EuroVis 2017 Machine Learning Methods in Visualization for Big Data 2017

12 June 2017, Barcelona, Spain

Methods for text data

Jaakko Peltonen³

3 University of Tampere, Aalto University

Topic models

- Topic models are a prominent way to model the content of documents
- Each document is represented as a bag of words: counts of how many times each different word has appeared
- Topic models represent documents as a mixture of underlying latent topics, where each topic has a probability table to generate different words over the vocabulary.
- Topic models are a kind of dimensionality reduction for count data: instead of representing a document by the vector of all counts, the document can be represented as a vector of inferred topic activities over a small number of topics.

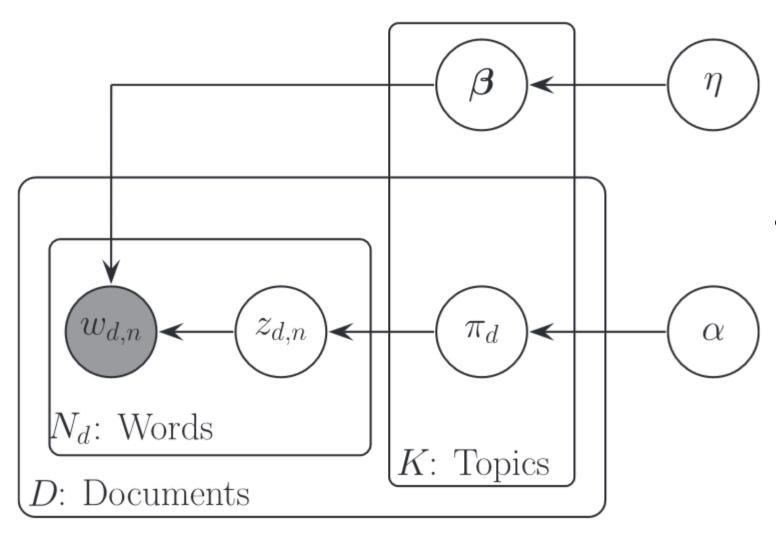
- Latent Dirichlet Allocation (LDA not to be confused with Linear Discriminant Analysis!) is a simple topic model
- Also called "discrete PCA".
- To generate a document d, a **topic distribution** π_d is drawn from a prior so that $\pi_d \sim Dirichlet(\alpha)$, and then the words are generated one by one.
- To generate the n:th word in the document:
 - a topic index $Z_{n,d}$ is drawn from the topic distribution:
 - the word is then drawn from a topic-wise word distribution:

$$Z_{n,d} \sim Multinomial(\pi_d)$$

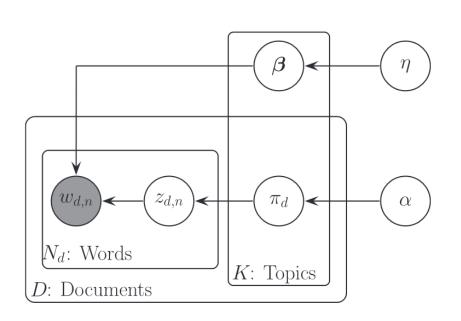
 $W_{n,d} \sim Multinomial(\beta_{Z_{n,d}})$

- $\beta_k = \{\beta_w | k\}_w$ are probabilities of each word w in the kth topic. The available topics are the same for all documents.
- Typically the topic-wise word distributions are drawn from a prior $\beta_k \sim Dirichlet(\eta)$, where η is the topic hyperparameter.

• Plate diagram model of the LDA generative process.

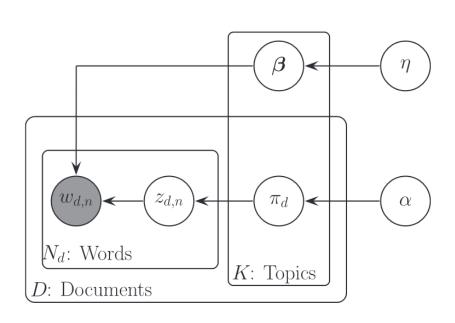


- In LDA each word is generated independently given the topic.
 The order of the word occurrences does not matter.
- LDA is suitable for count data such as bag-of-words representations of text, where only the overall count of each different word is observed

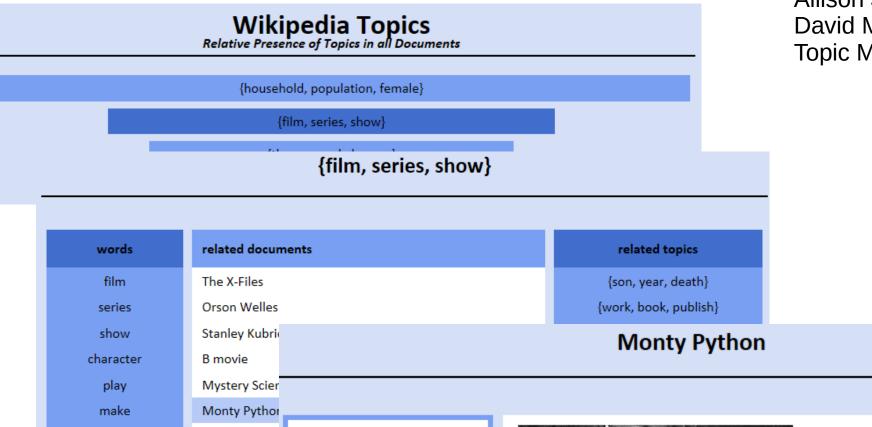


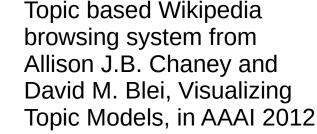
 Given a data set of documents, the LDA model can be fitted to the data by maximum a posteriori methods.

- When LDA is learned from a data set, the Dirichlet priors for the word distribution mitigate overfitting with large vocabularies: words that do not appear in the training set still get some probability to appear in future documents. (cf. PLSA: no priors)
- Why use Dirichlet priors? Convenient properties: finite dimensional sufficient statistics & conjugate to the multinomial distribution. Allows some parameters to be integrated out analytically when fitting an LDA model



- Given a fitted LDA model, data can be visualized in two ways.
- Each document has a vector π_d (probability distribution) whose elements are proportions over the available topics.
- Each topic is an **axis** of a low-dim. topic space,(simplex), and the π_d positions the document in this space.
- Each topic k has a word probability table $\beta_k = \{\beta_w | k\}_w$ whose top words can be listed.
- Sometimes researchers try to guess semantical meanings to the topics, and give them semantically meaningful names. These meanings and names do not arise from the topic model.





Mystery Science Theater 3000 Doctor Who Sam Peckinpah Married... with Children History of film The A-Team Pulp Fiction (film) Dubbing (filmmaking) Alfred Hitchcock

related documents



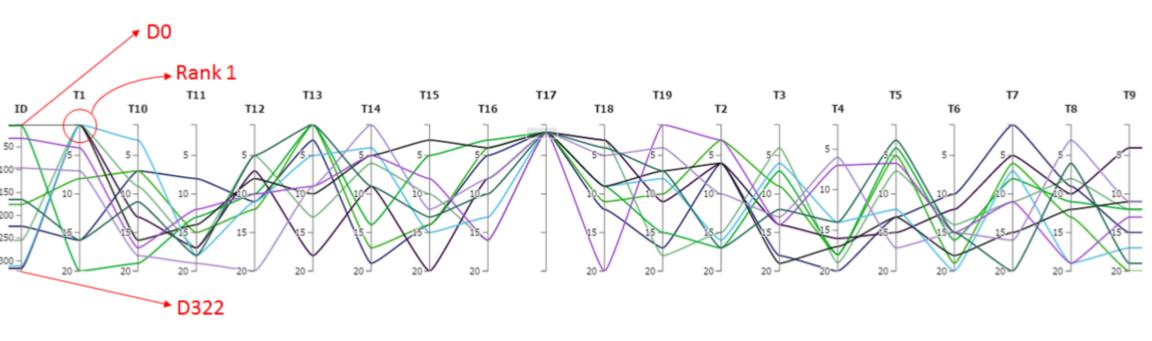
{film, series, show} {album, band, music} {theory, work, human} The Python troupe in 1969

Monty Python (sometimes known as **The Pythons**)^{[2][3]} was a British comedy group that created the influential *Monty Python's Flying Circus*, a British television comedy sketch show that first aired on the BBC on 5

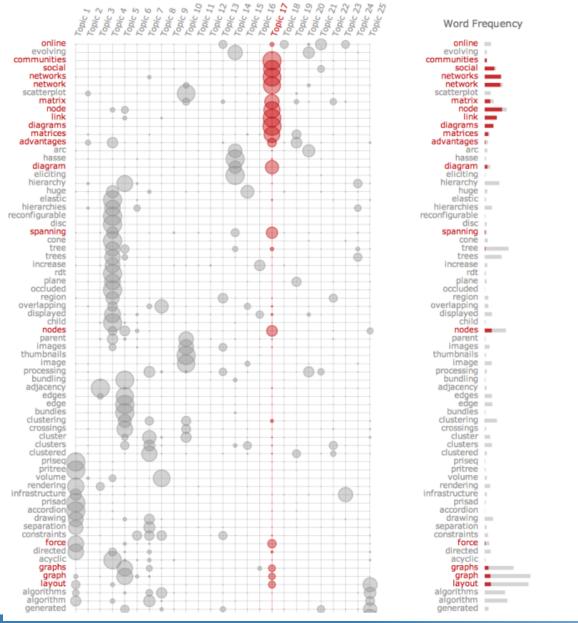
Machine Learning Method Visualization for Big Data

nada avar Duffu th

Parallel coordinate plot of several documents versus 20 topics, from Ashwinkumar Ganesan, Kiante Brantley, Shimei Pan and Jian Chen, LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation, IUI workshop on text analytics 2015



Matrix representation of several documents versus their words, using seriation. From Jason Chuang, Christopher D. Manning, Jeffrey Heer, Termite: Visualization Techniques for Assessing Textual Topic Models, in proc. AVI '12, 2012.

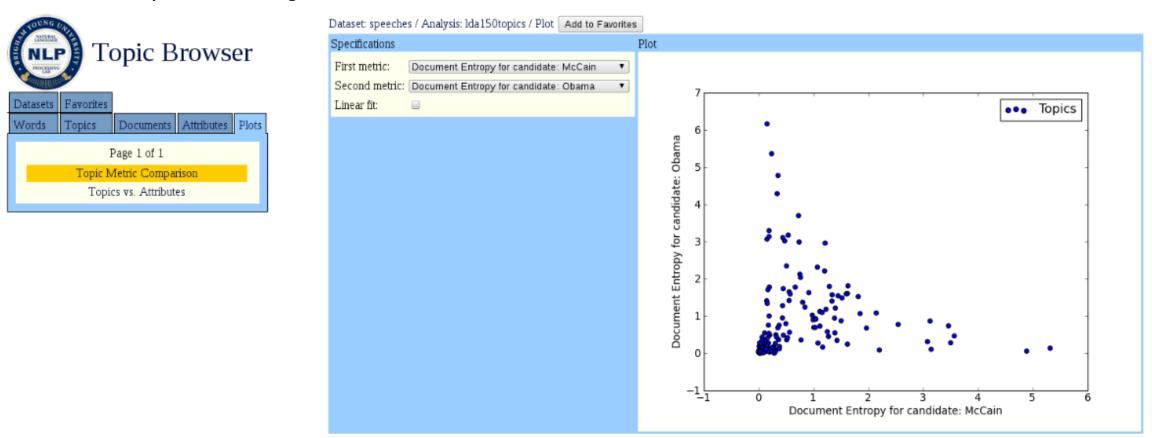


Representative Documents

A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations Mohammad Ghoniem Jean-Daniel Fekete Philippe Castagliola Using Multilevel Call Matrices in Large Software Projects Frank van Ham Improving the Readability of Clustered Social Networks using Node Duplication Nathalie Henry Anastasia Bezerianos Jean-Daniel Fekete MatrixExplorer: a Dual-Representation System to Explore Social Networks Nathalie Henry Jean-Daniel Fekete NodeTrix: a Hybrid Visualization of Social Networks Nathalie Henry Jean-Daniel Fekete Michael J. McGuffin The need to visualize large social networks is growing as hardware capabilities make analyzing large networks feasible and many new data sets become available. Unfortunately, the visualizations in existing systems do not satisfactorily resolve the basic dilemma of being readable both for the global structure of the network and also for detailed analysis of local communities. To address this problem, we present NodeTrix, a hybrid representation for networks that combines the advantages of two traditional representations: node-link diagrams are used to show the global structure of a network, while arbitrary portions of the network can be shown as adjacency matrices to better support the analysis of communities. A key contribution is a set of interaction techniques. These allow analysts to create a NodeTrix visualization by dragging selections to and from node-link and matrix forms, and to flexibly manipulate the NodeTrix representation to explore the dataset and create meaningful summary visualizations of their findings. Finally, we present a case study applying NodeTrix to the analysis of the InfoVis 2004 coauthorship dataset to illustrate the capabilities of NodeTrix as both an exploration tool and an effective means of communicating results. Visualizing Causal Semantics using Animations Nivedita R. Kadaba Pourang P. Irani Jason Leboe **Balancing Systematic and Flexible Exploration of Social Networks** Adam Perer Ben Shneiderman Social network analysis (SNA) has emerged as a powerful method for understanding the importance of relationships in networks. However, interactive exploration of networks is currently challenging because: (1) it is difficult to find patterns and comprehend the structure of networks with many nodes and links, and (2) current systems are often a medley of statistical methods and overwhelming visual output which leaves many analysts uncertain about how to explore in an orderly manner. This results in exploration that is largely opportunistic. Our contributions are techniques to help structural analysts understand social networks more effectively. We present SocialAction, a system that uses attribute ranking and coordinated views to help users systematically examine numerous SNA measures. Users can (1) flexibly iterate through visualizations of measures to gain an overview, filter nodes, and find outliers, (2) aggregate networks using link structure, find cohesive subgroups, and focus on communities of interest, and (3) untangle networks by viewing different link types separately, or find patterns across different link types using a matrix overview. For each operation, a stable node layout is maintained in the network visualization so users can make comparisons. SocialAction offers analysts a strategy beyond opportunism, as it provides systematic, yet flexible, techniques for exploring social networks. Causality Visualization Using Animated Growing Polygons Niklas Elmqvist Philippas Tsigas

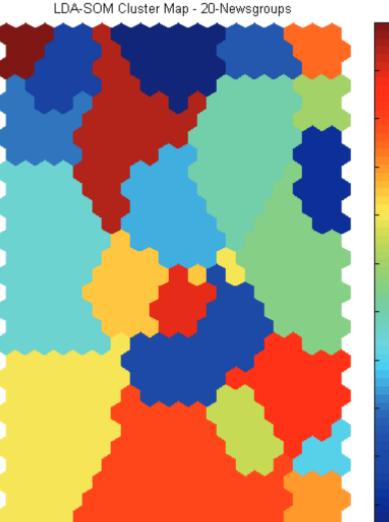
SpicyNodes: Radial Layout Authoring for the General Public Michael Douma Grzegorz Ligierko Ovidiu Ancuta Pavel Gritsai Sean Liu

Comparison of topics' activity (summarized by entropy over documents = speeches) between two different document subgroups (here subgroups = authors who are political candidates). From Matthew J. Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi, The Topic Browser: An Interactive Tool for Browsing Topic Models, in proc. NIPS 2010 workshop on Challenges of Data Visualization.



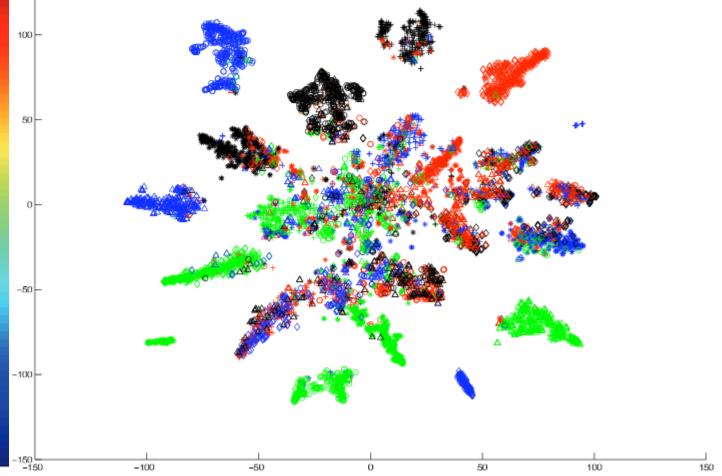
150

Self-organizing map trained on topic model outputs (topic distributions of each document). From Jeremy R. Millar and Gilbert L. Peterson and Michael J. Mendenhall, Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps, in proc. FLAIRS 2009.

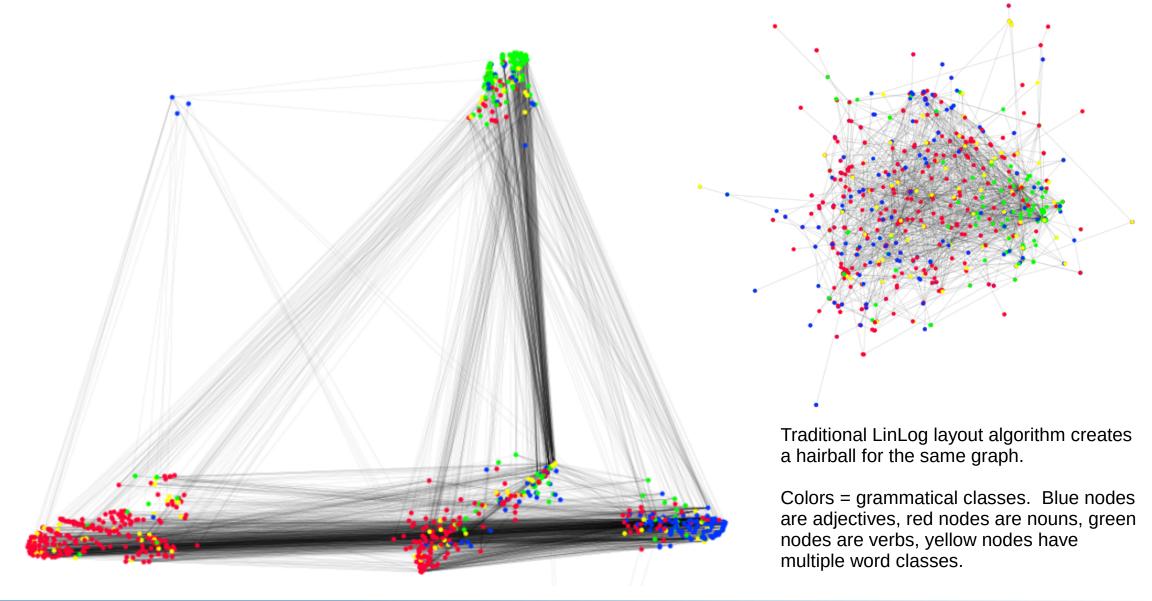


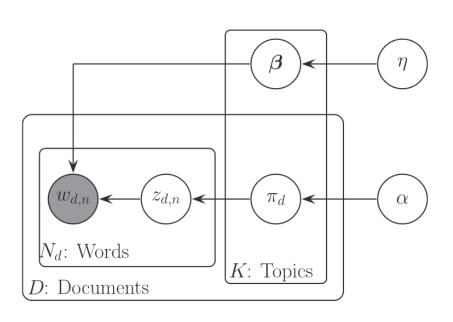
T-distributed Stochastic Neighbor Embedding trained for documents represented by topic model outputs (topic distributions of each document).

From Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan, DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification, in proc. NIPS 2008.

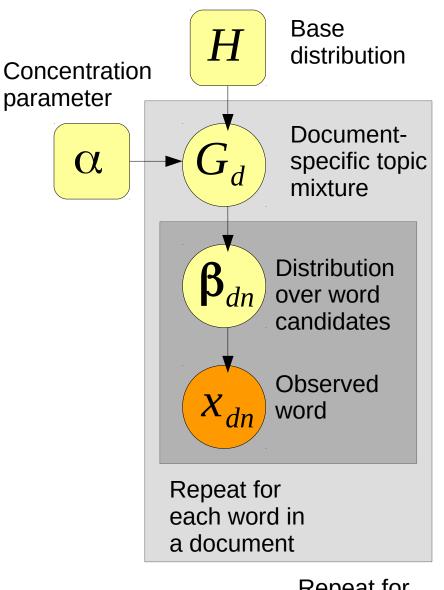


Graph layout created by Neighbor Retrieval Visualizer applied to topic model outputs (topic distributions of each graph node), when inputs to the topic model are nodes represented as a "bag of links". In this case the graph arises from text data: nodes are different words, and links are adjacencies in the novels of Jane Austen. From Juuso Parkkinen, Kristian Nybo, Jaakko Peltonen, and Samuel Kaski. Graph Visualization With Latent Variable Models. In Proceedings of MLG 2010, the Eighth Workshop on Mining and Learning with Graphs, 2010.



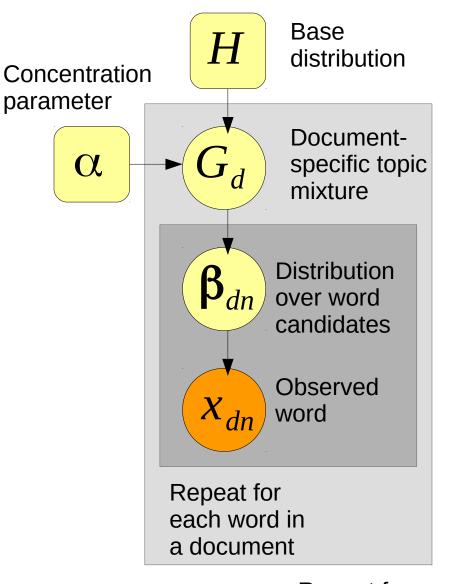


- The LDA model assumes the number K of available topics to be specified in advance.
- Problematic when the number of actual underlying topics can be large, and expert knowledge for choosing the correct number of topics may not be available.
- If the number of topics is set too small, it forces the model to merge some of the real topics
- If the number of topics is set too large, maximum likelihood fitting will overfit and split some real topics according to artifacts in the observed data.



Repeat for each document

- "Nonparametric" topic model where the number of topics does not need to be chosen beforehand
- A document is assumed to come from a mixture of topics.
- The mixture is drawn from a **prior over possible mixtures:** all possible numbers of topics up to infinity, and all possible proportions over topics
- The **Dirichlet process** is a suitable prior
- The model is described in three ways:
 - 1. A "theoretic way" to generate data
 - 2. An iterative process to generate data
 - 3. Inference equations



Repeat for each document

• Theoretic way:

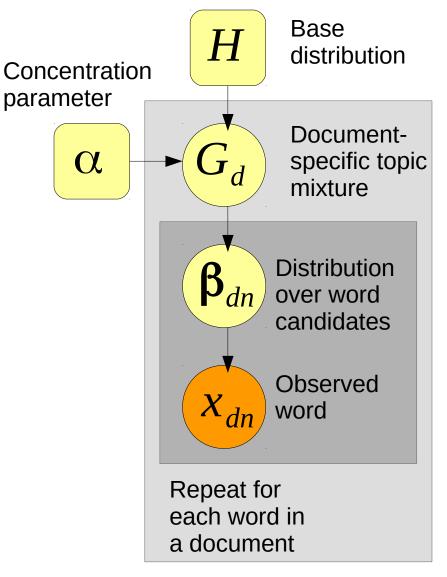
- generate a mixture from the Dirichlet process prior, $G_d\!\sim\!DP(H\,,alpha)$
- For each word n, draw a mixture component (=word distribution = word probability table) from the mixture

$$\boldsymbol{\beta}_{dn} \sim \boldsymbol{G}_d$$

and draw the observed word from the word distribution

 $x_{dn} \sim Multinomial(\boldsymbol{\beta}_{dn})$

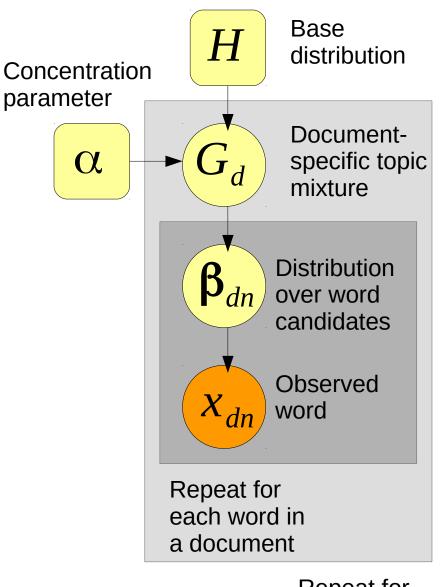
Problem: it is hard to draw an entire mixture (potentially an infinite-size object) over all possibilities



 Iterative way: chinese restaurant process



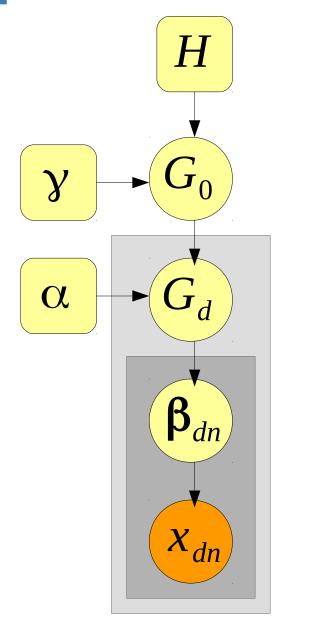
Repeat for each document



Repeat for each document

- Iterative way: chinese restaurant process
- Observed words X_{dn} are food
 servings experienced by customers
- Each customer comes in and chooses a **table** where to sit.
- Customers prefer to sit at popular tables where others already sit, but sometimes choose to take a new table
- Topics are **dishes**. Each table serves a particular dish: all customers sitting at the table get servings from that dish.
- Because customers prefer to sit at popular tables, the dish β_{dn} of a new customer is likely to be the same as a dish already being served

Hierarchical Dirichlet Processes



- Hierarchical extension of the Dirichlet Process Mixture Model
- Instead of drawing all documents in the collection from an uninformative generic base distribution:
- Sample a collection-specific base distribution
- Allows the model to use the same topics over multiple documents, helps against overfitting
- Inference is similar as in the DPMM, but slightly more complicated.
- Visualization similar as in DPMM