Benjamin Limoges
Assignment 10

1. One important domain where pairwise instance-level constraints would be important would be porting any of our clustering ideas to time series. We haven't dealt with time series in this course, but if we were interested in classifying financial securities as growth versus value it would be very important to keep all of the time elements of one asset (Salesforce.com for instance) together. If securities do not switch in their lifetime from one type to another, then it is important to keep 2009 Salesforce.com and 2013 Salesforce.com together.

   Another domain would be in computational biology. For certain characteristics in populations there need to be certain combinations of genes that are turned on. Must link constraints could help study other effects of gene pairings by mandating that these sets of genes be in the same cluster (otherwise desired effects may not be seen - if any one of the genes isn't active then the whole effect disappears).

2. It is important to see how the algorithm deals with random constraints being imposed on it for two reasons. One is that it gives a benchmark on how the algorithm performs. If we assume that domain knowledge of constraints only improves clustering accuracy, then the random constraints merely shows us a sort of "worst case" scenario for the algorithm as the number of constraints is increased.

   Another reason the random constraints is important is that it brings no domain knowledge into play. It can be seen as a check to make sure that the algorithm works rather than clever constraints that makes COP-KMEANS converge.

3. Even with the change to the horizontal line, the unrestricted k-means lacks the two fundamental constraints that are necessary to create the lanes. Unrestricted k-means does not have to include all of one trace into a single lane; the COP-KMEANS algorithm assures that all of one trace is in a single lane since the data was preprocessed to make sure that the vehicles did not change lanes. Even with the restriction, k-means still prefers roughly hyperspherical clusters.

   The other crucial constraint that is not binding in k-means is that data more than 4 meters apart are not forced to be other lanes. This means that the k-means clustering can span over multiple real world lanes. Since we're trying to cluster by these real world lanes, the k-means output is not good data.

4. To make COP-MEANS the same as AGGLOM, we get rid of the progression through space, and instead collapse it to one dimension (the horizontal axis that spans across the lane). So instead of seeing the jitter of the car as it travels down the road, we just look at the x-axis jitter. This is just a simple projection from a 2 dimensional problem to a 1

dimensional problem.

Agglom also does not have the same constraint set that COP-KMEANS does. The Agglom algorithm does not have the 4 meter rule, which places all points more than 4 meters apart into different clusters. Instead it changes its terminating condition to incorporate this information; Agglom stops when “the two closest clusters were more than a given distance apart.” This threshold is presumably 4 meters. If we shoehorn the COP-KMEANS to eliminate the 4 meter constraint and instead stop when the Aggglom algorithm stops, then the two problems become identical.

5. One way that we could get around this would be to relax the definition of the constraints. For instance in the problem above, the trace constraint needs to be hard fixed since the same car will occupy the same lane for the experiment, but the 4 meters constraint does not need to be hard fixed. Around interchanges and construction the lanes may narrow or widen - the hard fixed 4 meter rule may fail in these cases (and therefore the clustering would return {}). You could get around this in a couple different ways. One way was described in question 4 (changing the threshold and setting the problem up as agglom problem). The authors claim though that this isn’t particularly good in non-normal settings (interchanges for instance.)

   Another way would be to soften this constraint to be a probabilistic constraint. This would in effect change the COP-KMEANS problem into more of an expectation maximization problem - two points that are 4 or more meters apart are most probably in different lanes, but not definitely. This approach would allow for there to be a little bit more flexibility in these circumstances.