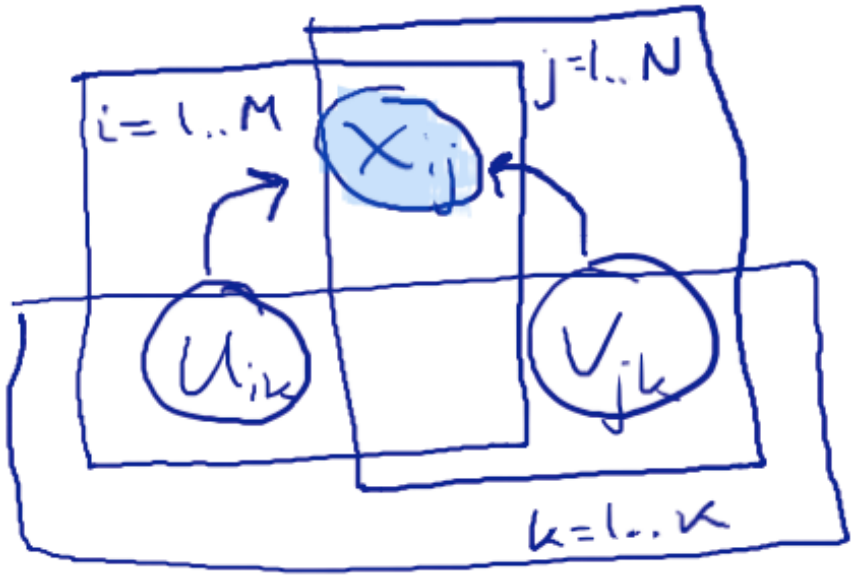
Review

- Gibbs sampling
  - MH with proposal
    - $Q(\mathbf{X} \mid \mathbf{X}') = P(\mathbf{X}_{B(i)} \mid \mathbf{X}_{\neg B(i)})\ I(\mathbf{X}_{\neg B(i)} = \mathbf{X}'_{\neg B(i)}) / \#B$
  - failure mode: "lock-down"
- Relational learning (properties of ***sets*** of entities)
  - document clustering, recommender systems, eigenfaces

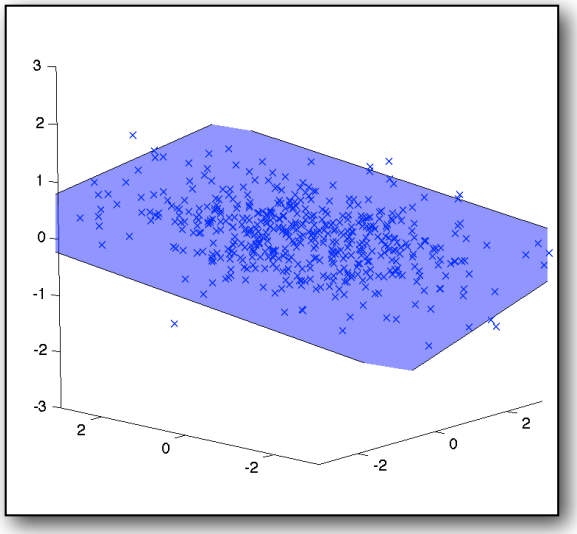
Review



- Latent-variable models
- PCA, pPCA, Bayesian PCA
  - everything Gaussian
  - $E(X \mid U,V) = UV^T$
  - MLE: use SVD

$$E(x_{ij} \mid u, v) = \sum_k u_{ik} v_{jk}$$

- Mean subtraction, example weights

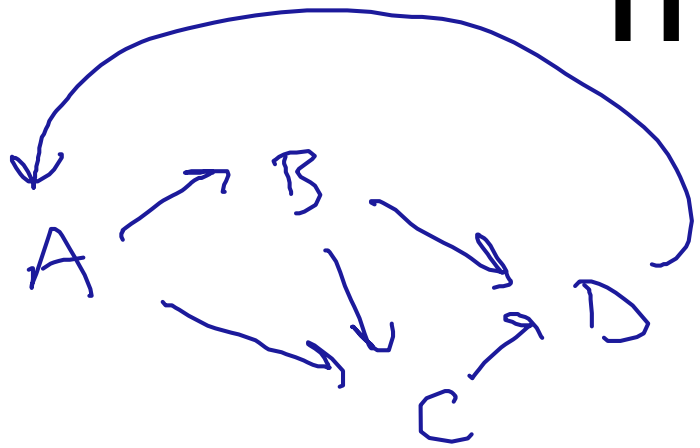


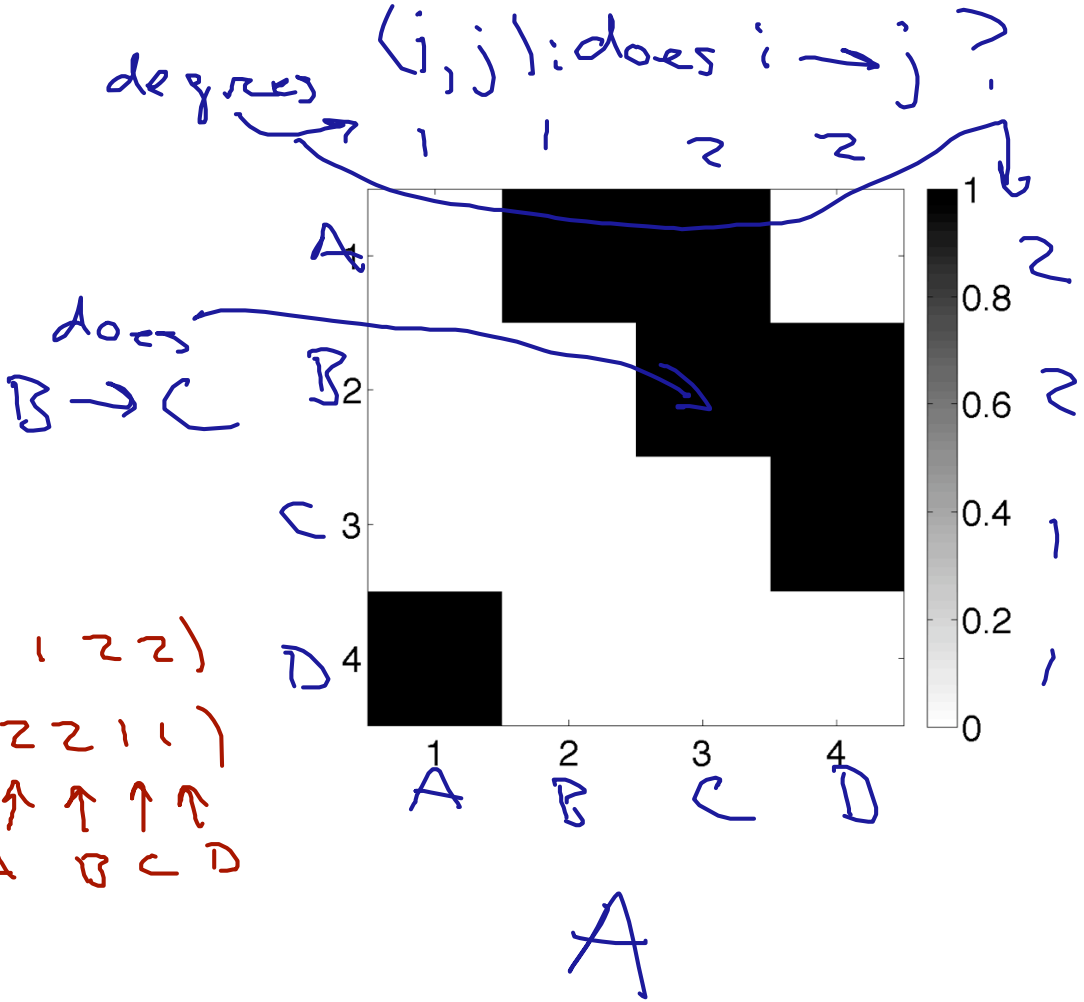
PageRank

Eigen-decomposition

- ~~SVD~~ is pretty useful: turns out to be main computational step in other models too
- A famous one: PageRank
  - Given: web graph (V, E)
  - Predict: which pages are important

PageRank: adjacency matrix





in-degrees = col sums = (1 1 2 2)

out-degrees = row sums = (2 2 1 1)

for A B C D

Random surfer model

- W. p. α: jumps uniformly to new page
- W. p. (1–α): follow random link

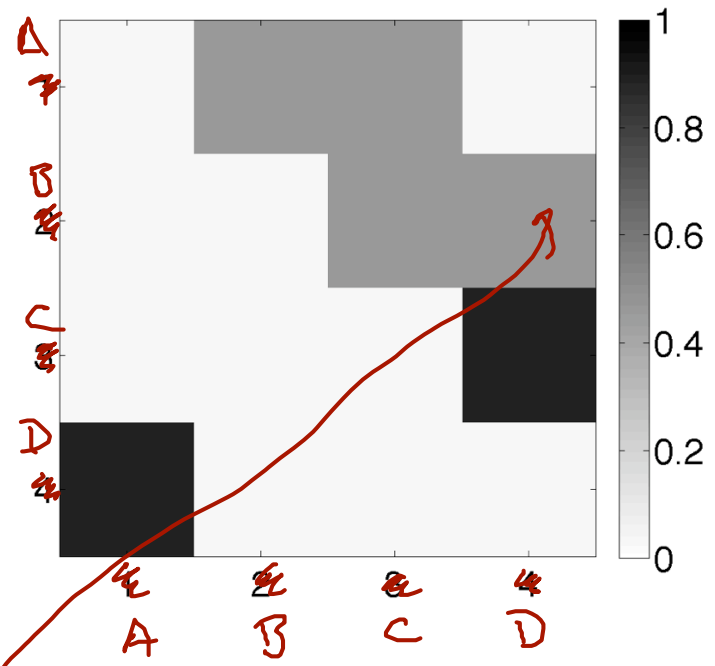
$T_{ij} = P(j \mid i) = \alpha \frac{1}{N} + (1-\alpha) A_{ij}/d_i$

$d_i = \sum_j A_{ij}$

- Intuition: page is important if a random surfer is likely to land there

$\alpha = 0.15$

T matrix



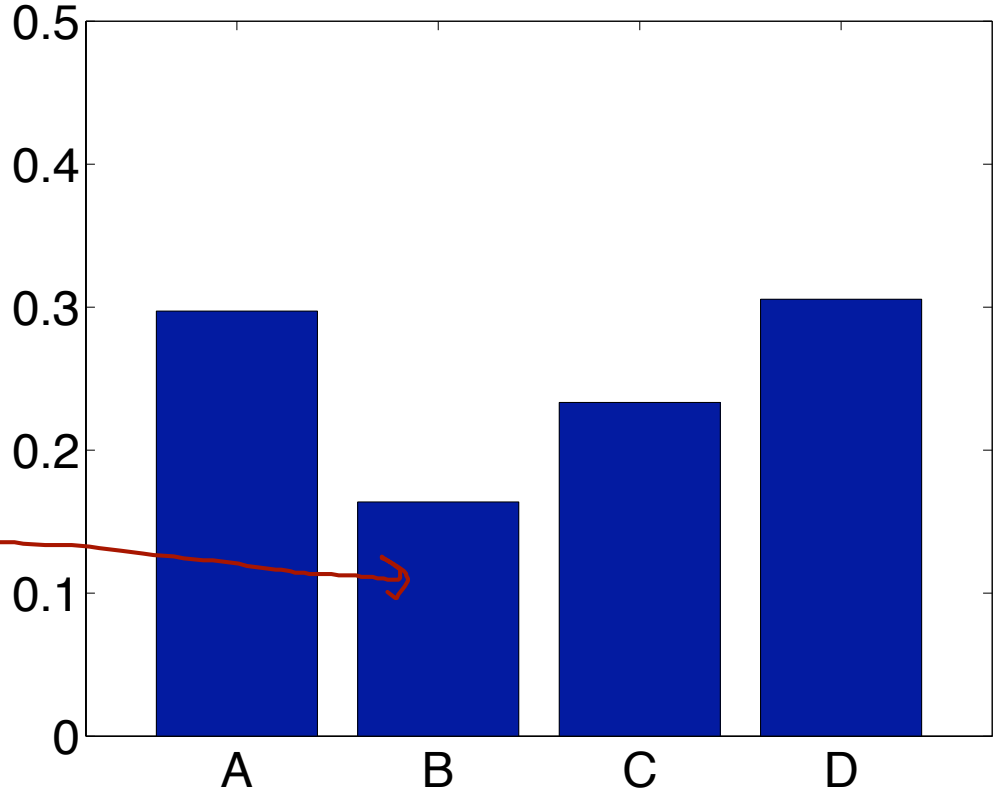
$T_{BD} = \frac{\alpha}{4} + (1-\alpha)\frac{1}{2}$

Stationary distribution

$$\pi_j = \sum_i \pi_i T_{ij}$$

$$\pi^T T = \pi^T$$

π is left eigenvector of T w/ eigenvalue 1



random surfer visits B least

$\Rightarrow$ B is least important

Thought experiment

- What if A is symmetric?
  - note: we're going to stop distinguishing A, A'

$T_{ij} = A_{ij} / d_i$ $\quad d_i = \sum_j A_{ij}$ $\quad A_{ij} = A_{ji}$

$D = \mathrm{diag}(d)$ $\quad T = D^{-1} A$

$\pi^T D^{-1} A = \pi^T$

suppose $\pi = d \Rightarrow \pi^T D^{-1} = 1^T$

$\Rightarrow \pi^T D^{-1} A = 1^T A = d^T = \pi^T$

- So, stationary dist'n for symmetric A is: d boring!!
- What do people do instead?

2nd eigenvector of T

Spectral embedding

- Another famous model: spectral embedding (and its cousin, spectral clustering)
- Embedding: assign low-D coordinates to vertices (e.g., web pages) so that similar nodes in graph $\Rightarrow$ nearby coordinates
  - A, B similar = random surfer tends to reach the same places when starting from A or B

Where does random surfer reach?

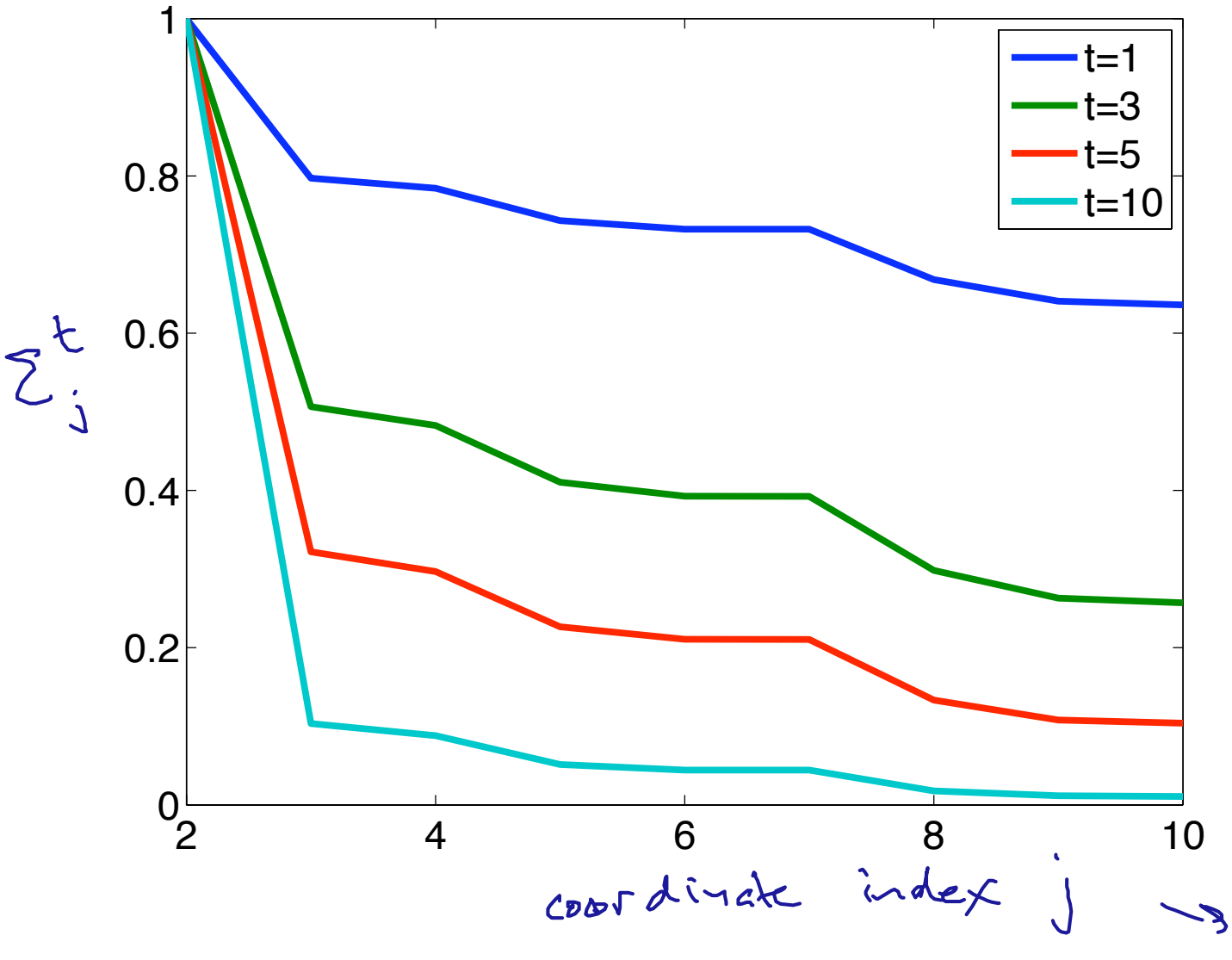
$T = D^{-1}A$

- Given graph: adjacency A degree $= d$ $D = \deg(d)$
- Start from distribution π
  - after 1 step: $P(k \mid \pi, 1\text{-step}) = \sum_i \pi(i) P(k \mid i) = \sum_i \pi(i) T_{ik} = \pi^T T$
  - after 2 steps: $P(k \mid \pi, 2\text{-step}) = \pi^T T T$
  - after t steps: $\pi^T T^t$

Similarity

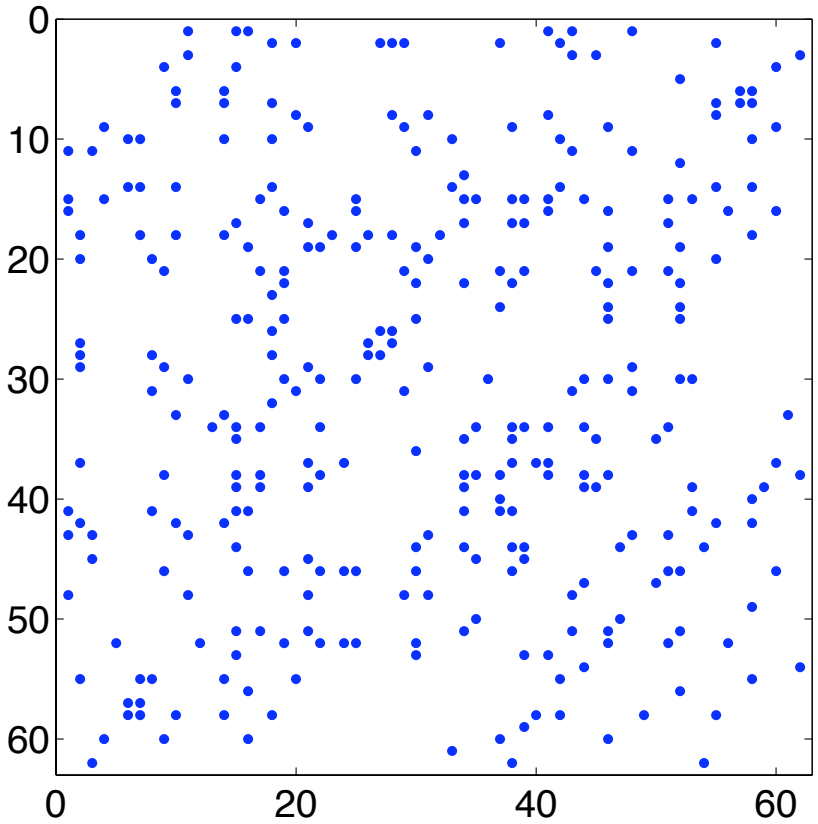
- A, B similar = random surfer tends to reach the same places when starting from A or B
- P(k | π, t-step) =
  - If π has all mass on i:
  - Compare i & j:
  - Role of Σ^t: see next slide

Role of Σ^t (real data)



initial entries of u_i more important particularly if $t >> 1$

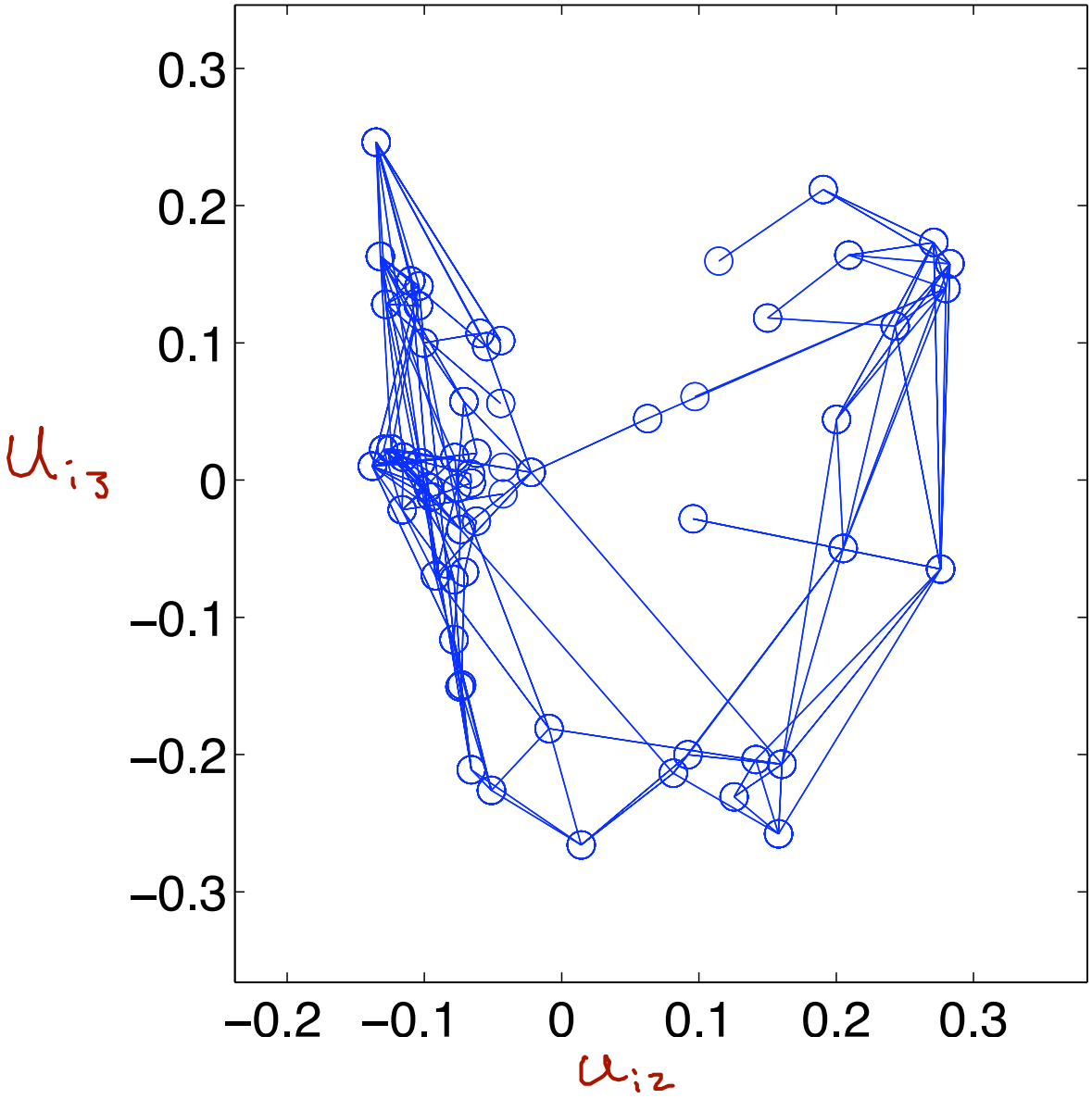
Example: dolphins



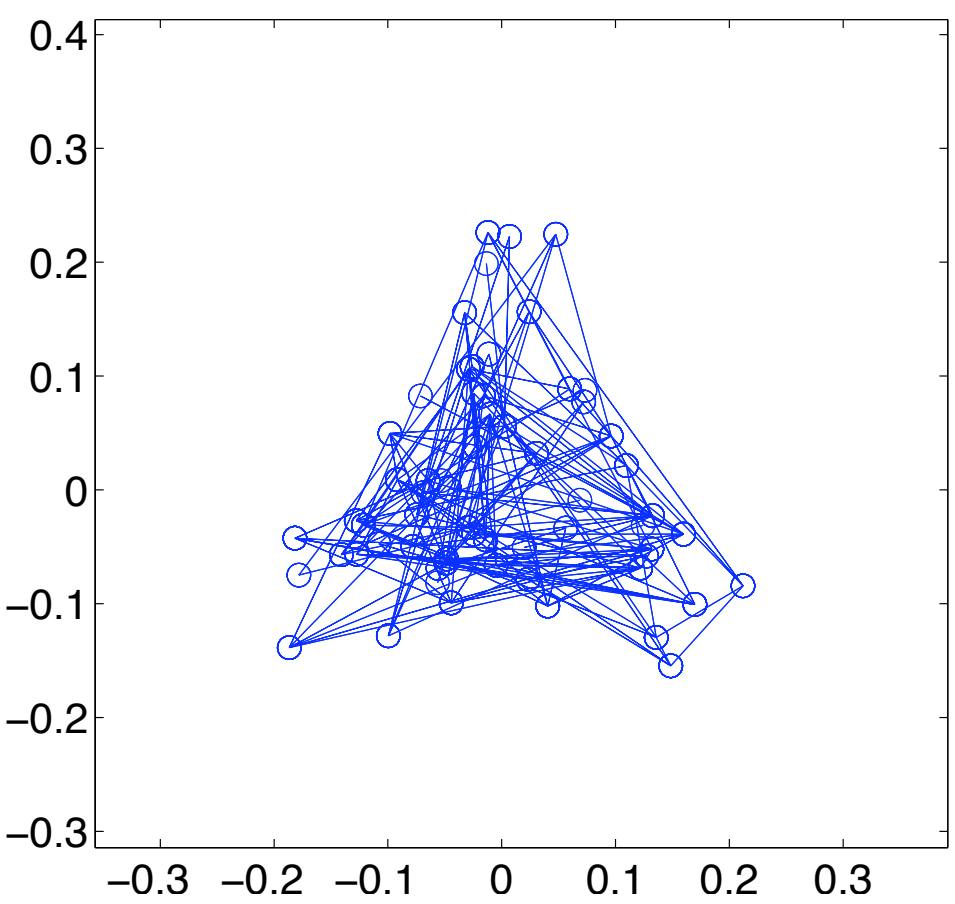
- 62-dolphin social network near Doubtful Sound, New Zealand
  - A_{ij} = 1 if dolphin i friends dolphin j

(Lusseau et al., 2003)

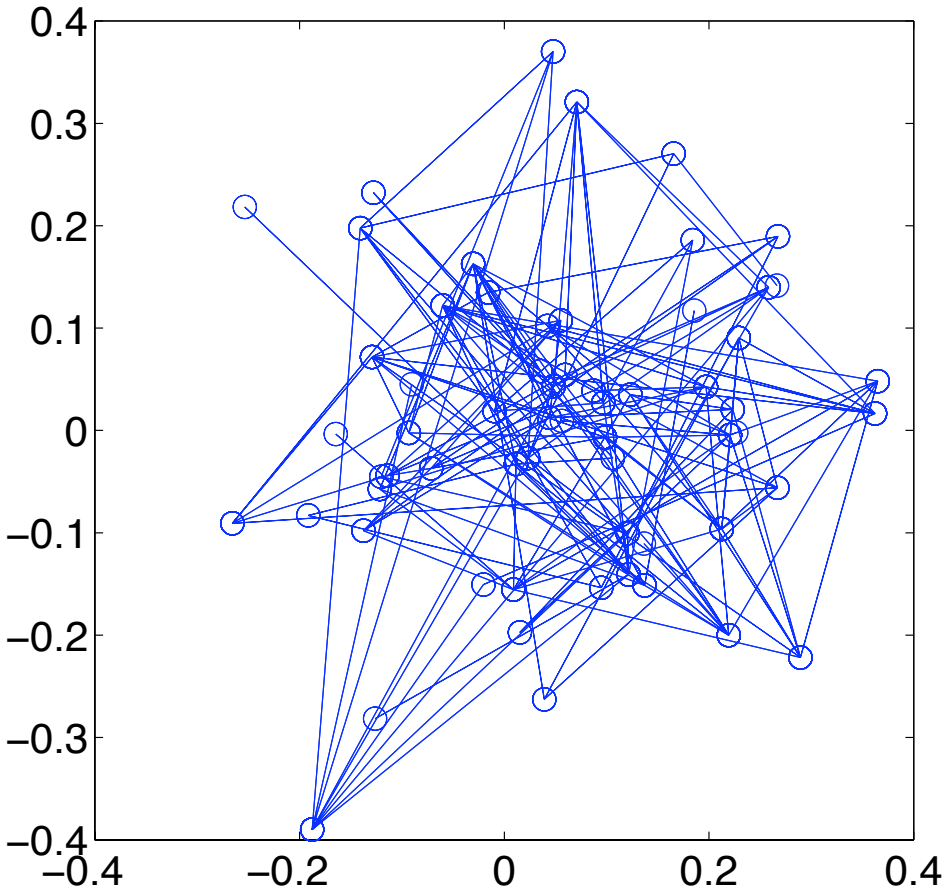
Dolphin network



Comparisons

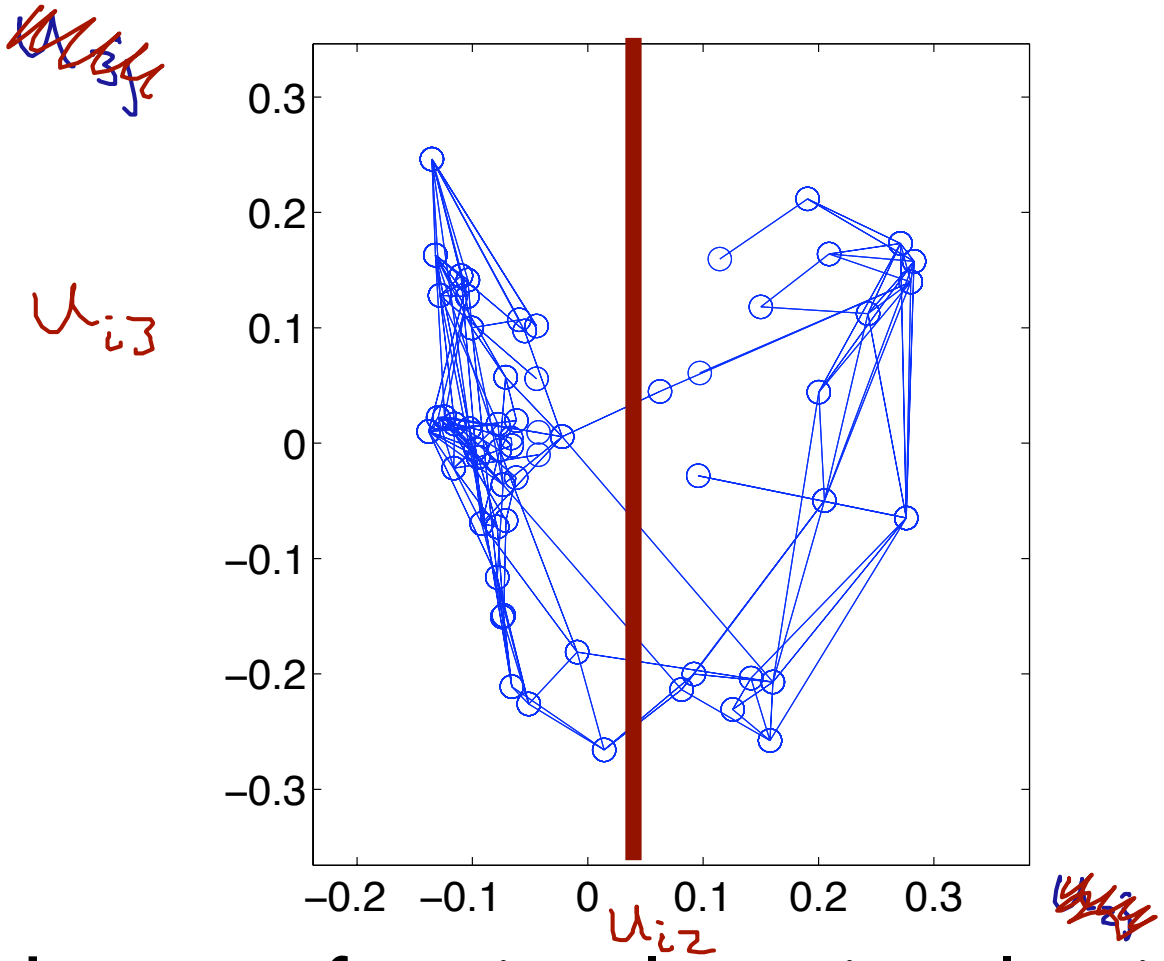


spectral embedding of random data



random embedding of dolphin data

Spectral clustering



- Use your favorite clustering algorithm on coordinates from spectral embedding

PCA: the good, the bad, and the ugly

- The good: simple, successful
- The bad: linear, Gaussian
  - $E(X) = UV^T$
  - X, U, V ~ Gaussian
- The ugly: failure to generalize to new entities

fold-in problem

Consistency

(convergence is w.p. 1)

Consistent = as |data| → ∞ est. params → true params.

- Linear & logistic regression are ***consistent***
- What would consistency mean for PCA?
  - forget about row/col means for now → $E(X|U,V) = UV^T$
- Consistency:
  - #users, #movies, #ratings (= nnz(W)) ← data
  - numel(U), numel(V) ← parameters
  - consistency = want $\hat{U} \to U_{true}$ ← infinite size

all increasing

can never happen!!

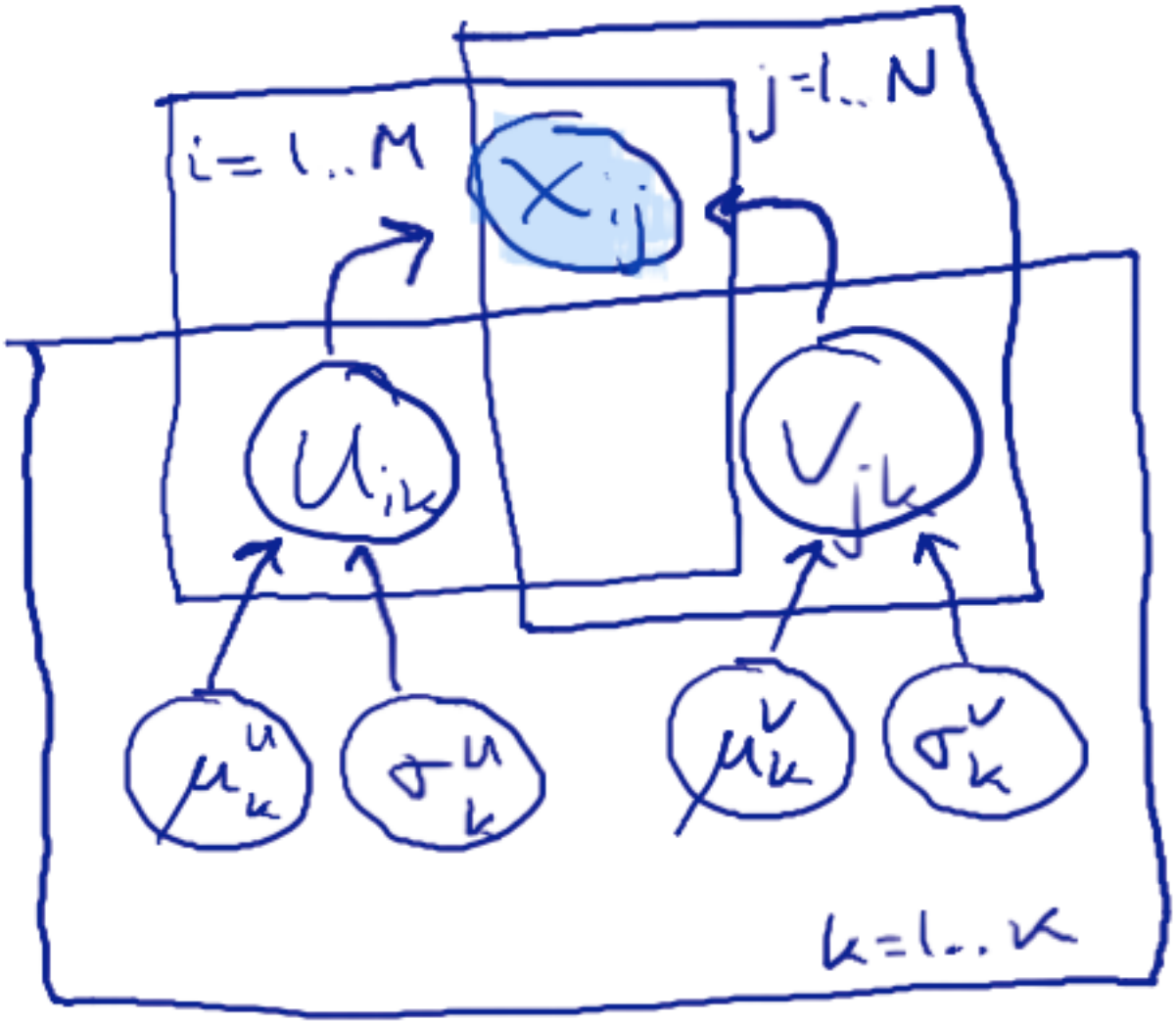
(some entries U_{ij} not seen yet)

Failure to generalize

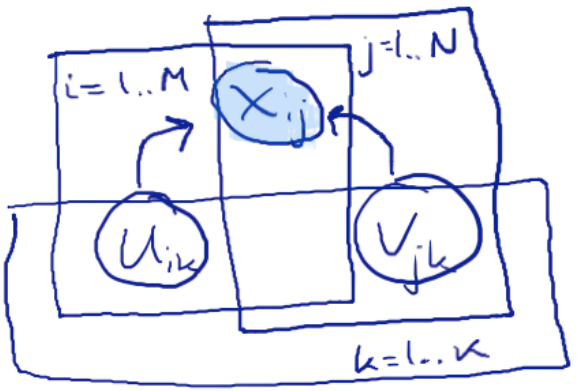
- What does this mean for generalization?
  - new user's rating of movie$_j$: only info is nothing!
  - new movie rated by user$_i$: only info is nothing!
  - all our carefully-learned factors give us: nothing!
- Generalization is:

only to new entries in existing rows/cols

Hierarchical model



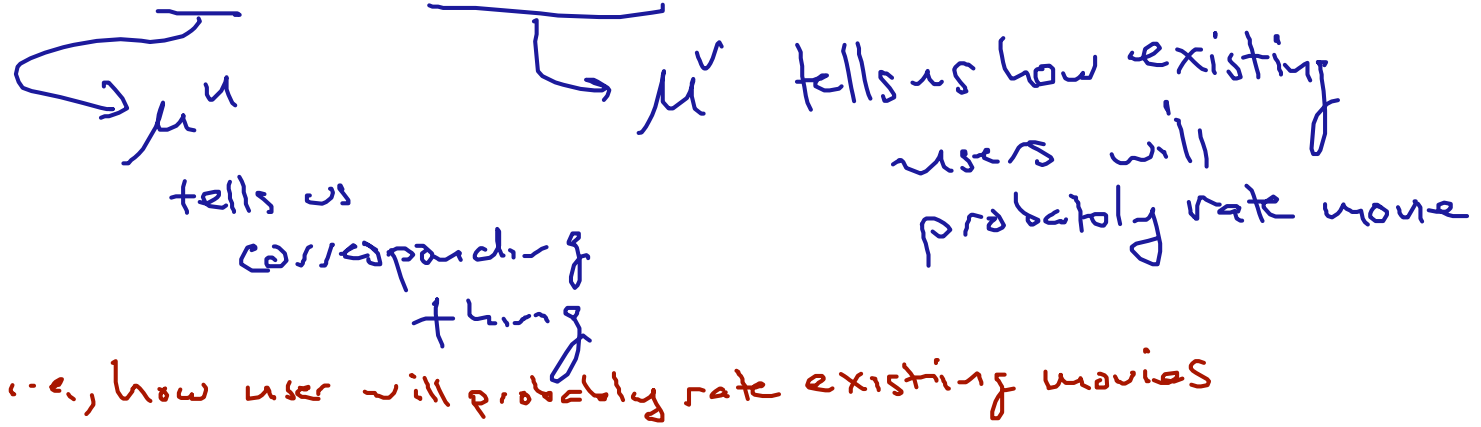
old, non-hierarchical model



Benefit of hierarchy

- Now: only k μ^U latents, k μ^V latents (and corresponding σs)
  - can get consistency for these if we observe more and more X_{ij}

- For a new user or movie:



Mean subtraction

- Can now see that mean subtraction is a special case of our hierarchical model
  - Fix $V_{j1} = 1$ for all j; then $U_{i1} =$ row mean
  - Fix $U_{i2} = 1$ for all i; then $V_{j2} =$ col mean
  - global mean: unnecessary — include in row or col mean