PRESENTER: YANG HAIYAN

DEP. ECE HKUST

HIERARCHICAL TOPIC MODELS

CONTENTS

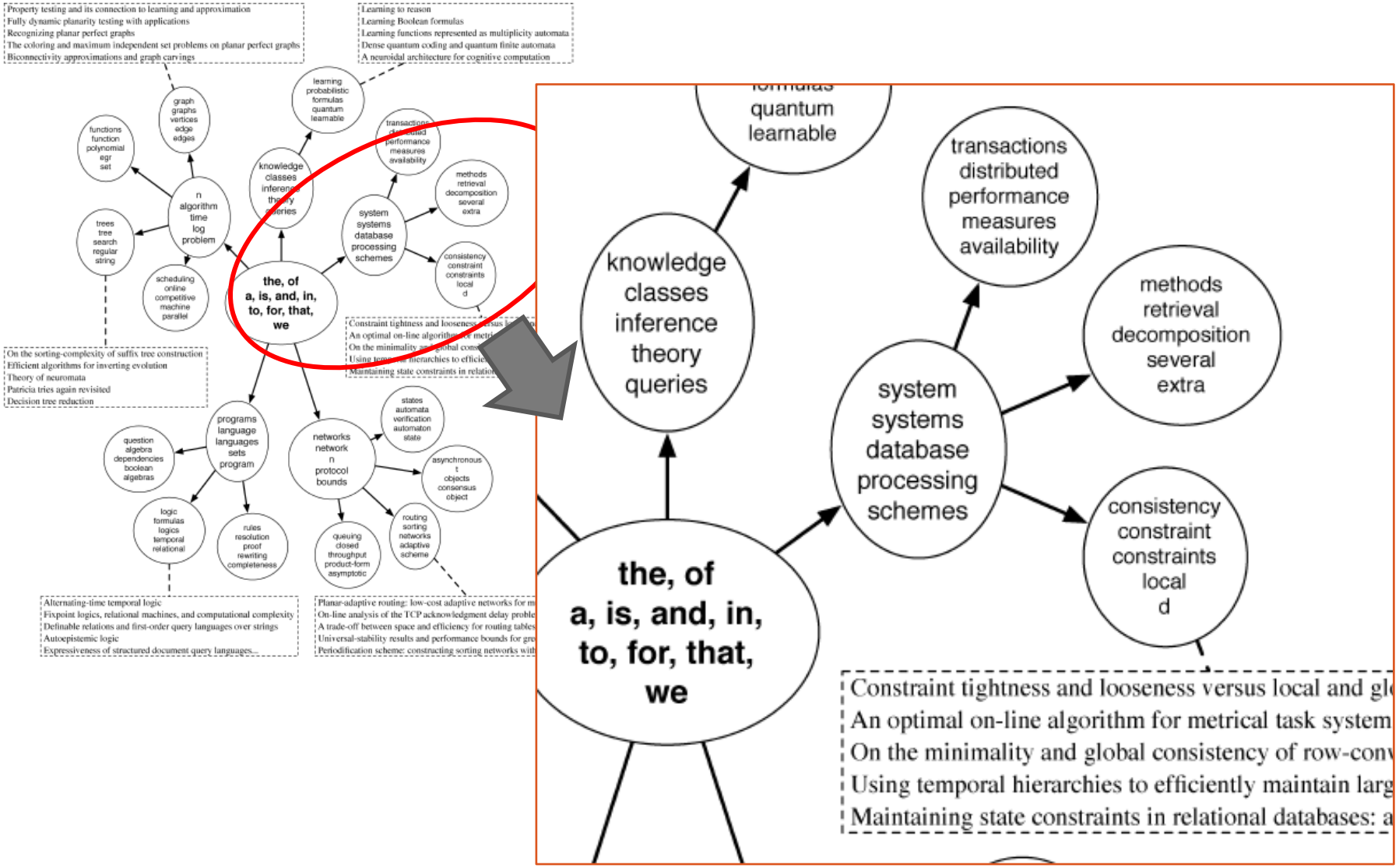
1. INTRODUCTION

The problems of document classification

- There is no reason to assume that each document could only have limited topics which really happens at LDA and CTM.
- In LDA, we pick a mixture distribution over K topics and generate words from it.
- In CTM, they extend the LDA by considering the relationship between different topics but the topic number is still limited.
- If we increase the topic number from finite to infinite, how should these topics be organized?
- One possible way is to organize the infinite topics in a hierarchy.

MOTIVATION

- LDA fails to draw the relationship between one topic and another.
- CTM considers the relationship but fail to indicate the level of abstract of a topic.
- In order to better model the real world data behavior, we want an algorithm to both find useful sets of topics and learn to organize the topics according to a hierarchy in which more abstract topics are near the root of the hierarchy and more concrete topics are near the leaves.



D. Blei, T. Griffiths, and M. Jordan, “The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies,” Journal of the ACM, vol. 57, no. 2, pp. 7:1–30, 2010.

CONTENTS

2.1 CRP

What is the Chinese Restaurant Process?

- The CRP is a single parameter distribution over partitions of integers.
- It has been used to represent the uncertainty over the number of components in a mixture model.
- Suppose we have a collection of observations, and we want to partition them into groups.
- Every possible group corresponds to a table in an infinitely large Chinese restaurant .
- Each observation corresponds to a customer entering the restaurant and sitting at a table.

2.1 CRP

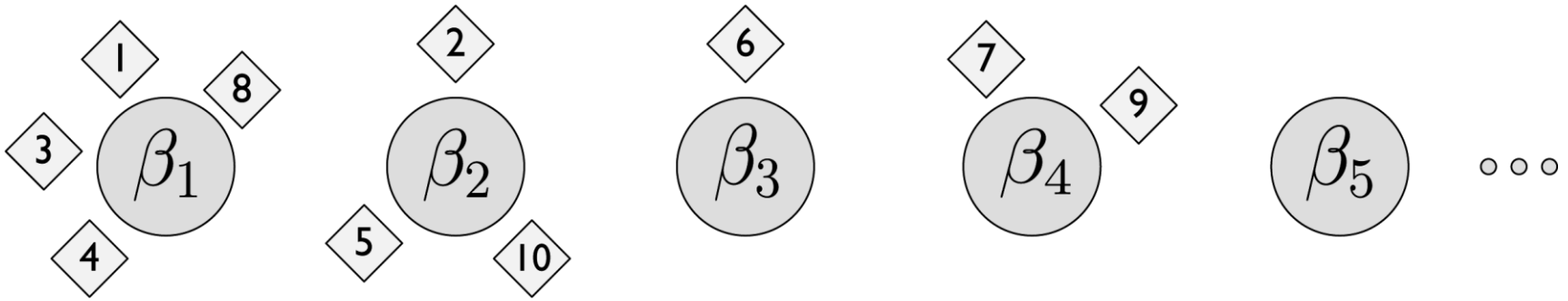
Generating from the CRP:

- A sequence of *N* customers arrive, labeled with the integers {1, . . . , *N*}.
- First customer sit at the first table.
- The *n*th customer sit at:

  Table *i* with probability $\frac{n_i}{\gamma+n-1}$ where n_i is the number of customers currently sitting at table *i.*

  A new table *i+1* with probability $\frac{\gamma}{\gamma+n-1}$.

2.1 CRP



$$p(\text{occupied table } i \mid \text{previous customers}) = \frac{n_i}{\gamma+n-1}$$

$$p(\text{next unoccupied table} \mid \text{previous customers}) = \frac{\gamma}{\gamma+n-1}$$

2.1 CRP

CRP to Topic Model:

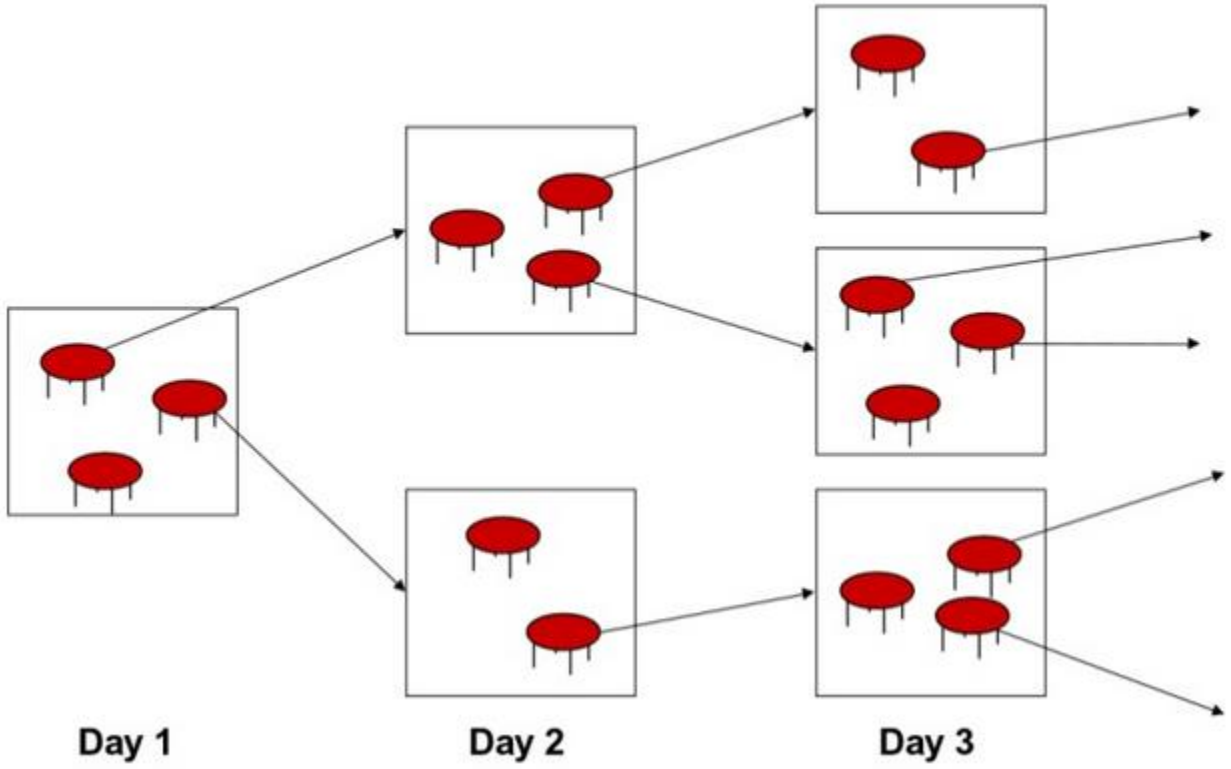
- Consider a restaurant as a document.
- Consider a customer as a word.
- Consider a table as a topic.

2.2 NESTED CRP

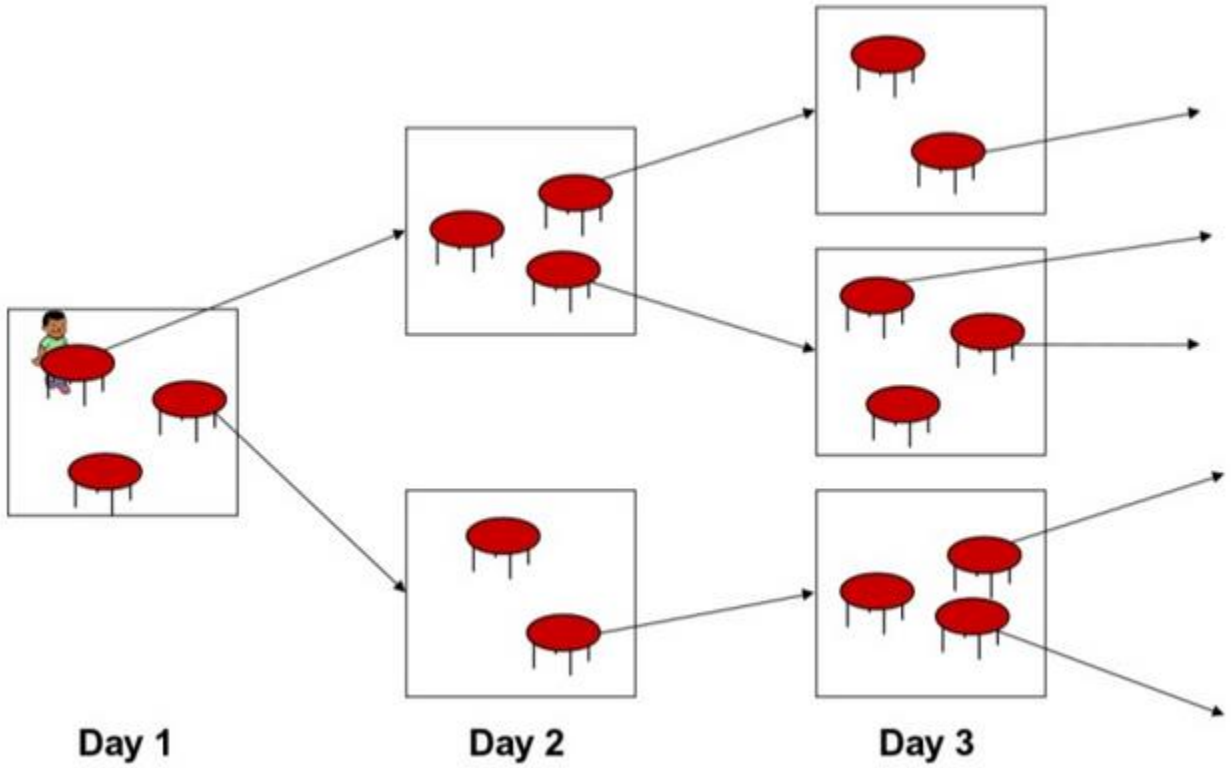
Nested Chinese Restaurant Process:

- Suppose there are an infinite number of infinite-table Chinese restaurants in a city.
- One restaurant is identified as the root restaurant, and on each of its infinite tables is a card with the name of another restaurant. this structure repeats infinitely many times
- Each restaurant is referred to exactly once.
- the restaurants in the city are organized into an infinitely branched, infinitely-deep tree.
- Note that each restaurant is associated with a level in this tree. The root restaurant is at level 1, the restaurants referred to on its tables. cards are at level 2, and so on.

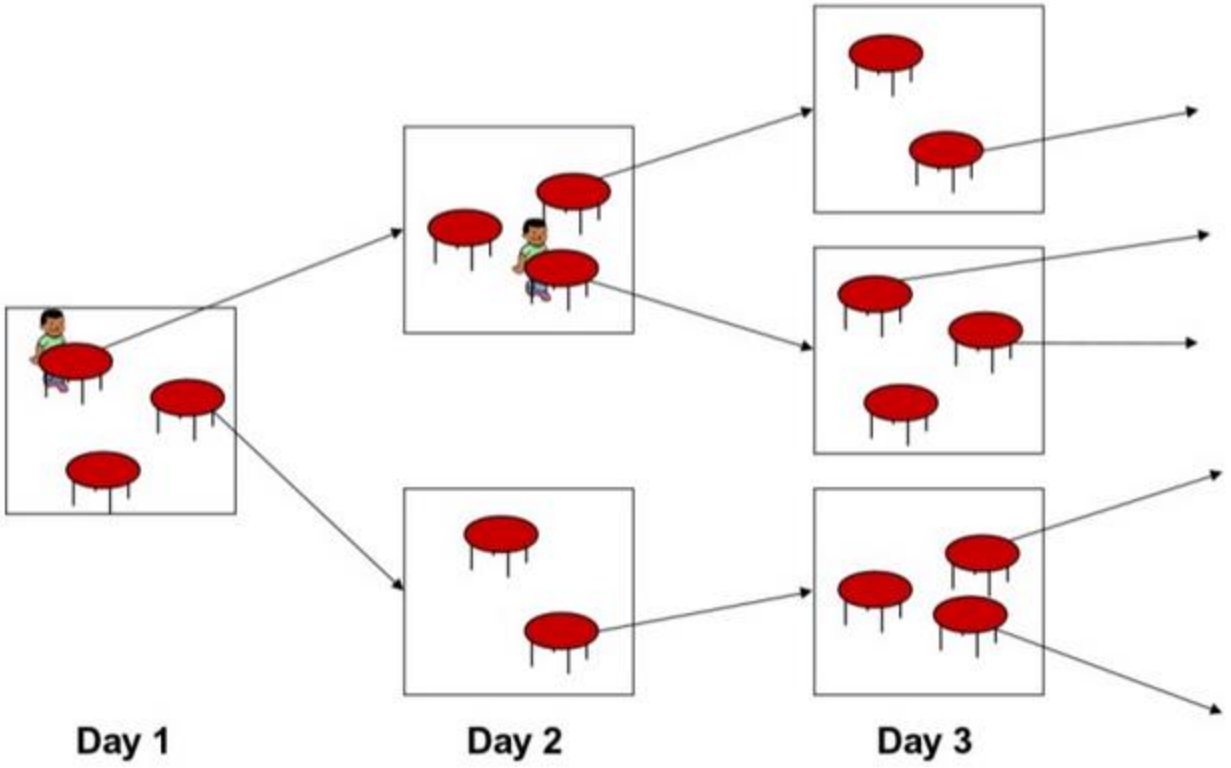
2.2 NESTED CRP



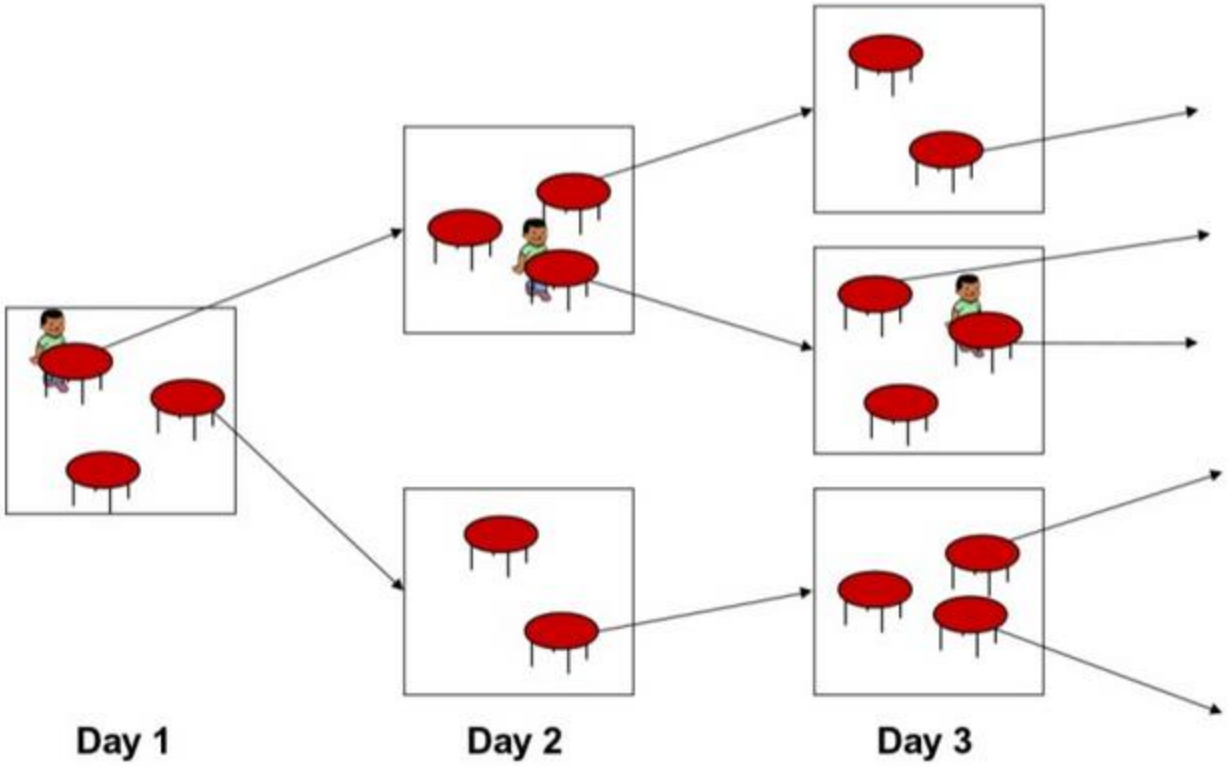
2.2 NESTED CRP



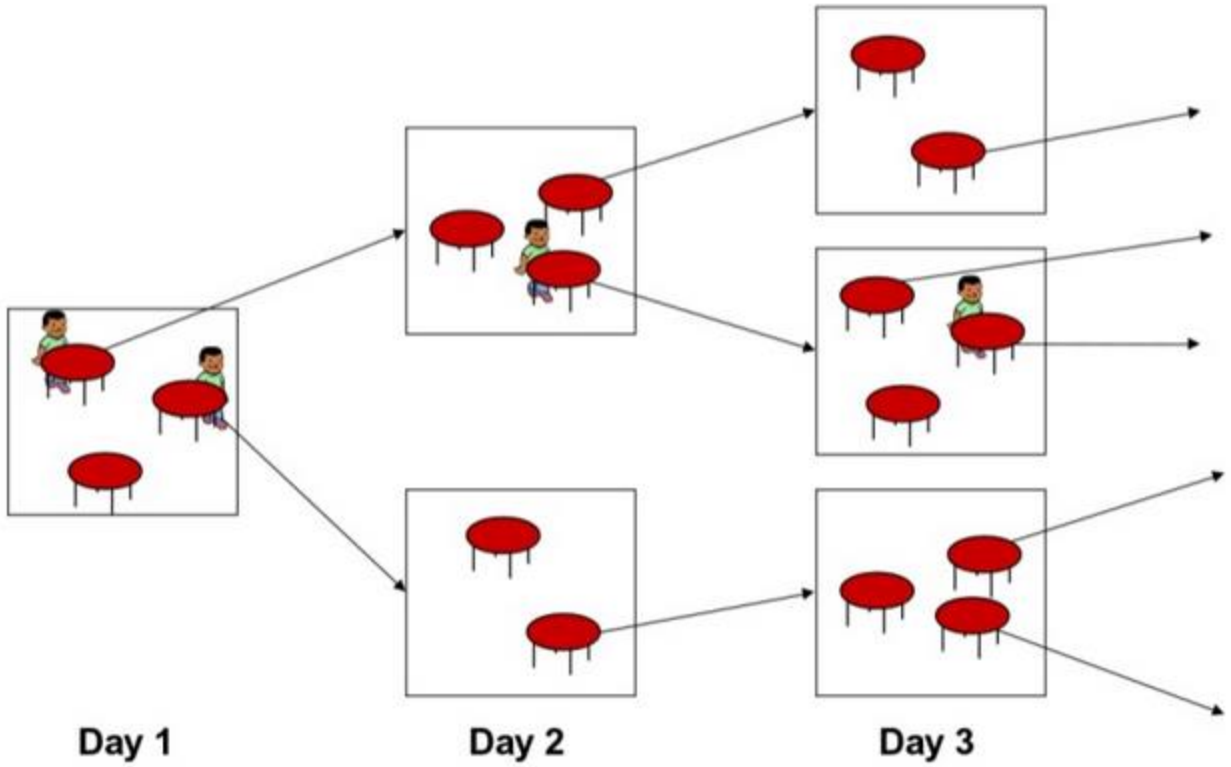
2.2 NESTED CRP



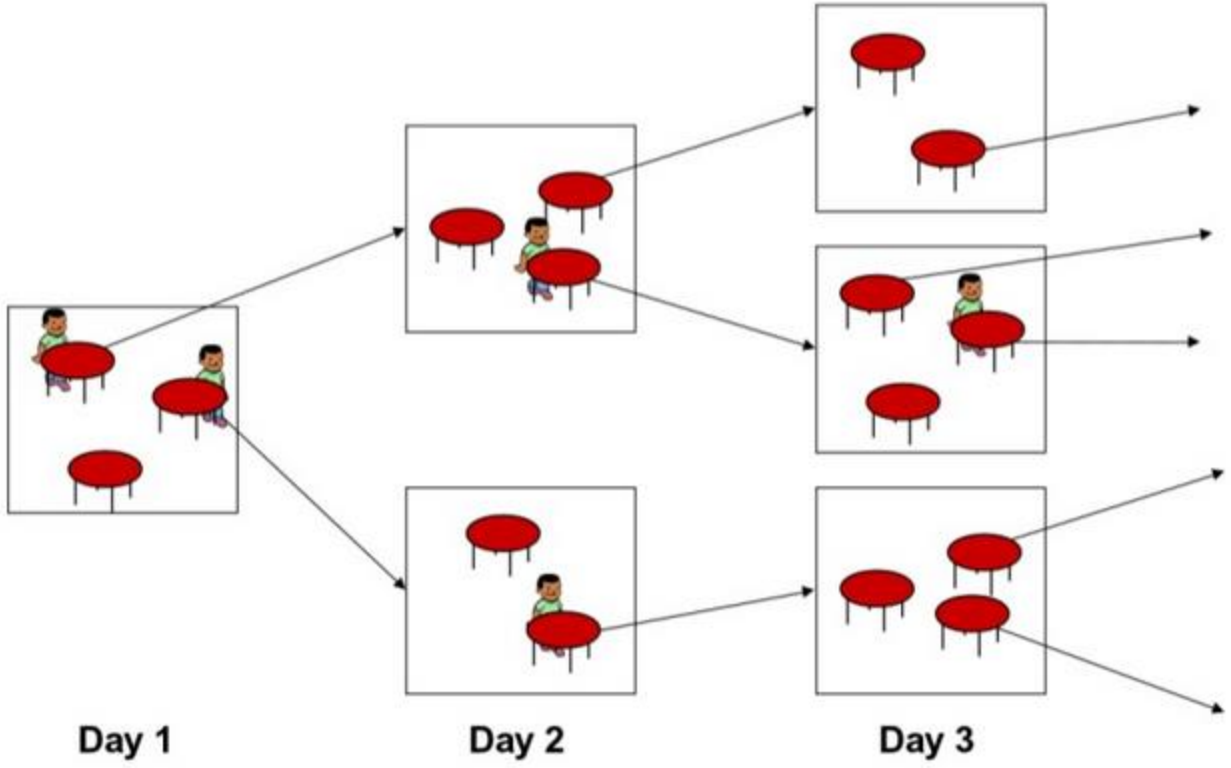
2.2 NESTED CRP



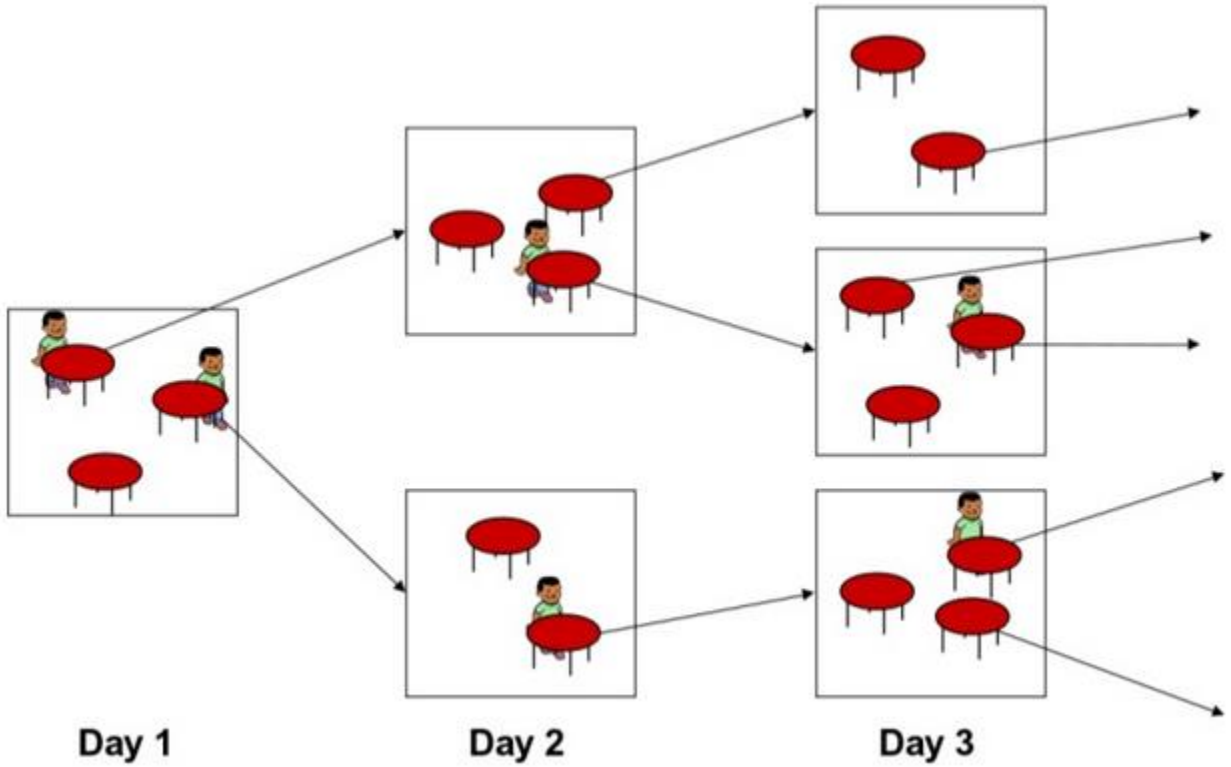
2.2 NESTED CRP



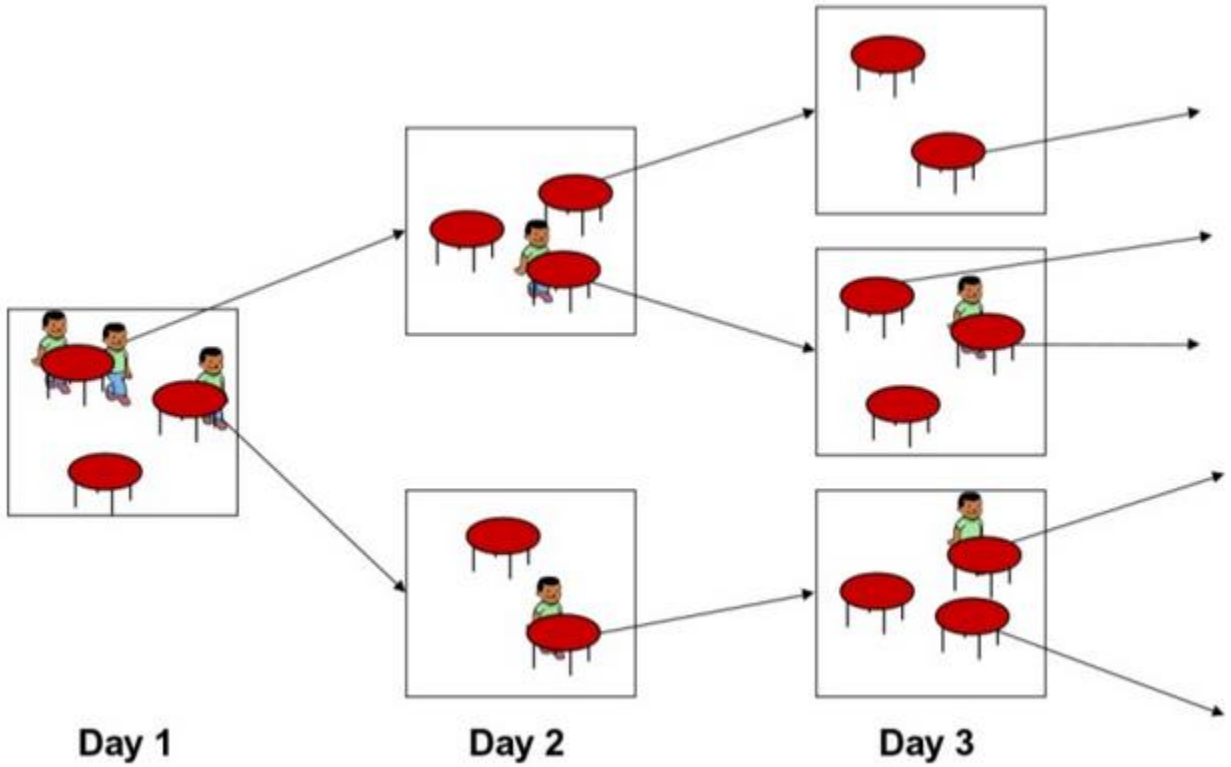
2.2 NESTED CRP



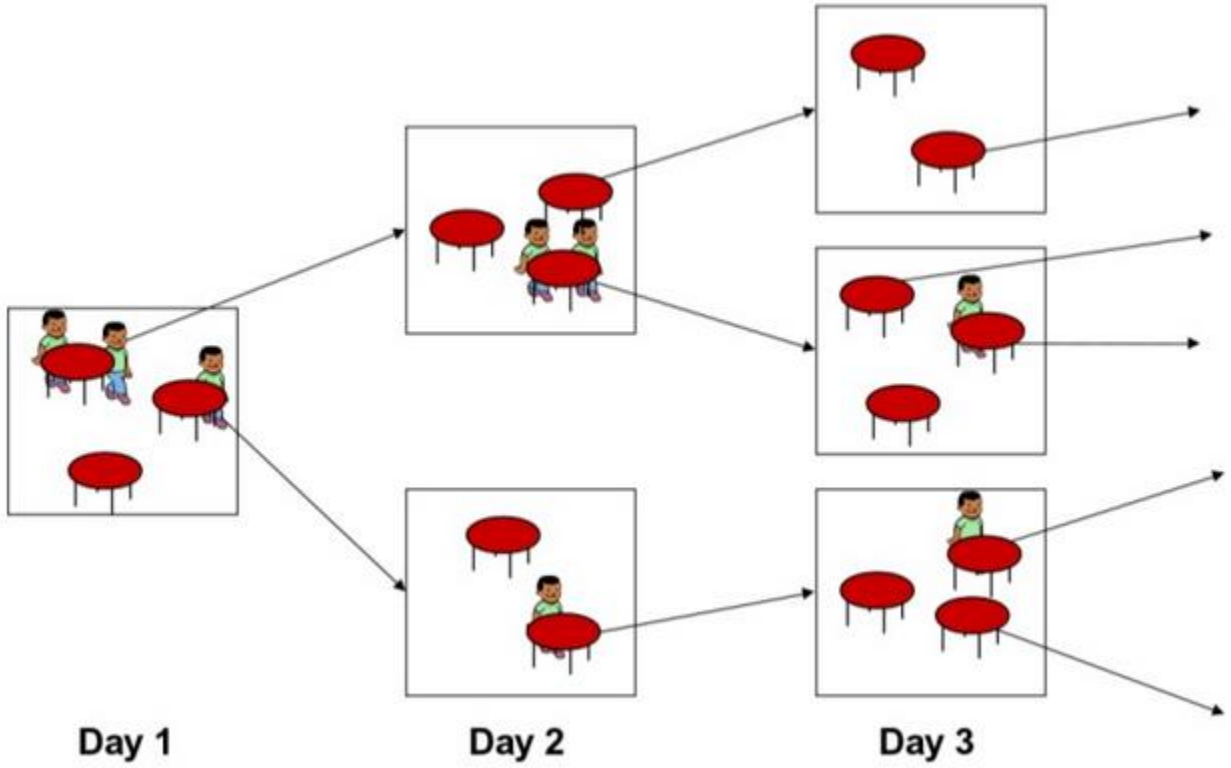
2.2 NESTED CRP



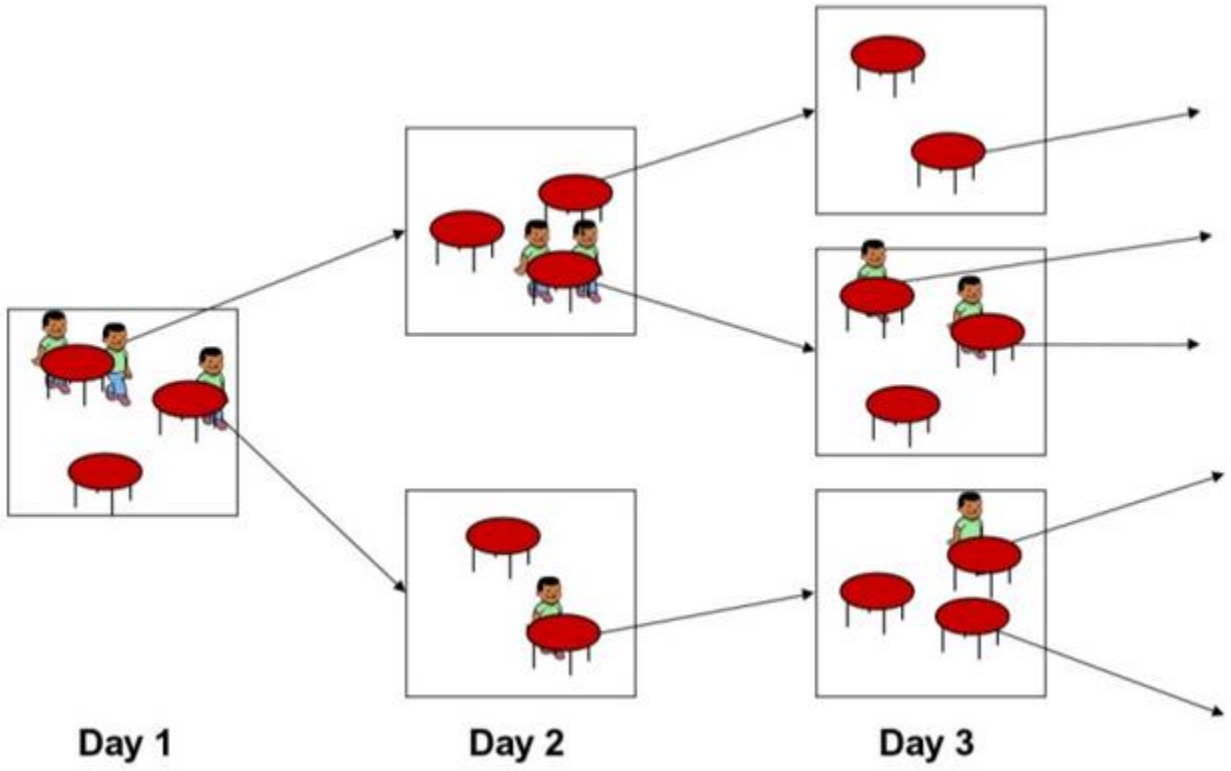
2.2 NESTED CRP



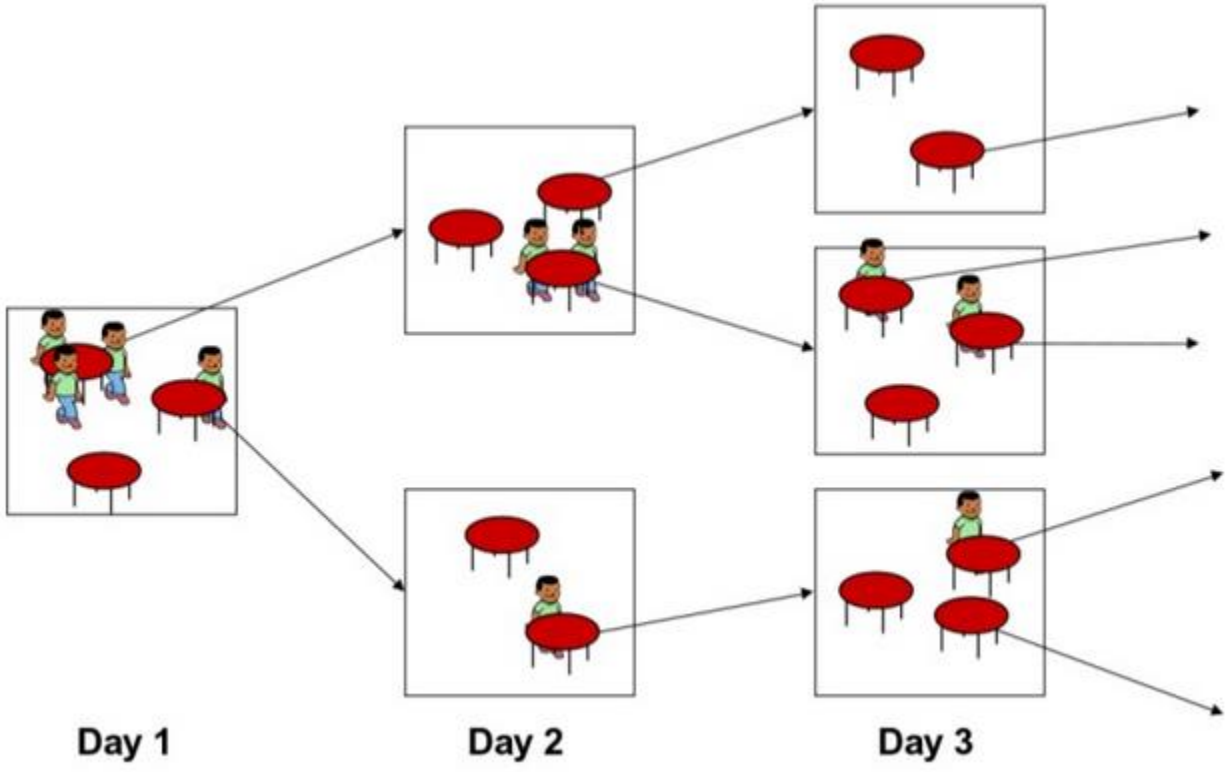
2.2 NESTED CRP



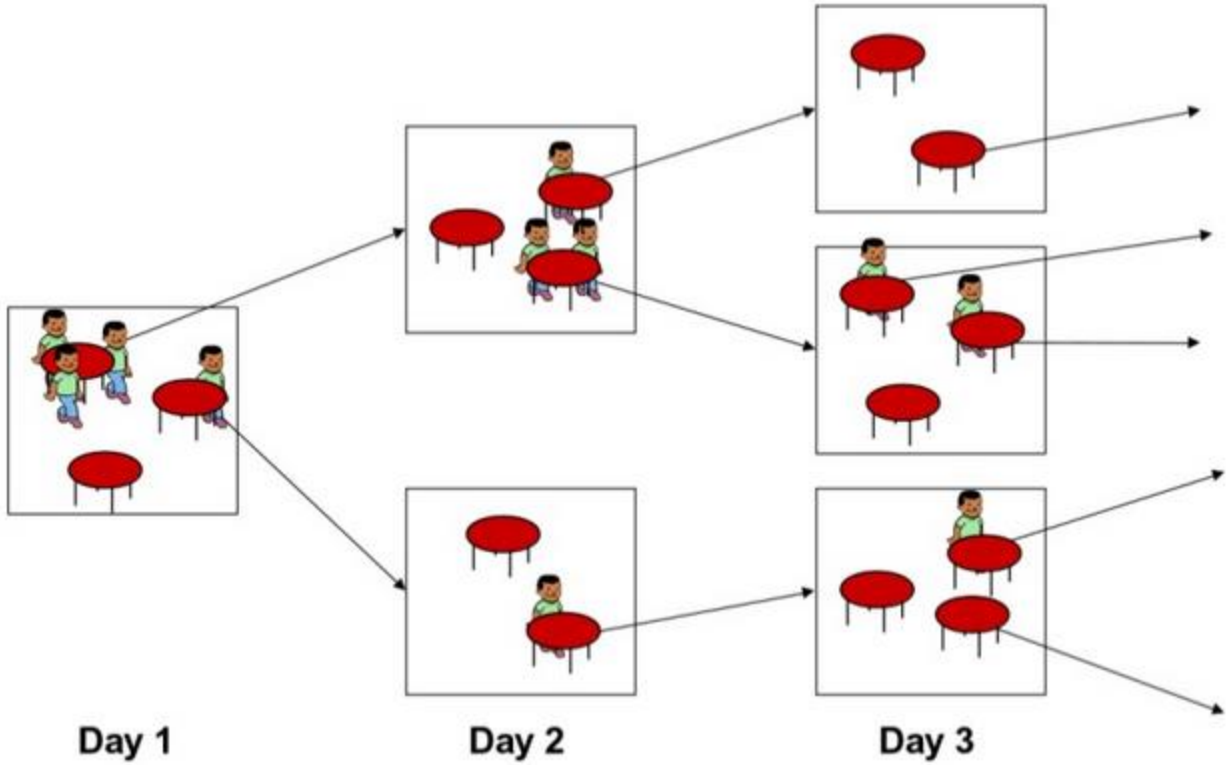
2.2 NESTED CRP



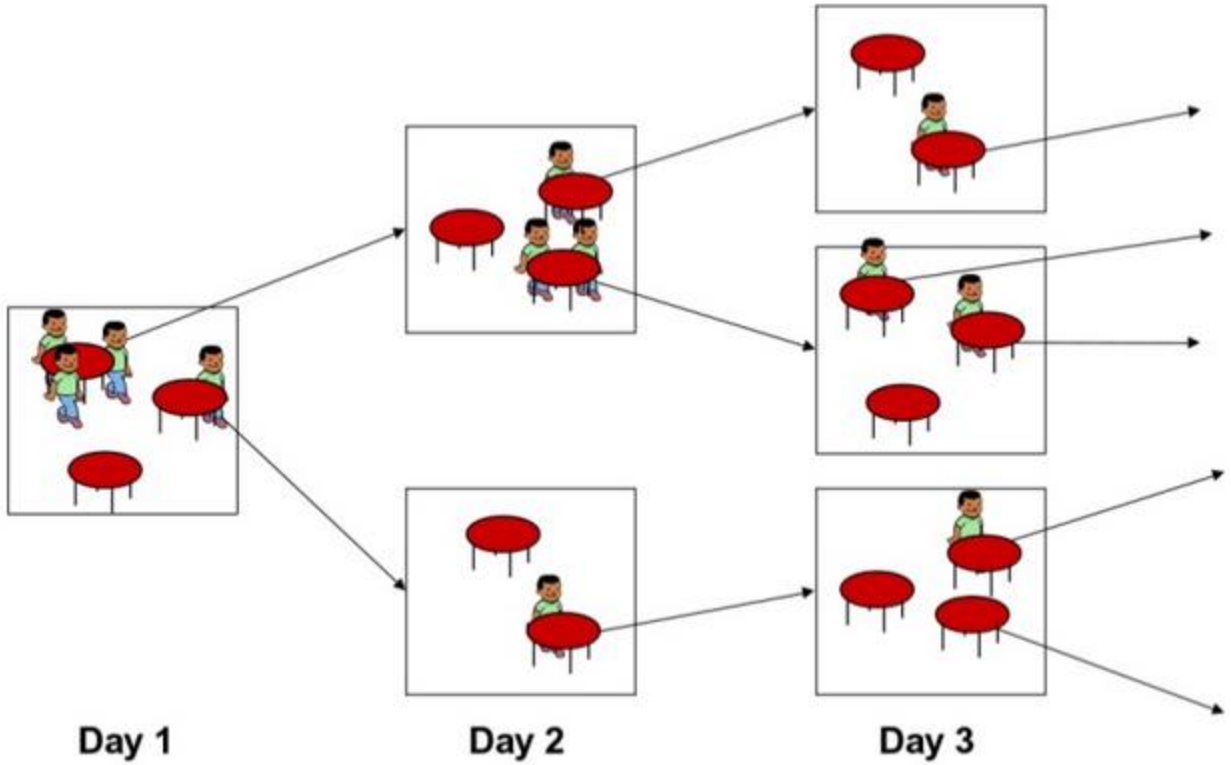
2.2 NESTED CRP



2.2 NESTED CRP

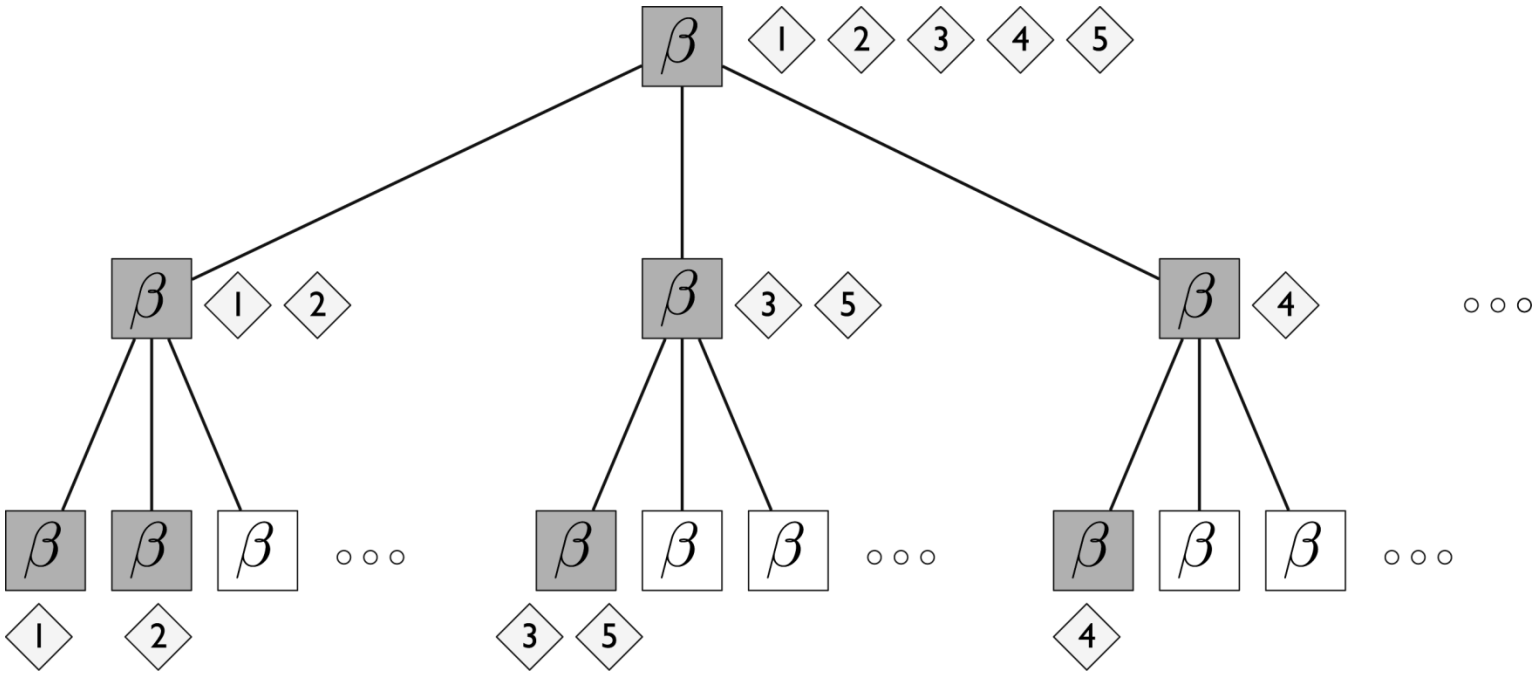


2.2 NESTED CRP



2.2 NESTED CRP

Nested CRP with level L=3:



2.2 NESTED CRP

Generate a document given a tree with L levels:

- Choose a path from the root of the tree to a leaf.
- Draw a vector θ of topic mixing proportions from a L-dimensional Dirichlet.
- Generate the words in the document from a mixture of the topics along the path, with mixing proportions θ .

CONTENTS

3. A HTM

The hierarchical LDA (hLDA):

- Generate model of multiple-topic documents.
- Generate a mixture distribution on topics using a Nested CRP prior. Topics are joined together in a hierarchy by using the nested CRP.
- Pick a topic according to their distribution and generate words according to the word distribution for the topic.

3. A HTM

Observation

- Topics are not independent.

Example

- The topic of CS consists of AI, Systems, Theory, etc.
- AI consists of NLP, Machine Learning, Robotics, vision, etc.

Question

- How to model dependencies between topics.

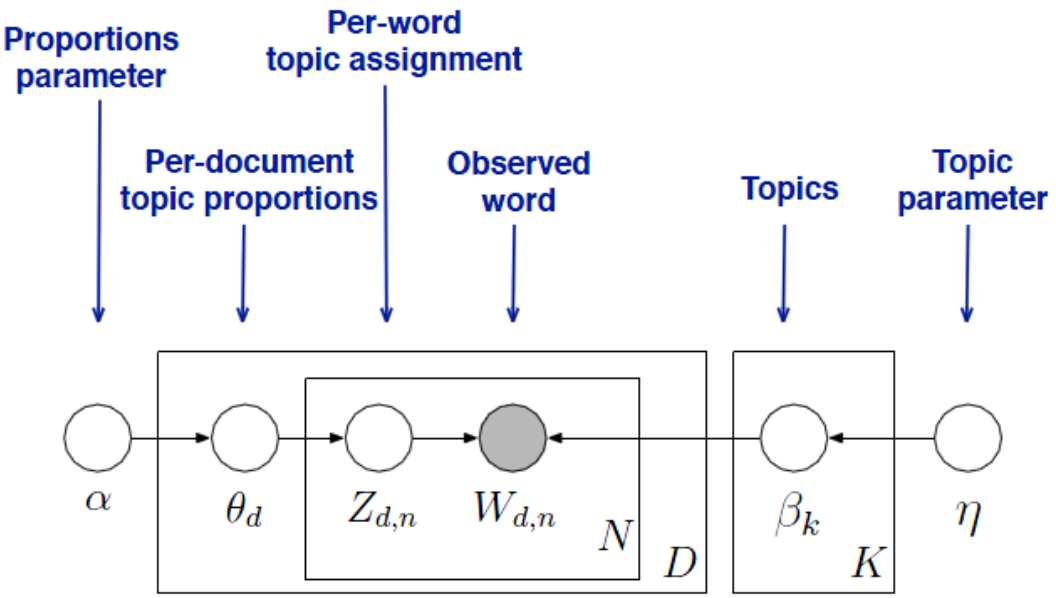
REVIEW OF LDA

Word: the basic unit from a vocabulary of size **V** (includes V distinct words).

Document: a sequence of **N** words. $W = [w_1, w_2, \cdots, w_N]$

Corpus: a collection of **M** documents. $D = [W_1, W_2, \cdots, W_N]$

α, β : hyperparameters, specifying the nature of the priors on θ and ϕ



$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i \mid \eta)\right)\left(\prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n})\right)$$

REVIEW OF LDA

The Dirichlet distribution

- The Dirichlet distribution is a distribution over distribution.

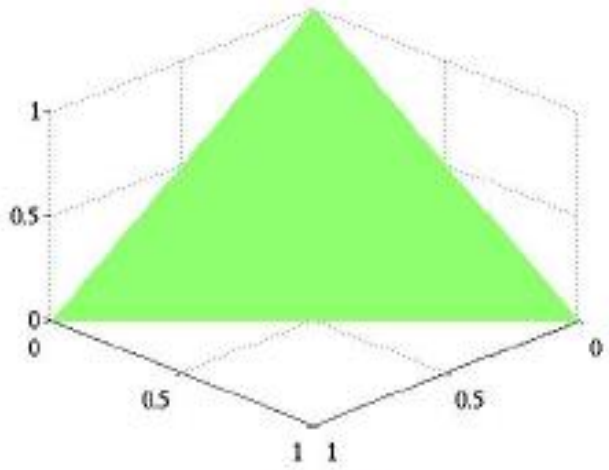
- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one.
  in other words: a draw from a Dirichlet distribution is a vector of positive real numbers that sum up to 1.

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$

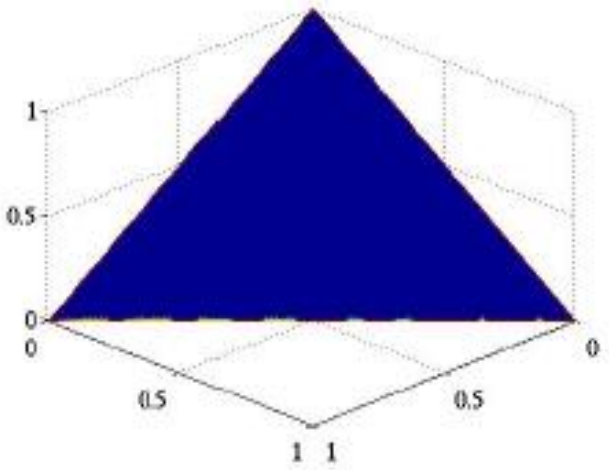
- It is conjugate to the multinomial.

- The parameter α controls the mean shape and sparsity of θ .

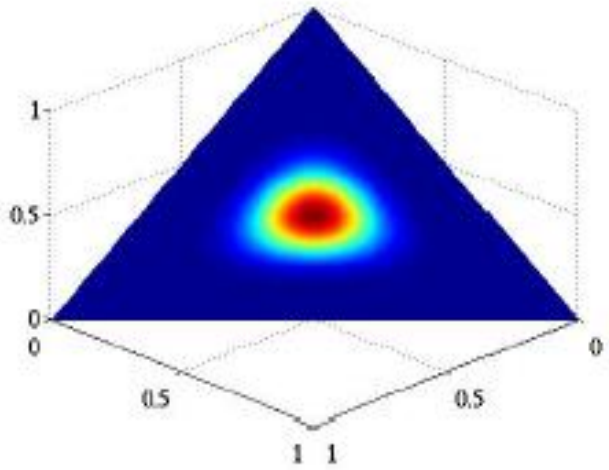
REVIEW OF LDA



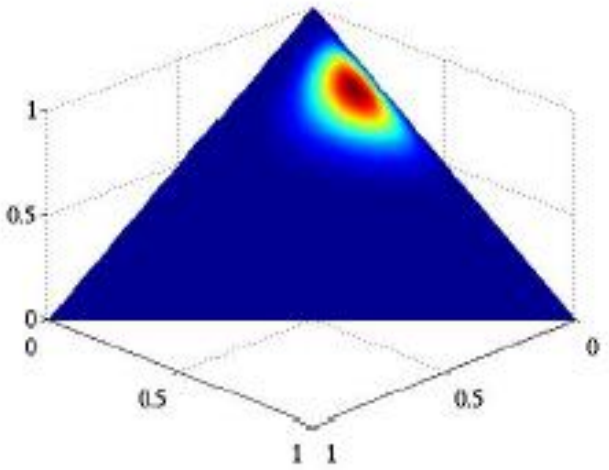
$\alpha = [1, 1, 1]$



$\alpha = [.1, .1, .1]$

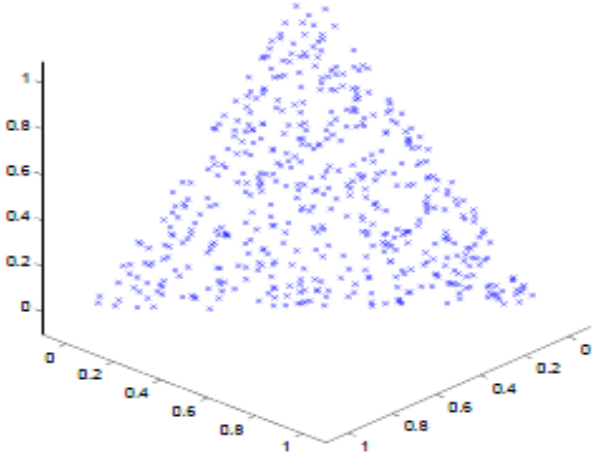


$\alpha = [10, 10, 10]$

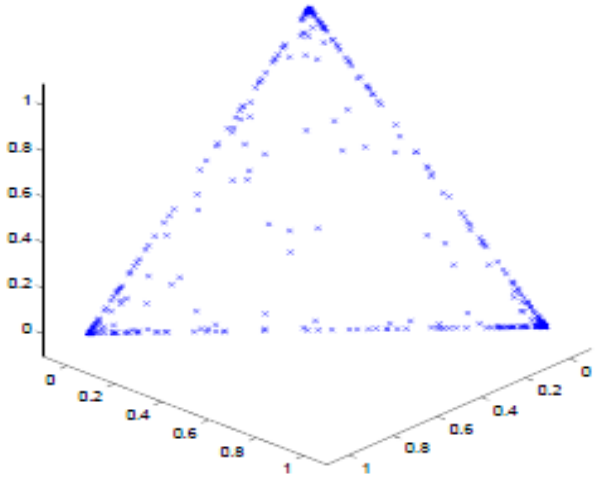


$\alpha = [2, 5, 15]$

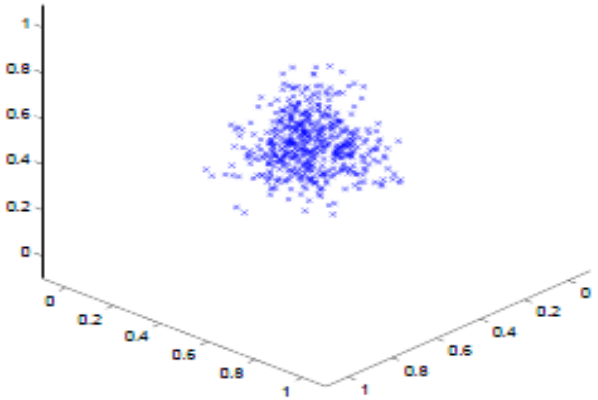
REVIEW OF LDA



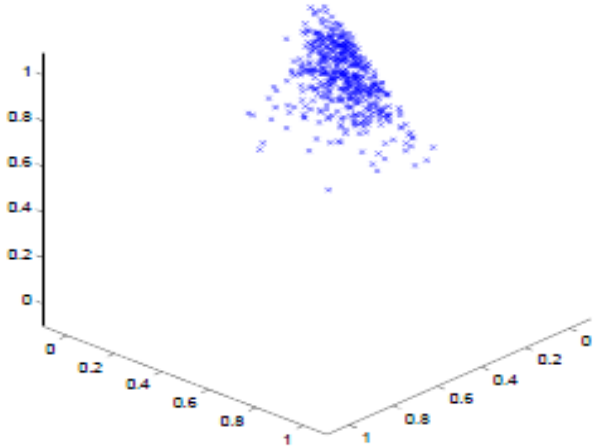
$\alpha = [1, 1, 1]$



$\alpha = [.1, .1, .1]$



$\alpha = [10, 10, 10]$



$\alpha = [2, 5, 15]$

HLDA

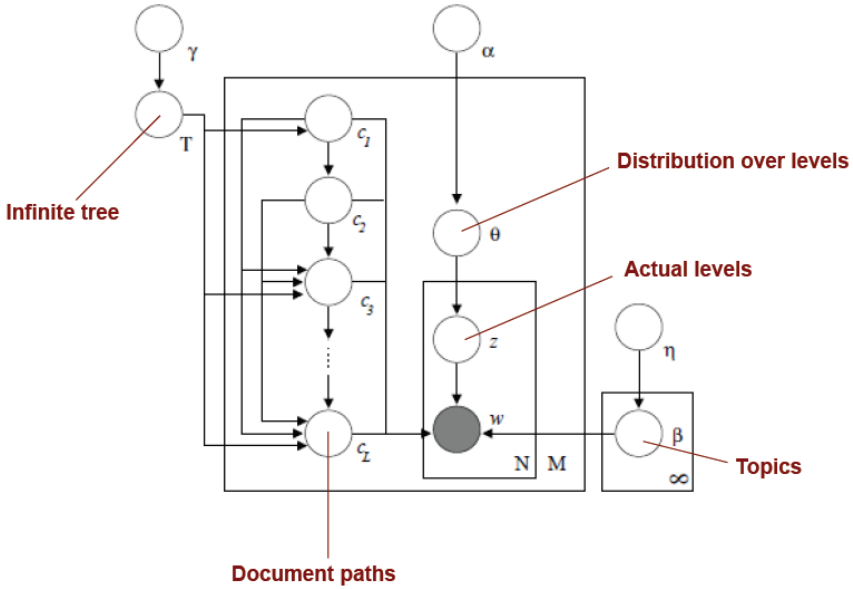


Figure from Blei, et al 2003

Simple representation of hLDA:

- Let c_1 be the root restaurant.
- For each level $l \in \{2, \cdots, L\}$:
  Draw a table from restaurant c_{l-1} .
  Set c_l to be the restaurant referred to by that table.
- Draw an *L*-dimensional topic proportion vector θ from *Dir(α).*
- For each word $n \in \{1, \cdots, N\}$:
  Draw $z \in \{1, \cdots, L\}$ from *mult(θ)*.
  Draw w_n from the topic associated with restaurant c_z.

HLDA

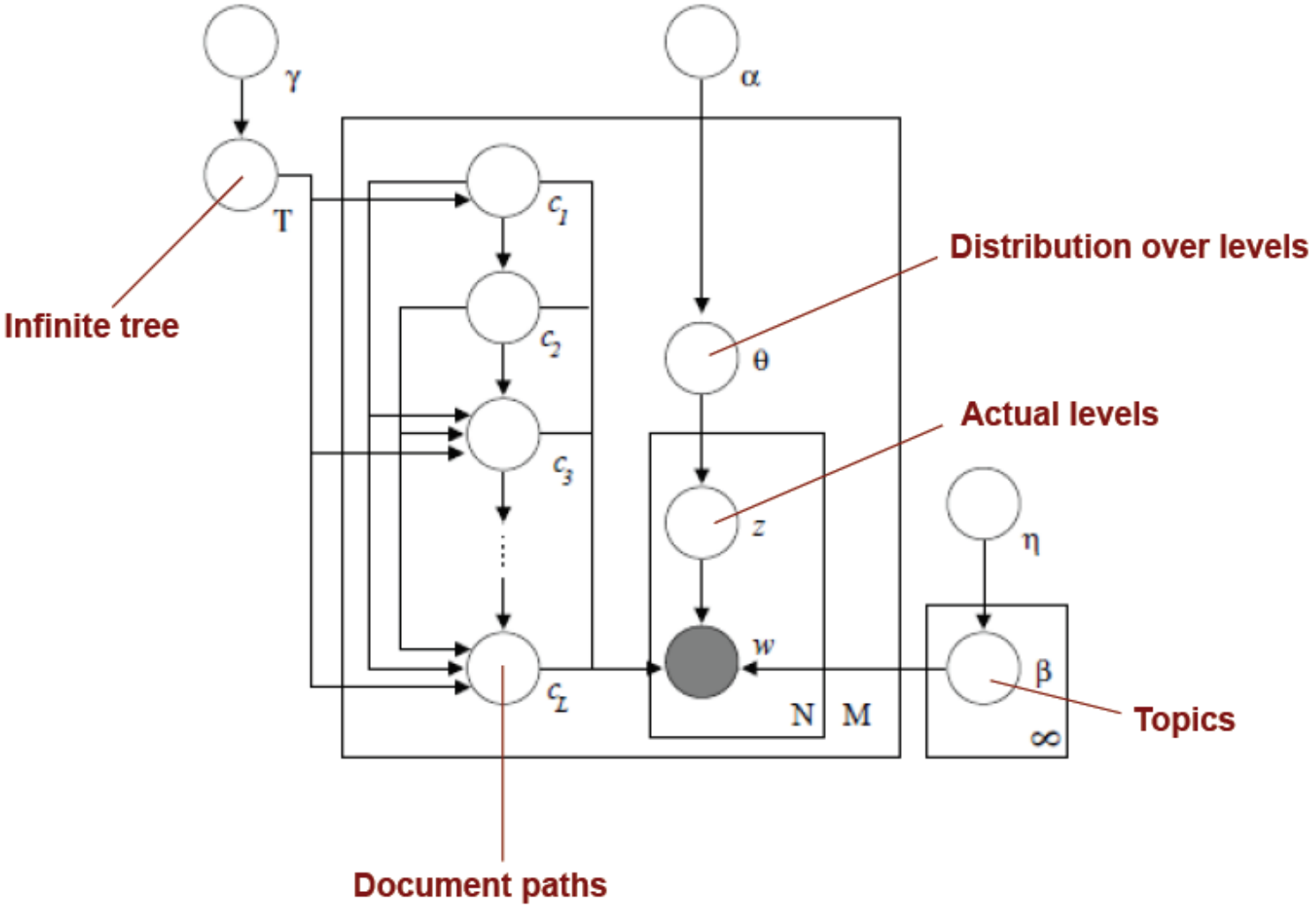


Figure from Blei, et al 2003

CONTENTS

4. PROBABILISTIC INFERENCE

Gibbs sampling:

- Gibbs sampling is commonly used for statistical inference to determine the best value of a parameter. (e.g., θ in LDA)

- The standard step for Gibbs sampling over a space of variables a, b, c.
  Draw a conditioned on b, c
  Draw b conditioned on a, c
  Draw c conditioned on a, b

4. PROBABILISTIC INFERENCE

The variables needed by the sampling algorithm:

- $w_{m,n}$: the *n*th word in the *m*th document (the only observed variable in the model)
- $c_{m,l}$: the restaurant corresponding to the *l*th topic in document m
- $z_{m,n}$: the assignment of the *n*th word in the *m*th document to one of the L available topics.
- All other variables in the model (θ and β) are integrated out.

4. PROBABILISTIC INFERENCE

The conditional posterior distribution for z_i :

- z_{-i} : the assignment of all z_k, $k \neq i$.
- $n_{-i,j}^{(w_i)}$: the number of words assigned to topic j that are same as w_i.
- $n_{-i,j}^{(\cdot)}$: the total number of words assigned to topic j.
- $n_{-i,j}^{(d_i)}$: the number of words from document d_i assigned to topic j.
- $n_{-i}^{(d_i)}$: the total number of words in document d_i.
- α, β : free parameters that determine how heavily these empirical distributions are smoothed.

$$P(z_i = j | z_{-i}, W) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i}^{(d_i)} + T\alpha}$$

4. PROBABILISTIC INFERENCE

The conditional distribution for c_m (the L topics associated with document *m*) :

- $p(w_m|c, w_{-m}, z)$: the likelihood of the data given a particular choice of c_m
- $p(c_m|c_{-m})$: the prior on c_m implied by the nested CRP

$$p(c_m|w, c_{-m}, z) \propto p(w_m|c, w_{-m}, z)p(c_m|c_{-m})$$

The likelihood is obtained by integrating over the parameter β :

- $n^{(w)}_{c_{m,l},-m}$: the number of instances of word w that have been assigned to the topic indexed by $c_{m,l}$, not including those in the current document.
- W : the total vocabulary size.

$$p(w_m|c, w_{-m}, z) = \prod_{l=1}^{L} \left(\frac{\Gamma(n^{(\cdot)}_{c_{m,l},-m} + W\eta)}{\prod_w \Gamma(n^{(w)}_{c_{m,l},-m} + \eta)} \frac{\prod_w \Gamma(n^{(w)}_{c_{m,l},-m} + n^{(w)}_{c_{m,l},m} + \eta)}{\Gamma(n^{(\cdot)}_{c_{m,l},-m} + n^{(\cdot)}_{c_{m,l},m} + W\eta)} \right)$$

CONTENTS

5. DISCUSSION

Good points:

- The nested CRP allows a flexible family of prior distributions over arbitrary tree structures; definitely could be useful for more than just topic models.
- Nice qualitative results for topic hierarchies.

Bad points:

- The restriction that documents can only follow a single path in the tree is a possibly limiting one.
- Quantitative evaluation is not extensive enough.

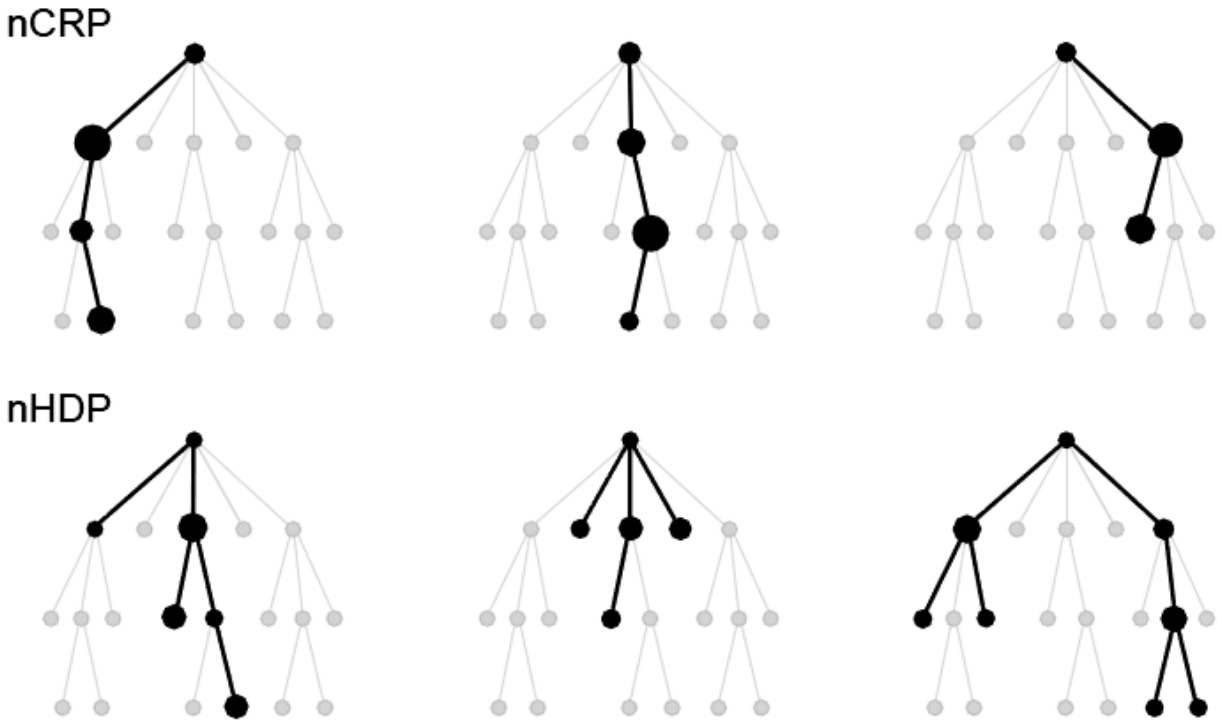
5. DISCUSSION

Further thinking:

- Using this model, we can break down the documents into topics, but
- Relationship among the document?
- Relationship among the topics? (combine HTM and CTM ??)

5. DISCUSSION

Next stage of nCRP ⟶ nHDP:



Nested Hierarchical Dirichlet Processes. 2012
John Paisley, Chong Wang, David M. Blei and Michael I. Jordan1;

THE END

Thanks!!!