Shalini Kurian
Serene Kosaraju
Zavain Dar

Supervised Random Walks on Homogeneous Projections of Heterogeneous Networks

Introduction:

Most of the complex systems in the real world can be represented as networks that define interactions between the various components. Examples are the communication networks such as the graphs of the Internet, Information networks such as the connections between political blogs, road networks, organizational and economic networks along with biological networks and the recently introduced, widely popular social networks such as Facebook, Twitter,etc. One interesting observation of most of these networks that we come across in the real world is that they have entities of different kinds and they can't be categorized into a specific unique type. Thus most of these real world networks call for the analysis of a new breed of 'heterogeneous' networks.

The problem of prediction of links and recommendation is an important problem in most of these networks especially in social and co-authorship networks. We define the Link Prediction Problem (LPP) as follows: Given a snapshot of a dynamic network G at some time t0, how can we best determine how G will evolve to be at time t1 using only data available from G at t0. Although this problem has been extensively studied, the challenge of how to effectively combine network structure information with node and edge attribute data to make good predictions still remains largely open. Literature [1] presents a study of modeling link prediction for a heterogeneous network based on topological features of a network such as normalized path count and random walks by empirical testing on the Collaboration of Computer Scientists (DBLP) dataset. We leverage a variant of both the Supervised Random Walks algorithm[3], an algorithm that naturally combines information from the aggregate network structure with edge and node level attributes, and the Meta-Path based topology as presented in [4], to model the link prediction problem. Concretely we encapsulate information from the heterogeneous structure of graphs using meta paths topolgoy and test it on the DBLP dataset using our redesign of the Supervised Random Walks algorithm. We choose the DBLP simple because it serves as a non-trivial and widely available specimen to study heterogeneous networks.

We test a novel algorithmic framework to improve the current recommendation/link prediction algorithms so as to generate effective recommendations of link formation for arbitrary data sets based on network and other structural properties which is a very important problem in most of the commercial networks such as books (Amazon), movies(Netflix) and music (Spotify) industries in addition to the most popular problem of friend prediction/recommendation in social networks such as Facebook. As networks evolve and grow through the addition of new edges, the link prediction problem offers insights into the factors behind creation of individual edges and into network formation in general. Moreover, the link-prediction and the link-recommendation problems are relevant to a number of interesting current applications of social networks. First, for online social networking websites, like Facebook and Myspace, being able to predict future interactions has a direct impact on their businesses. Other large organizations always desire a promotion of interactions within the company wide (intra) social network and link-prediction methods can be used to suggest such possible collaborations and interactions. Outside the model of relevance of links to predict future collaborations , link prediction has a myriad of applications. For instance, research in security has recently recognized the role of social network analysis for this domain (e.g., terrorist networks). In this context link prediction can be used to suggest the most likely links that may form in the future. Similarly, link prediction can also be used for prediction of missing or unobserved links in networks [9] or to suggest which individuals may be working together even though their interaction has yet been directly observed. Link Prediction techniques can be used to predict unobserved links in protein-protein interaction networks in systems biology or give suggestions to bloggers about which relevant pages on the Web to link to.

Prior Work:

For a proper and contextual understanding of the current problem formulation and status we found the following papers and results to be most relevant.

1a. The anatomy of a large-scale hypertextual web search engine [1]

The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International World Wide Web Conference, 1998 PageRank was introduced in [1] to improve the quality of search results by using random walks to leverage the link structure of the world wide web. In the original formulation, only one PageRank vector was calculated for the whole web, which meant that each page had a single score corresponding to its significance in the web as a whole. The work of [2] aimed to improve search results even further, by recognizing that topics were an important factor in determining a web page's importance For instance, a popular pet website with a high PageRank would be all but useless to someone researching an electronics purchase. Hence, in [2] multiple PageRank vectors are computed each biased to give more importance to a specific topic, in order to generate query specific importance scores for pages.

While initial PageRank algorithms were very successful in web information ranking, the success of using PageRank as a feature in other network tasks has yielded mixed results. One of the main reasons for this is that if one fixes a network, the PageRank or even the topic-sensitive PageRank is fixed regardless of what prediction task is underway. It will be a good node feature for some tasks and not so great for others. For our project, we used PageRank as a metric for predicting links from a source node s to destination nodes which are ranked based on their page rank scores in order to infer which nodes are the most likely ones that the current source node would connect to in the future. The page rank scores are calculated as a linear combination of features that are considered to contribute maximally to link formation between the two nodes.

1b. Supervised Random Walks: Predicting and Recommending Links in Social Networks [3]

Recent work from Backstrom and Leskovec [3] has changed the above discussed PageRank scenario, allowing traditional random walks done in PageRank to be biased towards a broader class of predictive tasks on networks. In particular, the authors use what they call Supervised Random Walks for solving the link prediction problem in networks. Their approach uses supervised learning to learn a random walk transition probability on each edge from a set of both topological and categorical features in order to guide the random walk to parts of the network where links are likely to form. This algorithm naturally combines the information from the network structure with node and edge level attributes and uses these attributes to guide a random walk on the graph. The supervised learning task has the goal to learn a function that assigns strengths to edges in the network such that a random walker is more likely to visit the nodes to which new links will be created in the future. We develop an efficient training algorithm to directly learn the edge strength estimation function. This supervised random walk could further be used as a method for predicting links in the coauthor network as it demonstrates a good generalization and overall performance of supervised random walks. This is a good algorithm to follow for our problem as it automatically takes into consideration the network structure, node and edge attribute information through the page rank scores and requires no network feature generation separately.Supervised Random Walks are not limited to link prediction, and can be applied to many other problems that require learning to rank nodes in a graph, like recommendations, anomaly detection, missing link, and expertise search and ranking.

As evidenced by the immediate success of Google over its competitors, the PageRank idea had many strengths. By using the link information in the web, the authors leveraged the trust that is implied when one page links to another. Also, since page rank uses only the information contained in the link structure, the algorithm is very fast but it is more inclined to old pages than new ones and hence time-activity information should be integrated into the PageRank scores along with users behavior.The random surfer model at the core of PageRank, however, does make some unrealistic assumptions. In particular, this random surfer has no agenda unlike an actual user of the web. What this means is that the PageRank score for a page is an unweighted average over all possible browsing agendas.

1c. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks [4]

The problem of predicting links or interactions between objects in a network, is an important task in network analysis. Along this line, link prediction between co-authors in a co-author network is a frequently studied problem. In most of these studies, authors are considered in a homogeneous network, i.e., only one type of objects (author type) and one type of links (co-authorship) exist in the network. However, in a real bibliographic network, there are multiple types of objects(e.g., venues, topics, papers) and multiple types of links among these objects. This paper studies the problem of co-author relationship prediction in the heterogeneous bibliographic network, by systematically extracting topological features from the network. A supervised model is used to learn the best weights associated with different topological features in deciding the co-author relationships and experiments are presented on a real bibliographic network, the DBLP network, which show that heterogeneous topological features can generate more accurate prediction results as compared to homogeneous topological features. In addition, the level of significance of each topological feature can be learned from the model, which is helpful in understanding the mechanism behind the relationship building.

This paper used meta-paths based topology and measured functions on meta paths such as path count, normalized path count, random walks and symmetric random walk. In our project, we used meta paths in order to encapsulate the heterogeneous structure of the DBLP coauthorship network and we used path-based features such as shortest path count and edge betweenness centrality of two kinds of meta paths namely author -article-author and author-article-venue-article-author. We do not consider the other meta-paths discussed in this paper in our project as the DBLP database did not have a significant number of meta-paths of other kinds and hence learning features based on those did not contribute much in arriving at page rank scores that were good indicators of link prediction.

But it has been shown by our experiments that by considering heterogeneous topological features, the relationship prediction accuracy can be significantly improved, and the model using hybrid features that have combined different meta paths and different measures gives the best overall performance. Furthermore, the learned significance for each topological feature can provide better understanding of the relationship building mechanism in such networks.

1d. Multi-Relational Link Prediction in Heterogeneous Information Networks

In this paper, a novel probabilistically weighted extension of the Adamic/Adar measure for heterogenous information networks has been introduced called the Multi-Relational Link Prediction(MRLP) which is only one of many ways to formulate or extend a topological link predictor for multi-relational data, and is used to demonstrate the potential benefits of diverse evidence, particularly in cases where homogeneous relationships are very sparse. Most significant real-world networks, modeled naturally as complex networks, have heterogeneous interactions and complicated dependency structures. Thus link prediction in such networks must model the influences between heterogenous relationships and distinguish the formation mechanisms of each link type, a task which is beyond the simple topological features commonly used to score potential links. In accordance with previous research on homogeneous networks, it was demonstrated that a supervised approach to link prediction can enhance performance and is easily extended to the heterogeneous case.Developing supervised methods and features which efficiently capture the richness of heterogeneous information networks is perhaps the next logical step from the observations in this paper.

In our project, we incorporated the essence of this paper by extending link prediction to the heterogeneous case in DBLP databases. We also tried to run an unsupervised K-means algorithm to learn the types of the entities in the network and thereafter performing homogeneous link prediction on particular types of entities/nodes. Here we have an incentive to integrate heterogeneous data as homogeneous data (for example, the venue nodes) are sparse in our network.

2.Extraction of data:

The DBLP dataset which is being used for the purposes of this project is referenced from http://dblp.uni-trier.de/xml/ . The data is in XML format .Looking at the dtd and our needs, we are most interested in the elements ‘article’ , ‘author’ and ‘journal’ . These form the heterogeneous nodes of our network.

After parsing the xml we have the following nodes and attributes:

Node	Attribute 1	Attribute 2	Attribute 3

article	type	title	year
author	type	name	
venue	type	venueName	

As seen , the attribute ‘type’ associated with each node identifies the type of the node in the heterogeneous dblp network. The three types are ‘author’, ‘article’ and ‘venue’.

We preprocess the data set and divide it into sub graphs to serve as part of our training stage and testing set. Time stamps till 2007 are being used as the time stamp range in our training set and time stamps 2007 onwards (inclusive of 2007) are being used as the time stamp range in our testing stage.

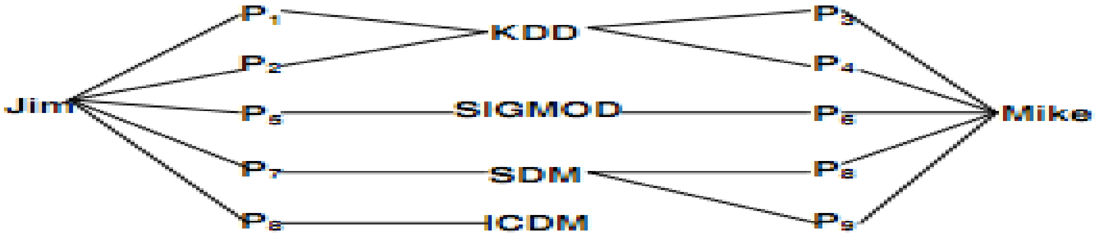
3.Model/Algorithm/Method:

We first describe our use Meta Path Based Topological [4] features in utilizing DBLP’s heterogeneous format. We then shift our discussion to our variant of Random Supervised Walks for our designed learning platform.

Topological Features in Heterogeneous Networks

The general norm of topological features is to have characteristics targeted at deducing connectivity based interactions between nodes of a network. Certain topological features are structured to serve as link prediction features as they leverage a network’s present structure to form deductions about the same network’s future structure. For a heterogeneous networks like the DBLP network under consideration , we resolve to follow a similar structure as presented in [4] , which uses meta paths to define the topology between objects(nodes) in the network and then define measurement metrics on the topologies. For the purposes of our algorithm , we consider the following two meta paths:

Meta Path	Semantic Relationship
Auth -> A - > Auth	two authors co-authored a paper
Auth - > A -> V -> A ->Auth	two authors published in the same venue



An Example of A-P-V -P-A Paths Between Two Authors

On these given meta paths, we use four linkage features out of which two are node based and two are
path based features.

- **Normalised Common Neighbors :** In a graph , the number of common neighbors of two objects (nodes) of x and y $\Gamma(x)I\Gamma(y)$ potentially represents the similarity between those two objects. This serves as an incentive to lean towards using common neighbors to predict future links between two objects in a network .
- **Normalised Jaccard's Coefficient :** Jaccard similarity coefficient is a statistic used to measure the similarity and diversity between two objects . Mathematically it is formulated as the intersection divided by the union of the sample sets under observation.

  $$J(A, B) = \frac{|\, A \; intesect \; B \,|}{A \bigcup B}$$
- **Normalised shortest path :** Normalised shortest path is essentially a measure of the closeness of two objects in a network and is used as a link prediction features.
- **Normalised edge betweenness centrality :** Following the normalised shortest path feature, the choice of normalised egde betweenness centrality was a natural metric to serve as a link prediction feature . The betweenness centrality for an edge e is the sum of the fraction of all-pairs shortest paths that pass through e .

  $$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)}$$

Random Walks Learning Platform

In designing our learning algorithm we were heavily influenced by the Supervised Random Walks algorithm as presented in [2]. As such, we suggest viewing our learning process as a variant of the original Supervised Random Walks algorithm. Crucial in understanding the allure of Supervised Random Walks is realizing that the algorithm is not a supervised learning algorithm per say, but rather more akin to a reinforcement learning algorithm. Concretely, in the training stage the algorithm is not given a set of positive and negative examples as popularly conceived in the literature [4]. To be fully clear, most popular link prediction algorithms are given two instances of a graphs for training, one at time t_0 and one at time t_1. The rational being that the learning mechanism can see how edges evolve going from time t_0 to time t_1 of the network and the prediction phase take an instance of the network at time t_2 and predictions the evolution of the graph from time t_2 to time t_3 . Explicitly, such leaning algorithms train on two instances of a graph in order to predict the future given one instance of the same graph. This differs from Supervised Random Walks which trains on a single snap shot of a graph at time t_0

and immediately tests on its predictions for new links from a source node s at time t_1. We find this to be a very appealing property of Supervised Random Walks as it inherently requires less training data to be able to make the same types and quality of predictions. Moreover, this sort of learning seems more applicable to real life situations where we may not always have access to a graph over time from which to learn, but rather only an instantaneous snap shot from which we have to make predictions.

However, our algorithm is not Supervised Random Walks as explicated in [4], but rather a simplified variant of Supervised Random Walks. We decided to use this variant for a number of reasons including ease of implementation and it's ability to loosely proxy results of the actual Superwised Random Walks algorithm. Note that we decided to use this alternate implementation only after realizing the intractable nature of attempting to explicitly run the original Supervised Random Walks algorithm on DBLP which consists of over 1,300K nodes and 2,800K edges. Our algorithm is defined only after formalizing the same constraint satisfaction problem as solved for by Supervised Random Walks.

The constraint satisfaction problem is as follows. Given a graph G at time t_0, a source node: *s*, a set of neighbors of *s*: D, and the remaining nodes that s does not connect to: L, we aim to generate a function f: edges →[?][?][?] such that we can use this function to generate, for the particular source node *s*, a page rank vector *p*, such that for any *l* in L and any *d* in D, *p[d] > p[l]* . We then predict the highest scoring nodes in *p* that *s* is not already connected to at time t_0 and predict those to be the new nodes which *s* connects to at the next time step t_1. Our algorithm differs from the initial Supervised Random Walks mainly as follows: Supervised Random Walks first iteratively solves for an optimal *p* vector then solves for the optimal weights with which to scale the features. Conversely, our algorithm assigns equal weighting to each of the features and only then solves for the optimal page rank vector to satisfy the initial constraint.

We use a convergence algorithm to solve for our optimal page rank vector p as follows:

```
1) Initialize matrix Q such that Q[u][v] = featurescores (u,v)/ sum_over_j(featurescores(u, j))
2) for each u,j:
        Q[u][j] = (1 - alpha)Q[u][j] +alpha 1{j == s}, where alpha is our restart probability for
the random walks
3) For each node j in V, initialize p^0[j] = 1/|V|
4) While not converged:
        for each node j in V
                p^t[j] = sum_over_u(Q[u][j]*p^(t-1)[u])
```

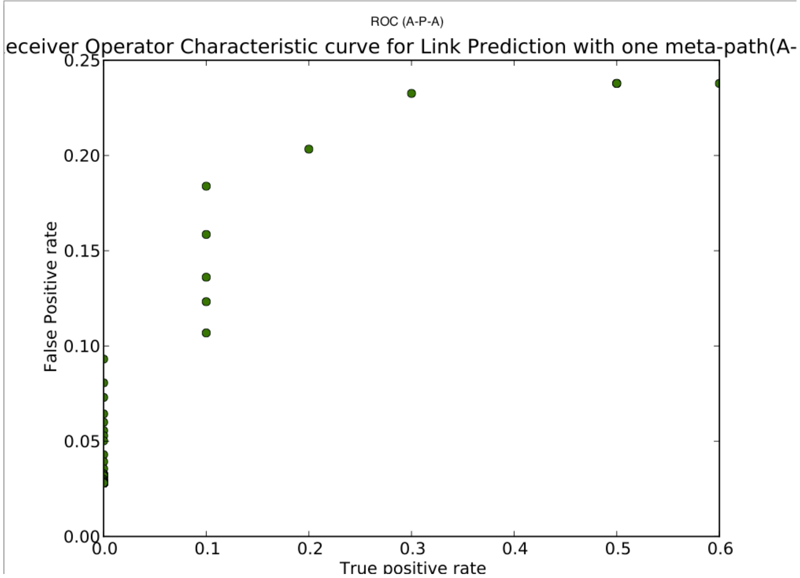
Only after p converges can we make predictions as to the new neighbors of s.

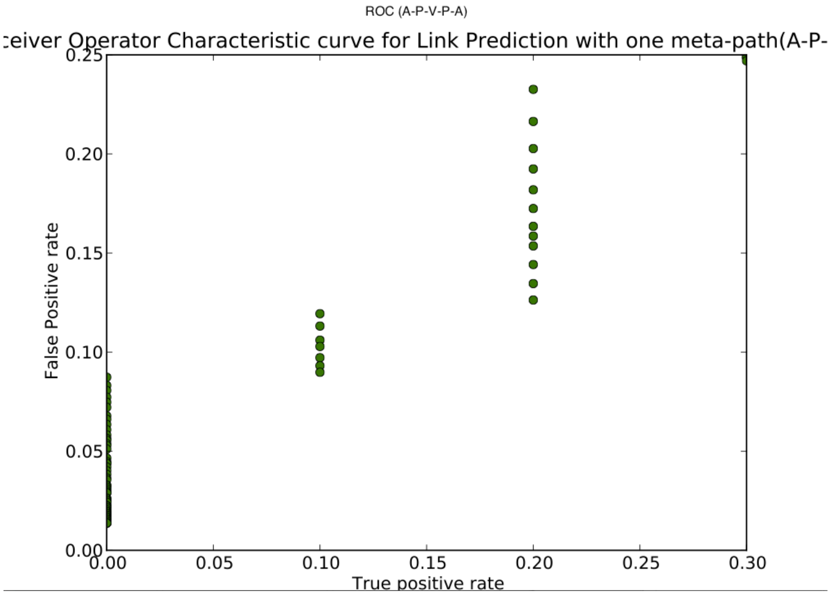
Experimental Setup

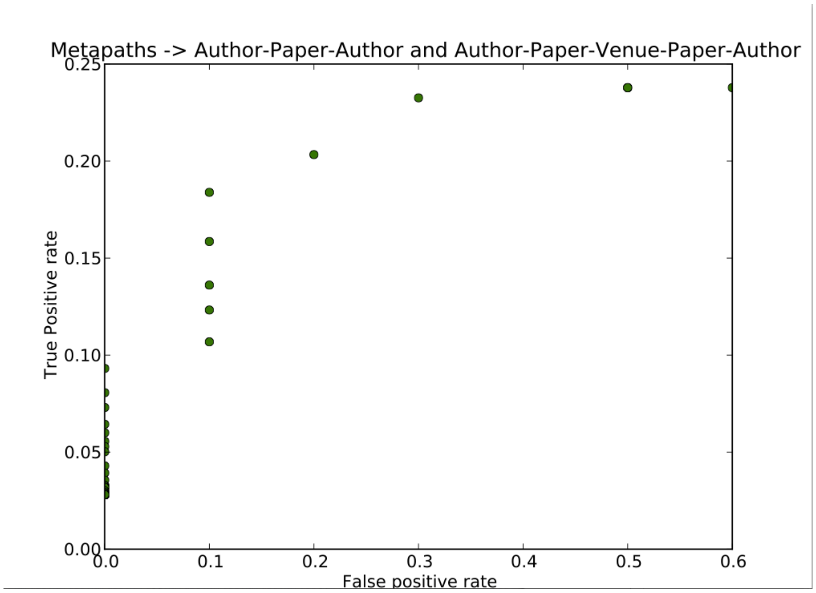
While the initial DBLP graph had over 1,300K nodes and 2,800K edges we quickly realized, given our limited computing power on available machines, that we would have to run experiments on much smaller sub graphs. As such we ran our link prediction algorithm on an arbitrary sub graph with 2K edges. Given our subgraph of size 2000 we generated two types of edges: author - paper - author type edges, and author - paper - venue - paper - author type edges.

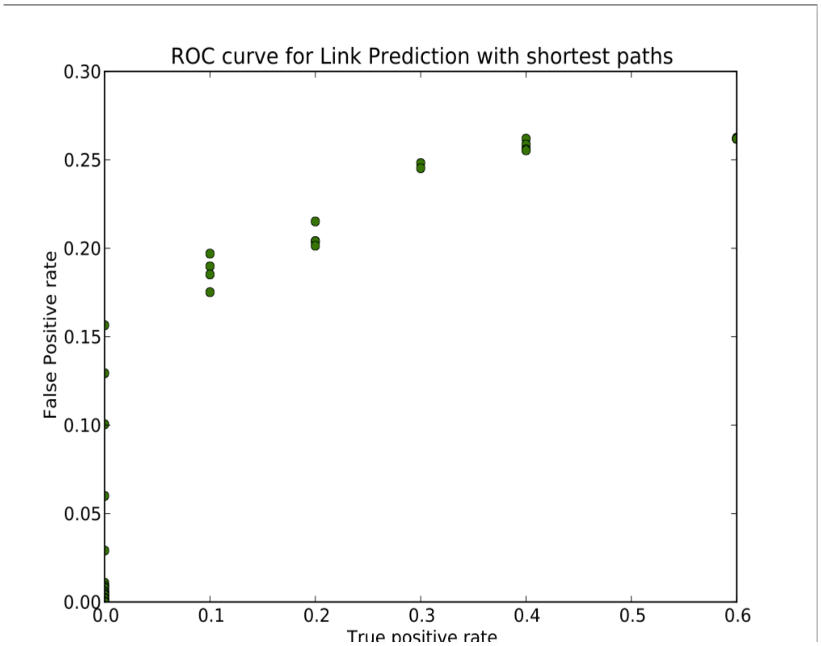
For each of these type of edges we calculated the following four topological features: common neighbors, jaccard's coefficient, edge betweenness centrality, shortest path length. Only once this had been done could we run our proxy algorithm to approximate Supervised Random Walks on the graph for particular source nodes. We chose the network before 2007 to train on and the network during or after 2007 to test on as empirically this provided the most even split of number of edges.

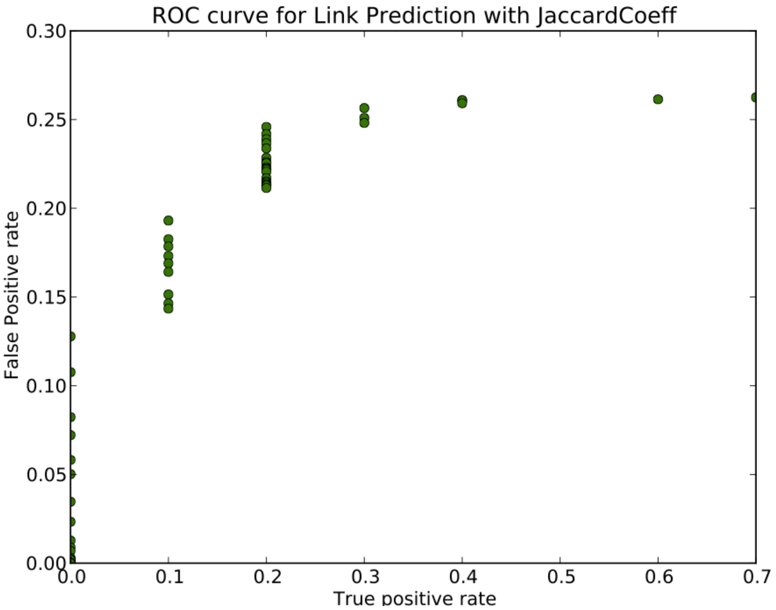
Results and Findings:

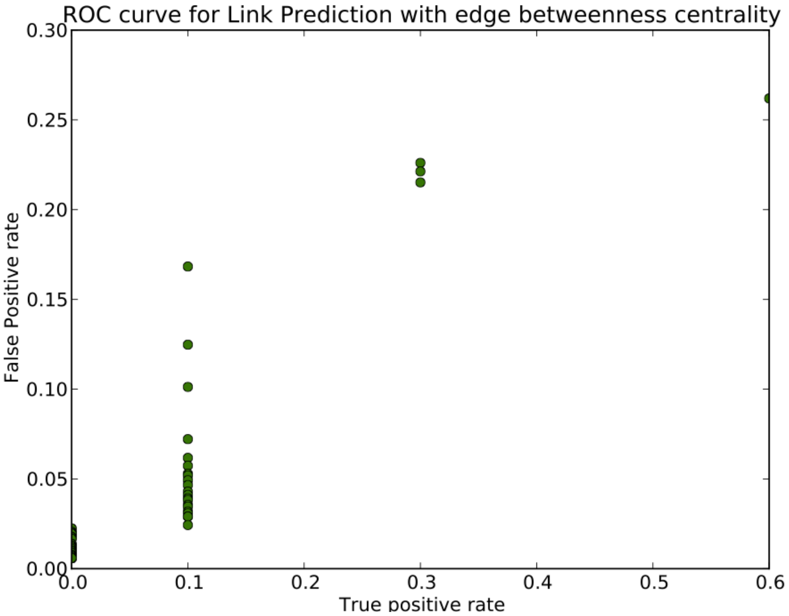


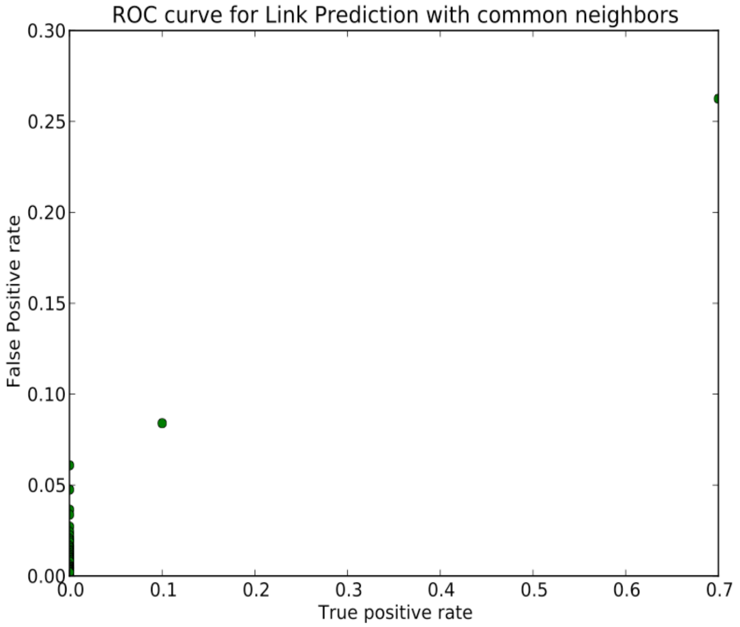



Metapaths -> Author-Paper-Author and Author-Paper-Venue-Paper-Author
True Positive rate
False positive rate
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.1
0.2
0.3
0.4
0.5
0.6


ROC curve for Link Prediction with shortest paths
False Positive rate
True positive rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.1
0.2
0.3
0.4
0.5
0.6


ROC curve for Link Prediction with JaccardCoeff
False Positive rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
True positive rate


ROC curve for Link Prediction with edge betweenness centrality
False Positive rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.1
0.2
0.3
0.4
0.5
0.6
True positive rate



We plotted Receiver Operating Characteristics graphs when the data that we had trained on had the following three different combinations of metapaths:
1. The meta-path set consisted of only Author-Article-Article links.
2. The meta-path set consisted of only Author-Article-Venue-Article-Author links.
3. The meta-path set consisted of a linear combination of Author-Article-Author and Author-Article-Venue-Article-Author links.

In signal detection theory, a receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the sensitivity, or true positive rate, vs. false positive rate, for a binary classifier system as its discrimination threshold is varied. The ROC can also be represented equivalently by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate). Also known as a Relative Operating Characteristic curve, because it is a comparison of two operating characteristics (TPR & FPR) as the criterion changes. The true positive rate represents the ratio of the number of links that have been predicted to have links and truly have do links to the total number of links that a particular node has - it is representative of the recall rate of the particular node. The false postivie rate represents the ratio of the number of nodes that are predicted to have links with the given node but which do not in reality have links to the node(False positives) to the total number of nodes that a node does not connect to(False Positive + True Negative). In short, TPR = TP/(TP+FN) and FPR = FP/(FP+TN)

Our ROC curves show that some metapaths are more important than others in link prediction of the source node under consideration. We found that Author-Article-Venue-Article-Author is a much weaker meta-path compared to Author-Article-Author metapath for predicting links in the heterogeneous network due to the much weaker ROC curve as the Area under the ROC curve is representative of the strength of accurate link prediction. This adheres to our initial intuitions as the link types that we are hoping to predict are exactly those of type Author-Paper-Author. Certainly there’s nothing surprising that in predicting the formation of a new type of link the strongest indicator will be the preexisting links of the same type.

Also, we plotted the ROC curves when only one feature is considered individually such as Jaccard coefficient, common neighbors, edge betweenness centrality and length of shortest paths with the linear combination of the two metapaths (A-P-A and A-P-V-P-A) and found that Jaccard Coefficient had a higher Area under the ROC curve as opposed to common neighbors which seems to have a much more ill defined area under the curve as most of the true postive rates are concentrated at zero when the threshold is varied and there are only two points(or thresholds) at which the true positive rate goes beyond zero. This was a little bit surprising. We hypothesize that this phenomena occurs due to the fact that the DBLP seems to follow a scale free exponential pattern and thus there are some authors that a disproportionally high number of authors co-author with.

Our analysis shows the effect of particular features and their contribution to the problem of link prediction and how these features when combined with the metapath based features of heterogeneous networks have a cumulative effect on the accuracy and recall rate of predictions. We used page rank scores which took into account the effect of the above four described features and applied a random walk algorithm with restarts from a particular seed node to calculate the page ranks of all the other nodes in the network with respect to the given seed node. These page ranks were predictive of the probability of link formation between the two nodes. We found that some of the features had a more favorable recall rate as compared to the others and thus arrived at a metric to determine the relative importance of various features for link prediction. Also we found that using a combination of all possible metapaths results in a better recall rate from the area under the ROC curves thus validating our earlier claim that incorporating heterogeneous features (in this case metapaths) results in better results for the problem of link recommendation.

Lastly, we hope to have taken a small step in pushing forward both the study of link prediction in networks and heterogeneous vs homogeneous graphs. In this particular study we explored homogeneous projections of heterogeneous maps for topological feature extraction with Meta Based Paths. We also worked with Supervised Random Walks via an invented proxy algorithm which we found to have not as good results but more efficient run time. Future research directions could be to work on the entire DBLP data set via a distributed high performance implementation of both the feature generation and the Supervised Random Walks algorithms. However, we hope that our smaller case studies shed reliable light on the outcome and potential of such future projects.

References

[1] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International World Wide Web Conference, 1998

[2] T. H. Haveliwala. Topic-sensitive PageRank. Proceedings of the 11th InternationalWorld Wide Web Conference (WWW), 2002.

[3] L.Backstrom and J.Leskovec.Supervised Random Walks: Predicting and Recommending Links in Social Networks. Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM), pages 635-644, 2011

[4] Yizhou Sun,Rick Barber, Manish Gupta,Charu C. Aggarwal,Jiawei Han.Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks.IBM T. J. Watson Research Center, Hawthorne, NY.

[5] Darcy Davis, Ryan Lichtenwalter, Nitesh V. Chawla. Multi-Relational Link Prediction in Heterogeneous Information Networks