flexibility by analyzing a set of *networked documents* with a distance dependent CRP mixture model. Networked data induces an entirely different distance function, where any data point may link to an arbitrary set of other data. We emphasize that we can use the same Gibbs sampling algorithms for both the sequential and networked settings.

Specifically, we analyzed the CORA data set, a collection of Computer Science abstracts that are connected if one paper cites the other (McCallum et al., 2000). One natural distance function is the number of connections between data (and ∞ if two data points are not reachable from each other). We use the window decay function with parameter 1, enforcing that a customer can only link to itself or to another customer that refers to an immediately connected document. We treat the graph as undirected.

Figure 6 shows a subset of the MAP estimate of the clustering under these assumptions. Note that the clusters form connected groups of documents, though several clusters are possible within a large connected group. Traditional CRP clustering does not lean towards such solutions. Overall, the distance dependent CRP provides a better model. The log Bayes factor is 13,062, strongly in favor of the distance dependent CRP, although we emphasize that much of this improvement may occur simply because the distance dependent CRP avoids clustering abstracts from unconnected components of the network. Further analysis is needed to understand the abilities of the distance dependent CRP beyond those of simpler network-aware clustering schemes.

We emphasize that this analysis is meant to be a proof of concept to demonstrate the flexibility of distance dependent CRP mixtures. Many modeling choices can be explored, including longer windows in the decay function and treating the graph as a directed graph. A similar modeling set-up could be used to analyze spatial data, where distances are natural to compute, or images (e.g., for image segmentation), where distances might be the Manhattan distance between pixels.

5.4 Comparison to the traditional Gibbs sampler

The distance dependent CRP can express a number of flexible models. However, as we describe in Section 2, it can also re-express the traditional CRP. In the mixture model setting, the Gibbs sampler of Section 3 thus provides an alternative algorithm for approximate posterior inference in DP mixtures. We compare this Gibbs sampler to the widely used collapsed Gibbs sampler for DP mixtures, i.e., Algorithm 3 from Neal (2000), which is applicable when the base measure G_0 is conjugate to the data generating distribution.

The Gibbs sampler for the distance dependent CRP iteratively samples the customer assignment of each data point, while the collapsed Gibbs sampler iteratively samples the cluster assignment of each data point. The practical difference between the two algorithms is that the distance dependent CRP based sampler can change several customers' cluster assignments via a single customer assignment. This allows for larger moves in the state space of the posterior and, we will see below, faster mixing of the sampler.

Moreover, the computational complexity of the two samplers is the same. Both require computing the change in likelihood of adding or removing either a set of points (in the distance dependent CRP case) or a single point (in the traditional CRP case) to each cluster. Whether adding or removing one