Distance Dependent Chinese Restaurant Processes

David M. Blei

Peter I. Frazier

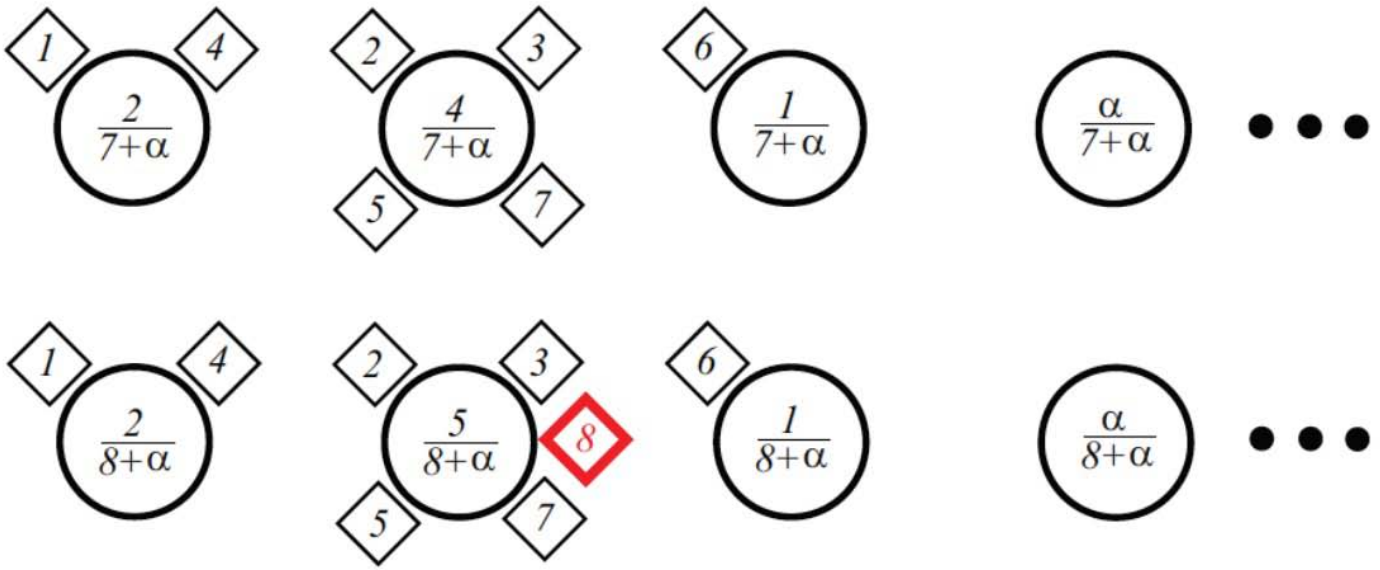
Outline

- Chinese Restaurant Processes - CRP
- Distance Dependent Chinese Restaurant Processes – ddCRP
- Marginal Invariance
- Language Modeling & Mixture Modeling
- Posterior Inference & Prediction
- When is ddCRP Marginally Invariant?
- Related Work
- Empirical Studies
- Discussion

Chinese Restaurant Processes

- A sequence of customers sitting at randomly chosen tables, their configuration represents a random partition.

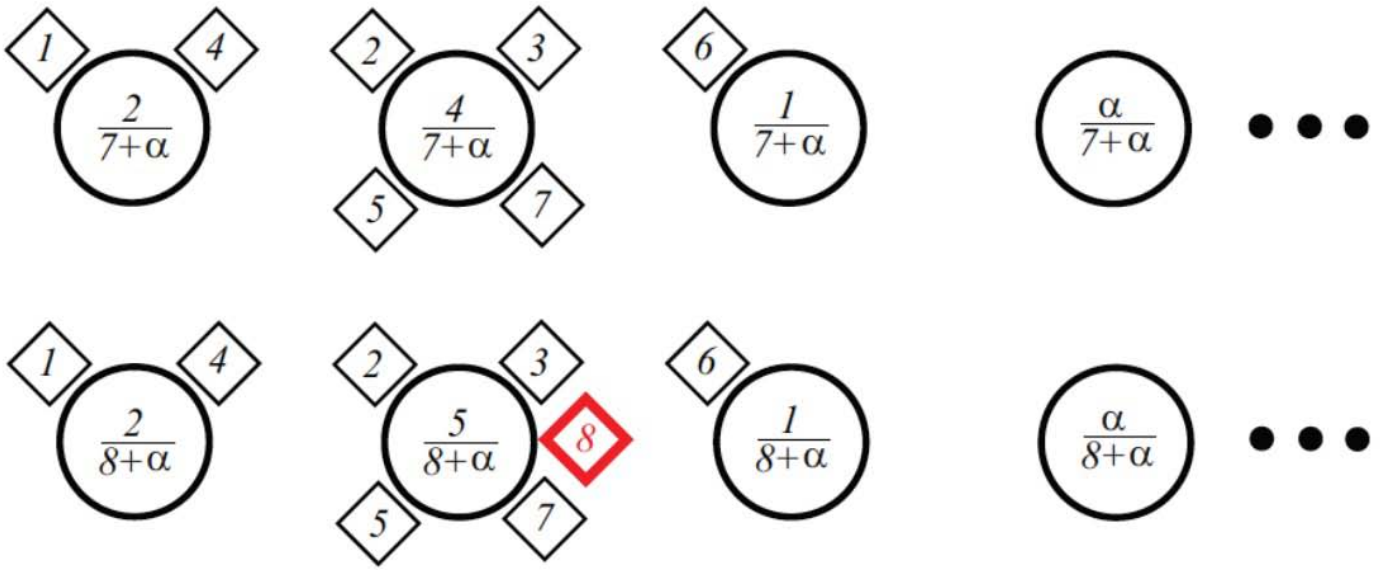
$$p(z_i = k \,|\, z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k & \text{for} \quad k \leq K \\ \alpha & \text{for} \quad k = K+1 \end{cases}$$



Chinese Restaurant Processes

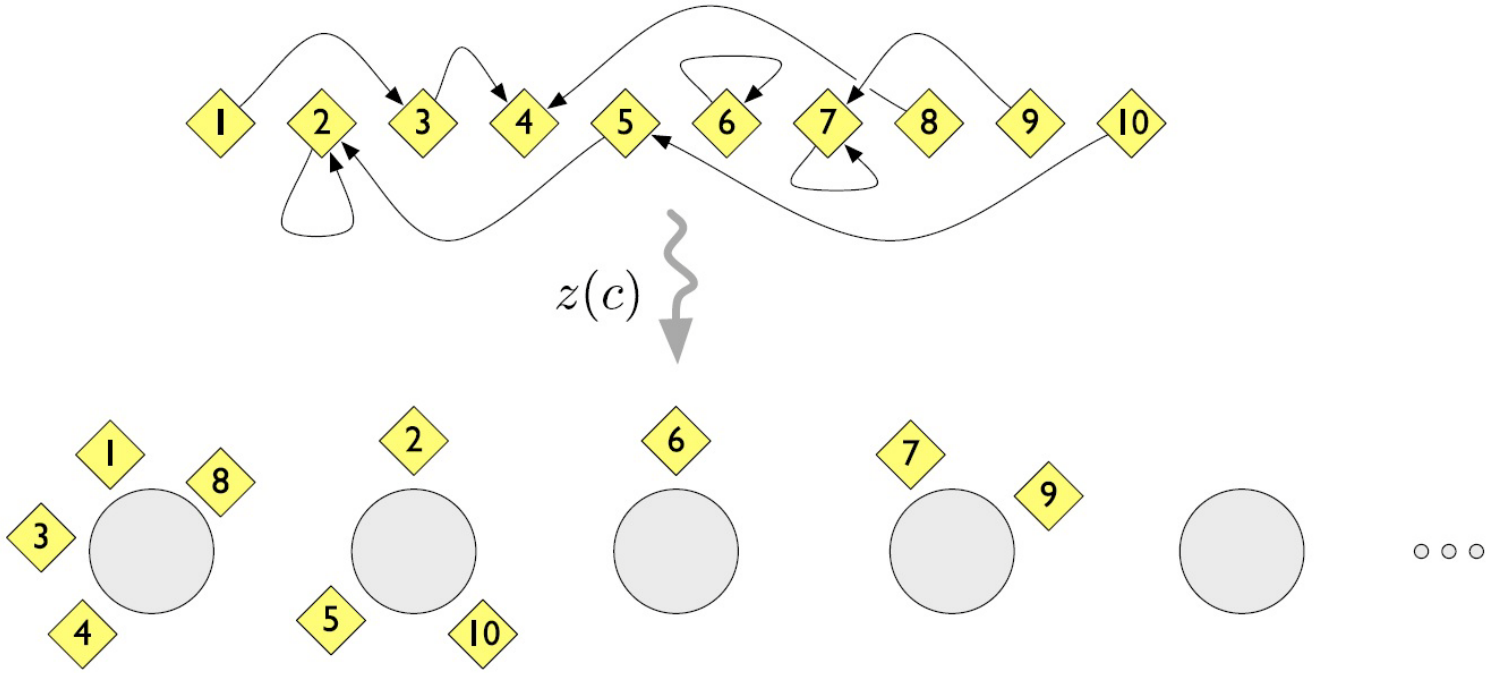
- A sequence of customers sitting at randomly chosen tables, their configuration represents a random partition.

$$p\left(z_i = k \mid z_{1:i-1}, \alpha, G_0\right) = \begin{cases} \dfrac{n_k}{i-1+\alpha}, & k \le K \\ \dfrac{\alpha}{i-1+\alpha} G_0, & k = K+1 \end{cases}$$



Distance Dependent CRP

- The random seating assignment of the customers depends on the distances between them, in terms of the probability of a customer sitting with each of the other *customers*



- Non-exchangeable

Distance Dependent CRP

- Denotation: c_i - index of the customer the *i*th customer sits with

  d_{ij} - distance measure between customer *i* and *j*

$$p(c_i = j \mid D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if} \quad j \neq i \\ \alpha & \text{if} \quad i = j \end{cases}$$

- Decay Functions: non-increasing, takes non-negative finite values, and satisfies $f(\infty) = 0$

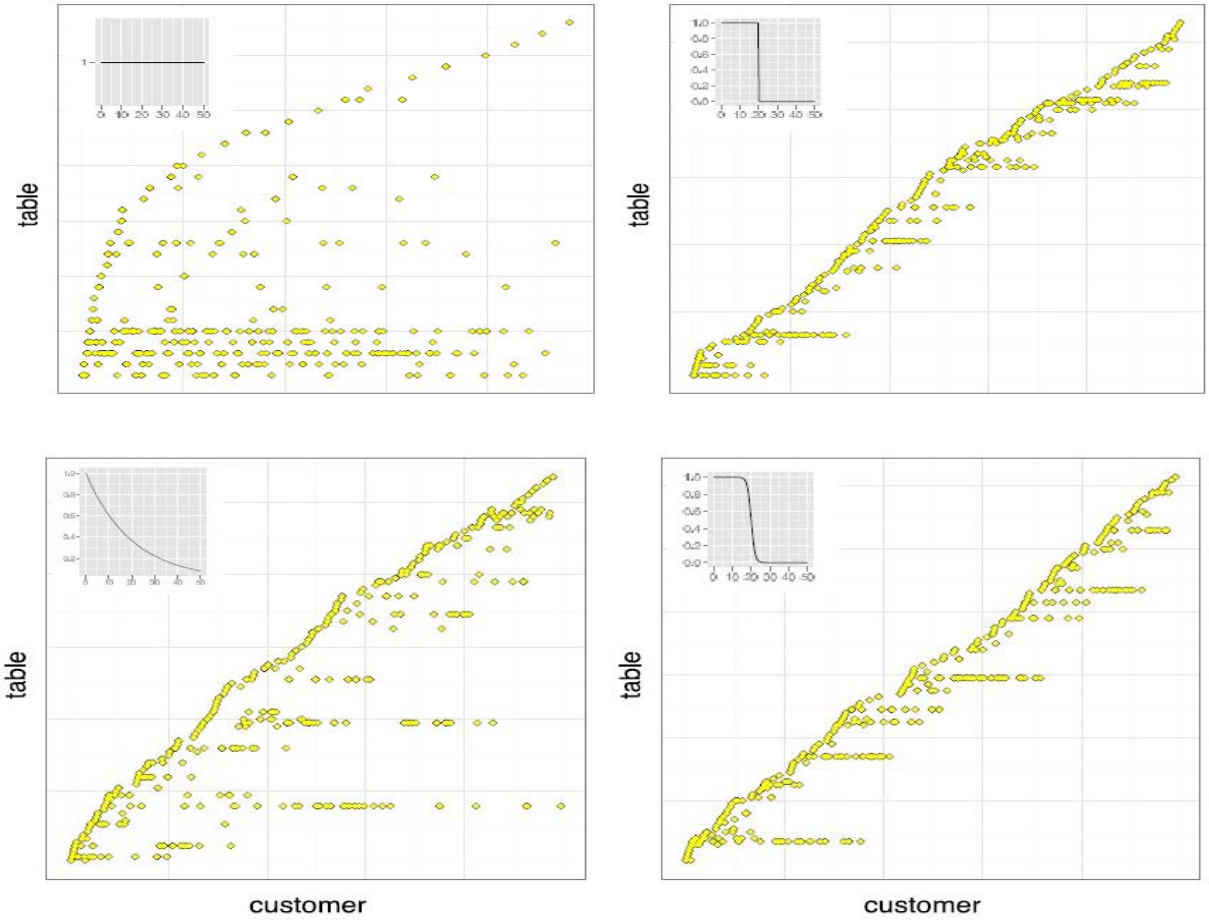
  - Window Decay $f(d) = 1(d < a)$

  - Exponential Decay $f(d) = \exp(d/a)$

  - Logistic Decay $f(d) = \dfrac{\exp(-d + a)}{1 + \exp(-d + a)}$

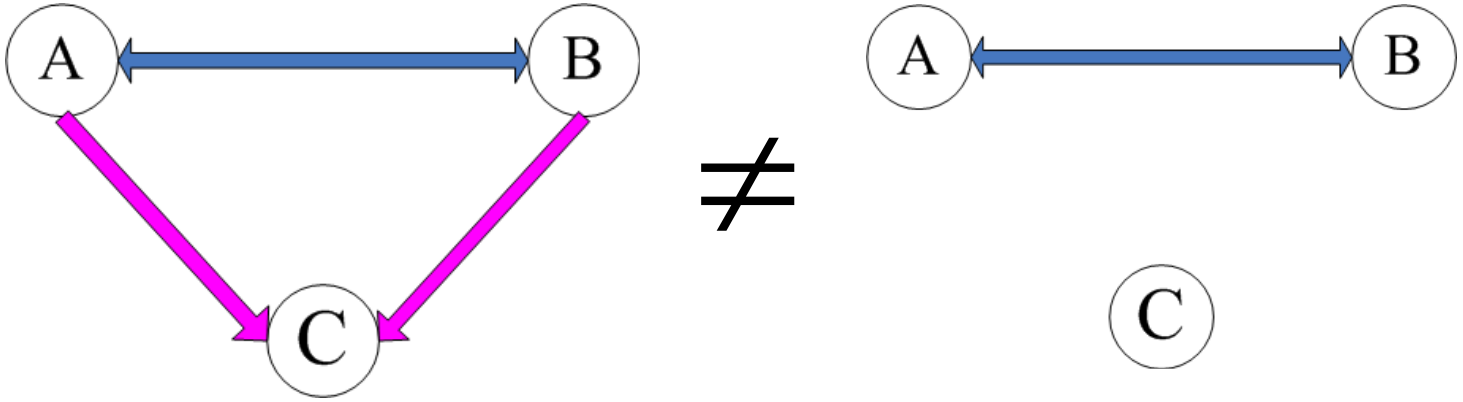
Distance Dependent CRP

- Sequential CRP: constructed by assuming $d_{ij} = \infty$ for $j > i$
- Traditional CRP: $f\left(d_{ij}\right) = 1$ for $d_{ij} \neq \infty$, $d_{ij} < \infty$ for $j < i$



Marginal Invariance

- A missing observation does not affect the joint distribution
- Convenient factorization and easier computation
- Distance Dependent CRPs generally doesn't have the property, influence might be transmitted from one point to another. eg. social network, spread of disease



Language Modeling

- A document is associated with a distance dependent CRP, each table is embellished with iid draws from a distribution over words.
- The data are first placed at tables via customer assignments, and then assigned to the word associated with their tables.
- $z(\mathbf{c})_i$ - the table assignment of the *i*th customer

1. For each word $i \in \{1, \ldots, N\}$ draw assignment $c_i \sim \text{dist-CRP}(\alpha, f, D)$.
2. For each table, $k \in \{1, \ldots\}$, draw a word $w^* \sim G_0$.
3. For each word $i \in \{1, \ldots, N\}$, assign the word $w_i = w^*_{z(\mathbf{c})_i}$.

Mixture Modeling

- Similar to CRP mixture, but the mixture component for data points depends on mixture components for nearby data.
- Observations are documents, instead of individual words.
- G_0 is typically a DP distribution over distribution of words

1. For each document $i \in [1, N]$ draw assignment $c_i \sim \text{dist-CRP}(\alpha, f, D)$.
2. For each table, $k \in \{1, \ldots\}$, draw a parameter $\theta_k^* \sim G_0$.
3. For each document $i \in [1, N]$, draw $w_i \sim F(\theta_{z(\mathbf{c})_i})$.

Posterior Inference

- no exchangeability, so exact posterior computation needs likelihoods of exponential number of assignment vectors
- "fortunately", we have MCMC methods like Gibbs sampling (assuming conjugacy)
- proposed sampler does updates on individual customers' seating assignments

- likelihood: $$p(\mathbf{x} \mid z(\mathbf{c}), G_0) = \prod_{k=1}^{|z(\mathbf{c})|} p(\mathbf{x}_{z^k(\mathbf{c})} \mid G_0).$$

- $z^k(\mathbf{c})$: set of indices assigned to table k

Posterior Inference, cont.

- we want the marginal distribution of a data point over seating assignments
- this gives us a collapsed sampler
- marginal likelihood:

$$p(\mathbf{x}_{z^k(\mathbf{c})} \mid G_0) = p(x_{z^k(\mathbf{c})_1} \mid G_0) \prod_{i \in z^k(\mathbf{c})} 1(x_i = x_{z^k(\mathbf{c})_1})$$

$$= \int \left(\prod_{i \in z^k(\mathbf{c})} p(x_i \mid \theta) \right) p(\theta \mid G_0) d\theta$$

- but wait…don't collapsed samplers mix super-slow?

Posterior Inference, cont.

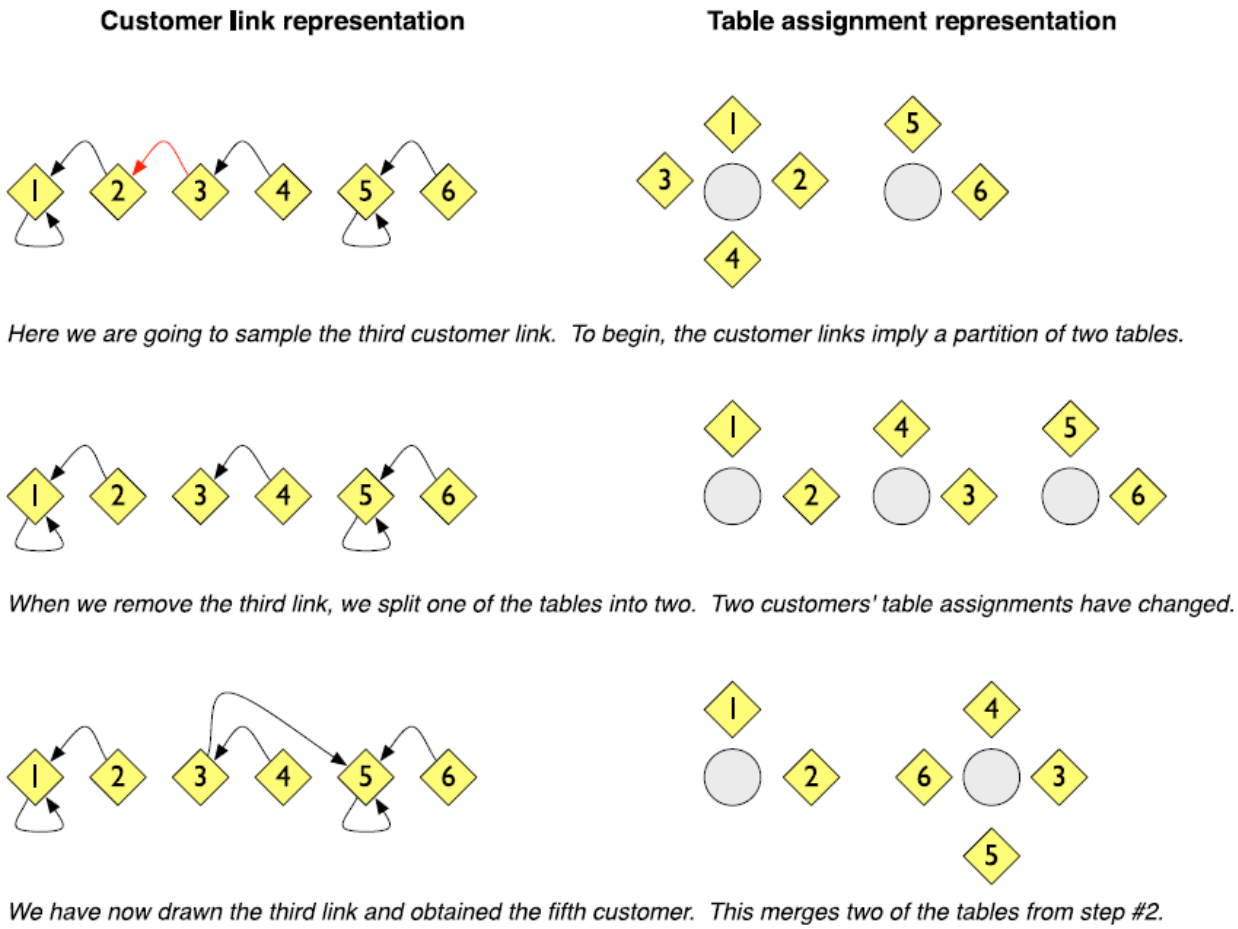
- not necessarily: the updates change the linkage index of a customer, (who they chose to sit with), which in turn changes the linkage index of all customers pointing to that customer (illustration next slide)
- update equation:

$$p(c_i^{(\text{new})} \mid \mathbf{c}_{-i}, \mathbf{x}, \eta) \propto p(c_i^{(\text{new})} \mid D, \alpha) p(\mathbf{x} \mid z(\mathbf{c}_{-i} \cup c_i^{(\text{new})}), G_0).$$

$$\propto \begin{cases} \alpha & \text{if } c_i^{(\text{new})} \text{ is equal to } i. \\ f(d_{ij}) & \text{if } c_i^{(\text{new})} = j \text{ does not join two tables.} \\ f(d_{ij}) \dfrac{p(\mathbf{x}_{z^k(\mathbf{c}_{-i}) \cup z^\ell(\mathbf{c}_{-i})} \mid G_0)}{p(\mathbf{x}_{z^k(\mathbf{c}_{-i})} \mid G_0) p(\mathbf{x}_{z^\ell(\mathbf{c}_{-i})} \mid G_0)} & \text{if } c_i^{(\text{new})} = j \text{ joins tables } k \text{ and } \ell. \end{cases}$$

$$\eta = \{D, \alpha, f, G_0\}$$

Gibbs Sampler example



- we can not only change several table assignments at once, but split and merge tables as well, giving a wider range of possible updates
- authors' experiments suggest that this property causes faster mixing

Predictive Distribution

$$p(x_{\text{new}}|\mathbf{x}, D, G_0, \alpha) = \sum_{c_{\text{new}}} p(c_{\text{new}}|D, \alpha) \sum_{\mathbf{c}} p(x_{\text{new}}|c_{\text{new}}, \mathbf{c}, \mathbf{x}, G_0) p(\mathbf{c}|\mathbf{x}, D, \alpha, G_0)$$

$$\sum_{\mathbf{c}} p\left(x_{new} \mid c_{new}, \mathbf{c}, \mathbf{x}, G_0\right) p\left(\mathbf{c} \mid \mathbf{x}, D, \alpha, G_0\right) \approx \frac{1}{N_{samples}} \sum_{j=1}^{N_{samples}} p\left(x_{new} \mid c_{new}, \mathbf{c}^{(j)}, \mathbf{x}^{(j)}, G_0\right)$$

- if distances are sequential and x_{new} is a data point in the future, the customer assignment vector (**c**) is independent of its time, so we can reuse the previously generated samples of customer assignment vectors
- otherwise, the prior (and as a result, the posterior) of **c** is changed, and we'd have to re-generate samples

When is ddCRP Marginally Invariant?

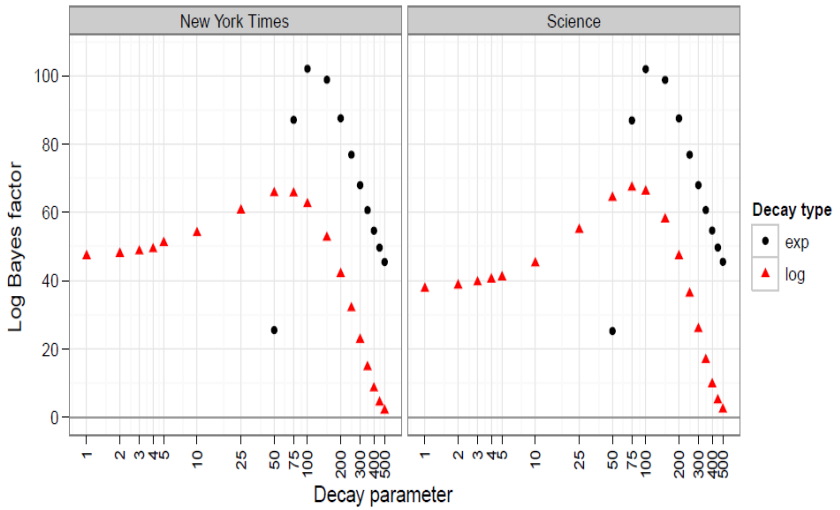
- Marginal Invariance: all sub-vectors formed by removing one data point (customer) from a given vector of seating assignments have same probability
- Sequential Distances:
  - easy to construct example showing sequential ddCRP is not marginally invariant without "all or nothing" property of window decay function
  - for any partition over K tables, sequential distances identify a group by the linkage index of the first customer to sit at that table (ie. the lowest one)
  - with window decay function, resulting ddCRP is marginally invariant but **identical to that of K independent traditional CRP's**
- General (sequential and non-) Distances:
  - from before, must use window decay function
  - solving for marginal invariance gives zero decay for all distances, so every customer self-links, ie. **each customer sits at its own table**

Related Work

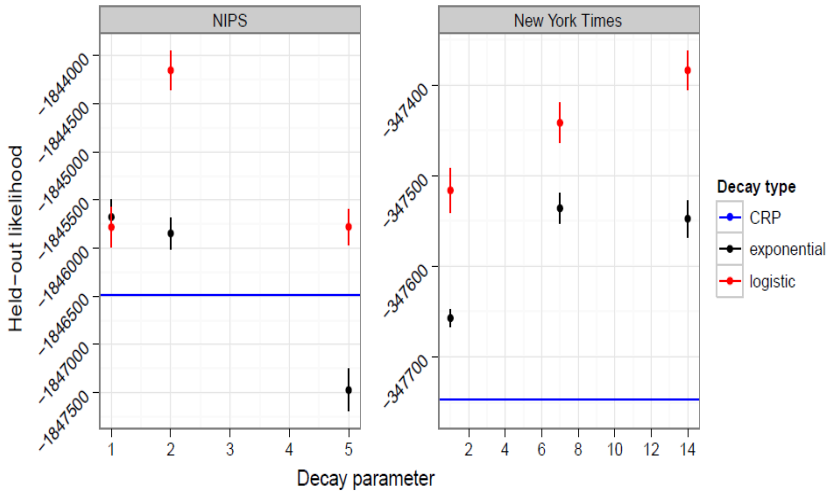
- the authors do not classify the ddCRP as a bayesian non-parametric model, because it is not a mixture model generated from a random measure
- these “Random-Measure” mixture models (such as the dependent DP) place priors on corresponding pairs of covariates (customer assignments) and observations (customer values) that assume observations are conditionally independent of each other given covariate distributions
- the prior distribution on observations is independent of all non-corresponding covariates, giving marginal invariance
- again, the ddCRP is generally not marginally invariant, making it more flexible and definitive (network data example presented in experiments section)

Empirical Studies

Language Modeling (Science, NYT)

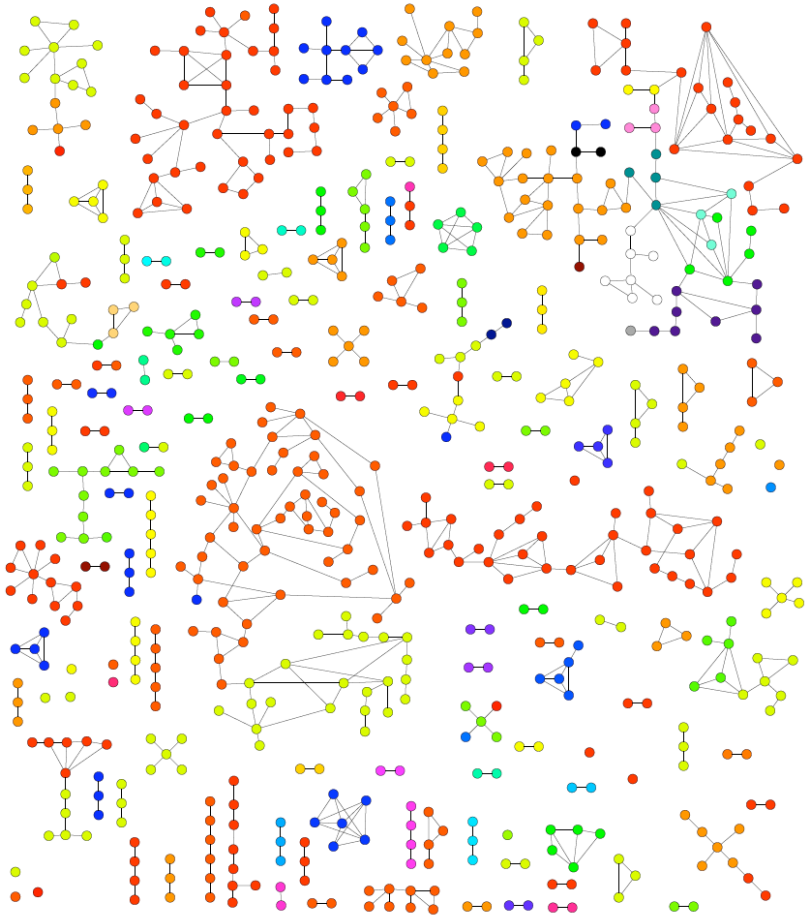


Mixture Modeling (NYT, NIPS)



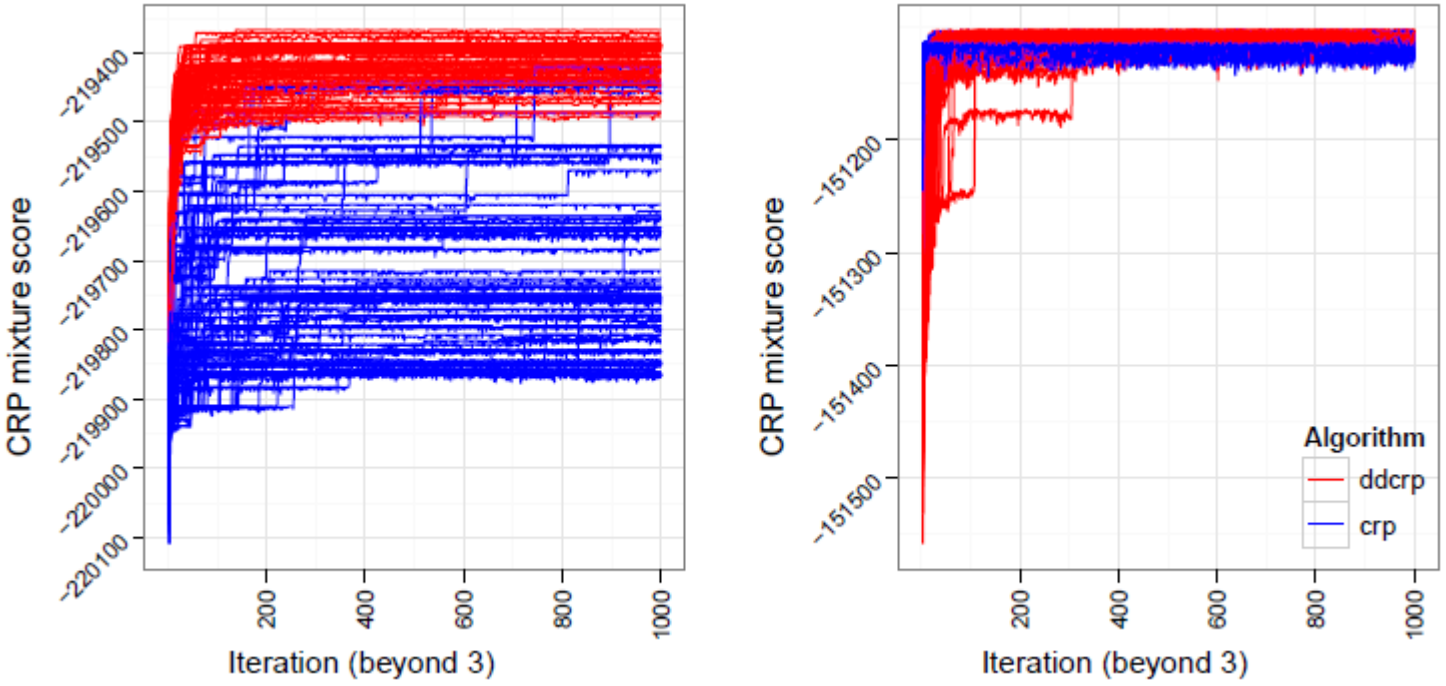
bayes factor: $BF_{f,\alpha} = p(w_{1:N} \mid \text{dist-CRP}_{f,\alpha}) / p(w_{1:N} \mid \text{CRP}_{\alpha})$.

Network Data (CORA)



- traditional CRP does not favor such “sub-clusters”

Gibbs Sampler Mixing: ddCRP vs. CRP



- compared to Algorithm 3 in Neal (2000), on same dataset as first experiment
- claim: this faster mixing is caused by updating several customers at once, breaking/merging tables

Talking Points

- extension to hierarchical clustering
- hyperparameter choices (concentration parameter, decay function; paper suggests “Griddy Gibbs” for sampling)
- general modeling with ddCRP
- any other questions?