Probabilistic Topic Models

Angus Scott

3rd March 2015

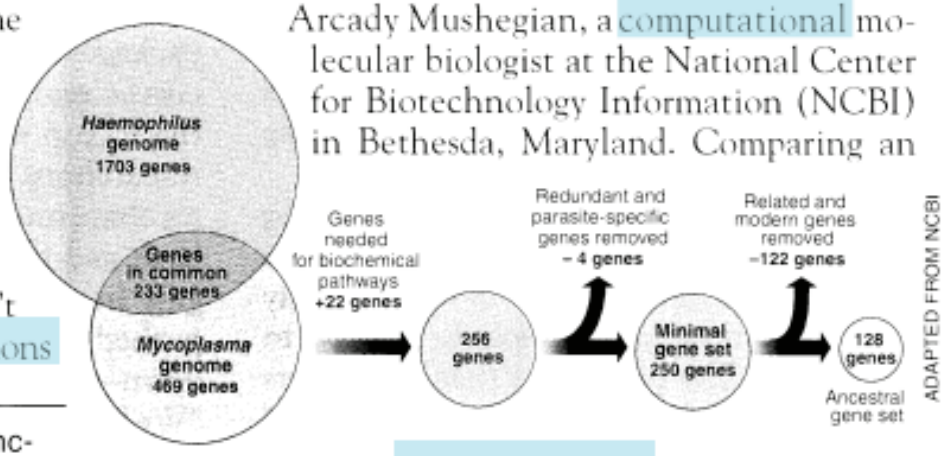
Motivation

- Given a collection of documents, we want to find themes which connect them
- Supervised topic classification is unfeasible:
  - Many topics, so many examples required!
  - Topics are fluid, so need to keep providing examples
- Problem is effectively a clustering task, clustering words and documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



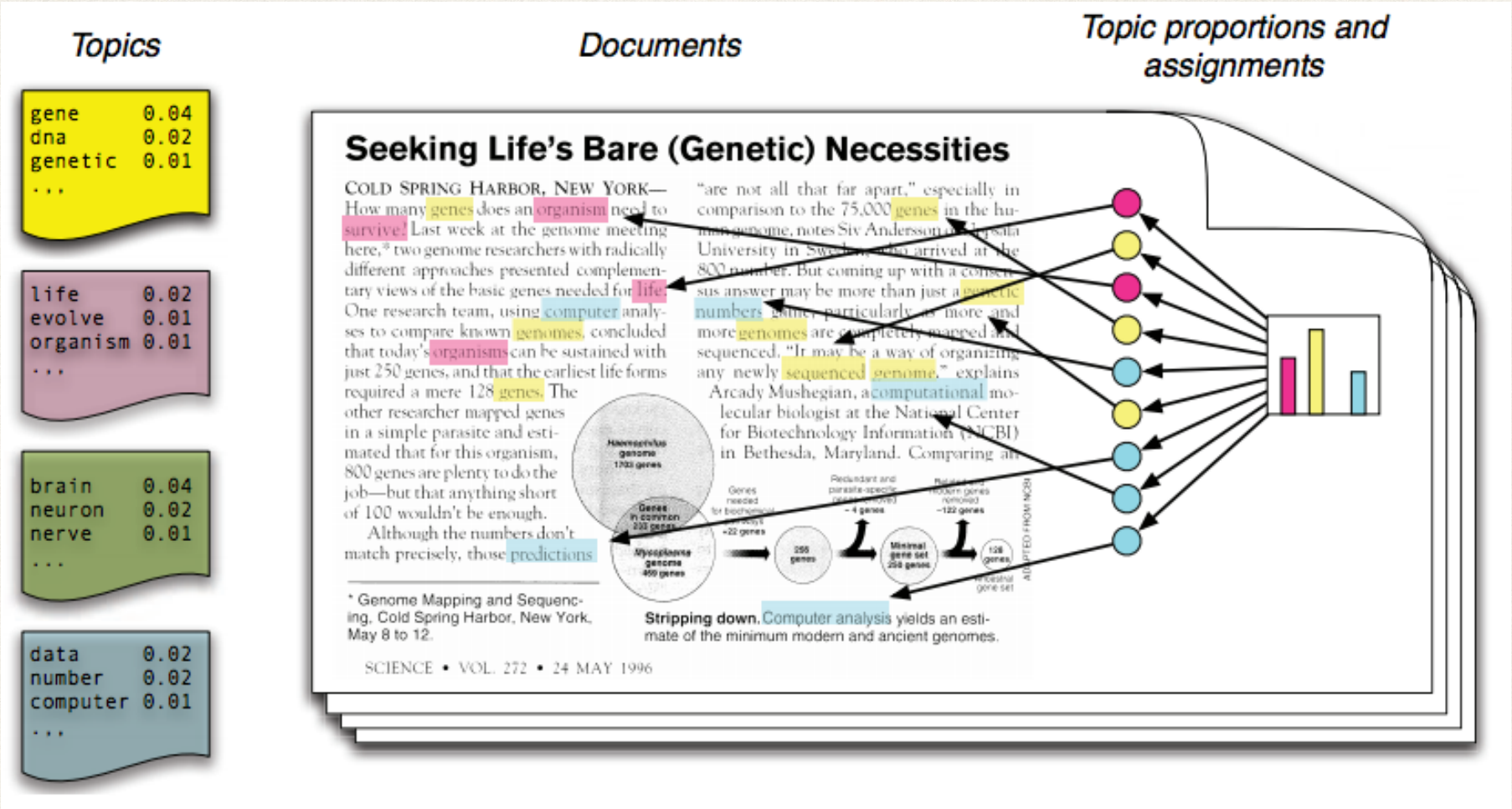
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Genes Evolution Computers

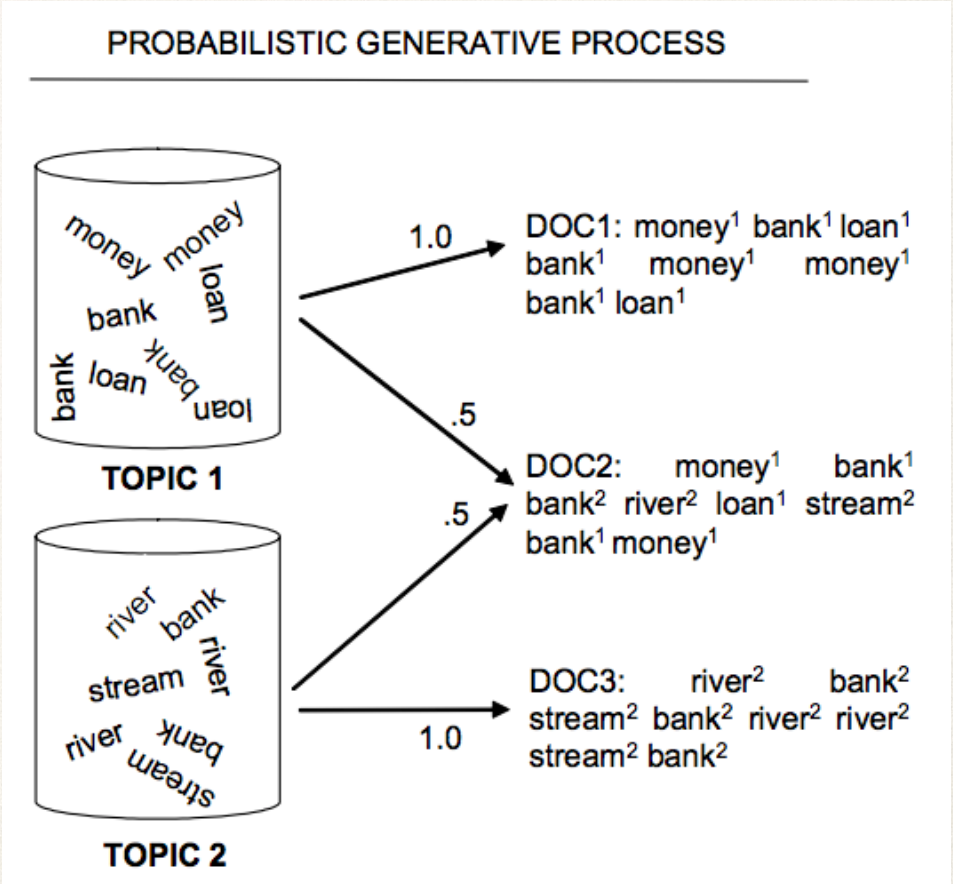
Documents are a collection of themes / topics



- Documents are a collection of topics
- Topics are a distribution over words

Probabilistic Generative Process

- Topic Models are a generative model for documents
- Documents are treated as bags of words
- Can create documents by creating a distribution of topics
- Sample words from the topics distribution to generate document



Probabilistic Topic Models

- All Topic Models have a fundamental assumption - document is a mixture of topics - but make different statistical assumptions

$P(z)$ - Distribution over topics z in particular document

$P(w|z)$ - Distribution over word w given topic z

$P(z_i = j)$ - Probability j-th topic was sampled for i-th word token

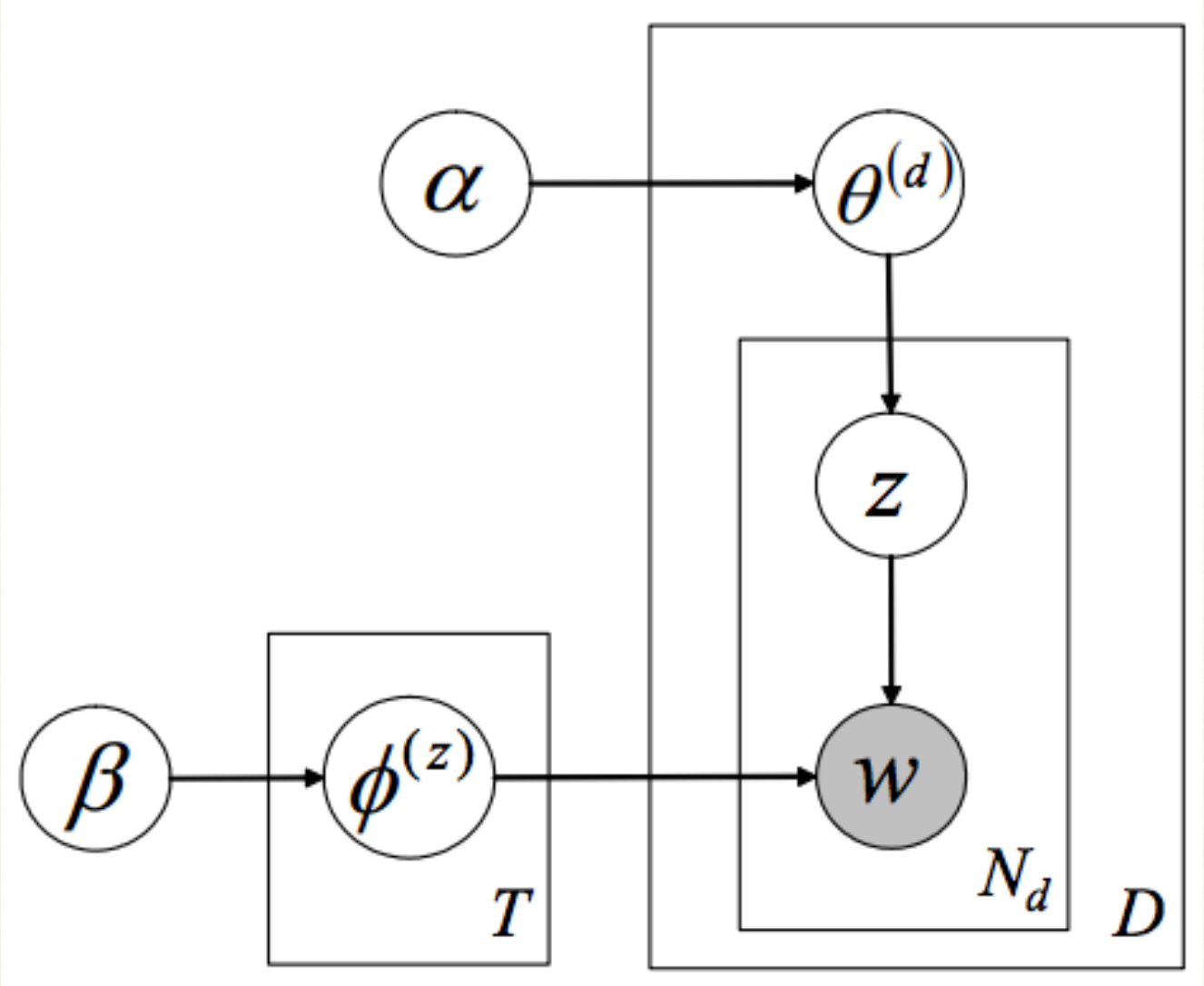
$P(w_i|z_i = j)$- Probability of i-th word token under topic j

- Model can be defined as:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

$$\theta^{(d)} = P(z) \qquad \phi^{(j)} = P(w|z = j)$$

Latent Dirichlet Allocation (LDA)



$$\theta^{(d)} = P(z) \qquad \phi^{(j)} = P(w|z = j)$$

Latent Dirichlet Allocation

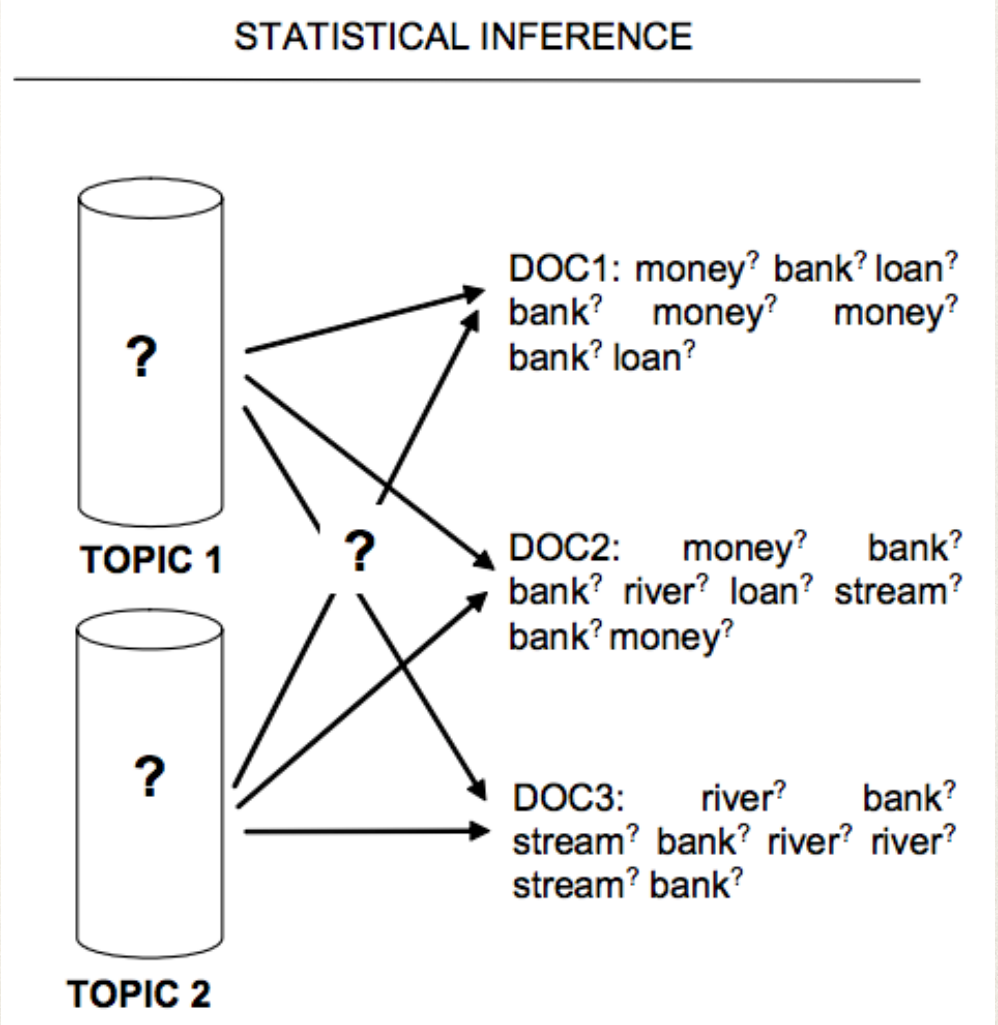
Generative LDA Algorithm

For $j = 1...T$ topics :
  Choose $\phi^{(j)} \sim Dirichlet(\beta)$
For $d = 1...D$ documents :
  Choose $\theta^{(d)} \sim Dirichlet(\alpha)$
  For $i = 1...N_d$ words in doc d :
    Choose $z_i \sim Multinomial(\theta^{(d)})$
    Choose $w_i \sim Multinomial(\phi^{(z_i)})$

$\theta^{(d)} = P(z)$ $\qquad$ $\phi^{(j)} = P(w|z = j)$

Statistical Inference

- Only given observed documents, want to know which topic generated data
- Infer probability distribution of topics over words
- Infer distribution of documents over topics
- Also need to know the topic responsible for generating each word



Algorithm for Extracting Topics

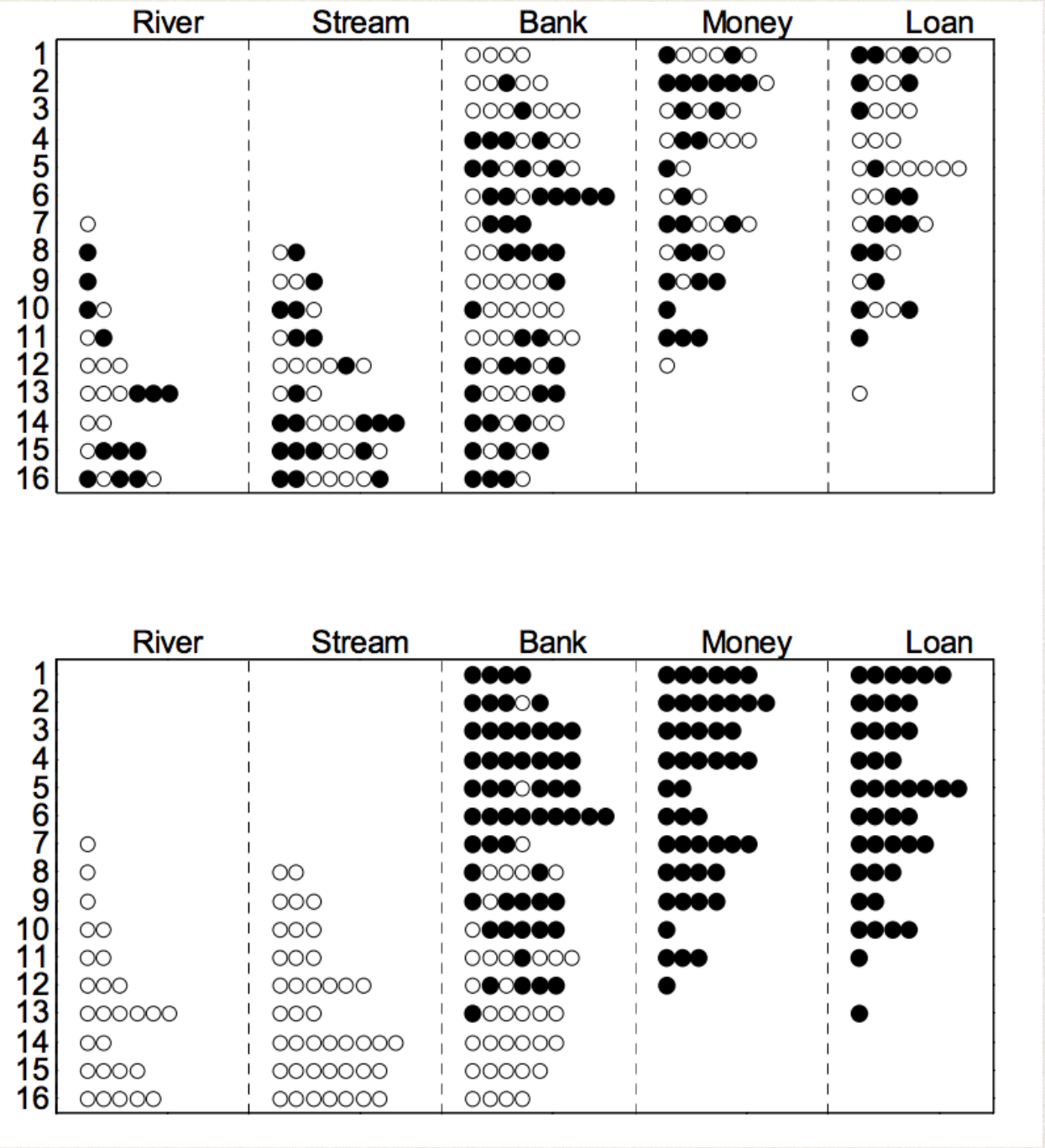
- Need to estimate the topic-word distribution ϕ and topic distribution θ
- Early work used Expectation Maximisation to estimate ϕ and θ
- But suffered from local maxima in likelihood function
- Alternative approaches have used Gibb Sampling

Gibb Sampling

- Iterative process
- Start by randomly assigning topics to each word
- Per iteration, for each word in the collection:
  - Assume you know (from the prev. iteration) the topics of all other words. (pretend they are correct)
  - Determine the probabilities of each topic-assignment given the rest of the data.
  - Choose the most probable assignment.
- Iterate until convergence.

Example Inference

- Iteration 0:
  - Topic1: $\phi^{(1)}_{\text{money}} = \phi^{(1)}_{\text{loan}} = \phi^{(1)}_{\text{bank}} = 1/3.$
  - Topic2: $\phi^{(2)}_{\text{river}} = \phi^{(2)}_{\text{stream}} = \phi^{(2)}_{\text{bank}} = 1/3.$
- Iteration 64 (convergence):
  - Topic1: $\phi^{(1)}_{\text{money}} = 0.32$ $\phi^{(1)}_{\text{loan}} = 0.29$ $\phi^{(1)}_{\text{bank}} = 0.39$
  - Topic2: $\phi^{(2)}_{\text{river}} = 0.25$ $\phi^{(2)}_{\text{stream}} = 0.4$ $\phi^{(2)}_{\text{bank}} = 0.35$



Applications and Conclusion

- We have discussed probabilistic topic models, set of algorithms that allow us to manage large collections of documents
- Find topics or themes across a collection of documents
- Interesting applications including JSTOR's discipline browser, a search tool (appears to be broken currently)
- Applications across disciplines, political/poll sampling on Twitter