An Integrated Link Prediction Model for Co-Authorship Networks

Sojeong Song
School of Electrical Engineering
KAIST
Daejeon, South Korea
akqjq4985@kaist.ac.kr

Kanghoon Yoon
School of Industrial and Systems Engineering
KAIST
Daejeon, South Korea
ykhoon08@kaist.ac.kr

Yeseul Choi
Graduate School of Future Strategy
KAIST
Daejeon, South Korea
ustii93@kaist.ac.kr

ABSTRACT

How can one accurately and efficiently predict a co-authorship network? This project proposes using an integrated methodology of preferential attachment and rejection sampling to create a prediction model for co-authorship networks. The results showed that our proposed model produced greater accurate prediction of links between authors in comparison to the traditional prediction models used in previous studies.

Keywords: co-authorship network, link prediction, hypergraph, scale-free network, preferential attachment, rejection sampling

1. INTRODUCTION

Research collaborations have been found to bring about innovative ideas and achieve groundbreaking scientific work.[4] Researchers are affiliated in social communities such as government, public and private institutions. They collaborate for many reasons, but mostly for the following: (1) to cooperate with productive researchers to secure their position within their respective scientific communities and (2) to help each other to achieve success on common topics of interest. However, it is not an easy task for researchers to build cooperative relationship within their filed or even when conducting interdisciplinary research. Then, how can one predict collaborative relationships in a network of researchers? In the current paper, we want to address the following questions: (1) How can one accurately predict a co-authorship network? (2) How can one efficiently predict a co-authorship network?

Research on link prediction models of various network models including co-authorship networks have been a popular topic of interest. IN previous studies, various types of similarity such as common neighbors, Jaccard's coefficient, Katz similarity index, and Adamic/Adar are used for high accuracy on link prediction. [2] It has also been found that link predictions are more accurate in a highly productive author group. Nevertheless, some have argued that out of all the traditional link prediction methods the Adamic/Adar method is the most accurate method.

While previous literature utilized traditional similiarity indexes and methodology, this project proposes a novel and more efficient algorithm of predicting co-authorship networks. We propose an integrated model of preferential attachment and rejection sampling for hypergraphs. In addition, data analysis on co-authorship networks using hypergraphs has been observed, but we want to find a more efficient algorithm. Moreover, we focus on preferential attachment (PA) because of the ability to calculate a global property. We decided to compensate the PA algorithm to create higher accuracy on link prediction by combining other local similarities together. The preferential attachment is a popular link prediction and has been modified to fit with different network topologies. As well, PA has been modelled for hypergraphs and scale-free networks. By combining this method with rejection sampling when sampling new nodes or edges, we can get a more accurate and efficient estimation of future co-authorship links using a novel algorithm. The next section reviews previous literature, and discusses our proposed model, the experiment results and conclusions.

2. LITERATURE REVIEW

2.1 Hypergraphs and Preferential Attachment

Hypergraphs are composed of edges that can be connected to any number of vertices. We suggest that hyperedges are more representative of collaborative networks as authors may collaborate on more than one paper with multiple authors. The first paper was the random PA hypergraph paper by Avin et al.(2015) [1]

- *Main idea*: Instead of using random PA on graphs, the researchers proposed non-uniform PA hypergraph evolution model based on the in-degree distribution of power law. The data was a co-authorship network of the computer science field extracted from DBLP. The algorithm choose randomly k-ary relation in each step of the graph evolution.
- *Use for our project*: It is extremely related to our integrated link prediction model, because we can use the model to evolve our hypergraph and learn the maximum likelihood parameters of the model on our

.

data. Instead of random sampling, rejection sampling is used.

- *Shortcomings*: Our data does not contain any time information, so using the model on simulation may lead to non-intuitive results. The model assume that the degree distribution of the hypergraph is a power law, which may not assure high accuracy on our data.

The second paper was the PA paradox by Sheridan and Onodera(2018) [5]

- *Main idea*: The authors point out the ARS citation network can be better described by a log-normal in-degree distribution, not a power law distribution. They defined the PA as the rate measured by corrected Newman's method, which increases linearly in k. Since the PA generates the networks of a power-law in-degree distribution, they argue using the PA evolution model on the citation network is a paradox. They solve this paradox by redefining and testing the attachment rate of the PA as a nonlinear function or log-linear function.
- *Use for our project*: It is highly related to our integrated link prediction model, because we can use the conclusion to generate more accurate hypergraphs. The attachment rate of PA may be a nonlinear function, and affect the hypergraph evolution.
- *Shortcomings*: This paper uses a citation network so the properties of hypergraphs may be different from our data. The paper does not suggest the distribution that describes the network best. There may be limitation of applying a nonlinear PA rate function.

2.2 Link Prediction

Link prediction in social networks is to calculate the probability of a new link based on the features of the networks. In a co-authorship network, link prediction has been studied continuously. The first paper was the PathPredict model by Sun et al.(2011) [6]

- *Main idea*: Instead of traditional meta-path based link prediction model, propose a new model, that is PathPredict, to improve the accuracy of link prediction considering up to 3-hop co-authors. They test it for four data sets divided by the productivity of authors in the DBLP heterogeneous bibliographic network. High productive author data set regarding 3-hop co-authors shows high accuracy of the link prediction
- *Use for our project*: It is extremely related to our link prediction model, because we will use local similarities that are used in this paper to determine if a sampled hyperedge is chosen when rejection sampling. The results support that our results are intuitive, because the accuracy is better when considering more co-authors in this paper. Moreover, we can compare the accuracy with this paper, since it only consider up to 3-hop co-authors, but our model consider the global features.
- *Shortcomings*: Although high productive authors have been shown that link prediction is more accurate, they do not control the number of data in both groups. Thus, the number of data can be affect to the accuracy, by not only productivity. Moreover, we use the local similarity on rejection sampling, not on directly link prediction.

The second paper was the link prediction in medical co-authorship networks by Yu et al.(2014) [7]

- *Main idea*: The paper tests two supervised classification models, that are LR and SVM, based on various structural topological features including preferential attachment in the biomedical research domain. The authors compared the accuracy of link prediction through the local similarities considering up to 2-hop co-authors.
- *Use for our project*: This research is extremely related and supports our link prediction model, because we will incorporate the preferential attachment algorithm within our model. In addition, the study showed that machine learning algorithms were just as accurate and reliable in link prediction in comparison to logistic regression analysis.
- *Shortcomings*: It uses the preferential attachment with the expected parameter not considering attachment rate, which may result in greater linear fit within the network. The researchers were unable to consider hyper-authorship or name ambiguity within the paper. However, in our data do not have the same issue and thus does not pose a problem.

The third paper was the link prediction with hybrid mechanism by Zhang J.(2017) [8]

- *Main idea*: The paper observes the meta-path based model on the research databases of the DBLP bibliographic network. He generates multiple predictors by various topological features, and proposes the hybrid mechanism that combines the predictors to improve the accuracy of the link prediction
- *Use for our project*: It is related to our link prediction model, because as our initial data set does not have very many topological features, we had to utilize multiple mechanisms as our predictors.
- *Shortcomings*: The limitations of the paper include geographical proximity of authors which may pose a problem with our model. But, distance or proximity between neighbors may not pose an issue in today's digital and information age.

2.3 Rejection Sampling

Sampling methods help us to get subgraphs closer to the desired graphs in general or specific networks by adding constraints. We suggest to use rejection sampling is one of the revised random sampling methods. The hypergraphs generated by PA describe the our network better by applying rejection sampling. The first paper was the sampling methods by Lu and Bressan(2012) [8] [3]

- *Main idea*: Instead of the traditional sampling methods, this paper propose a new algorithm, Neighbour Reservoir Sampling, to avoid sampling subgraphs that have high clustering coefficients. They evaluate the new algorithm with Acceptance-Rejection Sampling, Random Vertex Expansion, Metropolis-Hastings Sampling with synthetic graphs. The researchers compare how randomly the algorithms generate induced subgraphs preserving the topological features of graphs.
- *Use for our project*: This research is related to our model for generating more accurate hyperedges when evolving hypergraphs with preferential attachment.
- *Shortcomings*: When comparing the accuracy and efficiency of the algorithms, the researchers only consider the cases of when global average features are preserved.

Since we consider the local features of the hypergraphs, the algorithm may be applicable with a different accuracy and efficiency than what the researchers found.

3. PROPOSED METHOD

We propose a new model to predict co-authorship networks using hypergraph preferential attachment and rejection sampling. The dataset that was used in the current project consisted of a list of author IDs and their respective collaborative documents/papers. There is a total of 36, 949 authors (nodes) and a total of 137, 959 documents (links). We took an initial analysis of our dataset and found it follows a power-law distribution which is suitable for our model. The author IDs are represented as nodes of the graph and their link or edge is represented by the document they co-author.

3.1 Hypergraph PA and Rejection Samplinig

We propose using rejection sampling when adding the hyperedges and predicting links between the nodes. The preferential hypergraph model is formed according to the following guidelines outlined below:

- starting with an initial undirected graph and an estimated hypergraph edge (learning graph), $G =< V, E >$, of an author network and the assumption that edges can appear multiple times
- initial network evolves in every time, t, with one event occurring at each time step
- at each time step, the number of nodesY_t is determined at random and this pair of nodes are randomly chosen according to the preferential attachment rule (probability proportional to the degree of edges in the graph at time,t)
- The pairs are compared based on their similarity and the rejection threshold. If *True*, generate a hyperedge, if *False* no hyperedge is created and a new pair of nodes chosen
- the process is repeated until the similarity of the hypergraph is equal to or greater than the rejection sampling threshold (maximum likelihood estimation of the graph)

```
Input $G =< V, E >$ ;
Generate a sequence $(Y_1, Y_2, ..., Y_t)$;
for $Y_i$ to $Y_t$ do
    Choose $Y_i$ nodes ;
    Generate hyperedge E ;
    if $Similarity > threshold$ then
        add E to G;
    end
end
```

Algorithm 1: Rejection Hypergraph PA

3.2 Feature Extraction and MLP

We use multilayer perceptron to predict the co-authorship of queries. This prediction model has shown the high accuracy to solve classification problems. By introducing the MLP, we defined the co-authorship prediction as a binary classification problem using input features extracted from hypergraph. To apply the MLP, we need to convert each query to single vector. However, the number of authors in query vary in co-authorship network.Thus, we should find the graph features that is commonly defined on different length of query and can express sufficient information of query and graph.

4. EXPERIMENTS

We implemented our integrated our model to the author data set and found the following results. All results are averaged out with the results of 10 trials with the prediction model, MLP. The dropout probability was wet to 0.05, the epochs were set to 300, and the activation function was a sigmod function. When doing the rejection sampling, we use two similarity methods, that is Jaccard and Adamic/adar. Since the nodes are highly not the neighbor node of each other, we newly defined these functions that deal with more neighbors of some radius. Neighbors with a radius indicates the nodes that have shorter shortest path than the radius value. We control this radius value, threshold for rejection sampling, and iteration of PA for the two methods. Figure 1 displays the effects of rejection sampling threshold when the hypergraph is evolved by the iteration of 2000, and 5000. The set of threshold was determined by calculating the average of threshold with random candidate sets. Two graphs are the cases that Adamic/adar is used on rejection sampling where the radius was set to 3. The results displayed on the graph show the accuracy, precision, recall, and f1-score of the model. From the figure, the model shows overall a moderately high level of accuracy. There is no big change on f1-score, while the recall slightly changes.

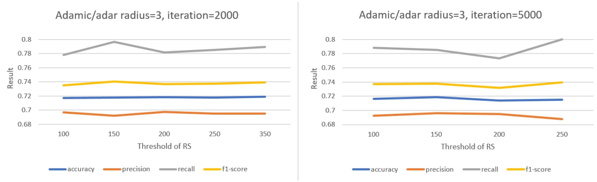

Figure 1: The effect of RS threshold

Next, we do experiments with many different conditions to figure out the effects of the iteration value of PA. We changes the iteration value from 2000 to 20000. Figure 2 shows the accuracy, precision, recall and f1-scores of our prediction model parameters. There is actually no noticeable change.

We tested various link similarity methods such as Adamic/Adar (varying radius) and Jaccard similarity. In order to compare these methods, the recall and f1-score results of them are shown in figure 3. Overall, the results tend to increase as the iteration value increases. When the iteration of PA is 5000 or 10000, Adamic/adar with the radius of 3 is the best to predict the co-authorship. However, Adamic/adar with the radius of 2, and Jaccard with the radius of 3 shows the highest prediction ability when the iteration of 15000.

5. DISCUSSION

We proposed an extended model of hypergraphs to predict co-authorship networks through preferential attachment and rejection sampling. The proposed method of PA and rejection sampling in hypergraphs has the following advantages:

- Useful in cases where there is very little information on the features of the network or nodes. By applying

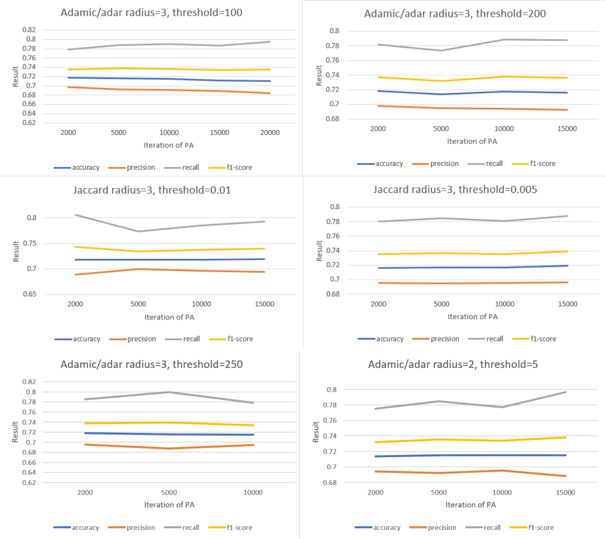

Figure 2: The effect of iteration of PA

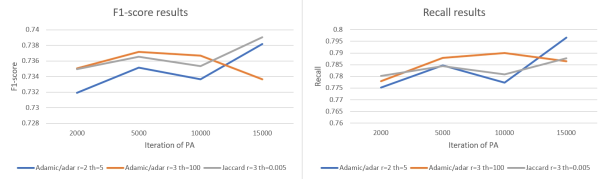

Figure 3: Recall depending on different similarity methods

rejection sampling methods such as Jaccard similarity and Adamic/Adar, able to further distinguish features of the edges and the graph.

- High accuracy and efficiency in comparison to the traditional link prediction models and methods. As we combined several tools into one prediction model, we were able to scope the focus on valued prediction methods rather than random estimation.

5.1 Limitations

The original dataset that was used in this project had very little information providing node features or to what sort of co-authorship network was portrayed. Nevertheless, with the use of rejection sampling methods using link similarity methodology (such as Jaccard similarity and Adamic/Adar) we were able to distinguish link pattern features within the network as well as find local clusters or communities. As well, by integrating and combining several methodology found to be useful in prediction methods, we were able to ascertain a greater degree of accuracy in comparison to previous literature.

Although our prediction model provided strong results in accurately predicting collaborative relationships, there were still some limitations or improvements that could be considered.

- Computational Time (Model Efficiency)
- Utilization of graph projection features for predicting links
- Consideration of the extent of the number of papers' an author will collaborate on

Firstly, the computational time of executing the model is not as efficient and the time it takes to go through the process is one of the model's limitations. Secondly, we utilized some features of the projected graph for predicting links as we had little information on the actual features of the network. This was a limitation because it affected the prediction accuracy of links between authors. Lastly, in a real network, authors may be limited to a certain number of papers to collaborate on, which we did not take into consideration in the current prediction model.

5.2 Conclusion

In summary, we proposed a prediction model using hypergraphs and rejection sampling through preferential attachment. Since some papers shows that the PA do not show highly improvement on link-prediction than other methods, starting from PA was a big challenge. However, fortunately, we found that out of all the similarity functions used in rejection sampling, Adamic/Adar showed greater accuracy and recall scores. In conclusion, the prediction model used in this project showed strong accuracy in predicting co-authorship networks in comparison to the more traditional models.

6. REFERENCES

[1] Chen Avin, Zvi Lotker, Yinon Nahum, and David Peleg. Random preferential attachment hypergraph. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 398–405, 2019.

[2] Albert-László Barabási et al. *Network science.* Cambridge university press, 2016.

[3] Xuesong Lu and Stéphane Bressan. Sampling connected induced subgraphs uniformly at random. In *International Conference on Scientific and Statistical Database Management*, pages 195–212. Springer, 2012.

[4] N Roopashree and V Umadevi. Future collaboration prediction in co-authorship network. In *2014 3rd International Conference on Eco-friendly Computing and Communication Systems*, pages 183–188. IEEE, 2014.

[5] Paul Sheridan and Taku Onodera. A preferential attachment paradox: How preferential attachment combines with growth to produce networks with log-normal in-degree distributions. *Scientific reports*, 8(1):1–11, 2018.

[6] Yizhou Sun, Rick Barber, Manish Gupta, Charu C Aggarwal, and Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 121–128. IEEE, 2011.

[7] Qi Yu, Chao Long, Yanhua Lv, Hongfang Shao, Peifeng He, and Zhiguang Duan. Predicting co-author relationship in medical co-authorship networks. *PloS one*, 9(7), 2014.

[8] Jinzhu Zhang. Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction. *Information Processing & Management*, 53(1):42–51, 2017.

APPENDIX

A. APPENDIX

A.1 Labor Division

The team performed the following tasks

- Presentation [Song, Yoon, Choi]
- Written work for progress report [Song, Choi]
- Code Implementation and Algorithm Design [Yoon]
- Data collection [all]
- Literature Review and Brainstorming[all]

A.2 Full disclosure wrt dissertations/projects

A.2.0.1 Song:.

She is not doing any project or dissertation related to this project.

A.2.0.2 Yoon:.

He is not doing any project or dissertation related to this project: his interest is on the topic of dynamic control model such as variational recurrent neural network and neural ordinary differential equation.

A.2.0.3 Choi:.

She is not doing any project or dissertation related to this project: her dissertation is on the topic of organizational learning and homophily in networks.