Visualizing Relationships between Long Tailed Random Variables

Art B. Owen

based on joint work with

Justin S. Dyer

Department of Statistics

Stanford University

Categorical variables

Internet data sets feature many categorical variables. For example

- IP address
- URL
- creative
- search query
- cookie
- etc. etc. $\cdots$ etc.

Not like 3 types of iris

These categorical variables have the following properties:

1) They take an enormous number of levels
2) New levels are to be expected on a regular basis
3) They have a heavy tail distribution:
   some levels hugely popular, others appear once if ever

N.B.: long tail $\equiv$ heavy tail

Heavy tailed distributions

Suppose the entities are $i = 1, 2, 3, \ldots$ in decreasing order of frequency.

For example X_i could be frequency of i'th most popular query string, IP, URL, . . .

Zipf's law is the most famous

$$X_i \propto i^{-\alpha}, \quad \alpha > 1$$

$\cdots$ but it usually doesn't quite fit

Zipf–Mandelbrot fits a bit better

$$X_i \propto (i + k)^{-\alpha}, \quad \alpha > 1, \quad k > -1$$

Question:

What’s better than a long tailed random variable?

Question:

What’s better than a long tailed random variable?

Answer:

Two long tailed random variables

because then we can study their dependence

Bivariate Zipf distribution

As late as 2008, a search showed no web pages with the term ’bivariate zipf’

So I added it to my web page

Suppose that Z and Y are both heavy tailed. How should we

- predict Y from Z?
- measure correlation?
- form clusters and biclusters?
- plot Y versus Z?

Here we look at plotting/visualizing the Z Y dependence

(saving X for later!)

Why visualize?

It is hard to beat ensemble learners for accuracy (e.g. Netflix).

But they're not interpretable.

Why interpret?

Ensemble learners work best when the data are in some kind of steady state.

But we might be anticipating changes in the system or planning to make such changes.

Notation

Data are $X_{ij} \geq 0$.

X may be binary valued, dollar valued, dwell times, counts, etc.

Row entities $1 \leq i < \infty$

Column entities $1 \leq j < \infty$

Examples

i	j	X_{ij}
IP address	URL	1 iff i visited j
Facebook page	user	# messages left
User	Page	seconds spent
Rater	Movie	# stars given
Rater	Movie	1 iff i rated j
Node	Node	1 iff i links to j
Sender	Recipient	Number of emails from i to j

Example: Netflix

To sidestep whether 5 stars are 25% better than 4 stars, take

$$X_{ij} = \begin{cases} 1 & i \text{ rated } j \\ 0 & \text{else.} \end{cases}$$

Marginal counts

$$X_{i\bullet} = \sum_{j=1}^{\infty} X_{ij}$$ # ratings by customer i

$$X_{\bullet j} = \sum_{i=1}^{\infty} X_{ij}$$ # ratings of movie j

$$X_{\bullet\bullet} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} X_{ij}$$ # ratings total $\approx 10^8$

Netflix margins

$480{,}189$ customers' $X_{i\bullet}$'s

17,653 17,436 16,565 15,813 14,831 9,822 9,767 9,740 9,064 8,880

...

1 1 1 1 1 1 1 1 1 1 1 1 1

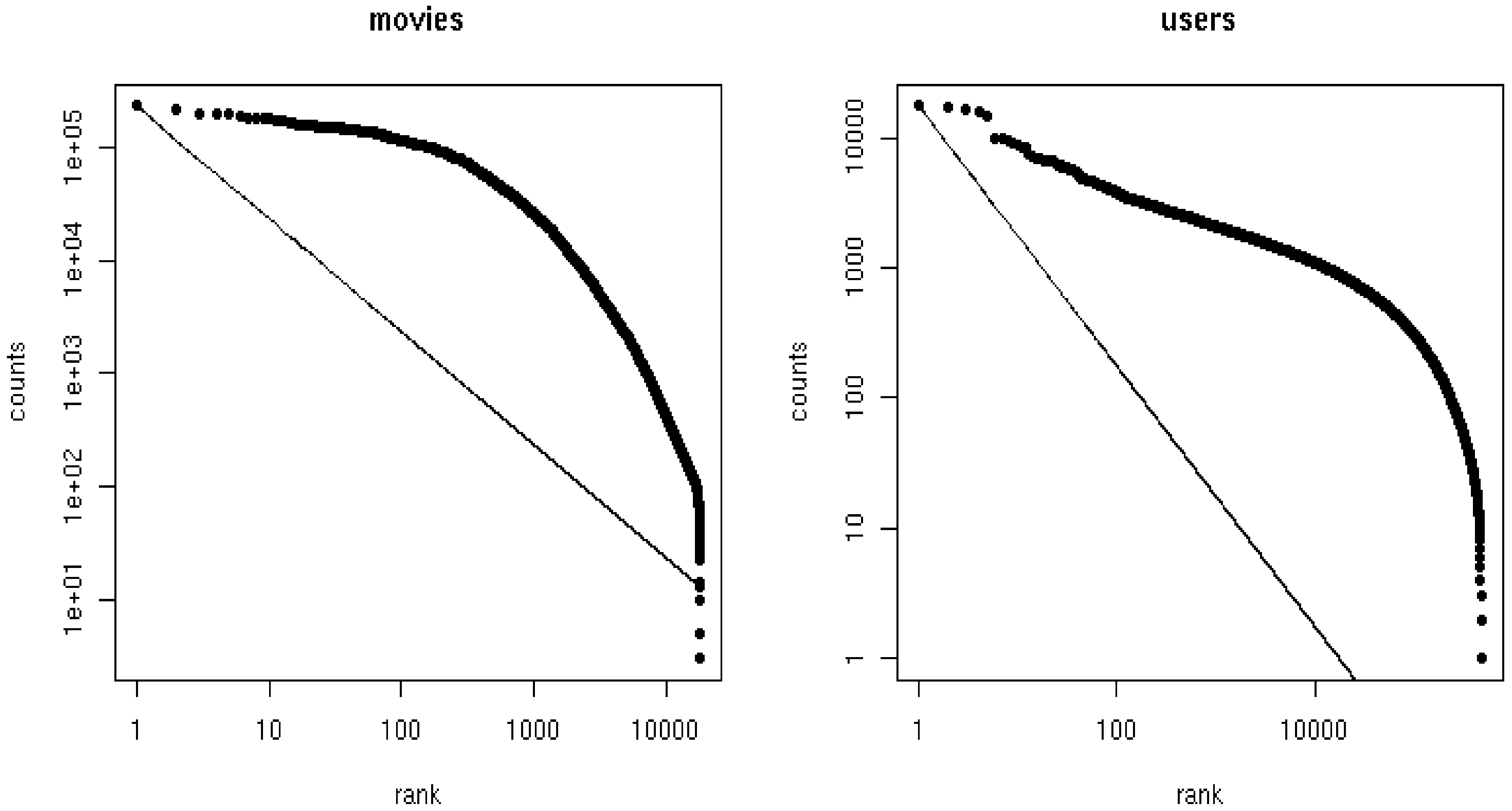
$17{,}770$ movies' $X_{\bullet j}$'s

232,944 216,596 200,832 196,397 193,941 193,295 181,508 181,426 178,068 177,556

...

25 23 22 22 14 13 10 10 5 3

Marginal counts



Both entities are long tailed, with curved Zipf plots

Joint behavior

We would like to know how movie popularity varies with customer popularity

Similarly:

- most read vs most emailed news stories
- in-degree vs out degree in networks
- citations vs references

How do these variables co-vary?

Graphic construction

Data preparation

1) Given data $X_{ij} \quad 1 \le i \le n \quad 1 \le j \le m$

2) Sort rows: $X_{1\bullet} \ge X_{2\bullet} \ge \cdots \ge X_{n\bullet}$

3) Sort cols: $X_{\bullet 1} \ge X_{\bullet 2} \ge \cdots \ge X_{\bullet m}$

The plot

Our plot is a gray scale image on $[0,1]^2$.

Row entities are sorted from small to large

They get horizontal space $X_{i\bullet}/X_{\bullet\bullet}$

Col entities get vertical space $X_{\bullet j}/X_{\bullet\bullet}$

Gray level of a rectangle $R = [a_0, a_1] \times [b_0, b_1] \subset [0,1]^2$ is proportional to

$$\frac{1}{(b_1 - b_0)(a_1 - a_0)} \sum_{i \in R} \sum_{j \in R} X_{ij}$$

So dark means positive association, light is negative and average is middle gray

Small example

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

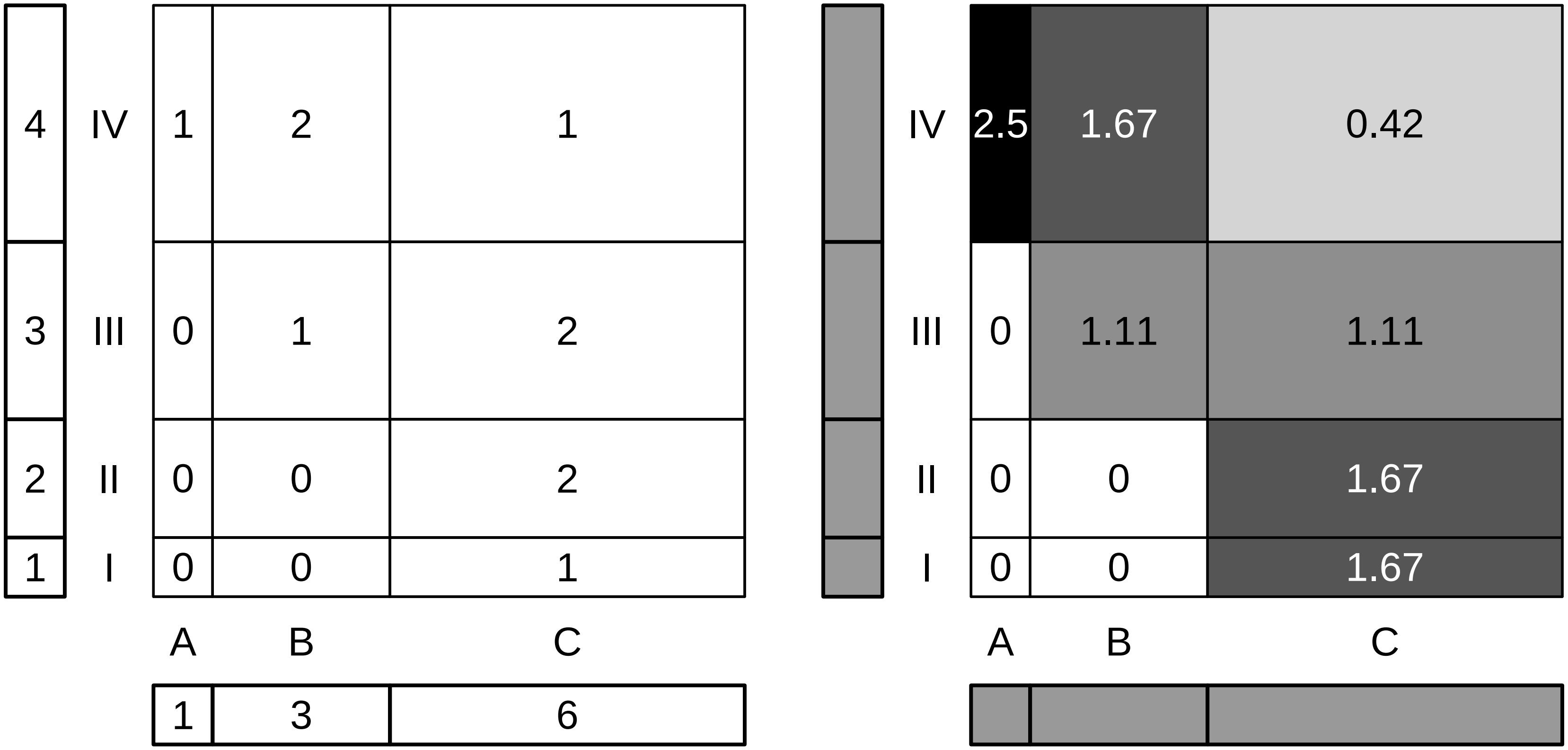
10 obs total. Col=B & Row=IV happened 2 times.

Row entities I, II, III and IV have relative frequencies 0.1, 0.2, 0.3 and 0.4.

Col entities A, B and C have relative frequencies 0.1, 0.3 and 0.6.

Small example ctd.

Copula plot for miniature example



		A	B	C
4	IV	1	2	1
3	III	0	1	2
2	II	0	0	2
1	I	0	0	1
		1	3	6

Raw counts

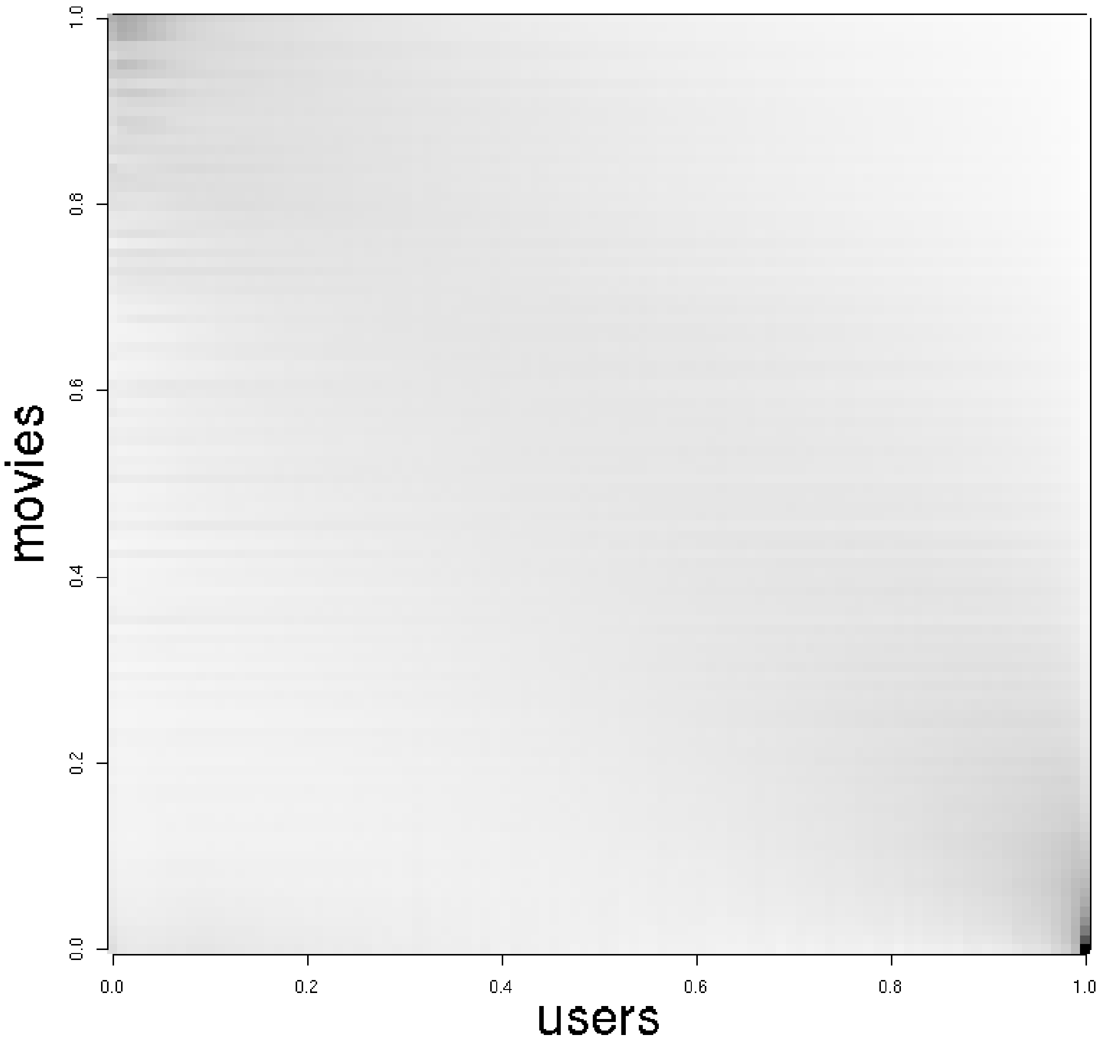
	A	B	C
IV	2.5	1.67	0.42
III	0	1.11	1.11
II	0	0	1.67
I	0	0	1.67

Relative counts

A-IV has area $0.1 \times 0.4 = 0.04$ but 0.1 of the data so it gets $0.1/0.04 = 2.5$

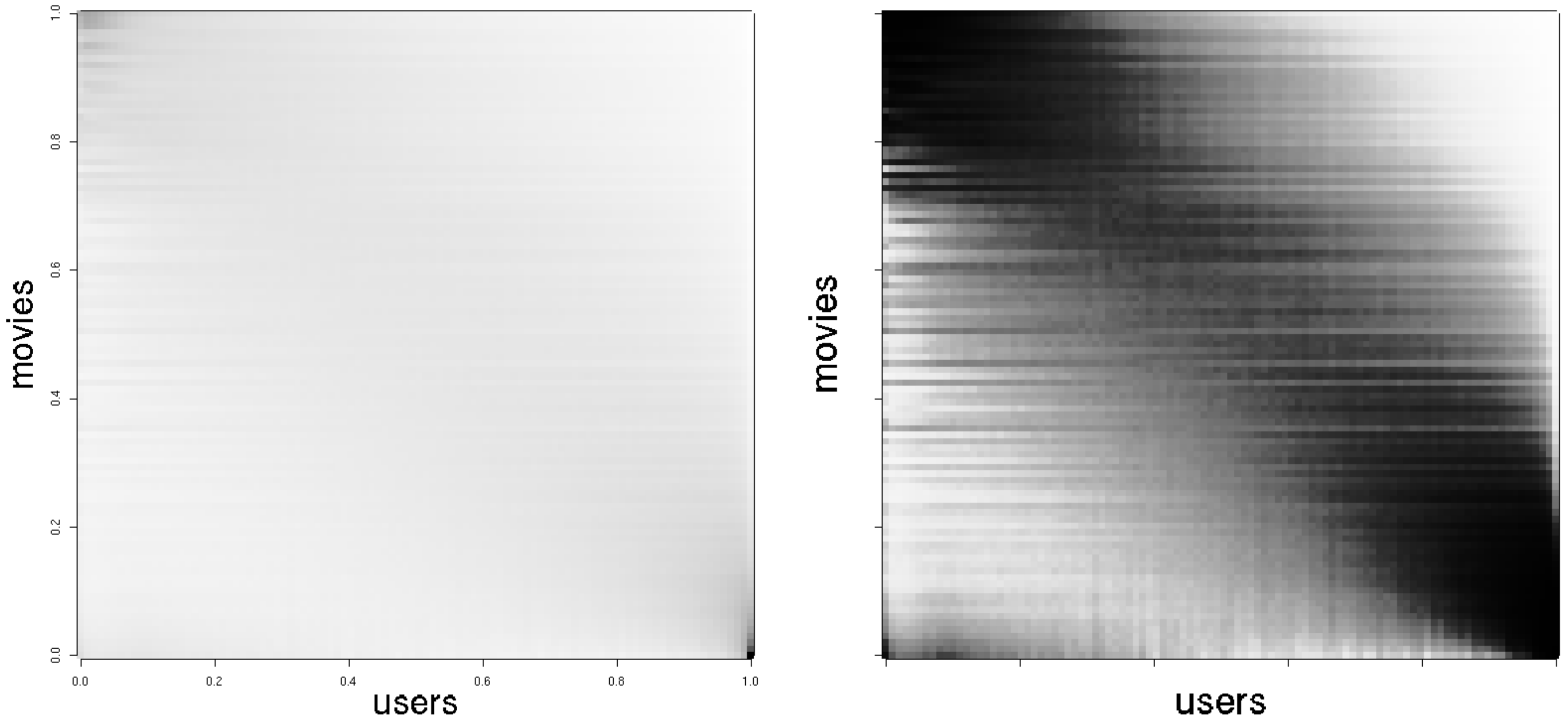
C-IV has area $0.6 \times 0.4 = 0.24$ but 0.1 of the data so it gets $0.1/0.24 = 0.42$

Netflix copula



Uses 100×100 grid of bins. Note head to tail affinity

Rescaled gray plot (Netflix)

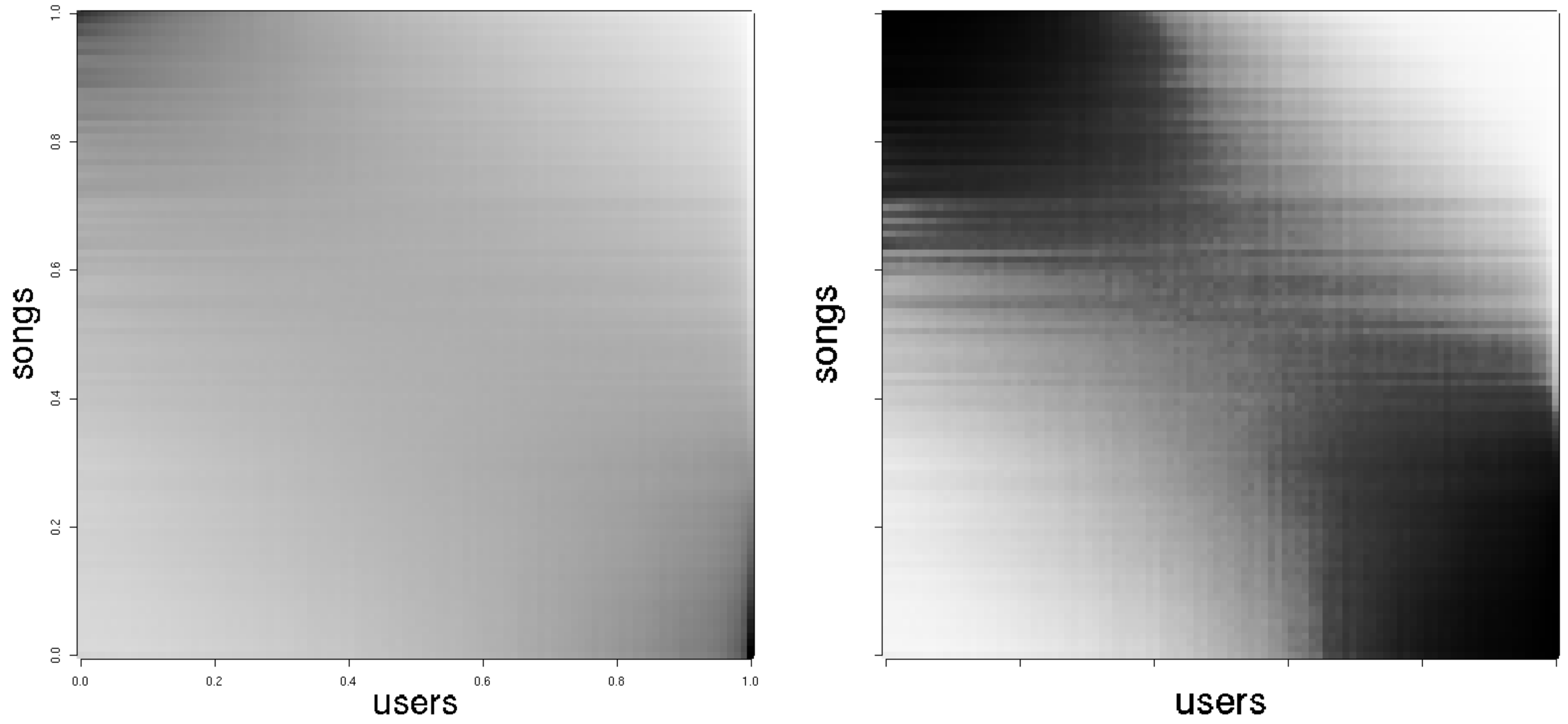


On the right 256 gray levels each get same # of pixels.

IE image rescaled to have a flat histogram

More features visible

Yahoo! songs from Webscope

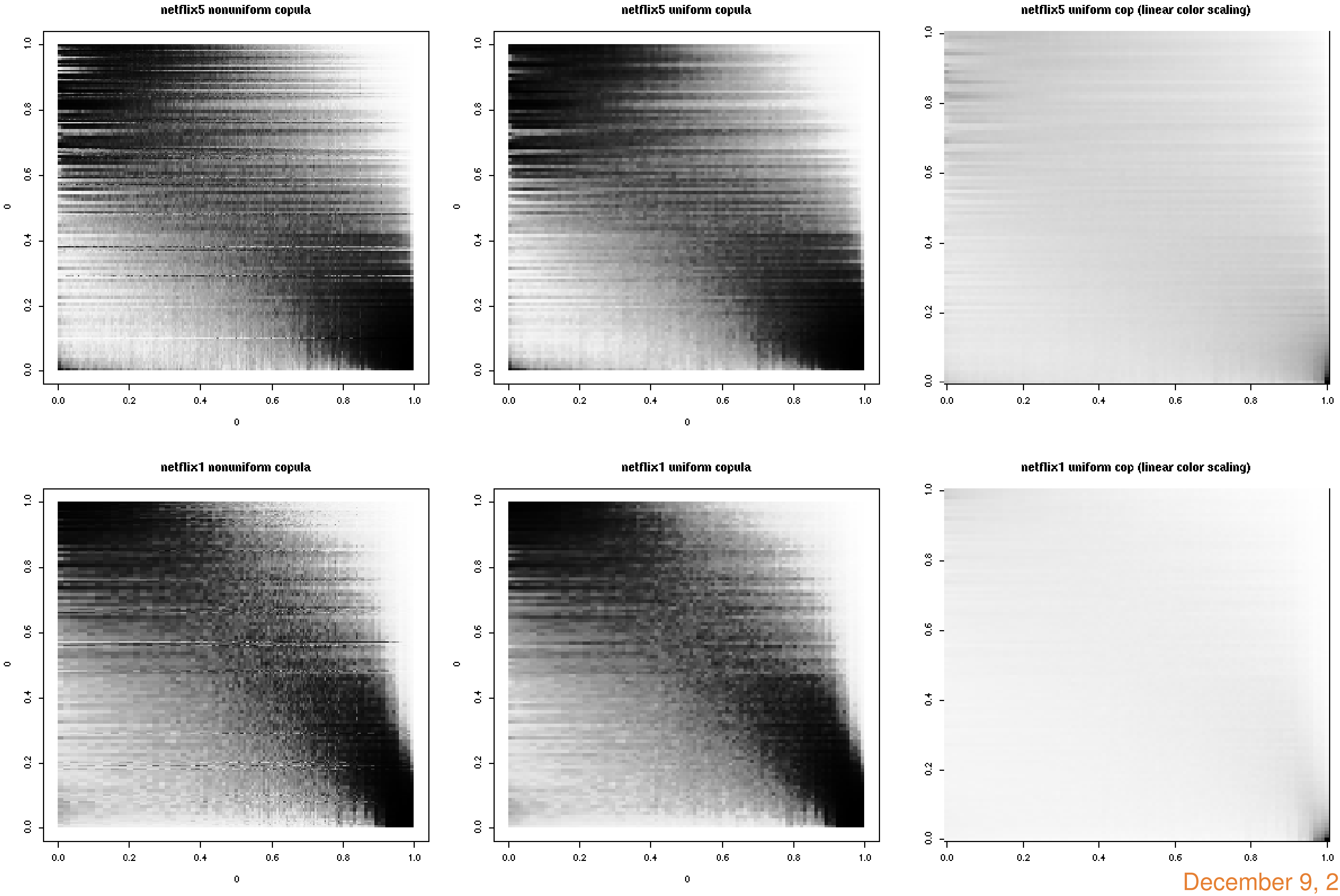


Same two figures for about $700,000,000$ song ratings

Songs have head to tail affinity too

Shape is different (flip back to Netflix to see)

Netflix extreme rankings (5s over 1s)



netflix5 nonuniform copula

netflix5 uniform copula

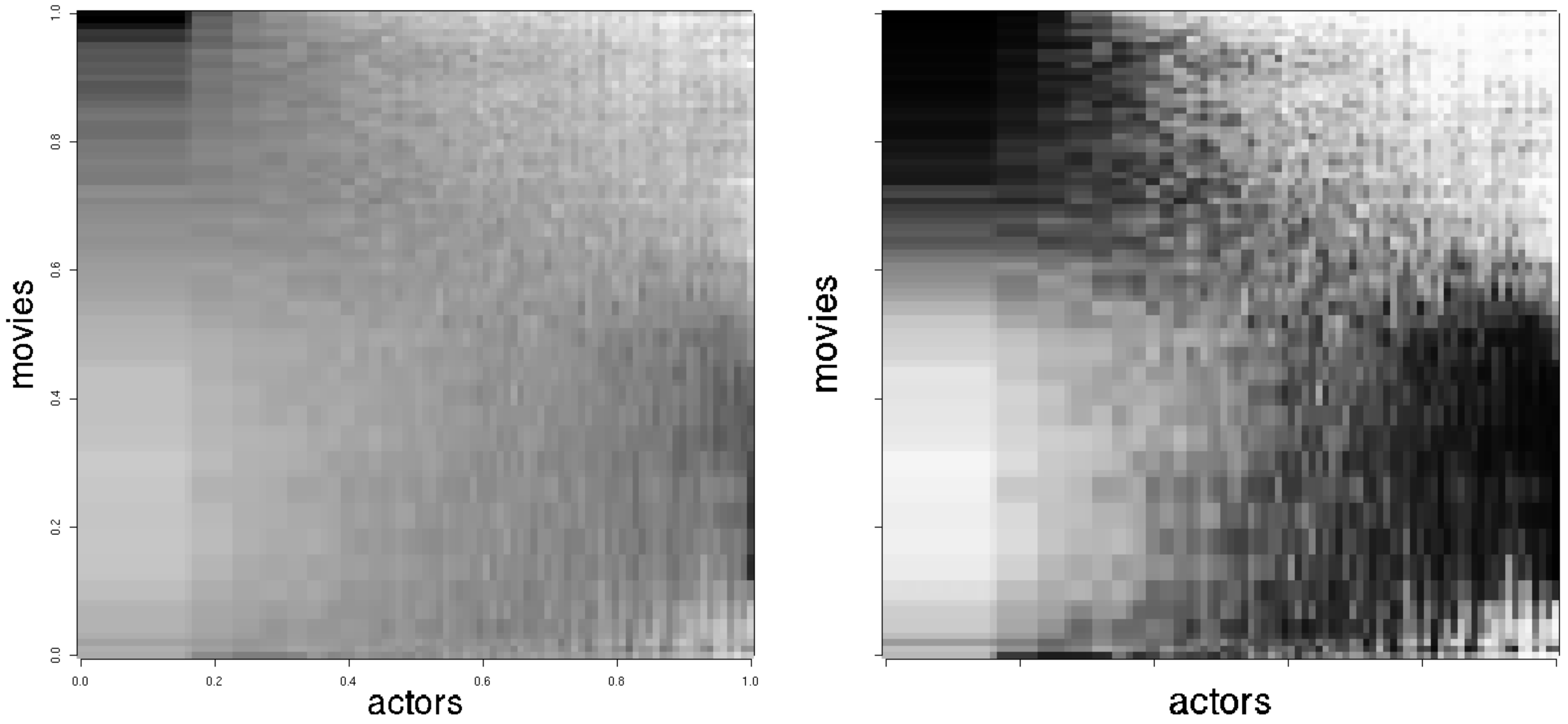
netflix5 uniform cop (linear color scaling)

netflix1 nonuniform copula

netflix1 uniform copula

netflix1 uniform cop (linear color scaling)

IMDB movies & actors

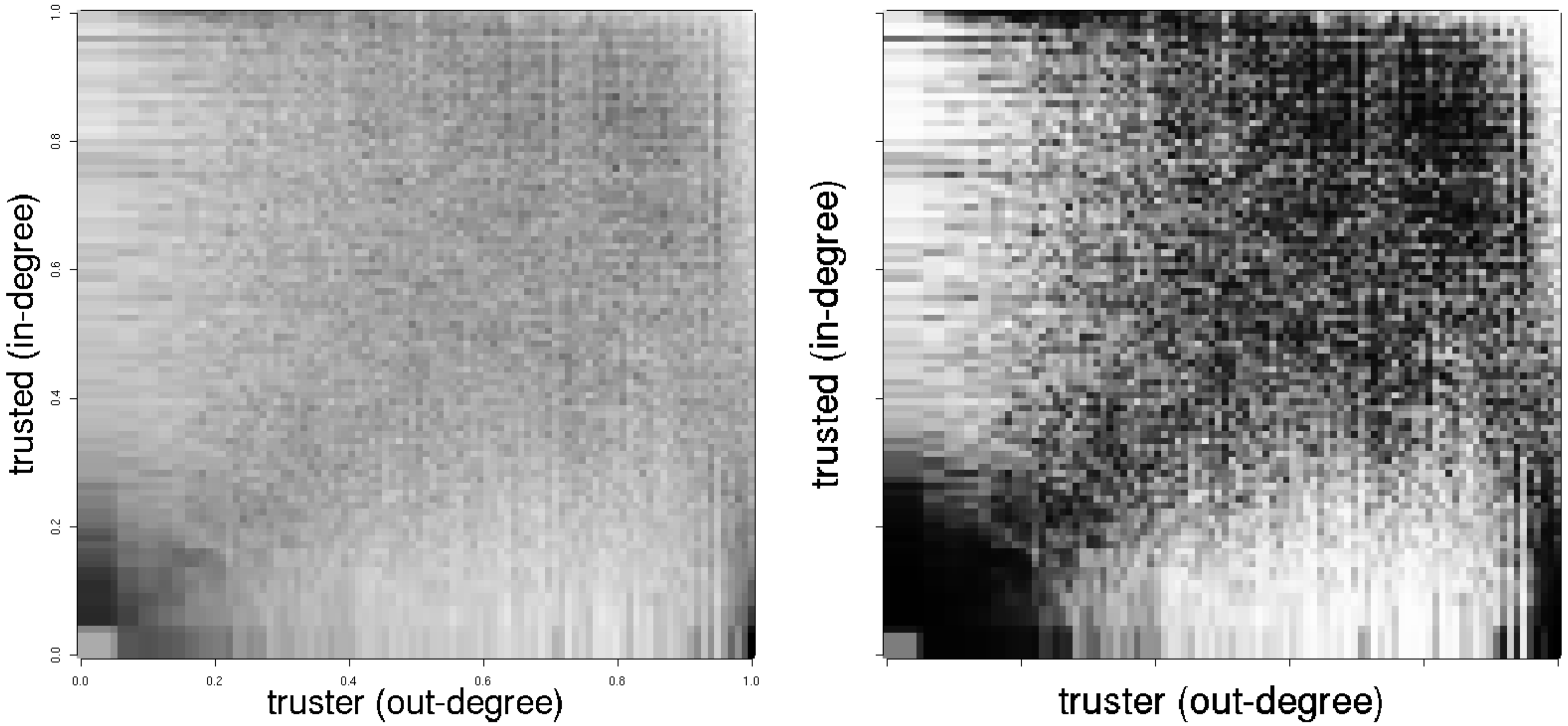


Data courtesy of Jure Leskovic

Two apparent modes

Actor distribution many 1s

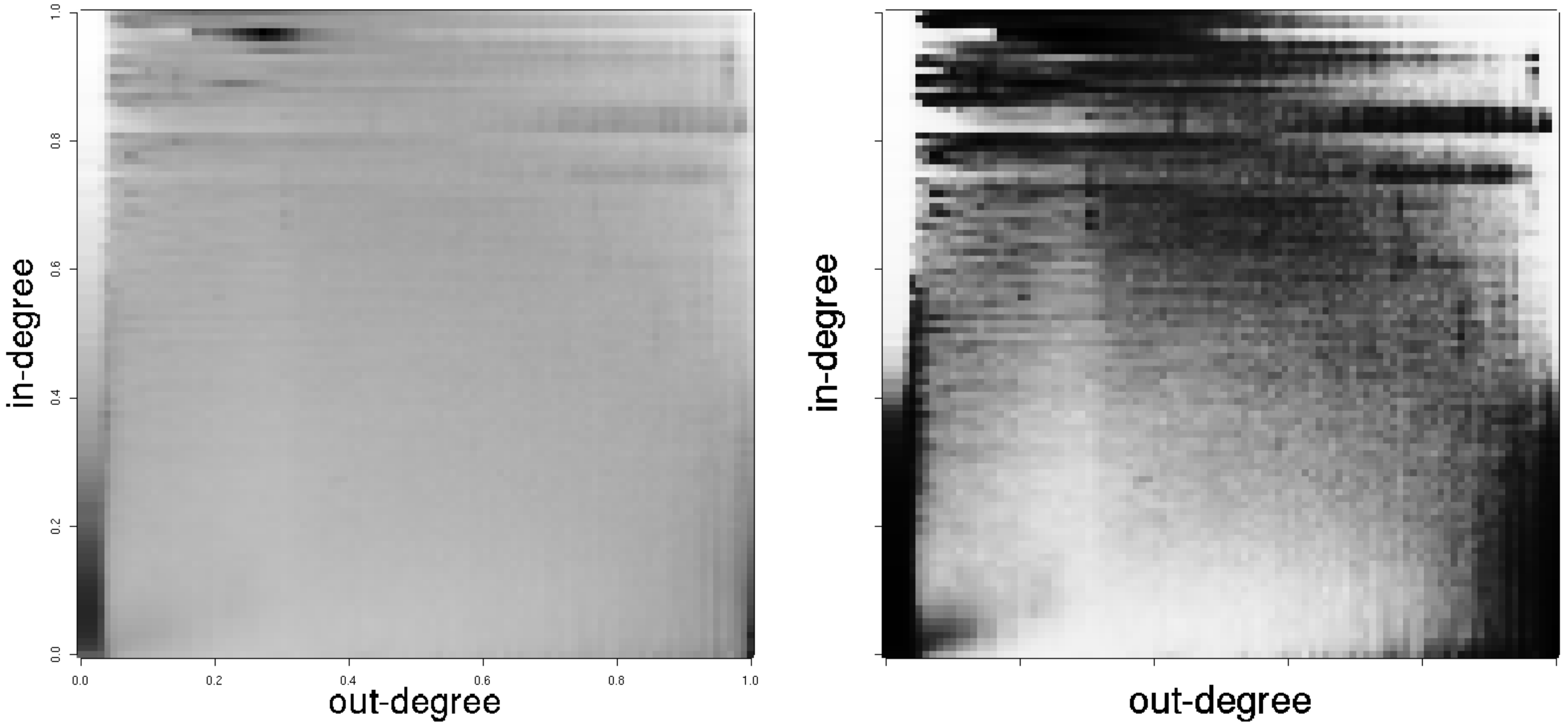
Epinions truster and trusted



Data courtesy of Jure Leskovic

Mode at lower right: those who trust many · · · trust others who are not widely trusted

Wikipedia in degree and out degree

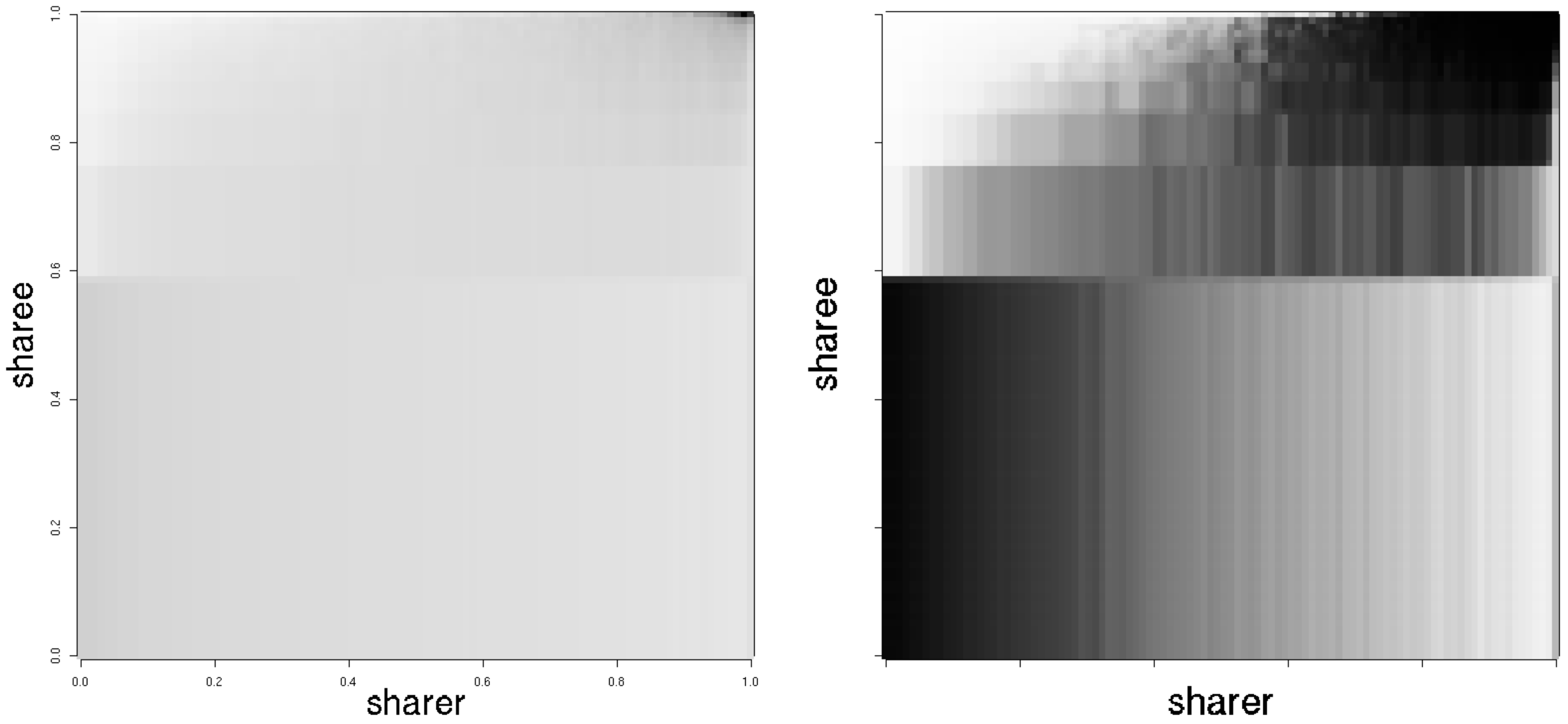


Data courtesy of David Gleich

Count of inlinks vs outlinks for Wikipedia pages

Modes correspond to hubs (eg lists), authorities (eg 1988 or France), stubs

Snapfish photo album sharing

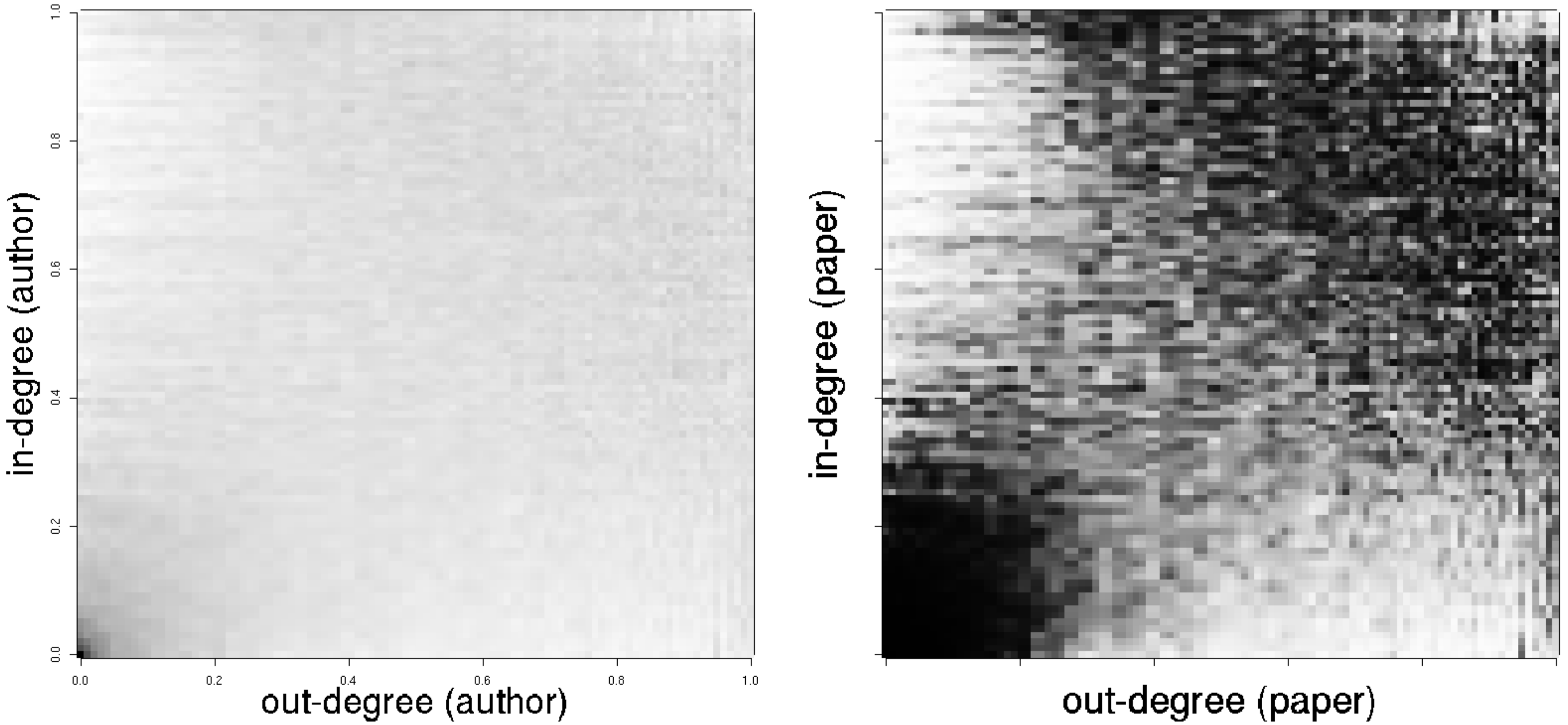


Data courtesy of Fereydoon Safai

Sharer shares photo album and sharee accepts

Note head to head affinity and lots of 1s

arXiv high energy physics citations

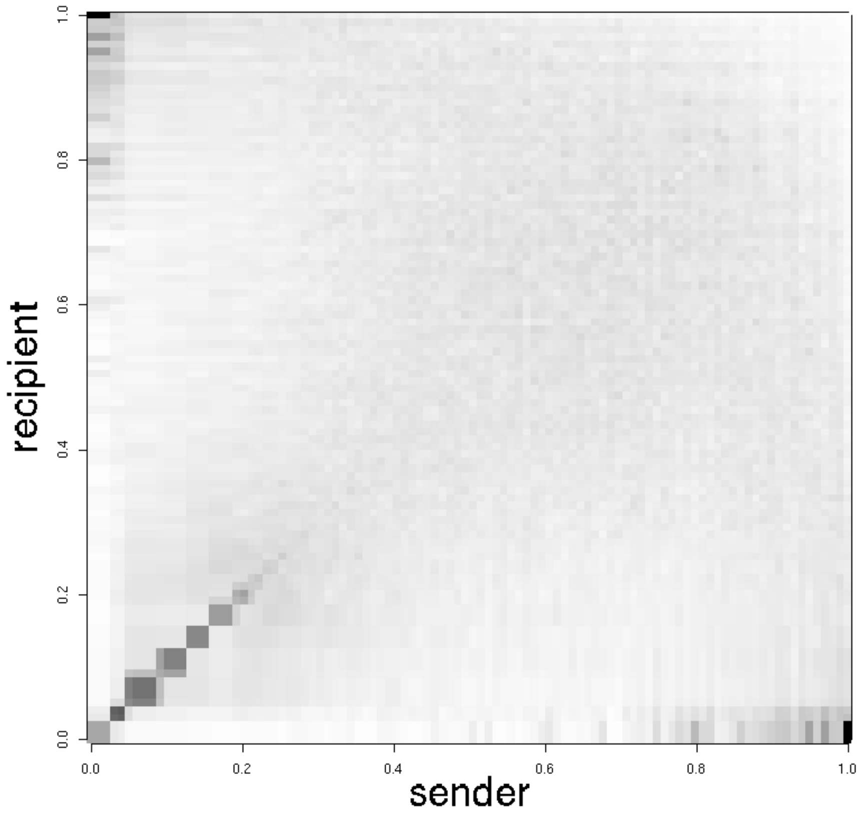


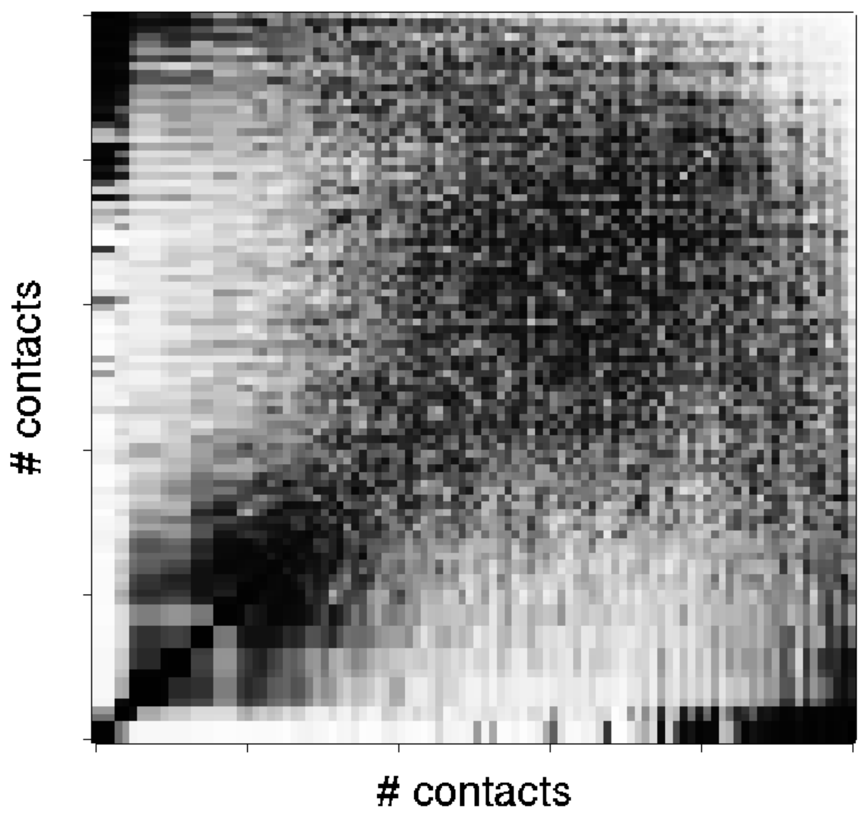
Data courtesy of Jure Leskovic

Paper citations

Strong tail to tail affinity

Enron email data



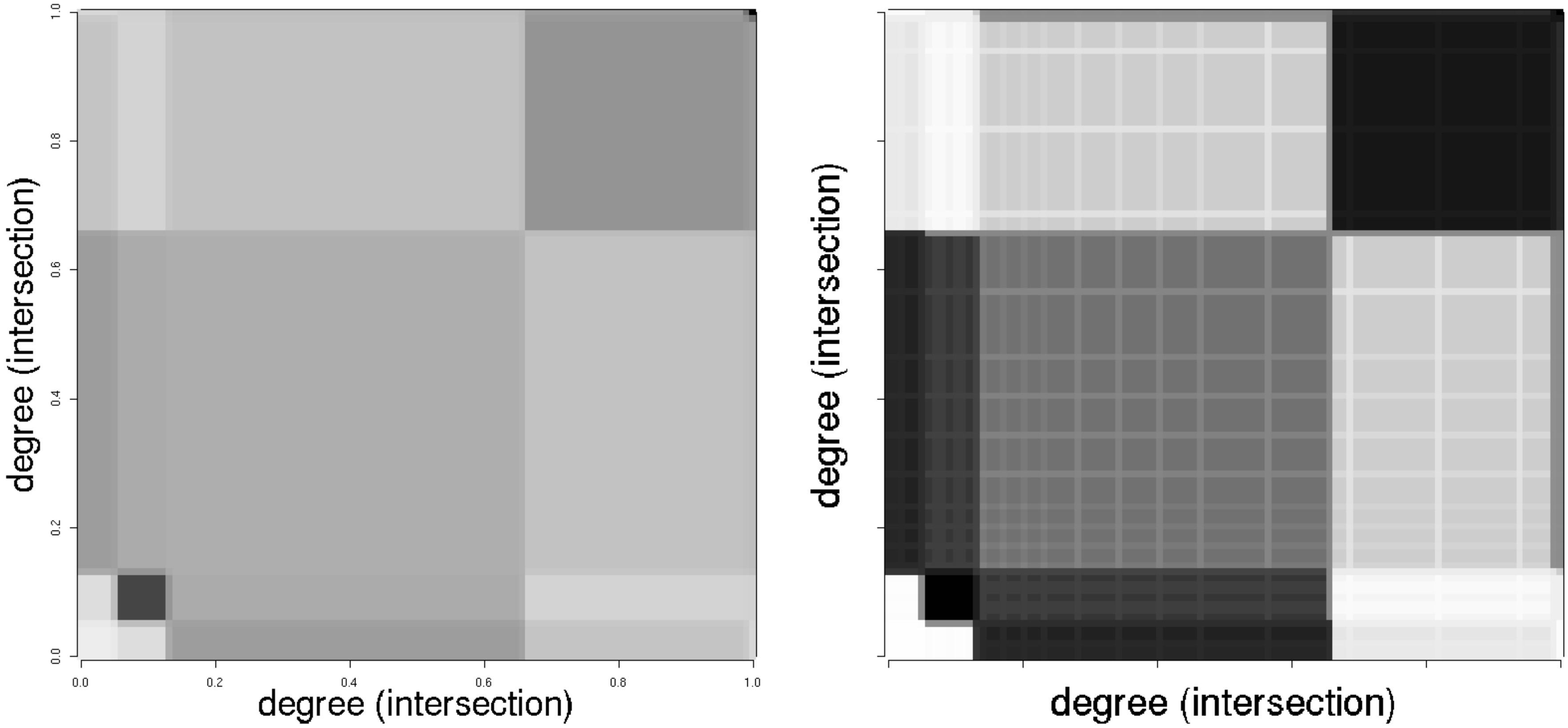


Data courtesy of Jure Leskovic

Symmetric: $X_{ij} = X_{ji} = 1$ if i and j ever emailed (either way)

Head-tail affinity, head-head aversion, tail-tail strangeness

California road networks



Data courtesy of Jure Leskovic

i and j are intersections

$X_{ij} = 1$ if any road connects i to j

Summary table

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 1: Numerical summary of corner affinities from plotted copulas. 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. Uses corners of size 0.05×0.05.

Other plots

No one plot will show all possible structure. These plots split apart communities and join them together in different ways. The view complements what we would see from clustering.

More generally, we could

1) sort rows by one feature
2) sort columns by another
3) plot gray or color to depict a third

Proper ordering

The plot shows a snapshot of the data X_i.

We want it to represent an underlying phenomenon like $\mathbb{E}(X_i)$.

Each data point contributes a drop of gray somewhere in $[0, 1]^2$.

We want it to be near its proper place.

If it is right horizontally and vertically then it is right spatially.

So it is enough to look at each entity separately.

Zipf–Poisson ensemble

The model is

$$X_i \sim \text{Poi}(Ni^{-\alpha}), \quad 1 \le i < \infty$$

independently, where $\alpha > 1$.

True pattern is Zipf, we get a noisy sample, and $N \to \infty$. We have

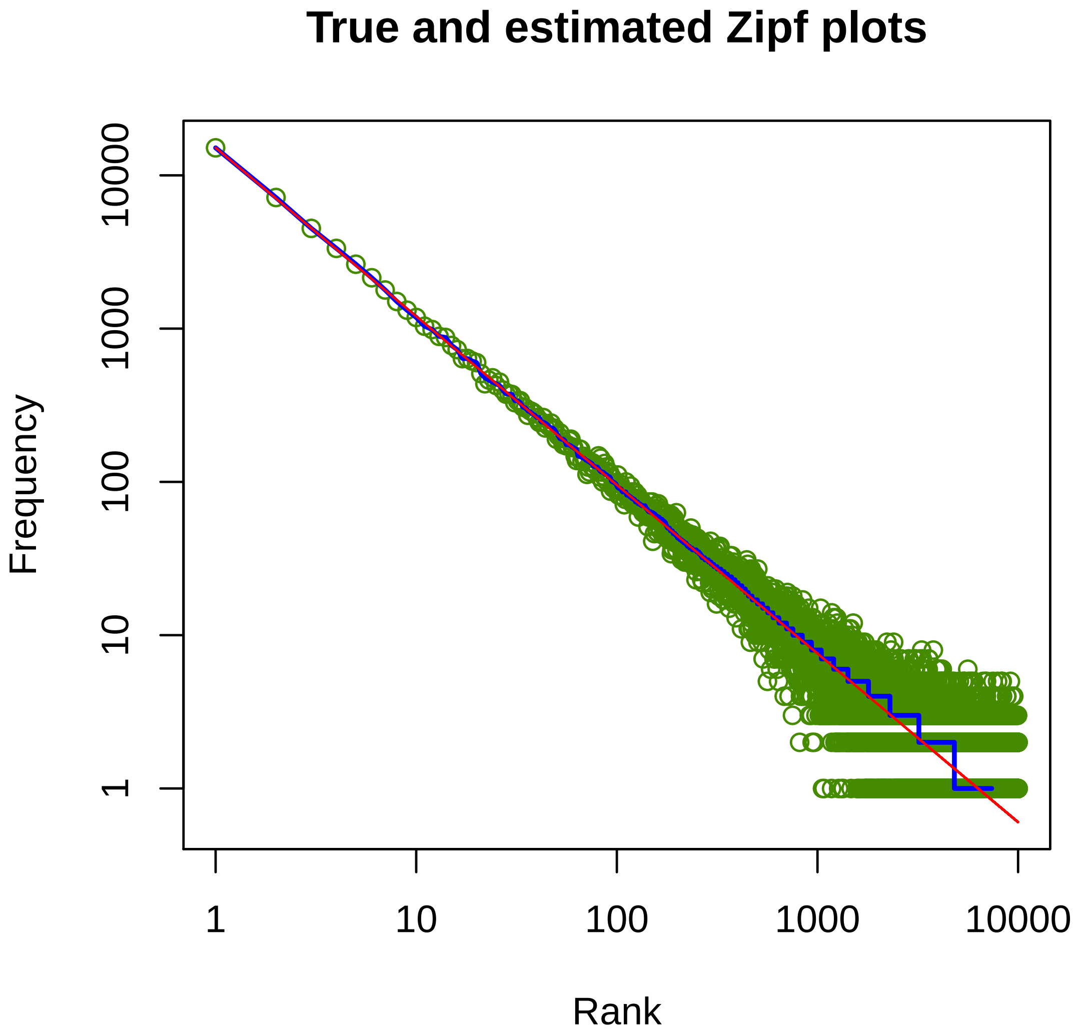
$$\mathbb{E}(X_1) > \mathbb{E}(X_2) > \mathbb{E}(X_3) > \cdots > \mathbb{E}(X_n) > \cdots$$

but maybe some X_i are out of order.

Zipf–Mandelbrot–Poisson ensemble

$$X_i \sim \text{Poi}(N\theta_i), \quad \theta_i = (i+k)^{-\alpha}$$

A sample from Zipf–Poisson



$N = 10^6 \quad \alpha = 1.1$

$X_i \sim \text{Poi}(N i^{-\alpha})$

Green X_i vs i

Red $\mathbb{E}(X_i)$ vs i

Blue $X_{(i)}$ vs i

Large entities correctly ordered.

Medium have small errors.

Small ones all over the place.

Correct ordering in Zipf-Poisson

Theorem 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*

$$\lim_{N\to\infty} \Pr\bigl(X_1 > X_2 > \cdots > X_n\bigr) = 1. \tag{1}$$

Upshot

The first $n(A) = (AN/\log(N))^{1/(\alpha+2)}$ entities have correct relative order.

Furthermore: For any $B < A$, the first $n(B)$ are correctly ordered with no interlopers.

Finally: if we take out the $\log$ factor then

$$\lim_{N\to\infty} \Pr\bigl(X_1 > X_2 > \cdots > X_n\bigr) = 0. \tag{2}$$

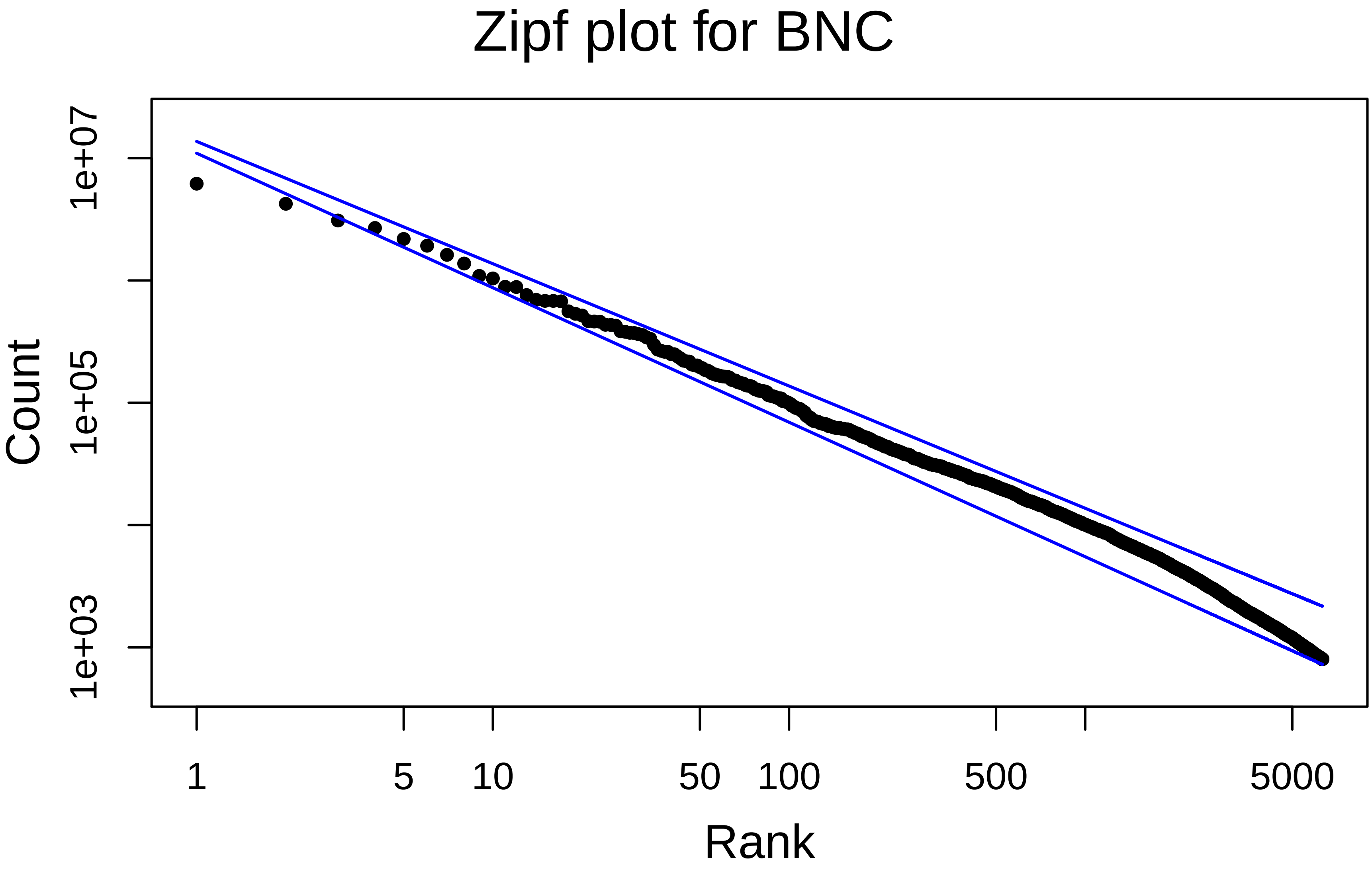
British National Corpus

Rank	Word	Count
1	the	6,187,267
2	be	4,239,632
3	of	3,093,444
4	and	2,687,863
5	a	2,186,369
6	in	1,924,315
7	to	1,620,850
8	have	1,375,636
9	it	1,090,186
10	to	1,039,323

Top ten words from the BNC, via Kilgarrif. Total count was $\approx 10^8$. Item 7 is the word ‘to’, used as an infinitive marker, while item 10 is ‘to’ used as a preposition.

“I went to (#10) the library to (#7) read.”

If we took another large sample ’the’ would win again. We would never see 0 or 2 such popular words.



Reference lines have slopes $\alpha = -1$ and $\alpha = -1.1$

The theorem estimates we'd get about 72 words in the right order. (Close match to simulations.)

Top 72 words account for about half of the total corpus.

More generally

$$X_i \sim \text{Poi}(N\theta_i), \qquad \theta_i > \theta_{i+1}$$

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

This is non-asymptotic. You can plug in your favorite hypothesized θ_i and then find n with

$$\Pr(X_1 > X_2 > \cdots > X_n) \geq 0.99.$$

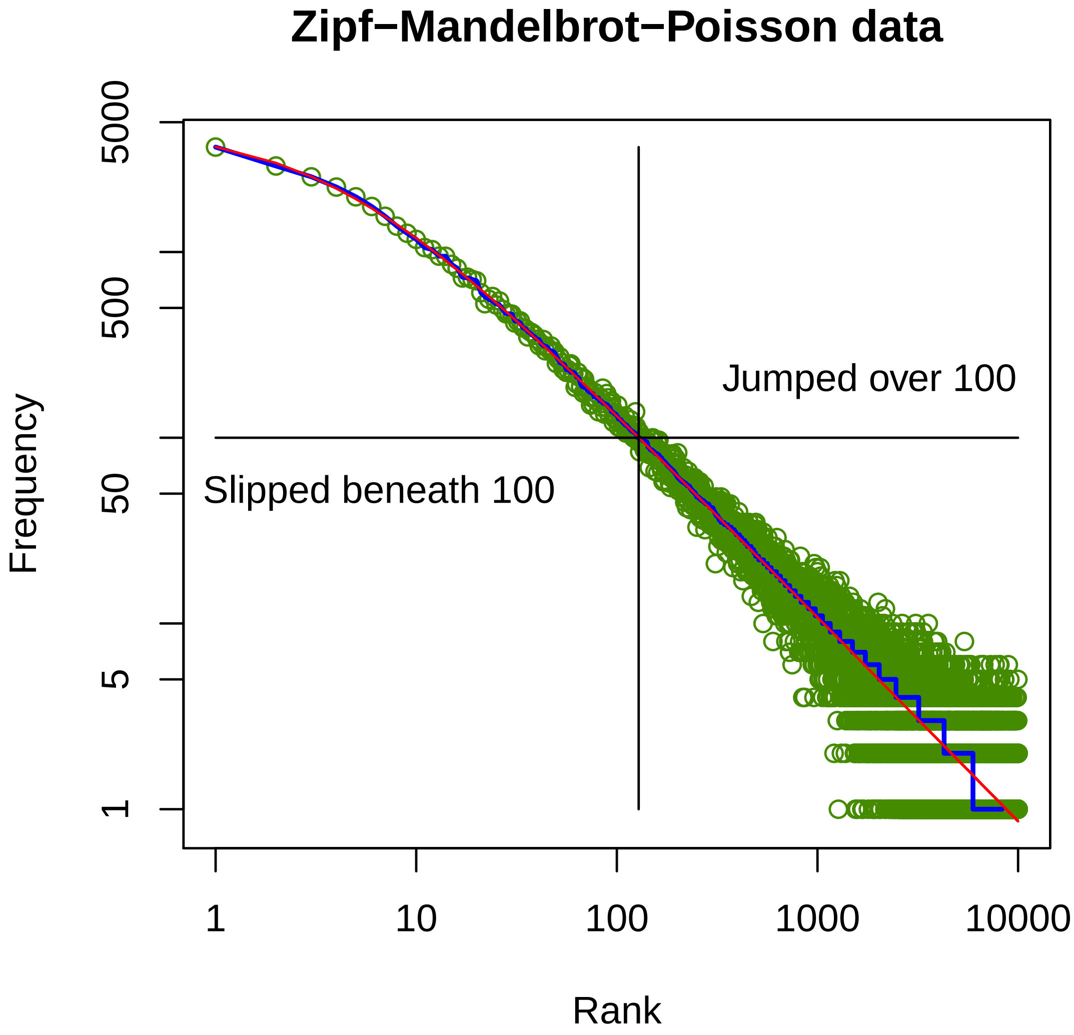
Bulk ordering

For large N, the sample ordering should be pretty good because

- The top entities get put in the right order.
- The middle ones don't move far, and,
- The bottom ones don't matter. (too rare)

We want to show that not much data lands from from where it should.

Jumpers and slippers



$N = 10^6 \quad \alpha = 1.1 \quad k = 4$

$X_i \sim$ Poi$(N(i+k)^{-\alpha})$

Green X_i vs i

Red $\mathbb{E}(X_i)$ vs i

Blue $X_{(i)}$ vs i

Some entities jumped over 100

Some entitiles slipped below 100

Most neither jumped nor slipped

Little data moved

Jumping and slipping

Even less data can jump from below σ to over τ where $\tau > \sigma > 0$:

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

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

We find

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

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

Accuracy of the plot

100 bins require 99 thresholds

β_ℓ for $\ell = 1, \dots, 99$.

$$\sum_{\ell=1}^{99} J(\beta_\ell, \beta_\ell - \epsilon_\ell) = O\Bigl(\frac{1}{N}\Bigr)$$

Markov's inequality: $\quad \Pr(J > \varepsilon) \le \mathbb{E}(J)/\varepsilon$

Not much can jump or slip for large N and 100 bins

Probably $o(\sqrt{N})$ bins could be used

Head to tail affinities

Maybe newcomers primarily rate the blockbusters while experienced movie raters know many rare gems. That provides a **taste based** explanation for head to tail affinity.

Another explanation: **saturation**. Nobody rates a movie twice and so busy raters quickly run through the popular movies and necessarily rate less known ones.

Saturation model

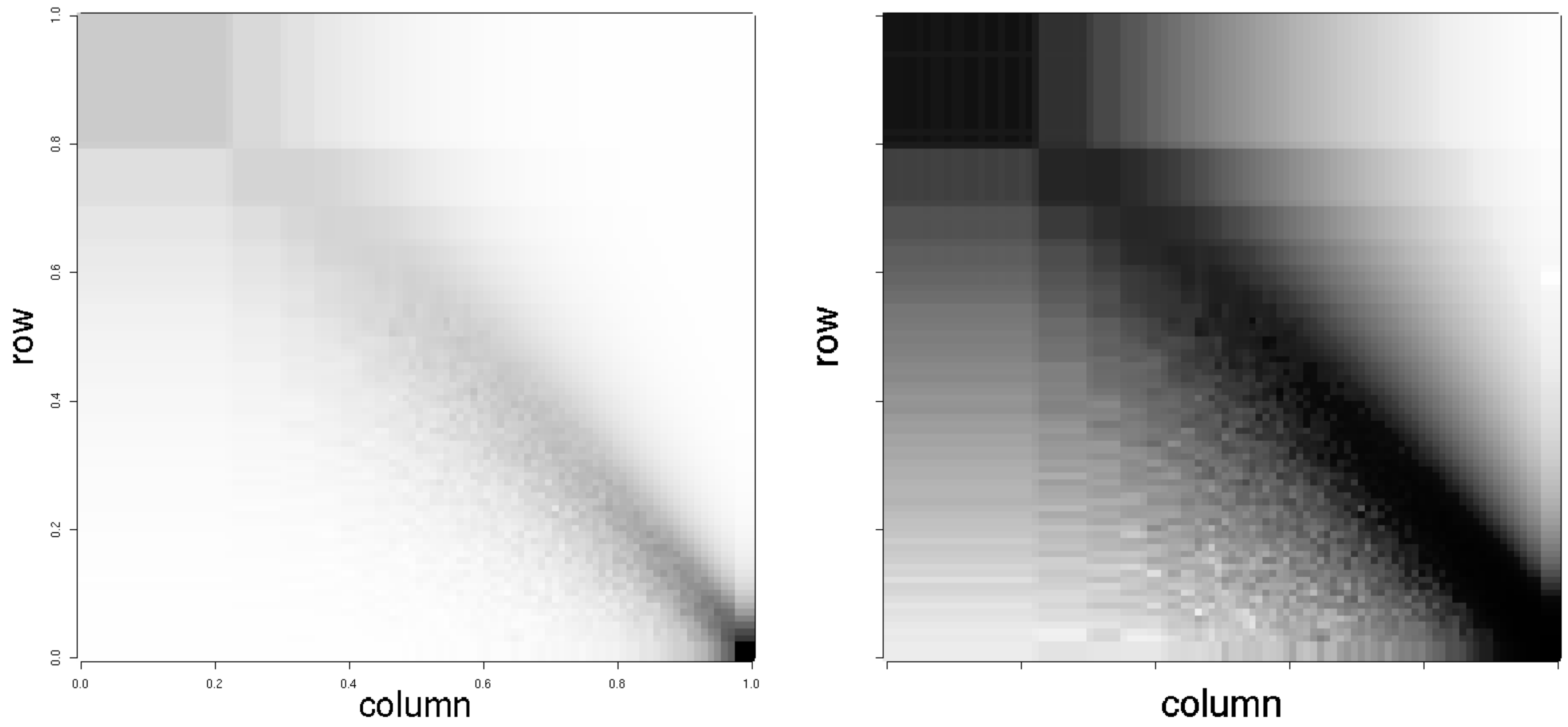
$$Y_{ij} \sim \mathrm{Poi}(N i^{-\alpha} j^{-\beta}) \quad \text{latent ratings}$$

$$X_{ij} = \begin{cases} 1 & Y_{ij} \geq 1 \\ 0 & \text{else.} \end{cases} \quad \text{observed data}$$

Entities i and j are independent in the latent model.

For small i and j we get $\mathbb{E}(Y_{ij}) \gg 1$ but $X_{ij} \leq 1$. The head to head matches are depleted bringing a head to tail affinity.

Saturation model



Left is raw, right is histogram equalized. Saturation brings a head to tail affinity.

It also brings strange marginal distributions.

When $\mathbb{E}(Y_{ij}) = Ni^{-a}j^{-b}$, then one margin decays at slope a/b and the other at b/a.

Bipartite preferential attachment

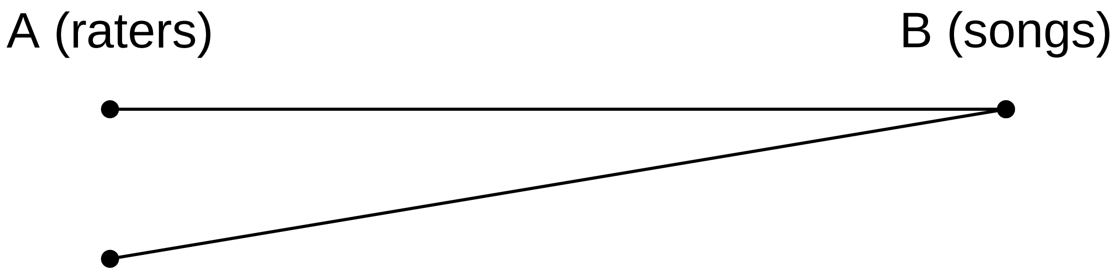
Graph at time t connects $m(t)$ nodes of type A to $n(t)$ nodes of type B.

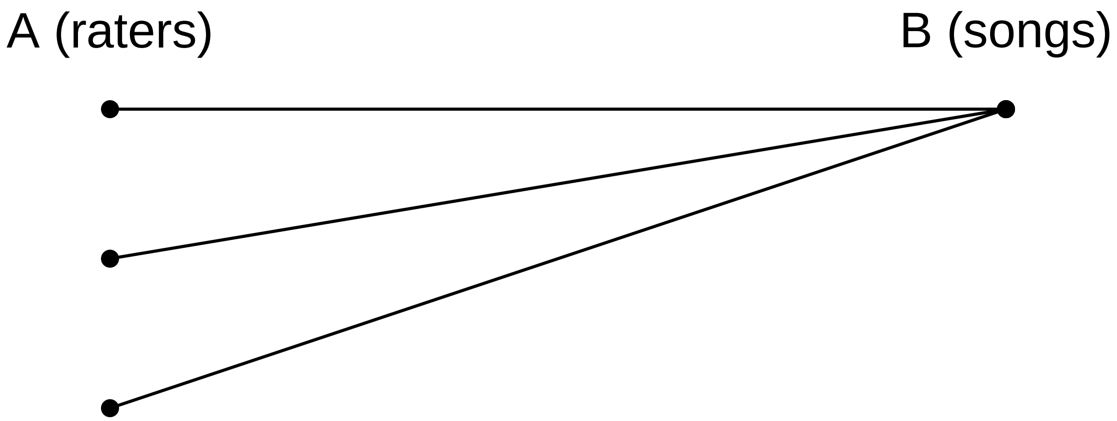
Start out with $n(1) = m(1) = 1$ (one node of each type) connected.

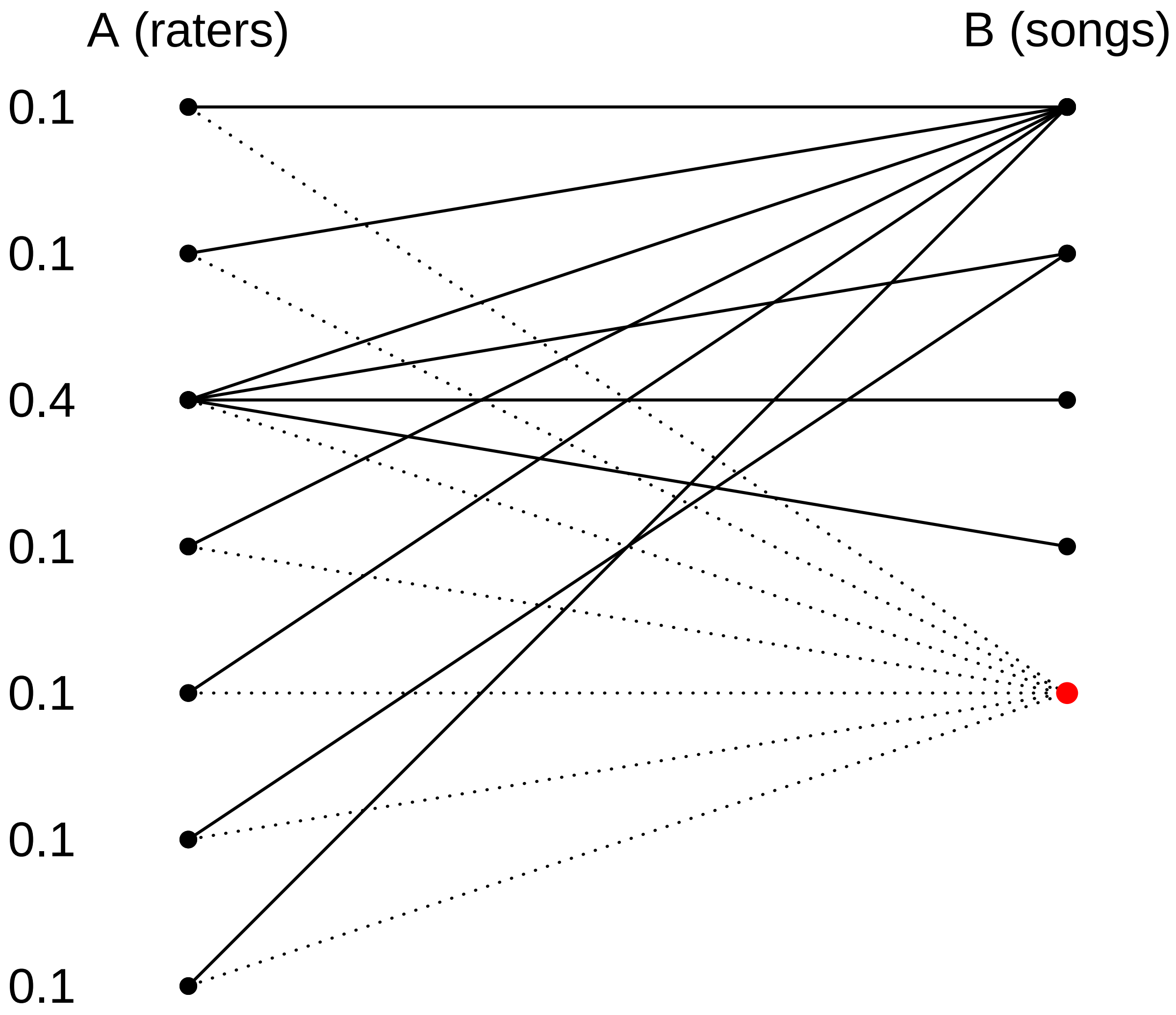
At time t toss a p coin. If heads add a new A node connected to random B node sampled with probability proportional to degree. For tails, add a B node connected by preferential attachment to an old A node.

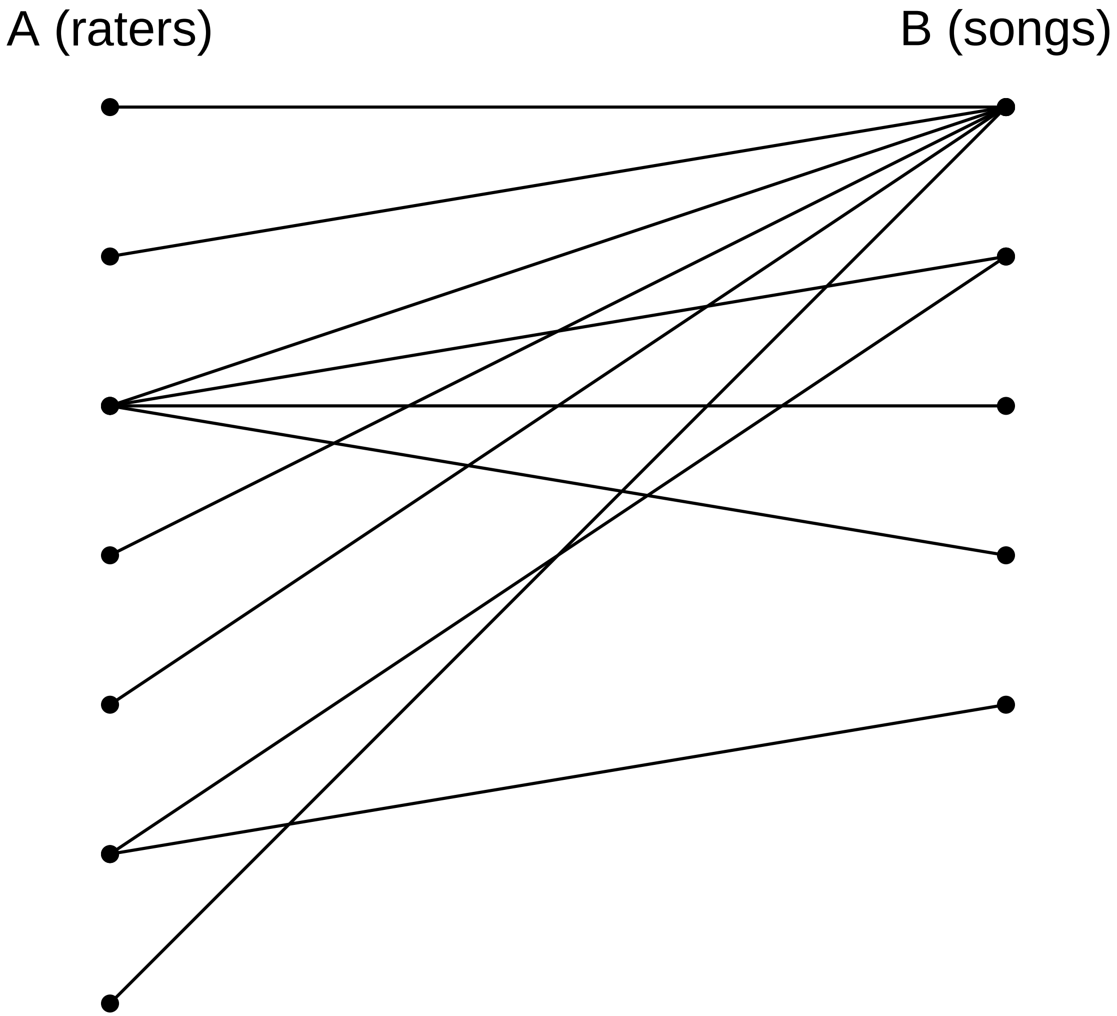
A (raters)

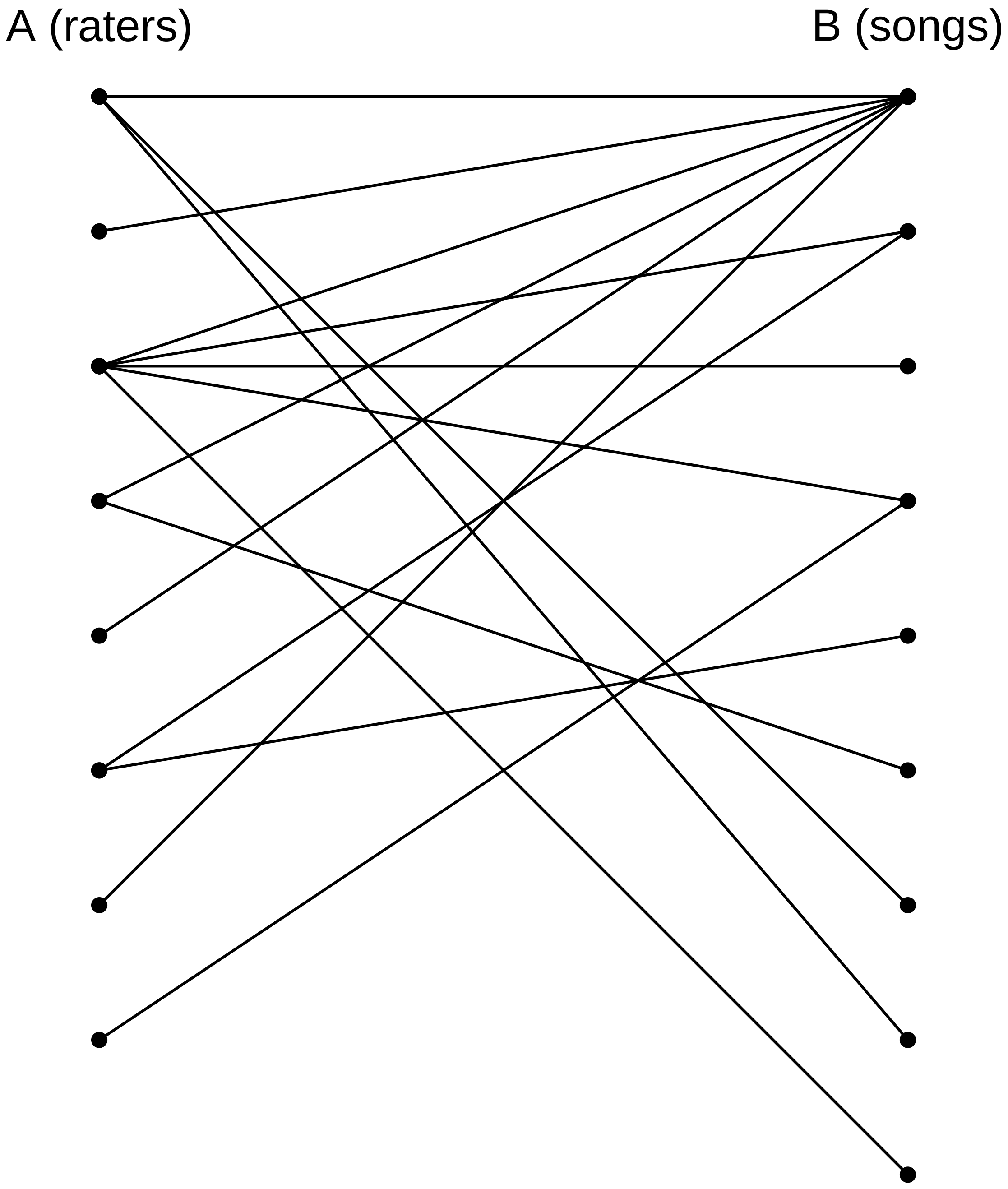
B (songs)


A (raters)
B (songs)

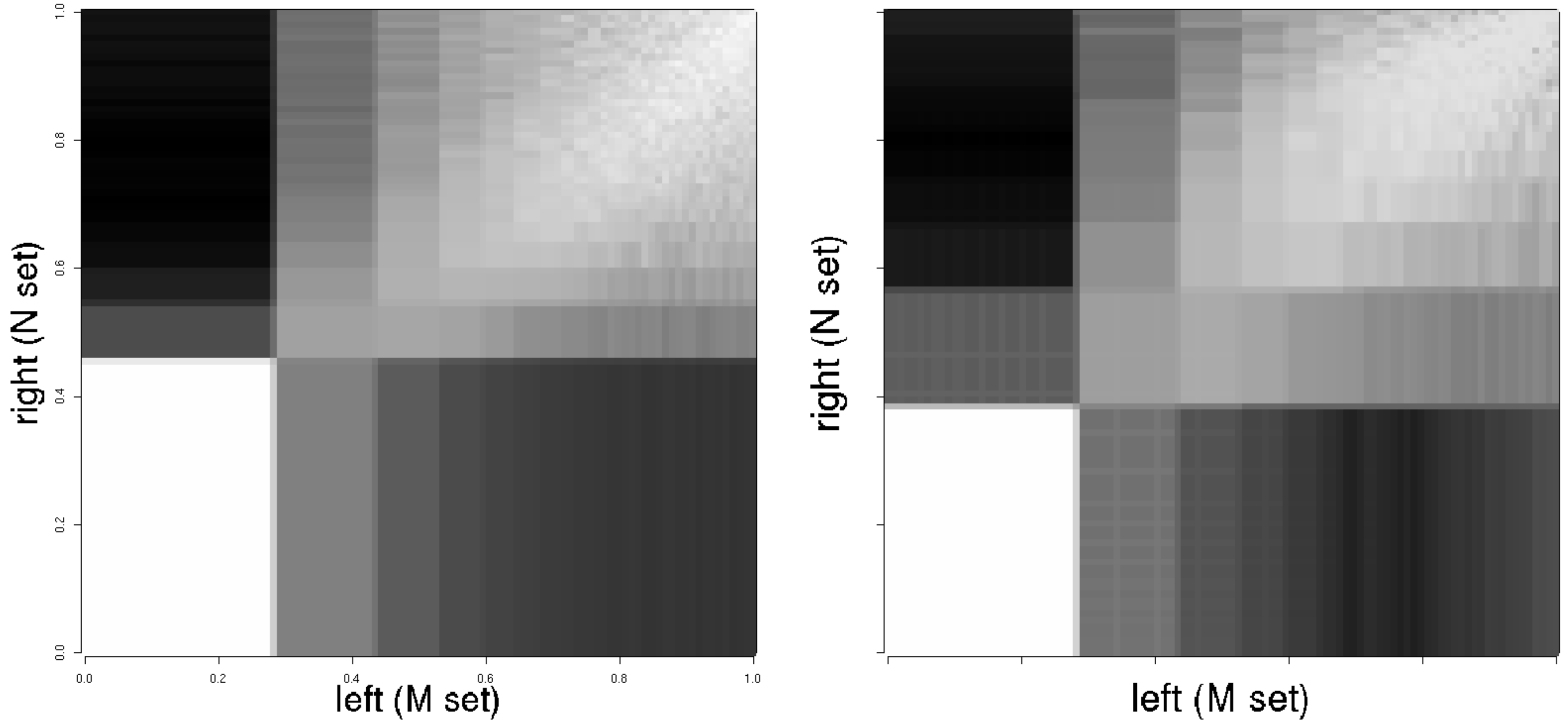

A (raters)
B (songs)


A (raters)
B (songs)
0.1
0.1
0.4
0.1
0.1
0.1
0.1


A (raters)
B (songs)


A (raters)
B (songs)

Bipartite pref. attachment copula



One margin has a tail rate $\alpha \in (2,3]$ the other has a tail rate $\beta \in [3,\infty)$. Like unipartite preferential attachment's slope of 3.

Conclusions

1) We can learn from plots of bivariate long tailed data

2) Ordering errors are not too severe

3) We can mimic head to tail affinities with simple generative models

Thanks

1) Deepak Agarwal for the invitation

2) Tao Qin, Deepak Agarwal, James Shanahan, Tie-Yan Liu, organizers

3) Microsoft Research, workshop sponsor

4) Church and Duncan group, workshop support

5) Jure Leskovic, Fereydoon Safai, David Gleich, data

6) NSF DMS-0906056, funding

These slides are slightly edited version of the ones presented in a keynote talk at NIPS 2010 workshop MLOAD: Machine Learning in Online Advertising.