Task 2

In this task, we are trying to divide the nodes into 10 groups each containing the movies, persons, and places of 10 consecutive years. This is a difficult task as this is not a natural division which separate nodes with certain equivalency. Therefore, we decided to use random walk algorithm to solve this problem. The steps are described below.

1. Starting at every time node, take 10 steps randomly. Repeat 10000 times
2. Calculate intersection rate for every time node pairs, i and j.

Where v(i) is the visited nodes list of random walk algorithm starting from time node i.

1. Starting from the node pairs with the highest intersection rate, assign the node pairs to the clusters with following rules.

* If node iis assigned a cluster already, node jwill be assigned to the same cluster.
* If neither nodes are assigned, calculate the average intersection rate of node i and node j with nodes in existing clusters, C. Assign the nodes to cluster with highest intersection rate.

Where n is the number of nodes in cluster C.

* If a cluster has more than 10 nodes, calculate the intersection rate of all nodes with the rest of the nodes. Nodes with smallest intersection rates will be removed from the cluster and reassigned to other clusters.
* Repeat steps above until all nodes are assigned.

1. Add neighbors 2 levels away from the time node to the clusters.

After running the algorithm, we have noticed that at the end there are a few hundred nodes that do not get assigned to any cluster. The list of nodes that are left out is consistent from run to run. We believe it is due to faulty time node assignments. Also, the clusters do not look the same from run to run. This is possible due to the random nature of the random walk algorithm.

With more accurate time nodes classification from task A, we believe we can achieve better NMI.