced by uy and xv.

Consider the tiny graph of Figure 2. Our measure takes the value 0 for connections between the letters of lowest degree (A and B) and numbers (1,2,3,4) of lowest degree, because there are no such connections in the graph. Inspecting that graph, we see that none of the edges connected to C can be removed. As a result, the ensemble of random graphs has only two members, the one shown and the one obtained by rewiring $A1$ and $B2$ to $A2$ and $B1$. That rewiring leaves all degrees unchanged and so their ratio takes the value 1 everywhere, and hence does not show the lack of affinity between $\{A, B\}$ and $\{1, 2, 3, 4\}$.

Their measure displays connectivity relative to that under random rewiring. Ours is relative to independence of row and column entities. This allows us to use the copula interpretation, described in Section 2.2.

3 Examples

Here we show some examples of the copula plots. The first set of examples come from bipartite graphs showing relations between two types of entity. The second set show directed graphs where the two quantities being displayed are in-degree and out-degree. The third group of examples are for symmetric matrices.

3.1 Bipartite graphs

Figure 3 shows the discrepancy plot for the Netflix data. The raw data take the form X_{ij} if user i rated movie j. We are visualizing the rating events themselves, not the ratings. We comment briefly below on how the ratings themselves look.

The left panel of Figure 3 shows that the Netflix data has a strong head-to-tail affinity. Users who rate few movies are over-represented at movies that received many ratings. Conversely, movies that are rarely rated get the majority of their ratings from users who rate many movies. This finding suggests a taste-based explanation: novices primarily rate blockbusters, while the cognescenti have also searched out some rare gems.

The head-to-tail spikes so dominate the plot that we cannot easily see other smaller structures. In the right panel of Figure 3, the data are replotted with a different gray scale. We have used 256 gray levels chosen so that each level is used (within rounding error) for the same number of pixels. As a result, the histogram of gray levels from the image is uniform.

The uniform histogram view of the Netflix data reveals that there is a small tail-to-tail affinity but no head-to-head affinity at all, in the Netflix data. We see that the copula contours in the two head-to-tail corners have different shapes, more rectangular for big movies but rounder for big raters. There are also prominent horizontal stripes corresponding to consecutive blocks of movies that go against the grain compared to the bulk of the data. Similar vertical stripes are fainter. This may be because there are more raters than movies. Because these features are small compared to the head-to-tail affinities, they do not show up in the left panel. Combining both views shows more about the data than either on its own.

Figure 3 displays where the ratings come from, but not how high or low those ratings are. To do that, we could define X_{ij} to be some increasing function of $\{1, 2, 3, 4, 5\}$ and redo the plot. We have done this, with some extreme

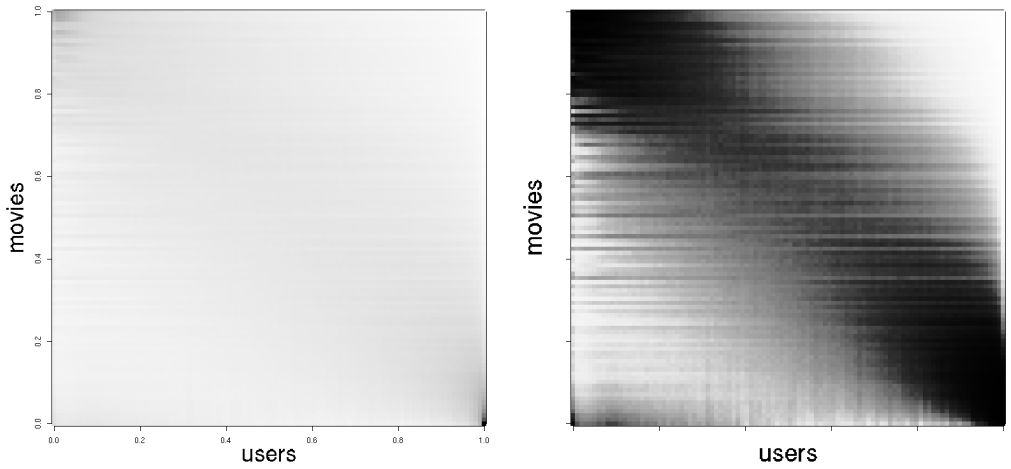


Figure 3: The left panel shows the discrepancy plot for the Netflix data. Dark spots in the upper left and lower right corners show a head-to-tail affinity for this data. The right panel using a uniform gray level reveals finer structure: some affinity among small movies and users, depletion for large movies and users, horizontal stripes and some weaker vertical stripes.

scores. The first score takes $X_{ij} = 1$ if the rating was 5, and 0 otherwise. It just looks at particularly fortuitous movie-customer combinations, the kind that a recommender would most like to have made. A second score took $X_{ij} = 1$ if the rating was 1 and $X_{ij} = 0$ otherwise. This score looks at the combinations that a recommender would most regret. 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.

Figure 4 shows the Yahoo! song ratings data. Here X_{ij} is 1 if user i rated song j. Once again we see head-to-tail affinity, but it takes a different shape than it did for the Netflix data. The busy users dominate the bottom 10% or more of songs, much more than the busy movie raters dominated the least viewed movies.

Further differences between the two data sets show up in comparing the figures. While the movie data has some tail-to-tail overlap, the song data has very little. Also, the shape of the contours in the lower right corner (active raters on unpopular items) is more rectangular for the songs than for the movies. For movies there is a progression where ever more rare movies get

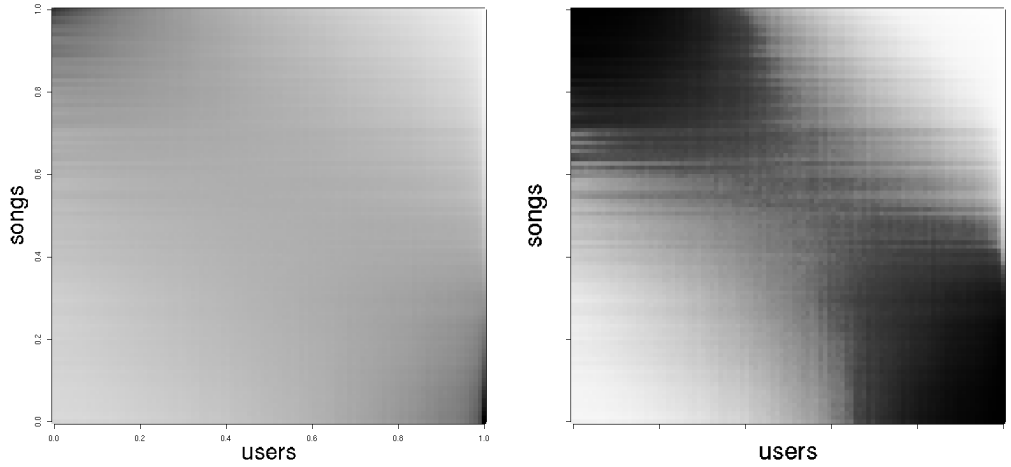


Figure 4: This figure plots the Yahoo! song ratings data in the same manner that Figure 3 shows the Netflix data.

their ratings from ever more busy raters. For songs, the contour is much straighter.

We can extract numerical assessments of head-to-tail affinities from this data. The busiest raters rate the popular movies only about 22.5% as often as they would under independence. This and other summaries appear in Table 2. From that data we also see that the corner associations are stronger for movies than for songs. Also the two head-to-tail affinities are roughly equally strong in the song data, but, for movies, the affinity between busy users and rare movies is stronger than the reverse.

Figure 5 shows data from the Internet Movie Database (IMDB). The variable $X_{ij} = 1$ if and only if actor i was in movie j. Actors who worked rarely (e.g. only once) are overrepresented in movies with the largest casts and nearly absent from movies with smaller casts. On the other hand, the busiest actors are overrepresented in movies with small casts, but not the very smallest casts.

This display also provides an example where a single level accounts for a large proportion of the data: actors appearing in only one movie account for about 16% of the data. This explains the wide column of cells on the left hand margin of the display which has not been split due to ties.

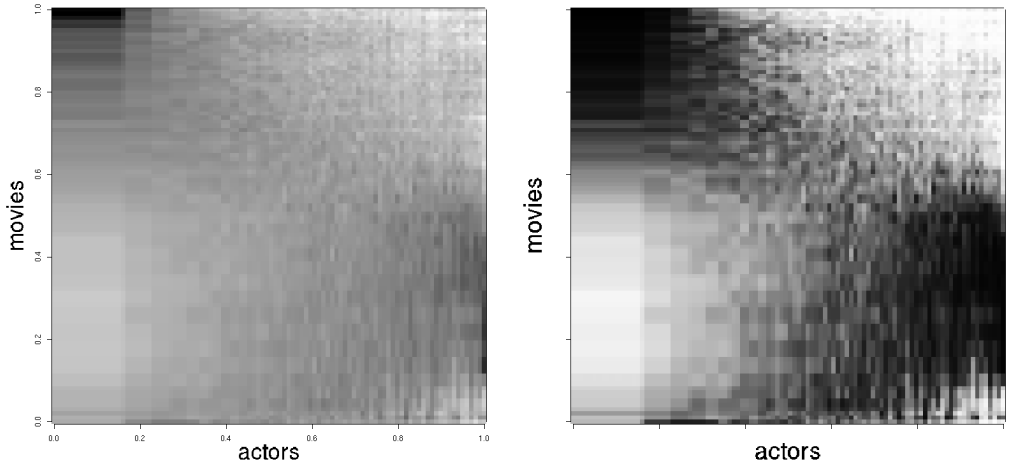


Figure 5: This figure plots data representing which actors appeared in which movies, in the IMDB. The ordinary copula is on the left, the uniformly scaled one is on the right.

3.2 Directed graphs

In a directed graph, each entity has both an in-degree and an out-degree. We are interested in visualizing the joint distribution of these two quantities.

Copula plots for some publicly available directed graphs appear in Figure 6. The upper left panel shows trust relations at the consumer review site Epinions. Here $X_{ij} = 1$ if user i trusts user j. We see two dark clusters. At the lower right we see that there are lots of edges from users who trust many people to users who are trusted only by a few others. We might see that pattern in users with computer generated trust out-links. If many of the links are of that type, then it might be wise to discount such potentially spammy links. Near the lower left we see that there are lots of edges from users who trust only a few people to users trusted only by a few others. Such links might arise from infrequent users trusting their friends, though there are other possibilities (they could also be computer generated).

The upper right panel of Figure 6 shows a snapshot of the Wikipedia graph as of February 2007. Here $X_{ij} = 1$ if the page for topic i links to the one for topic j. Each topic included had at least one in-link and at least one out-link. We might expect to see hubs and authorities (Kleinberg, 1999) in this graph. There is a cluster of topics with large out-degree and small

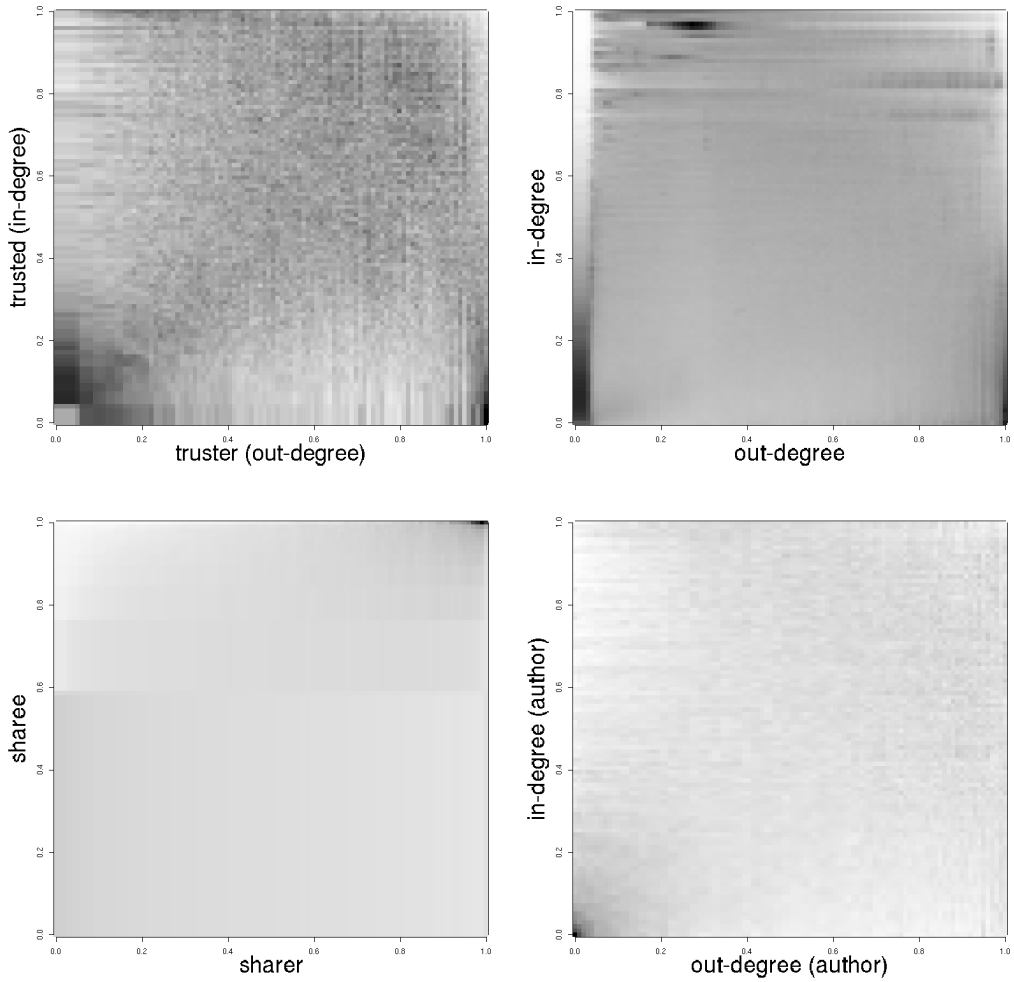


Figure 6: This figure shows copula plots for four directed graphs: Epinions, Wikipedia, Snapfish and arXiv hep-th, as described in the text.

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 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.

The lower left panel of Figure 6 shows some data from Hewlett-Packard's Snapfish photo sharing service. This is a directed graph where $X_{ij} = 1$ if user i shares a photo album with user j. Curiously, there is a fairly strong, though asymmetric, head-to-head affinity where very active sharers tend to share with the most active sharees. A large proportion of people that share photos do so only once, and, as indicated by the somewhat darker lower left corner of the plot, they tend to share with those that also accept very few sharing requests. The normalized view of this data in Figure 7 brings out some of these features.

The lower right panel of Figure 6 shows a snapshot of the arXiv hep-th paper citation network. The normalized version of this plot is in Figure 7. Here $X_{ij} = 1$ if paper i cites paper j. We see affinities among papers with few 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.

3.3 Symmetric matrices

Undirected graphs have symmetric incidence matrices, so $X_{ij} = X_{ji}$. In graphs based on social network links we have seen positive associations: members with the most out-links get the most in-links and conversely. This is the opposite of the head-to-tail affinities we saw in ratings data. Such a pattern is not pre-ordained by the symmetry. It is possible for graphs to contain a strong hub and spoke pattern, as for example protein networks do.

Figure 8 shows two publicly available data sets. For the Enron email data, X_{ij} is 1 if one or more emails were exchanged between addresses i and j. This version of the data is symmetric by construction. There are some head-to-tail affinities that we would expect if some email is broadcast to all or most accounts. There are also some affinities among smaller participants of equal size, giving dark squares along the main diagonal.

The second data set depicted in Figure 8 is based on a network of roads

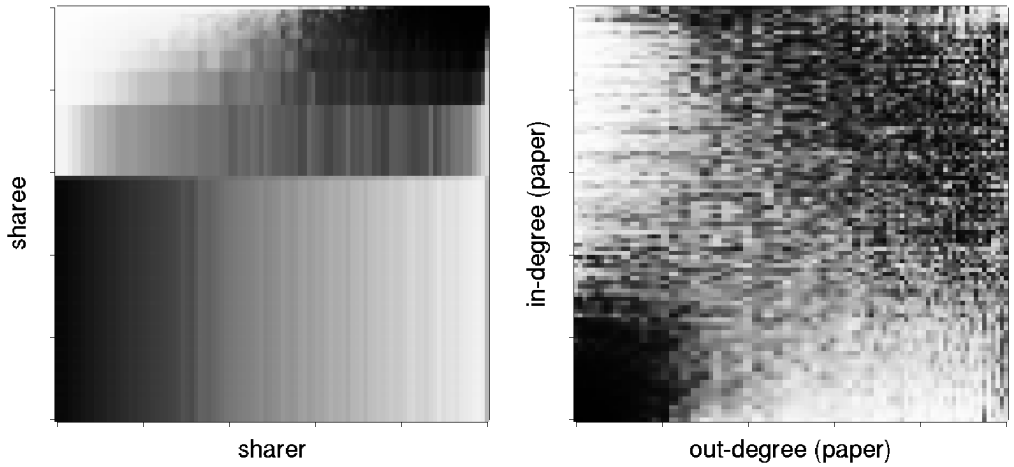


Figure 7: This figure shows normalized copula plots for the Snapfish and arXiv hep-th directed graphs of Figure 6.

in California, where $X_{ij} = 1$ if intersection i connects to intersection j. That is, vertices correspond to intersections and edges to road segments. The structure of this network is quite different from those involving people as entities. 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. The graph is nearly planar (as would be expected) and the maximum degree is only 12.

3.4 Numerical summaries

To compare the plots we make a numerical summary. For any rectangle, we can sum the values of X_{ij} in it and divide the sum by $X_{\bullet\bullet}$ times the area of the rectangle. This ratio gives us a lift statistic with a value of 1 corresponding to neutral affinity. To study corner affinities we have found it useful to measure the lift over small squares in the corners. Table 2 displays the affinities for squares of size 0.05×0.05 in each of 4 corners of the 11 copulas shown in Figures 3 through 8.

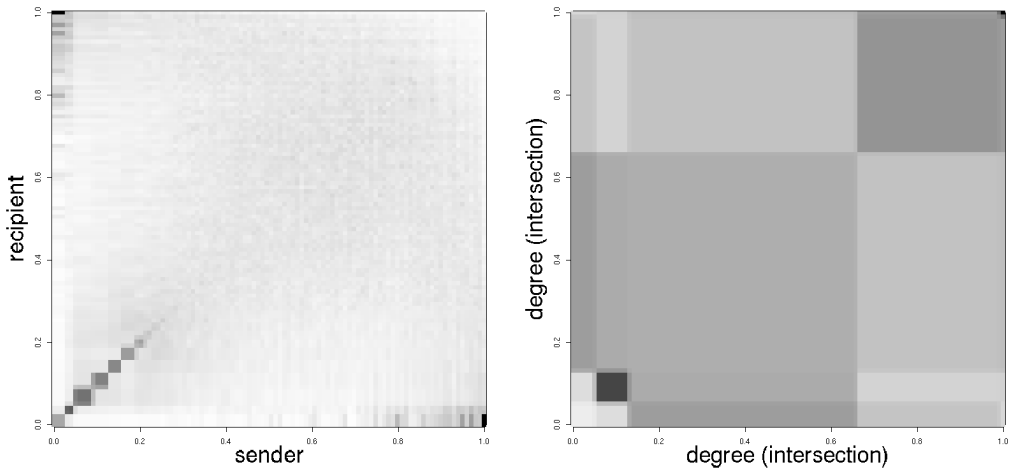


Figure 8: The left panel shows the data display for the Enron data set and the right panel corresponds to the California road-network data.

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: Numerical summary of corner affinities from Figures 3 to 8. Independence corresponds to a value of 1. For example, infrequent movie raters rate popular movies 2.776 times as often as they would under independence.

4 Proper ordering

The gray scale images we present depict bivariate copula densities. In forming the copula, we have replaced a categorical variable on many levels, such as a movie customer by one single number, that customer's rank. It is an enormous convenience to replace categorical variables such as customer, query string, credit card number, IP address and so on by a single nonnegative integer. But it is possible that some of those variable levels will be given the wrong rank, because the amount of data for each entity is random.

Dyer and Owen (2010) study the extent to which the biggest entities are placed in the correct order in a random sample. They suppose that entity i appears X_i times where $X_i \sim \mathrm{Poi}(N\theta_i)$ and θ_i are decreasing values. Entities i and $i+1$ are in the correct order if $X_i > X_{i+1}$. From their Lemma 2, we find that

$$\mathbb{P}(X_1 > X_2 > \cdots > X_n) \geq 1 - \sum_{i=1}^{n-1} \exp(-N(\sqrt{\theta_i} - \sqrt{\theta_{i+1}})^2).$$

They show that for the Zipf-Poisson ensemble in which $\theta_i = i^{-\alpha}$ for $\alpha > 1$ the top $(BN/\log(N))^{1/(2+\alpha)}$ are correctly ordered with probability tending to 1 as $N \to \infty$, when $B < \alpha^2(\alpha+2)/4$. Because $\alpha > 1$, we can take $B = 3/4$ and so we anticipate getting the top $(3N/(4\log(N)))^{1/(2+\alpha)}$ entities in the proper order. Asymptotically those entities will account for all but a fraction $o(N^{-(\alpha-1)/(\alpha+2)+\epsilon})$ of the total data $\sum_i X_i$, for any $\epsilon > 0$.

While the top entities are properly ordered, our eye is also drawn to the corners defined by small entities. Small entities are not properly ordered because for them, the sampling fluctuations are relatively large. In the Zipf–Poisson ensemble, the top $CN^{1/(2+\alpha)}$ entities for $C > 0$ will include some ordering errors as $N \to \infty$.

Here we extend the analysis from Dyer and Owen (2010) to show that most of the entities of any size are nearly correctly placed.

Suppose that movie i appears X_i times and customer j appears Y_j times. That one rating contributes to the darkness of the pixel at a point given by the relative size of movie i among the movie data and customer j among the customer data. If both the customer and the movie are located near where they should be, then the data from that rating is near its proper location. If the bulk of the movies and customers, weighted by their data size, are properly located then the copula plot is descriptive of the process generating the data, not just the one data set at hand.

We will suppose that, marginally, one of the entity types (e.g. the movies) comes from a Zipf–Mandelbrot ensemble defined as follows. The number of times entity i is observed is $X_i \sim \mathrm{Poi}(N\theta_i)$, independently, where $\theta_i = (i+k)^{-\alpha}$ for $i \geq 1$, $k > -1$ and $\alpha > 1$. The Zipf–Mandelbrot form is more flexible than the plain Zipf law ($k = 0$) and it provides a qualitatively better description of the data we study.

We want to show that very little of the data will appear at any great distance from where it should. For $\tau > \sigma \geq 0$, let

$$J(\tau, \sigma) = \frac{1}{N} \sum_{i=1}^{\infty} X_i \mathbf{1}_{X_i \geq N\tau} \mathbf{1}_{\mathbb{E}(X_i) < N\sigma}.$$

The quantity J represents a normalized total of all data from entities with mean below $N\sigma$ that have sample values at or above $N\tau$. These entities have jumped ahead of their true location.

Let $T = \sum_{i=1}^{\infty} X_i$ be the total sample size. Then $\mathbb{E}(T) = N \sum_{i=1}^{\infty} \theta_i$ (and $\mathbb{V}(T) = \mathbb{E}(T)$) and so the proportion of data jumping from below σ to over τ is nearly $J(\tau, \sigma)/C$, for large N, where $C = \sum_{i=1}^{\infty} \theta_i$. The Zipf–Mandelbrot model has $C < \infty$.

Similarly, we define

$$S(\tau, \sigma) = \frac{1}{N} \sum_{i=1}^{\infty} X_i \mathbf{1}_{X_i \leq N\sigma} \mathbf{1}_{\mathbb{E}(X_i) > N\tau}.$$

The quantity S represents a normalized amount of data from entities that have slipped from a true location above $N\tau$ to a position at or below $N\sigma$.

Our graphic is based on a fixed partition, such as 100 bins, into which the X_i are placed. Suppose that the bin boundaries are β_k for $k = 1, \ldots, 99$. If all of $J(\beta_k, \beta_k - \epsilon_k)$ and $S(\beta_k + \epsilon_k, \beta_k)$ are small, for small $\epsilon_k > 0$, then the data within each bin are representative of the entities that belong in that bin.

Theorem 1. *Let X_i be from the Zipf–Mandelbrot–Poisson ensemble with parameters $k > -1$ and $\alpha > 1$. Let $0 < \sigma < \tau \leq \theta_1$ be given. Then*

$$\mathbb{E}(J(\tau, \sigma)) \leq \frac{1}{N} \frac{1}{\alpha \tau^{1/\alpha}} \frac{\tau}{\tau - \sigma} + o(N^{-1}). \tag{1}$$

as $N \to \infty$.

Proof. We prove this in the Appendix. □

Theorem 1 shows that very little of the total data counts can have jumped from below expectation $N\sigma$ to $N\tau$ or above. From Markov's inequality, the jump fraction satisfies

$$\mathbb{P}(J(\tau,\sigma) > \epsilon) = O\Big(\frac{1}{N}\Big)$$

as $N \to \infty$ for any $\epsilon > 0$, and $\tau > \sigma > 0$.

The formula for $\mathbb{E}(J)$ diverges as $\tau - \sigma \to 0$. We defined J with σ strictly smaller than τ, disallowing $J(\tau,\tau)$. The value $J(\tau,\tau)$ describes entities that have jumped from below $N\tau$ to $N\tau$ or higher. In the models we consider, there is an index m such that $\theta_m \geq \tau > \theta_{m+1}$. Then $J(\tau,\tau) = J(\tau,\sigma)$ for any $\sigma \in (\theta_{m+1}, \tau)$. Therefore, we can always avoid the case with $\sigma = \tau$.

Theorem 2. *Let X_i be from the Zipf–Mandelbrot–Poisson ensemble with parameters $k > -1$ and $\alpha > 1$. Let $0 < \sigma < \tau \leq \theta_1$ be given. Then*

$$\mathbb{E}(S(\tau,\sigma)) \leq \frac{1}{N}\frac{1}{\alpha\sigma^{1/\alpha}}\frac{\tau}{\tau-\sigma} + o(N^{-1}). \tag{2}$$

as $N \to \infty$.

Proof. We prove this in the Appendix. □

4.1 Data from the deep tail

In the Zipf–Mandelbrot–Poisson model there are an infinite number of entities. The very small entities comprise the deep tail of the ensemble. We can use Theorem 1 to get an estimate of the total amount of data that could jump from the deep tail into the near tail. We define the near tail by excluding the largest entities which comprise a proportion $1 - \epsilon$ of the expected data. The deep tail is similarly defined through η where $0 < \eta < \epsilon \ll 1$.

The bound in Theorem 1 does not depend on the parameter k of the Zipf–Mandelbrot distribution, and so we illustrate it with $k = 0$, corresponding to the Zipf distribution. To simplify the formulas, we approximate the Zipf distribution with probability mass function proportional to $i^{-\alpha}$ by the Pareto density $\alpha x^{-\alpha}$ on $1 < x < \infty$. The u'th quantile of the Pareto distribution

is $x_u = (1-u)^{-1/\alpha}$ and the Pareto density there is $\alpha x_u^{-\alpha} = \alpha(1-u)^{1+1/\alpha}$. Therefore the relevant thresholds are

$$\tau = \alpha\epsilon^{1+1/\alpha} \quad \text{and} \quad \sigma = \alpha\eta^{1+1/\alpha}.$$

The expected fraction of mass jumping from below σ to over τ is asymptotic to

$$\frac{\tau^{1-1/\alpha}}{\alpha(\tau-\sigma)}N^{-1} = \frac{1}{\alpha^{1+1/\alpha}N}\frac{\epsilon^{-1/\alpha}}{1-(\eta/\epsilon)^{1+1/\alpha}}.$$

For illustration, taking $\epsilon = 0.01$ and $\eta = 0.005$ and $\alpha = 2$ we get

$$\frac{1}{N}\frac{10}{2^{5/2}-1} \doteq \frac{2.14}{N}$$

so very little of the data for sample entities making up the top 99% will have come from population entities not among the top 99.5%.

It is not necessary to scale the thresholds proportionally to N. We can take $N\tau = 1$ and $N\sigma = \eta < 1$ and find by equation (7) (a non-asymptotic expression used in the proof of Theorem 1 in the Appendix) that

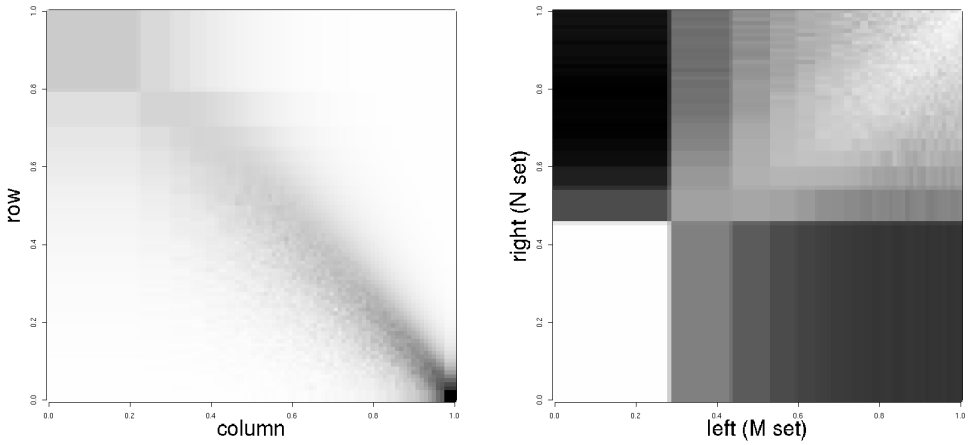
$$\mathbb{E}(J(\tau,\sigma)) \le N^{1/\alpha-1}\frac{\Gamma(1-1/\alpha)}{(1-\eta)\alpha} \le N^{1/\alpha-1}\frac{\alpha-1}{1-\eta}.$$

Thus variables X_i with $\mathbb{E}(X_i) < \eta < 1$ contribute a vanishing fraction of the total data, though there are infinitely many of them.

5 Affinity models

In this section we present some simple generative models for networks in which head-to-tail affinities arise. The first model is a saturation model in which a head-to-tail affinity appears as a simple consequence of no rater being allowed to rate any item more than once. The second model is a bipartite preferential attachment model in which each entity type exhibits a size based preference for the other entity type. Every entity starts out with a single edge and a preference for the pre-existing large entities of the opposite type.

These models show that head-to-tail affinities can arise from very simple processes. The models have one or two parameters each. Thus they provide only crude approximations to the empirical copulas we have seen which may require several degrees of freedom in each of four corners to describe well. Figure 9 shows example copulas from each of these models.



(a) Saturation model (b) Bipartite pref. attachment

Figure 9: Copula estimates for some simulated datasets. The left panel shows a saturation model with parameters $a = 1.5$ and $b = 2.5$. The right panel shows bipartite preferential attachment with $p = 4/9$.

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 Maslov et al. (2004) 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 let

$$Y_{ij} \sim \text{Poi}(Nci^{-a}j^{-b}), \quad \text{and,}$$
$$X_{ij} = \begin{cases} 1 & Y_{ij} \geq 1 \\ 0 & \text{else,} \end{cases} \tag{3}$$

where $c = c_a c_b = 1/(\zeta(a)\zeta(b))$ and $\zeta(x)$ denotes the Riemann-zeta function. The random variables Y_{ij} comprise a bivariate Zipf–Poisson ensemble. The Y_{ij} are latent and we only observe the truncated values X_{ij}. In the latent model, row and column entities are generated independently. The truncation that turns Y_{ij} into X_{ij} is more likely to deplete the head-to-head combinations than any others and this produces head-to-tail affinity.

Figure 9(a) shows an example simulated with $a = 1.5$, $b = 2.5$ and $N = 10^7$. 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. Figure 10(a) shows the marginal Zipf plot of the columns (the plot for the rows is similar). Instead of sorting the entities by observed counts, we have kept the original ordering. We see that these expected counts have curvature. Also the slope near the origin is not a, but has instead been altered by the sampling process. The bounds shown are those of the following theorem.

Theorem 3. *Let X_{ij} be sampled as the saturated bivariate Zipf–Poisson ensemble* (3)*. Let $X_{i\bullet}$ and $X_{\bullet j}$ be the marginal sums. Then*

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

and, as $i \to \infty$,

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

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

Proof. We prove this in the Appendix. □

From Theorem 3, 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 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.

5.2 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 and Albert (1999) model.

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 Newman et al. (2002). Graphs in which one kind of node has a prescribed degree distribution and the other kind is sampled by preferential attachment have also been considered Guillaume and Latapy (2006).

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 $i = 1, \ldots, m(t)$ from $\mathcal{M}$ and nodes $j = 1, \ldots, n(t)$ 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 $i = m(t) = m(t-1) + 1$ to $\mathcal{M}$ and connect it at random to one of the nodes $j \in \{1, \ldots, n(t-1)\}$ 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 $1, \ldots, m(t-1)$ 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 Barabási and Albert (1999) 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 Bollobás et al. (2001). Degree distributions in bipartite preferential attachment are fundamentally different from those in the unipartite case.

Theorem 4. *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*

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

where the arrows denote both convergence of the mean and convergence in probability as $t \to \infty$, and the asymptotic equivalence holds as $k \to \infty$.

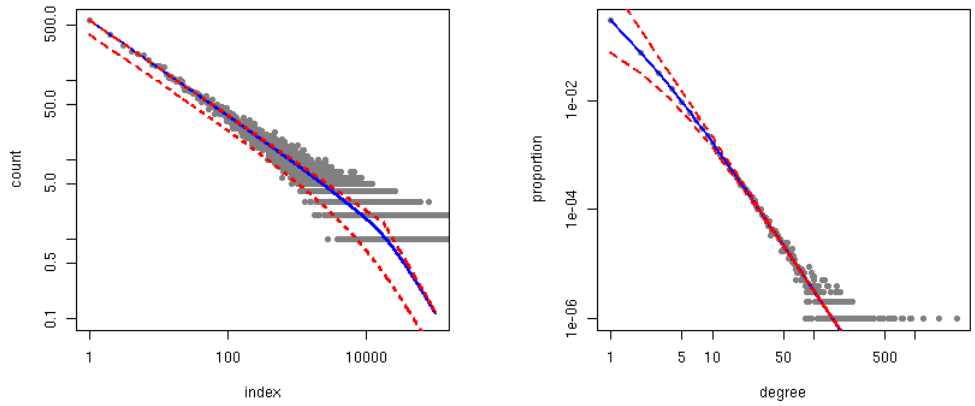
Proof. We prove this in the Appendix. □

Theorem 4 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 $(2,\infty)$ (Durrett, 2006, Chapter 4).

Like the saturation model, the bipartite model generates head-to-tail affinities, but they do not concentrate in the corners. This can be seen in Figure 9(b). In the appendix, we provide bounds in addition to the asymptotic statements in Theorem 4. Figure 10(b) shows the degree distribution of $\mathcal{M}$ for a simulation of one million (total) vertices with $p = 4/9$.

6 Discussion

We have investigated head-to-tail affinities for bivariate heavy tailed data arising from bipartite networks and directed networks. A graphical display



(a) Marginal plot of saturation model with $a = 1.5$, $b = 2.5$, $N = 10^7$.

(b) Degree distribution of $\mathcal{M}$ of bipartite pref. attachment ($p = 4/9$, $t = 10^6$).

Figure 10: Expected distribution and bounds for two reference models. Gray dots indicate sample degrees, the solid line is the expected distribution and the dashed lines are upper and lower bounds on the expected distribution.

reveals the locations and strengths of these affinities, along with other affinities that are initially surprising. Our display depends on ordering the entities by sample values of their magnitudes. Results on the bulk ordering of values show that for large sample sizes as we see in Internet applications, the graphs are indicative of properties of the underlying entities, not just the data at hand.

The copulas we have seen in real data rarely resemble classical parametric copula densities. As a result we advocate plotting the actual copula estimates instead of fitting parametric models. The Wikipedia plot in particular is distinct from all of the usual parametric copula densities.

Graphical displays cannot compete with sophisticated machine learning algorithms when the goal is to predict something like the rating a user will give an item. In those settings the best performing algorithms may be uninterpretable combinations of hundreds of predictions. The strength of graphical displays is that they can bring qualitative information to the attention of domain experts who may then interpret them and perhaps change the system somehow. We have found that people quickly start thinking about what the dense spots and voids in our copula density plots might mean in terms of the

underlying entities.

No single display can capture all of the structure in enormous data sets of this kind. The graphs we present do not show explicit community structure, as, for example, those of Newman and Girvan (2004) and Palla et al. (2005) do. By grouping row and column entities by size we merge members from different communities revealing patterns that hold across communities and not necessarily within communities.

The images we present can easily be generalized. The row entities can be sorted by one variable, the column entities by another, and a third quantity can be used to define the gray level. We have used the same quantity in all three roles to simplify exposition and interpretation.

Acknowledgments

We thank Jure Leskovec, Patrick Perry, Sarah Emerson, Eric Sun and Michael Mahoney for helpful discussions. We thank the reviewers at EJS for their helpful comments. We are grateful to those who shared data with us. Yahoo! Inc. provided the music ratings data under the Yahoo! Webscope program. Fereydoon Safai of Hewlett-Packard Labs facilitated our use of the Snapfish data. The Wikipedia data came from David Gleich. The other data came from Jure Leskovec.

This work was supported by the U.S. National Science Foundation under grant DMS-0906056 and by a National Science Foundation Graduate Research Fellowship.

References

E. Artin. *The Gamma Function*. Holt, Rinehart and Winston, New York, 1964.

A.-L. Barabási and R. Albert. The emergence of scaling in random networks. *Science*, 286:509–512, 1999.

J. Bennett and S. Lanning. The Netflix prize. In *Proceedings of KDD Cup and Workshop 2007*, 2007.

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. ISSN 1042-9832.

V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature physics*, 2:110–115, 2006.

R. Durrett. *Random Graph Dynamics*. Cambridge University Press, New York, 2006.

J. S. Dyer and A. B. Owen. Correct ordering in the Zipf–Poisson model. Technical report, Stanford University, Statistics, September 2010.

W. Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. Phys.*, 38:77–81, 1959.

J.-L. Guillaume and M. Latapy. Bipartite graphs as models of complex networks. *Physica A*, 371:795–813, 2006.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

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.

R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2nd edition, 2006.

M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67: 026126 1–13, 2003.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113 1–15, 2004.

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.

H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. S.I.A.M., Philadelphia, PA, 1992.

G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. Wiley, New York, 1986.

Yahoo! Webscope. http://research.yahoo.com/Academic_Relations.

7 Appendix: Proofs

Section 7.1 proves our two results on entities that either jump ahead or slip behind their proper place in the data ensemble. Section 7.2 proves Theorem 3 on the saturation model. Finally, Section 7.3 proves Theorem 4 on properties of our bipartite preferential attachment model.

7.1 Accuracy of the bulk ordering

We begin with a lemma that generalizes a bound of Shorack and Wellner (1986, page 485) on the Poisson tail probability, to certain tail factorial moments. For integers $p \geq 1$, define $X^{(p)} = X(X-1)\cdots(X-p+1)$. The p'th factorial moment of X is $\mathbb{E}(X^{(p)})$.

Lemma 1. *Let $X \sim \text{Poi}(\lambda)$ and let $p \geq 1$ be an integer. Then for integers $t > \lambda$,*

$$\mathbb{E}(X^{(p)}\mathbf{1}_{X \geq t}) \leq \frac{\max(t,p)^p}{1-\lambda/t}\mathbb{P}(X=t). \tag{4}$$

Proof. For $t \geq p$,

$$\mathbb{E}(X^{(p)}\mathbf{1}_{X \geq t}) = \frac{e^{-\lambda}\lambda^t}{t!}\sum_{j=t}^{\infty}\lambda^{j-t}\frac{t!}{(j-p)!} = \frac{e^{-\lambda}\lambda^t}{t!}\sum_{\ell=0}^{\infty}\lambda^{\ell}\frac{t!}{(t+\ell-p)!}.$$

Now $t!/(t+\ell-p)! \le t^{p-\ell}$ holds for integer $\ell \ge 0$, trivially for $\ell = p$ and by simple direct arguments in cases $\ell > p$ and $\ell < p$. Therefore

$$\mathbb{E}(X^{(p)}\mathbf{1}_{X\ge t}) \le \mathbb{P}(X=t)\sum_{\ell=0}^{\infty}\lambda^\ell t^{p-\ell} = \mathbb{P}(X=t)\frac{t^p}{1-\lambda/t}.$$

If $t < p$, then $\mathbb{E}(X^{(p)}\mathbf{1}_{X\ge t}) = \mathbb{E}(X^{(p)}\mathbf{1}_{X\ge p})$ and the result follows as before. □

Shorack and Wellner (1986) give the bound $\mathbb{P}(X \ge t) \le (1-\lambda/t)^{-1}\mathbb{P}(X = t)$ which can be interpreted as the $p = 0$ version of Lemma 1. The case with $p = 2$ is useful for bounding the variance of $J(\tau,\sigma)$ defined below. We omit that computation for reasons of space. The case of most interest to us has $p = 1$. Then $t > \lambda \ge 0$ implies $t \ge p$ and so equation (4) becomes

$$\mathbb{E}(X\mathbf{1}_{X\ge t}) \le \frac{t}{t-\lambda}\frac{e^{-\lambda}\lambda^t}{(t-1)!} = \frac{t}{t-\lambda}\frac{e^{-\lambda}\lambda^t}{\Gamma(t)}. \tag{5}$$

We will also use Gautschi's inequality (Gautschi, 1959) on the Gamma function,

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s} \tag{6}$$

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

Now we are ready to examine the amount of data jumping over thresholds. Let $\tau > 0$ be a threshold. Let $\sigma < \tau$ be a second threshold. Then

$$J(\tau,\sigma) = \frac{1}{N}\sum_{i=1}^{\infty} X_i\mathbf{1}_{X_i\ge N\tau}\mathbf{1}_{\mathbb{E}(X_i)<N\sigma}$$

is a normalized version of the total count from entities with mean below $N\sigma$ that have 'jumped' up to the threshold $N\tau > 0$.

The indices of the jumping observations are $i \ge n(\sigma)$ where $n(\sigma) = \min\{m \ge 1 \mid \theta_m < \sigma\}$. We may assume that $n(\sigma) \ge 2$ because X_1 has no other entity to jump ahead of. Now

$$\mathbb{E}(J(\tau,\sigma)) = \frac{1}{N}\sum_{i=n(\sigma)}^{\infty}\mathbb{E}(X_i\mathbf{1}_{X_i\ge N\tau}) \le \frac{\tau}{N(\tau-\sigma)}\sum_{i=n(\sigma)}^{\infty}\frac{e^{-N\theta_i}(N\theta_i)^{N\tau}}{\Gamma(N\tau)},$$

by equation (5).

Proof of Theorem 1, bounding $\mathbb{E}(J(\tau,\sigma))$.

For the Zipf–Mandelbrot–Poisson ensemble,

$$\begin{aligned}\mathbb{E}(J(\tau,\sigma)) &\le \frac{\tau}{N(\tau-\sigma)} \sum_{i>n(\sigma)} \frac{e^{-N(i+k)^{-\alpha}}(N(i+k)^{-\alpha})^{N\tau}}{\Gamma(N\tau)} \\ &\le \frac{\tau I}{N(\tau-\sigma)\Gamma(N\tau)}\end{aligned}$$

where

$$\begin{aligned}I &= \int_{n(\sigma)-1}^{\infty} e^{-N(x+k)^{-\alpha}}(N(x+k)^{-\alpha})^{N\tau}\,\mathrm{d}x \\ &\le \int_0^{\infty} e^{-y} y^{N\tau-1/\alpha-1} N^{1/\alpha}\alpha^{-1}\,\mathrm{d}y \\ &= \frac{N^{1/\alpha}}{\alpha}\Gamma(N\tau-1/\alpha),\end{aligned}$$

after making the substitution $y = N(x+k)^{-\alpha}$.

Applying the bound for I we get

$$\mathbb{E}(J(\tau,\sigma)) \le \frac{\tau N^{1/\alpha}}{N(\tau-\sigma)\alpha}\frac{\Gamma(N\tau-1/\alpha)}{\Gamma(N\tau)}. \tag{7}$$

Gautschi's inequality (6) gives $\Gamma(N\tau-1/\alpha)/\Gamma(N\tau) < (N\tau-1)^{-1/\alpha}$, once $N > 1/\tau$, and then

$$\mathbb{E}(J(\tau,\sigma)) \le \left(\frac{N}{N\tau-1}\right)^{1/\alpha}\frac{\tau}{N(\tau-\sigma)\alpha}. \quad \square$$

Now we turn to the amount of data from entities that have slipped below their proper places. Here

$$S(\tau,\sigma) = \frac{1}{N}\sum_{i=1}^{\infty} X_i \mathbf{1}_{X_i \le N\sigma}\mathbf{1}_{\mathbb{E}(X_i)>N\tau},$$

is a normalized version of the total count from entities with mean above $N\tau$ that have 'slipped' below the threshold $N\sigma$, where once again $\tau > \sigma \ge 0$. We will use the following result

$$\mathbb{P}(X \le t) \le (1-t/\lambda)^{-1}\mathbb{P}(X=t), \qquad X \sim \mathrm{Poi}(\lambda), \quad t < \lambda \tag{8}$$

which is equation (9) of Shorack and Wellner (1986, page 485). It then easily follows that

$$\mathbb{E}(X^p \mathbf{1}_{X \le t}) \le t^p \mathbb{P}(X \le t) \le (1 - t/\lambda)^{-1} \mathbb{P}(X = t) t^p. \tag{9}$$

A factorial moment version of (9) also holds.

For $t < \lambda$ with $p = 1$ we get

$$\mathbb{E}(X \mathbf{1}_{X \le t}) \le \frac{\lambda}{\lambda - t} \frac{e^{-\lambda} \lambda^t}{\Gamma(t)}. \tag{10}$$

Proof of Theorem 2 bounding $\mathbb{E}(S(\tau, \sigma))$

Let $n'(\tau) = \max\{m \mid \theta_m > \tau\}$. Then

$$\begin{aligned}
\mathbb{E}(S(\tau,\sigma)) &\le \frac{1}{N} \sum_{i=1}^{n'(\tau)} \mathbb{E}(X_i \mathbf{1}_{X_i \le N\sigma}) \le \frac{1}{N} \sum_{i=1}^{n'(\tau)} \frac{N\theta_i}{N\theta_i - N\sigma} \frac{e^{-N\theta_i} (N\theta_i)^{N\sigma}}{\Gamma(N\sigma)} \\
&= \frac{1}{N\Gamma(N\sigma)} \frac{\theta_{n'}}{\theta_{n'} - \sigma} \sum_{i=1}^{n'(\tau)} \exp(-N(i+k)^{-\alpha}) (N(i+k)^{-\alpha})^{N\sigma} \\
&\le \frac{I}{N\Gamma(N\sigma)} \frac{\tau}{\tau - \sigma},
\end{aligned}$$

where

$$I = \int_0^\infty \exp(-N(x+k)^{-\alpha}) (N(x+k)^{-\alpha})^{N\sigma} \, \mathrm{d}x \le \frac{N^{1/\alpha}}{\alpha} \Gamma(N\sigma - 1/\alpha),$$

by the same arguments used in Theorem 1. Therefore

$$\mathbb{E}(S(\tau,\sigma)) \le \frac{N^{1/\alpha - 1}}{\alpha} \frac{\Gamma(N\sigma - 1/\alpha)}{\Gamma(N\sigma)} \frac{\tau}{\tau - \sigma}, \tag{11}$$

so for $N\sigma > 1$,

$$\mathbb{E}(S(\tau,\sigma)) \le \frac{1}{N\alpha} \left(\frac{N}{N\sigma - 1} \right)^{1/\alpha} \frac{\tau}{\tau - \sigma}. \quad \square$$

7.2 Proof of Theorem 3

We will make use of the following integral.

Lemma 2. *Let $\alpha > 0$, $\beta > 1$. Then*

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

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 $\int_0^u e^{-w}\,\mathrm{d}w$, apply Fubini's theorem and simplify. □

The upper bound in Theorem 3 is now easy to obtain. Note that $\mathbb{E}(X_{i\bullet}) = \sum_{j=1}^\infty 1 - \exp(-Nci^{-a}j^{-b})$. By monotonicity, $\sum_j \mathbb{E}(X_{ij}) \le \sum_j \mathbb{E}(Y_{ij}) = Nc_a i^{-a}$. For the other part of the bound, $1 - \exp(-Nci^{-a}j^{-b})$ is decreasing in j, so

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

and an application of Lemma 2 with $\alpha = Nci^{-a}$ and $\beta = b$ gives the result.

For the lower bound we will use the following elementary inequality, valid for all real x,

$$1 - e^{-x} \ge \frac{x}{1+x}. \tag{12}$$

An application of (12) to $\mathbb{E}(X_{i\bullet})$ yields

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

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

$$\begin{aligned}
\mathbb{E}(X_{i\bullet}) &\ge \int_1^\infty \frac{Nci^{-a}y^{-b}}{1 + Nci^{-a}y^{-b}}\,\mathrm{d}y \\
&= \int_1^\infty \frac{Nci^{-a}}{u + Nci^{-a}} b^{-1}u^{-1+1/b}\,\mathrm{d}u, \qquad (u = y^b) \\
&= b^{-1}\int_1^\infty u^{1/b}\left(u^{-1} - (u + Nci^{-a})^{-1}\right)\,\mathrm{d}u \\
&\ge b^{-1}\int_1^\infty \left(u^{-1+1/b} - (u + Nci^{-a})^{-1+1/b}\right)\,\mathrm{d}u \\
&= (1 + Nci^{-a})^{1/b} - 1,
\end{aligned}$$

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 $i \to \infty$.

7.3 Proof of Theorem 4

The results will mostly be established via application of two lemmas in Durrett (2006) along with arguments adapted from Bollobás et al. (2001).

Lemma 3. *Let c and b be constants. Define the recurrence relation $x_{n+1} = c_n + (1 - b/n)x_n$. Then if $c_n \to c$, $x_n/n \to c/(1 + b)$.*

Proof. See Durrett (2006, Lemma 4.1.1) and Durrett (2006, Lemma 4.1.2). □

Lemma 4 (Azuma–Hoeffding inequality)**.** *Let X_t be a martingale with uniformly bounded increments. Then*

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

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 3 to $\mathbb{E}(M(1, t))$ with $c_t = c = p$ and $b = q$, we conclude that $\mathbb{E}(M(1, t))/t \to p/(2 - p)$.

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 3 yields

$$\frac{\mathbb{E}M(k,t)}{t} \to \frac{(k-1)q}{1+kq} \lim_{t\to\infty} \frac{\mathbb{E}M(k-1,t)}{t},$$

where the limit on the right-hand side exists by induction. Solving the recursion, we get

$$\lim_{t\to\infty} \frac{\mathbb{E}M(k,t)}{t} = \frac{p(k-1)!}{q\prod_{i=1}^{k}(i+1/q)}. \tag{13}$$

This establishes convergence of the mean. 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 Bollobás et al. (2001), $|X(k,s) - X(k,s-1)| \leq 2$ for all s. Noting that $X(k,0) = \mathbb{E}M(k,t)$, an application of Lemma 4 with $x = \sqrt{t\log t}$ gives the desired convergence in probability.

It remains to show that

$$\mu(k) \equiv \frac{p(k-1)!}{q\prod_{i=1}^{k}(i+1/q)} \sim \frac{p}{q}\Gamma(1+1/q)k^{-1-1/q} \tag{14}$$

as $k \to \infty$. We do this by constructing explicit bounds for $\mu(k)$ from (14), using some properties of the Gamma function from Artin (1964). First we introduce

$$\Gamma_n(x) \equiv \frac{n!n^x}{\prod_{j=0}^{n}(x+j)} \tag{15}$$

for real $x > 0$ and integer $n \geq 1$.

Lemma 5. *For $x \in (0,1]$ and integer $n \geq 2$,*

$$\Gamma_n(x) \leq \Gamma(x) \leq \Gamma_n(x)\frac{x+n}{n}\,.$$

Proof. This result is equivalent to an inequality at the top of page 15 in Artin (1964). It appears in the middle of his proof of Theorem 2.1, which is also known as the Bohr–Mollerup Theorem, or sometimes, the Bohr–Mollerup–Artin Theorem. □

Corollary 1. *For all $x > 0$ and all $n \geq 2$,*

$$\Gamma_n(x) \geq \Gamma(x)\Big(\frac{n}{n+x}\Big)^{x+1}.$$

Proof. By Lemma 5, for all $x \in (0, 1]$,

$$\Gamma_n(x) \geq \Gamma(x)\frac{n}{n+x} \geq \Gamma(x)\Big(\frac{n}{x+n}\Big)^{x+1}.$$

We extend the proof to $x > 1$ by induction on $\lfloor x \rfloor$, using

$$\begin{aligned}\Gamma_n(x+1) &= x\Gamma_n(x)\frac{n}{x+1+n} \\ &\geq \Gamma(x+1)\Big(\frac{n}{x+n}\Big)^{x+1}\frac{n}{x+1+n} \\ &\geq \Gamma(x+1)\Big(\frac{n}{x+1+n}\Big)^{x+2}.\end{aligned}$$

□

Lemma 6. *For $x > 0$,*

$$\Gamma(x) = \lim_{n\to\infty} \Gamma_n(x).$$

Proof. This result, due to Gauss, is in Artin (1964, page 15). □

Corollary 2. *For all $x \geq 1$ and every n, $\Gamma_n(x) \leq \Gamma(x)$.*

Proof. For $x \geq 1$,

$$\frac{\Gamma_{n+1}(x)}{\Gamma_n(x)} = \frac{\big(1+\frac{1}{n}\big)^x}{1+\frac{x}{n+1}} \geq \frac{1+\frac{x}{n}}{1+\frac{x}{n+1}} \geq 1.$$

So, $\Gamma_n(x) \uparrow \Gamma(x)$ for all $x \geq 1$. □

Now we are ready to establish the asymptotic order of $\mu(k)$. We begin by writing

$$\mu(k) = \frac{p}{q}\Gamma_{k-1}(1+1/q)(k-1)^{-1-1/q}.$$

From Corollary 1 we obtain the following lower bound, for $k \geq 3$,

$$\mu(k) \geq \frac{p}{q}\Gamma(1+1/q)\Big(\frac{k-1}{k+1/q}\Big)^{2+1/q}(k-1)^{-1-1/q}. \tag{16}$$

Next, we get an upper bound. Since $1/q > 1$, we can use Corollary 2, yielding for $k \geq 3$,

$$\mu(k) \leq \frac{p}{q}\Gamma(1+1/q)k^{-1-1/q}. \tag{17}$$

Now we conclude that $\mu(k) \sim (p/q)\Gamma(1+1/q)k^{-1-1/q}$ as $k \to \infty$ because both bounds (16) and (17) have that limiting behavior.