Matt Dickenson `mcd31`
STA571/CS590.01 Due: 5 February, 2014

This paper introduces an algorithm for Bayesian hierarchical clustering based on comparing marginal likelihoods, rather than distance-based methods. Starting with a set of trees each containing data sets of size one (i.e. a single data point), the model evaluates the likelihood that any candidate pair of trees contain data generated (iid) from the same model. (The exact form of that model, e.g. Gaussian or multinomial, can be determined by the researcher with a great deal of flexibility.) This hypothesis is compared with the likelihood that the candidate data set contains two or more clusters. Here a potential problem arises: there are exponentially many possible ways of partitioning the data into clusters. To simplify this, Heller and G limit the possibilities to "tree-consistent" clusters.

This model has several advantages, one of which has already been mentioned: the flexibility to choose different models for the data likelihood based on domain knowledge (and is fast when that model is conjugate and does not require sampling). The model also defines predictive distributions for new data points (not a hierarchical generative model). A third benefit is that the choice of the tree structure is based on Bayesian model comparison, rather than an arbitrary distance measure.

There are also some disadvantages associated with this model. First, if the assumed form of the generative model (eg Gaussian) is incorrect, the clustering will be poor. For example, average linkage outperforms BHC on the Glass data in the paper. Another limitation is the greediness of the algorithm, which could lead to getting stuck in local optima. Although faster than sampling-based clustering procedures, the model's computational complexity is quadratic in n. Finally, by directly comparing likelihoods as discussed above, this algorithm ignores uncertainty in the tree construction.