Griffin & Steel (2006), Dunson & Park (2008), Rao & Teh (2009), and Lin et al. (2010) define priors which encourage "close" data points to have similar allocation distributions.

In contrast, the hddCRP directly specifies distributions over partitions via a flexible set of user-specified affinity functions. This allows structural constraints on clusters, such as connectivity (Ghosh et al., 2011), to be directly enforced. The hddCRP does not require its "distance" functions to be true metrics or have any special properties, and thus provides an extremely flexible framework for modeling complex data. Alternative models based on latent Gaussian processes (Duan et al., 2007; Sudderth & Jordan, 2008) require appropriate positive-definite kernel functions, whose specification can be challenging in non-Euclidean spaces (e.g., of object shapes). By working directly with discrete partitions, rather than latent continuous measures, the hddCRP also allows more computationally efficient inference.

The hddCRP generative process defined in Section 2 is simple, but the data-level and cluster-level link variables are strongly coupled in the posterior. Section 3 develops a *Markov chain Monte Carlo* (MCMC) method that makes coordinated changes to links at both levels, and thus more effectively explores clustering hypotheses. This sampler is also a novel inference algorithm for the HDP that makes large changes to the partition structure, without needing to explicitly craft split or merge proposals (Jain & Neal, 2004; Wang & Blei, 2012). By reasoning about data and cluster links, our sampler changes cluster allocations at varying resolutions, perturbing both memberships of data instances to local clusters and clusters to global components.

In Section 4, we demonstrate the versatility of the hddCRP by applying it to the problems of video segmentation and discourse analysis. In addition to having diverse data types (video sequences versus text documents), these two problems exhibit very different kinds of relationships among data instances and latent clusters. Nevertheless, our hddCRP model and inference framework easily applies to both domains by selecting appropriate data and cluster-level affinity functions. In both domains, explicit modeling of dependencies between latent clusters boosts performance over models that ignore such relationships.

2 HIERARCHICAL DISTANCE DEPENDENT CLUSTERS

The distance-dependent CRP (Blei & Frazier, 2011) defines a distribution over partitions indirectly via distributions over links between data instances. A data point i has an associated link variable c_i which links to another data instance j, or itself, according to the following distribution:

$$p\left(c_i = j \mid A, \alpha\right) \propto \begin{cases} A_{ij} & i \neq j, \\ \alpha & i = j. \end{cases} \tag{1}$$

The *affinity* $A_{ij} = f(d(i,j))$ depends on a user-specified *distance* $d(i,j)$ between pairs of data points, and a monotonically decreasing *decay function* $f(d)$ which makes links to nearby data more likely. The resulting link structure induces a partition, where two data instances are assigned to the same cluster if and only if one is reachable from the other by traversing the link edges. Larger self-affinity parameters α favor partitions with more clusters.

2.1 THE HIERARCHICAL ddCRP

We propose a novel generative model that applies the ddCRP formalism twice, first for clustering data within each group into local clusters, and then for coupling the local clusters across groups. Like the ddCRP, our hddCRP defines a valid distribution over partitions of a dataset. It places higher probability mass on partitions that group nearby data points into latent clusters, *and* couple similar local clusters into global components. Examples of these data and cluster links are illustrated in Figure 1.

Consider a collection of G groups, where group g contains N_g observations. We denote the i^{th} data point of group g by x_{gi}, and the full dataset by $\mathbf{x}$. The data link variable c_{gi} for x_{gi} is sampled from a group-specific ddCRP:

$$p(c_{gi} = gj \mid \alpha_g, A^g) \propto \begin{cases} A^g_{ij} & i \neq j, \\ \alpha_g & i = j. \end{cases} \tag{2}$$

At this first level of link variables, we set the probability of linking observations in different groups to zero. The connected components of the links $c_g = \{c_{gi} \mid i = 1, \ldots, N_g\}$ then determine the local clustering for group g.

Data links $\mathbf{c} = \{c_1, \ldots, c_G\}$ across all groups divide the dataset into group-specific local clusters $T(\mathbf{c})$. The hddCRP then associates each cluster $t \in T(\mathbf{c})$ with a cluster link k_t drawn from a global ddCRP distribution:

$$p(k_t = s \mid \alpha_0, A^0(\mathbf{c})) \propto \begin{cases} A^0_{ts}(\mathbf{c}) & t \neq s, \\ \alpha_0 & t = s. \end{cases} \tag{3}$$

Here α_0 is a global self-affinity parameter, and $A^0(\mathbf{c})$ is the set of pairwise affinities between the elements of $T(\mathbf{c})$. We let $A^0_{ts}(\mathbf{c}) = f_0(d_0(t, s, \mathbf{c}))$, where $d_0(t, s, \mathbf{c})$ is a "distance" based on arbitrary properties of clusters t and s, and $f_0(d_0)$ a decreasing decay function. The connected components of $\mathbf{k} = \{k_t \mid t \in T(\mathbf{c})\}$ then couple local clusters into global components shared across groups. Let z_{gi} denote the component associated with observation i in group g, and $\mathbf{z} = \{z_{gi} \mid g = 1, \ldots, G; i = 1, \ldots, N_g\}$. Data instances x_{gi} and x_{hj} are clustered ($z_{gi} = z_{hj}$) if and only if they are reachable via some combination of data and cluster links.

Given this partition structure, we endow component m with likelihood parameters $\phi_m \sim G_0(\lambda)$, and generate observations $x_{gi} \sim p(x_{gi} \mid \phi_{z_{gi}})$. Let $M(\mathbf{c}, \mathbf{k})$ equal the number of global components induced by the cluster links $\mathbf{k}$ and data links $\mathbf{c}$. Because data links $\mathbf{c}$ are conditionally independent given $A^{1:G}$, and cluster links $\mathbf{k}$ are conditionally independent given $\mathbf{c}$ and the cluster affinities $A^0(\mathbf{c})$, the hddCRP