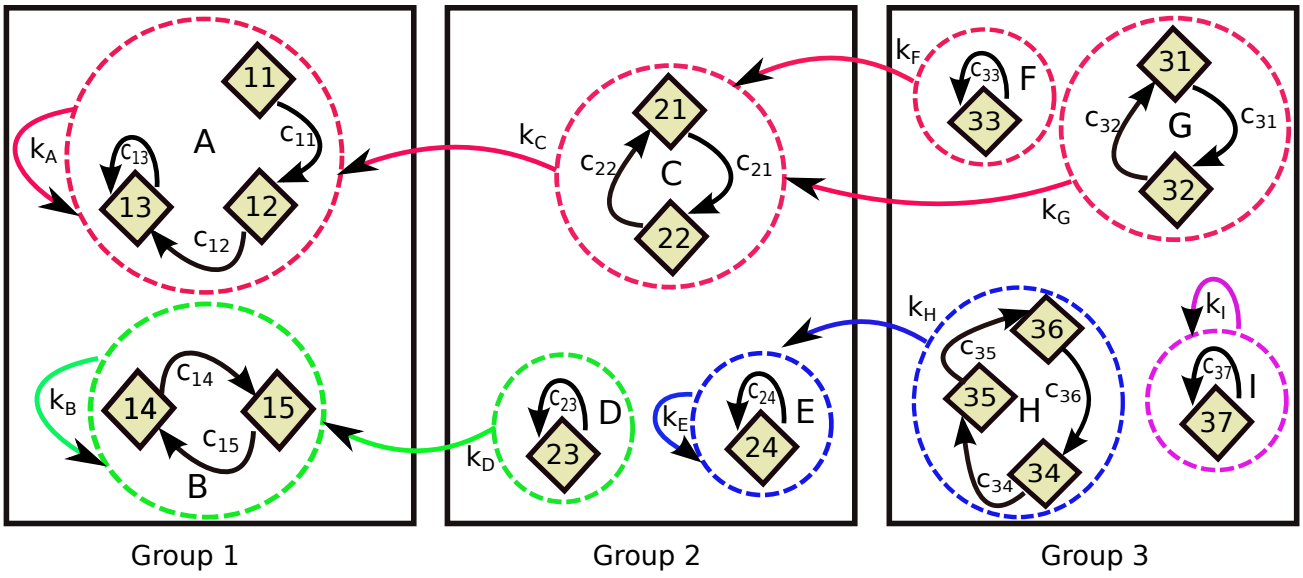

Figure 1: An example link variable configuration for a hierarchical ddCRP model of three groups (rectangles). Observed data points (customers, depicted as diamonds) link to other data points in the same group (black arrows), producing local clusters (dashed circles, labeled A through I). Cluster links (colored arrows) then join clusters to produce (in this case, four) global mixture components.

joint distribution on partitions and observations equals

$$p(\mathbf{x}, \mathbf{k}, \mathbf{c} \mid \alpha_{1:G}, \alpha_0, A^{1:G}, A^0, \lambda) = \prod_{m=1}^{M(\mathbf{c},\mathbf{k})} p(x_{\mathbf{z}=m} \mid \lambda) \prod_{g=1}^{G} \prod_{i=1}^{N_g} p(c_{gi} \mid \alpha_g, A^g) \prod_{k_t \in \mathbf{k}} p(k_t \mid \mathbf{c}, \alpha_0, A^0(\mathbf{c})) \quad (4)$$

The set of data in component m is denoted by $x_{\mathbf{z}=m}$, and

$$p(x_{\mathbf{z}=m} \mid \lambda) = \int \prod_{gi|z_{gi}=m} p(x_{gi} \mid \phi_m)\, dG_0(\phi_m \mid \lambda), \quad (5)$$

where λ are hyperparameters specifying the prior distribution G_0. Our inference algorithms assume this integral is tractable, as it always is when an exponential family likelihood is coupled with an appropriate conjugate prior. We emphasize that for arbitrary data and cluster affinities, the sequential hddCRP generative process defines a valid joint distribution $p(\mathbf{x}, \mathbf{k}, \mathbf{c}) = p(\mathbf{c})p(\mathbf{k} \mid \mathbf{c})p(\mathbf{x} \mid \mathbf{k}, \mathbf{c})$.

2.2 RELATED HIERARCHICAL MODELS

The hddCRP subsumes several recently proposed hierarchical extensions to the ddCRP, as well as the HDP itself, by defining appropriately restricted data affinities and local cluster affinities. Blei & Frazier (2011) show that the CRP is recovered from the ddCRP by arranging data in an arbitrary sequential order, and defining affinities as

$$A_{ij} = \begin{cases} 1 & \text{if } i < j, \\ 0 & \text{if } i > j. \end{cases} \quad (6)$$

Data points link to all previous observations with equal probability, and thus the probability of joining any existing cluster is proportional to the number of other data points already in that cluster. The probability of creating a new cluster is proportional to the self-connection weight α. The resulting distribution on partitions can be shown to be invariant to the chosen sequential ordering of the data, and thus the standard CRP is *exchangeable* (Pitman, 2002).

Hierarchical Chinese Restaurant Process (hCRP) The hCRP representation of the HDP, which Teh et al. (2006) call the "Chinese restaurant franchise", is recovered from the hddCRP by first defining group-specific affinities as in Eq. (6). We then arrange local clusters (tables, in the CRF metaphor) t sequentially with distances $A^0_{ts}(\mathbf{c}) = 1$ if $t < s$, and $A^0_{ts}(\mathbf{c}) = 0$ if $t > s$. Just as the two-level hCRP arises from a sequence of CRPs, the hddCRP is defined from a sequence of two ddCRP models.

Naive Hierarchical ddCRP (naive-hddCRP) The image segmentation model of Ghosh et al. (2011) clusters data within each group via a ddCRP based on an informative distance (in their experiments, spatial distance between image pixels). A standard CRP, as in the upper level of the HDP, is then used to combine these clusters into larger segments. Inference is substantially simpler for this special case, because cluster distances do not depend on properties of the data assigned to those clusters.

Distance Dependent Chinese Restaurant Franchise An alternate approach to capturing group-specific metadata uses a standard CRP to locally cluster data, but then uses the group labels to define affinities between clusters. Kim & Oh (2011) use this model to learn topic models of time-stamped documents. By constraining cluster affinities to depend on group labels, but not properties of the data assigned to within-group clusters, inference is simplified.