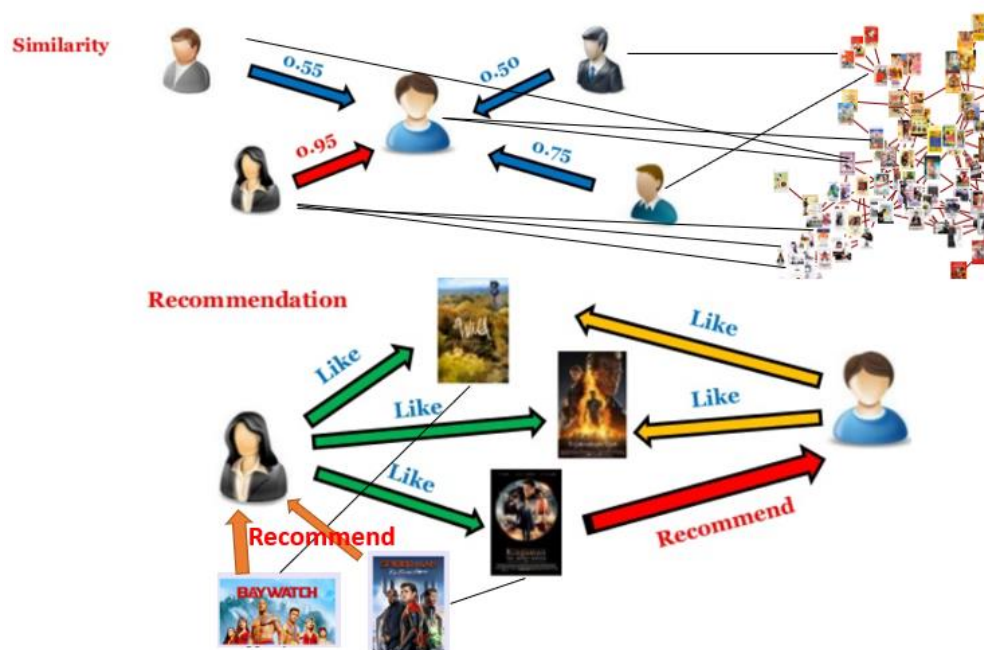


Introduction: Link Prediction is a well-known problem in different area of science. Link prediction means: Given a snapshot of a network (graph), can one predict the next most likely links to form in the network? For example, the friendship suggestion in Facebook, item purchase suggestion in Amazon, movie suggestion in Netflix are all different applications of link prediction. For example, in Netflix, we have a graph in which nodes present the users and movies. The links between movies present the similarity between movies (genre, actress, actor, ...), and the links between used represent the similarity between users (age, location, gender, ...). Links between movies and users shows if the user watched or liked the movie. Given the similarities between users and movies, the link prediction algorithms recommend movies to users (predict link between movie and users).



Problem. We want to apply the same concept in biology data, where the nodes are phosphorylated sites (phosphosite) and enzymes. The edges between phosphorylated sites represents biological similarity between them (co-evolution, same pattern of phosphorylation across cancer samples, ...) and edges between enzymes shows if they are in the same biological pathways (which means their biological functions are similar). The links between enzymes and phosphosites shows that enzyme is responsible to phosphorylate that phosphosite. The goal is using these information and link prediction methods to predict new edge between enzymes and phosphosites.

Approach. Split the known edges between enzymes and phosphosites into a training set and a test set. Find good link prediction algorithms. Run the algorithm on the training set, and test it on the test set. Check the accuracy. That would be great if you compare different link prediction algorithms.

Data. You can find the data at <https://github.com/msayati/HackResearch2019>. Each row in the files represents an edge.