Unsupervised Learning Final Project Notes

what the algorithm does now

this only works for data where there’s an adjacency matrix. (D not in algo)

* TO DO: Expand to graphs that don’t have an adjacency matrix.

until every point burnt:

chooses a point to start burning

starts burning at that point

until no more points burning:

check if each point is hot enough to start burning

if it is, set it on fire

how to choose a point to start burning (“flint”):

* if first fire, just choose one randomly
* else: for each point, calculate Euclidean distance between adjacencies for each point & each centroid (previous ignition point)
  + idea: we deem a point close to another if they share a lot of neighbors. Calculating the Euclidean distance is an okay way of doing this.
    - TO DO: use Smita’s “Potential Distance” metric from PHATE instead
  + Take minimum distance
  + Sample a point to be the new ignition point using the distances as probabilities
    - IDEA: if we want to have a fixed number of clusters k, initially sample k points to burn at, then burn all at once. Then, we can simulate this multiple times using the same initializations to calculate entropy scores for each point.

Determine whether or not to set a point on fire based on whether its head > its threshold

* Heat: fire\_temp \* proportion of points burning in this cluster that are neighbors of this point
  + If this point is likely in the cluster, then it will have neighbors in the cluster. And if a good chunk of the points on fire are neighbors of this point, then it’s probably in the cluster too.
* Threshold: flash\_point / degree. A point has more neighbors → need lower proportion of burning neighbors to ignite.

Evaluation right now is all pictorial. How to make quantitative evaluation:

* If points that are closer are likely to be neighbors (like the SBM), then silhouette score will be a good choice
  + This won’t use the actual cluster labels
* If we have the true cluster labels, want to match labels up (e.g. 1-->4, 3-->2, etc). This is a permutation mapping.
  + DONE. Only works for k>=10 if Forest Fire does a very good job.