# Prediction of Citation Counts and Citation Network Analysis

Srishti Majumdar

***Abstract*— Creating citation networks and collaboration networks is the traditional way of representing relationships between scholarly literature, researchers, and even publishing venues. The exploration and analysis of such networks can provide interesting insights and be leveraged to provide useful information. The applications of citation network analysis include the study of research impact, understanding citation patterns, and even, understanding the research developments in a field over time. Citation count prediction has been popularly solved using link prediction and regression methods. Some benefits of such predicting citation counts lies in technological forecasting as well as citation recommendation systems. This paper aims to explore citation network analysis and the task of predicting citation counts.**

***Index Terms*— citation network, citation count prediction, supervised learning**

## I. INTRODUCTION

While developing models for network analysis, graphical models aim to model both the data as well as the underlying dependencies in data. One way of visualising citation networks is that an edge between two papers, $P_i$ and $P_j$, exists if $P_i$ cites $P_j$. Citation networks have been widely researched over the last few decades for a variety of applications.

In this section, I will briefly outline some common research directions for citation networks.

1) Paper Classification: One application of network analysis is collective classification. This is the process of inferring class labels over a set of related nodes. This can be used in a citation graph to infer the field of study for various publications. Richardson and Domingos (2006) [1] propose the use of Markov Logic Networks (MLNs), an approach combining probabilistic modelling with First Order Logic. Crane and McDowell (2012) [2] further investigate and compare the use of MLNs for citation classification. MLNs provide one kind of approach used for the task in hand. Another approach includes approximation to find the best guess for the class label. One example of this approach includes the use of Graph Convolution Networks by Kipf and Welling (2016) [3] for this purpose.
2) Collaboration and/or Community Detection: Community detection helps visualize and understand the underlying structure of the network. This can prove especially useful in finding sub-fields in a research area and identifying trends in the communities obtained. Liu and Li (2015) [4] have developed a citation similarity based community detection model. In their work, Khan and Niazi (2017) [5] provide a review of various community detection methods and their comparisons.
3) Understanding Research Influence: Perhaps deemed one of the most important important tasks for citation networks, understanding and finding highly influential papers, authors and journals, is a widely researched topic. Page (1998)[6] proposed PageRank which remains one of the key methods for finding high impact papers. Kleinberg (1999)[7] proposed the HITS algorithm which makes use of hyperlinks between pages to determine ranking of pages. Ding et al. (2010) [8] use PageRank to rank authors rather than papers. Both PageRank and HITS are matrix based algorithms. Zhang et al. (2016) [9] propose a graph based approach using in-degree and out-degree counts to finding influential papers.
4) Citation Recommendation Systems: Appropriate citation recommendations are essential for conducting research. West, Smith and Bergstrom (2016) [10] utilise clustering on a citation network to provide varying levels of recommendations to users. Another work by Ebesu and Fang (2017)[11] utilises citation networks to provide context-aware recommendations.
5) Link Prediction: Link Prediction is an important task with real-world use case. The aim of a link prediction algorithm is to estimate the probability of links between nodes in a graph. In other words, given an observable portion of the network, link prediction tries to either infer missing links or predict future links for nodes in the graph.This can be used for citation recommendation systems. For example, if a new paper is published, its citations are low so understanding its correlations to other papers may prove useful in understanding its ranking. Similarly, another application of link prediction is estimating the projected citations for a paper to help understand how influential it may be in the future.

Link Prediction and Regression models for citation networks will be discussed in more detail in section II. Section III describes the initial data analysis conducted. Section IV outlines the methods, experiments conducted and results obtained. Finally, sections V and VI discuss the future work and conclusion of the project.

## II. LITERATURE REVIEW

In this section, some of the methods for modelling citation count are discussed.

Researchers around the world have explored and developed various strategies for link prediction in various applications.

Shibata et al. (2012) [12] express the citation network in the form of 11 features, including link-based Jaccard coefficient, difference in betweenness centrality, and cosine similarity of term frequency-inverse document frequency vectors to predict links between citations using a supervised learning model.

Many researchers have delved into the idea of link prediction as a time-series or sequential task. Gao et al. (2011) [13] propose the integration of node contents with structure for temporal link prediction. Jawed et al. (2015) [14] treat the task hand from time frame based approach and use an unsupervised model for link prediction. More recently, the use of neural networks, specifically graphical neural networks, has also been explored. Zhang and Chen (2018) [15] explore the use of graph neural networks to predict links by observing sub-graphs around each target link. Gu et al. (2019) [16] explore the use of Graph Attention Networks to learn appropriate node representations and learn from existing links.

The use of link prediction for benefiting applications has also been explored. Liu et al. (2019) [17] describe a link prediction model to construct a paper correlation graph in order to make sparsity in citation networks and improve recommendations. In their works, Pobiedina and Ichise (2016) [18] as well as Bütün and Kaya (2019) [19] describe the use of link prediction in citation counting.

Yan et al. (2011) [20] evaluate various regression models like linear regression and K-Nearest Neighbours to predict citation counts. In their paper, Li et al. (2015) [21] develop a trend- based model for this task.

Citation counts have also been studied from the view of the graph evolving over time. Graph embeddings are an important area of research in network analysis. Graph embeddings allow more comprehensive and detailed representations of nodes and labels. Perozzi et al. (2014) [22] presented DeepWalk for learning latent representations. Node2vec and metapath2vec by Grover et al. (2016) [23] and Dong et al. (2017) [24] are also popular embedding methods. Goyal et al. (2018) [25] provide an overview and comparison of graph embedding techniques. In a student project, Ramaprasad (2019) [26] describes the use of a graph embeddings to study citation counts over periods.

The above works prove the motivation and relevance of the task,citation count prediction.

## III. DATA ANALYSIS

The dataset used in this project is the AMiner Citation Network Dataset [27]. In particular, the DBLP v10 citation dataset has been used in this project. In this dataset, each paper is associated with an id, title, abstract, authors, year, venue, number of citations and references.

The data is stored in the form of JSON file where each line denotes a JSON object containing information about a particular paper. An example of such an object can be seen in Figure 1. A more detailed view of the dataset can be seen in Figure 2 to for the same paper.


```
{
"abstract":"The purpose of this study is to develop a learning
tool for high school students studying the scientific aspects of
information and communication net- works. More specifically, we
focus on the basic principles of network protocols as the aim to
develop our learning tool. Our tool gives students hands-on
experience to help understand the basic principles of network
protocols.",
"authors": ["Makoto Satoh", "Ryo Muramatsu",
            "Mizue Kayama", "Kazunori Itoh", "Masami Hashimoto",
"Makoto Otani","Michio Shimizu", "Masahiko Sugimoto"],
"n_citation": 0,
"references": ["51c7e02e-f5ed-431a-8cf5-f761f266d4be", "69b625b9-
ebc5-4b60-b385-8a07945f5de9"],
"title": "Preliminary Design of a Network Protocol Learning Tool
Based on the Comprehension of High School Students: Design by an
Empirical Study Using a Simple Mind Map",
"venue":"international conference on human-computer interaction",
"year": 2013,
"id": "00127ee2-cb05-48ce-bc49-9de556b93346"
}
```


Fig. 1. Example of JSON Structure for Paper in AMiner Citation Dataset

| Field Name | Field Type | Description | Example |
|---|---|---|---|
| id | string | paper ID | 00127ee2-cb05-48ce-bc49-9de556b93346 |
| title | string | paper title | Preliminary Design of a Network Protocol Learning Tool Based on the Comprehension of High School Students: Design by an Empirical Study Using a Simple Mind Map |
| authors | list of strings | paper authors | ["Makoto Satoh", "Ryo Muramatsu", "Mizue Kayama", "Kazunori Itoh", "Masami Hashimoto", "Makoto Otani", "Michio Shimizu", "Masahiko Sugimoto"] |
| venue | string | paper venue | international conference on human-computer interaction |
| year | int | published year | 2013 |
| n_citation | int | citation number | 0 |
| references | list of strings | citing papers' ID | ["51c7e02e-f5ed-431a-8cf5-f761f266d4be", "69b625b9-ebc5-4b60-b385-8a07945f5de9"] |
| abtract | string | abstract | This purpose of this study ... |

Fig. 2. Tabular View for Paper described in Fig. 1.

The dataset contains the information for 3,079,007 papers and has 25,166,994 citation relationships. Some initial analysis conducted on the citation dataset is listed below:

- A summary of dataset can be viewed as follows:
  - Number of Papers: 3,079,007
  - Number of Citation Relationships: 25,166,994
  - Number of Authors: 1762637
  - Number of Venues: 5074
  - Years: 83 (1936-2018)
- The top five venues with maximum paper count is shown in table I.
- Figure 3 shows an year-wise count of papers. This gives an insight on the publication rate and distribution of our dataset. The number for papers corrected for 2018 is very less, around 70 nodes, which result in the curve dipping at the end.
- Amongst the counts for missing information, it is important to know that 362865 papers have no citation

| Venue | Paper |
|---|---|
| lecture notes in computer science | 32137 |
| international conference on acoustics, speech, and signal processing | 26621 |
| international conference on robotics and automation | 18843 |
| international conference on image processing | 18336 |
| international conference on communications | 17679 |

TABLE I

"VENUES WITH MAXIMUM PUBLISHING COUNT"

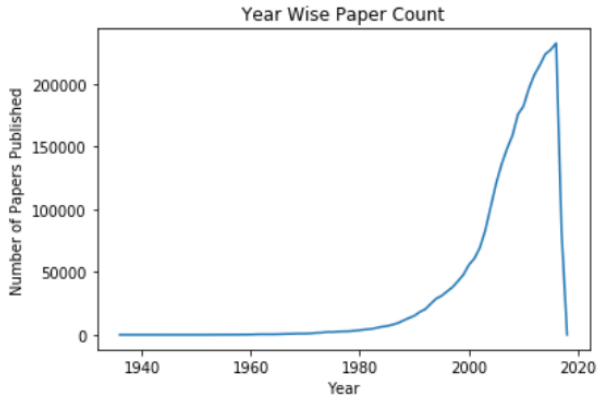


Fig. 3. Distribution of Papers from 1936-2018

information.

There are a couple of points that call attention. The values of venues in the table I is in lowercase to prevent duplication. One may notice that the years between mid to late 1990s and 2017 account for majority of the publications. This preliminary analysis of the dataset helps visualise the possible network as well as helps guide our model input and structure.

## IV. CITATION NETWORK AND COUNT PREDICTION

In this section, first I will discuss how the graph was created and the embedding strategy used to represent information in the graph. This will be followed by a small analysis report of the graph and the trends noticed in our citation graph. Finally, we discuss the model used, the motivation behind using this model and the experiments conducted on the citation network.

### A. Citation Graph Creation

The citation graph is created in the following manner:

1) For each paper, $P_i$, create a vertex corresponding to the paper, if such a mapping doesn't exist.
2) Obtain all references for $P_i$.
   a) For each reference paper, $P_j$, check if a vertex mapping exists. If not, create a vertex corresponding to $P_j$
   b) Create a directed edge from $P_i$ to $P_j$

A view of the citation network for papers published in 2018 is shown in Figure 4.

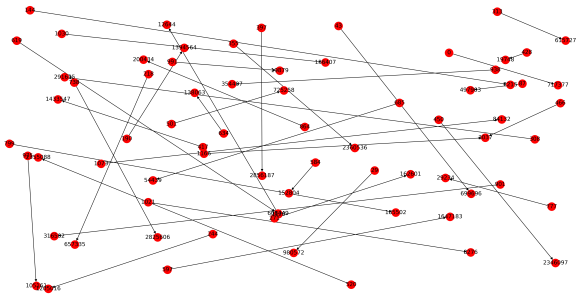

Fig. 4. Citation Relationships for Papers Published in 2018

In section IV. B. we use the entire citation network for the analysis conducted. However, for the implementation the citation network is divided for training and testing purposes.

### B. Citation Network Analysis

Once the citation network is built using the method mentioned in section IV.A, some analysis was conducted on the network. The number of edges was checked and was found to have the correct number of citation relationships, i.e. 25166994.

The network obtained is widely distributed. It has 2038092 strongly connected components and 7820 weakly connected components.

The in and out degree distribution can be seen in figures 5 and 6 respectively. For these plots, log scaling was performed both both the x and y axes. It can be seen that the trends for degree distribution are of the form of a power law distribution. A large majority of papers have low in/out degree and few papers have high in/out degree.

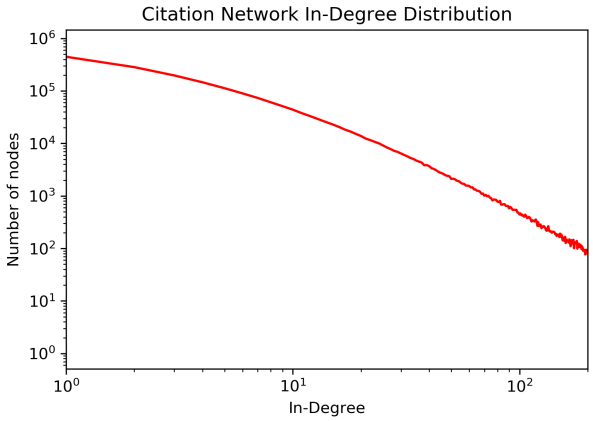


Fig. 5. In-degree Distribution of Citation Network

### C. Citation Prediction

*1) Objective:* Given the citation trends for a paper in the short term, predicting the trends in the long term. For this project, we take the first 5 years after publication of the paper as short term and the long term is considered the next 5 years

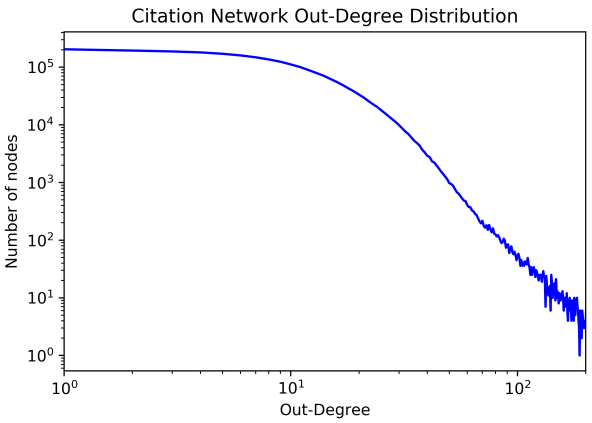


Fig. 6. Out-degree Distribution of Citation Network

after that. Hence we study citation trends over 10 years for each paper.

*2) Dataset Creation:* For creating the dataset, paper and venue information and well as citation counts over 10 years is used. In this project, a simplified view of the task is taken and the focus is on observing how well citation counts in the future can be predicted by knowing short term citation counts.

Two datasets are created for modelling- a large dataset containing papers published during 1998-2007 and smaller dataset with papers published during 1992-1998. The reason I consider publications only till 2007 is that the data of citations is available till 2017 and only 35 publications are listed for 2018. The larger dataset contains 1,102,241 papers and the smaller dataset contains 220,258 papers. A snapshot the data sets created can be seen in Figure 7. Y_0 refers to the citations in the same year as the paper for published. Y_i for i=1,2...,10 refers to citations 'i' years after publishing. The citation count over a 10-year period for each paper $P_i$ is calculated as follows:

1) Year of Publishing is P
2) All Y_i are initialised to 0 .
3) For each citation $C_i$:
   a) Year of citation is C
   b) The number of years after which $C_i$ was published is calculated by Equation 1 and represented as D
   c) Citation count for that Y_D is increased by 1 as in Equation 2.

$$D = C - P \tag{1}$$

$$Y_D += 1 \tag{2}$$

In future, to move from observing the effects of short term citations to making the predictions more holistic, I would like to include author information for a paper in terms of maximum, mean and minimum of all authors for the paper. However, adding this information is tricky as one has to

| | PID | VenueID | Y_0 | Y_1 | Y_2 | Y_3 | Y_4 | Y_5 | Y_6 | Y_7 | Y_8 | Y_9 | Y_10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1912 | 3 | 0 | 4 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 0 |
| 1 | 2009 | 31 | 0 | 1 | 1 | 3 | 1 | 5 | 2 | 2 | 3 | 0 | 1 |
| 2 | 2672 | 184 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3044 | 76 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | 3057 | 213 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Fig. 7. Dataset Created

account for citation till publishing year for each paper for each author.

*3) Model Description:* Regression models, especially Neural Networks, have been used for several citation prediction tasks. Neural Networks are high capacity models and allow the flexibility to model any kind of function or data. In recent years, there has been a lot of research into using neural network and/or deep learning methods for modelling various tasks.

I created a 3 layer neural network. However, this neural network treats each row or sample as independent. To introduce the concept of dependency, I then created a Recurrent Neural Network (RNN) for the dataset. I chose RNN as my model as it allows modelling for sequential and time series data by maintaining states between layers and time steps.

I used mini-batch training which ensures data is fed in batches of size k where k<m (m is the number of rows in training data) to improve performance. To introduce non-linearity to our neural network, Rectified Linear Unit (ReLU) is used as the activation function as shown in equation 3. The dropout is set to 0.2 as a way for regularising our model and preventing over-fitting.

$$f(x) = max(0, x) \tag{3}$$

The baseline for comparison was the assignment of the mean average of citations in the short term as predicted citation count for the following years.

*4) Implementation Details:* This project was conducted on a personal laptop with the following specifications:

- System: 64 bit operating system, x64 based processor
- Processor: Intel(R) Core(TM) i7-7500U CPU@ 2.7GHz 2.9GHz
- RAM: 8GB

The model was created using the Keras library [28], a Python deep learning library with Tensorflow backend. Themodels were trained and tested on first the smaller dataset and then the larger dataset and adjustments to the model were made accordingly.

*5) Results:* The implementation of 3 layer neural network and RNN were kept similar so that the models are comparable. In both, the input is encoded so the given features are projected into a higher dimensional vector, goes through a middle layer with a smaller number of units before taking the output in a 5 unit layer. Each unit in the output layer corresponds to a year from Y_6 to Y_10. The metric accuracy was used to evaluate the models.

| Dataset | Training Accuracy | Validation Accuracy |
|---|---|---|
| Smaller | 0.517 | 0.484 |
| Larger | 0.524 | 0.542 |

TABLE II

"3 LAYER NEURAL NETWORK"

| Dataset | Training Accuracy | Validation Accuracy |
|---|---|---|
| Smaller | 0.532 | 0.510 |
| Larger | 0.545 | 0.582 |

TABLE III

"USING RECURRENT NEURAL NETWORK"

The performances of the 3 layered NN and RNN are shown in table II and III. From table III, we can see that RNN performs better. Both models performed better than the baseline model which had accuracy between 30-42% in all cases.

In this project, I have simplified the interpretation of citation count prediction to observing just the short term citation patterns to predict long term citations. The distributed nature of the data and the simplified interpretation of the task explains the model performance. One can infer that more information is required for this task so as to create a better model. This may include more information like authors citations trends, venue based trends as well as content based analysis on the title and abstracts.

## V. FUTURE WORK

Citation Prediction is an interesting and significant task. Citation patterns vary across various fields of study, and even among publications within a field of study. To predict such differences, higher capacity and/or more generalised methods are required. The future work includes the incorporation of paper related information into the graph to build a heterogeneous graph for prediction. Additionally, the inclusion of abstract information may prove useful in identifying citation trends. A technique to include author information can also be included. Comparisons with other models may be conducted to understand performance differences and trade-offs.

## VI. CONCLUSION

Citation networks have been used to model various tasks like understanding academic influences, collaboration detection and classification. In this project, I studied and analysed a citation network to understand underlying trends within the network. Additionally, to understand the effects of early citation trends on future citation trends, I explored and developed neural network models to predict citation counts. This approach is proved to be a good way of understanding the citation network and its evolution. However, predicting citation counts is an intricate problem as trends vary not only among various fields but also among venues and authors. Hence, including such information may help not only make the predictions better but also help understand how each of these factors influence citation counts for a paper.

## REFERENCES


[1] M. Richardson and P. Domingos. "Markov Logic Networks". In: *Machine Learning Journal* (2006).

[2] R. Crane and L. K. McDowell. "Investigating Markov Logic Networks for Collective Classification". In: *International Conference on Agents and Artificial Intelligence (ICAART)* (2012).

[3] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *CoRR* abs/1609.02907 (2016). arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907.

[4] T. Liu and K. Li. "A citation similarity based community detection method in citation networks". In: *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. Dec. 2015, pp. 146–149. DOI: 10.1109/IAEAC.2015.7428536.

[5] Bisma S. Khan and Muaz A. Niazi. "Network Community Detection: A Review and Visual Survey". In: *CoRR* abs/1708.00977 (2017). arXiv: 1708.00977. URL: http://arxiv.org/abs/1708.00977.

[6] L. Page S. Brin R. Motwani and T. Winograd. "PageRank: Bringing Order To The Web". In: *published as technical report* (1998).

[7] J. Kleinberg. "Hubs, Authorities, and Communities". In: *Cornell University* (1999).

[8] Ying Ding et al. "PageRank for Ranking Authors in Co-Citation Networks". In: *Journal of the American Society for Information Science and Technology* 60.11 (Nov. 2009), pp. 2229–2243. ISSN: 1532-2882.

[9] S. Zhang et al. "Finding Influential Papers in Citation Networks". In: *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*. June 2016, pp. 658–662. DOI: 10.1109/DSC.2016.55.

[10] J. D. West, I. Wesley-Smith, and C. T. Bergstrom. "A Recommendation System Based on Hierarchical Clustering of an Article-Level Citation Network". In: *IEEE Transactions on Big Data* 2.2 (June 2016), pp. 113–123. ISSN: 2372-2096. DOI: 10.1109/TBDATA.2016.2541167.

[11] Travis Ebesu and Yi Fang. "Neural Citation Network for Context-Aware Citation Recommendation". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. Shinjuku, Tokyo, Japan: Association for Computing Machinery, 2017, pp. 1093–1096. ISBN: 9781450350228. DOI: 10.1145/3077136.3080730. URL: https://doi.org/10.1145/3077136.3080730.

[12] Naoki Shibata, Yuya Kajikawa, and Ichiro Sakata. "Link prediction in citation networks". In: *JASIST* 63 (2012), pp. 78–85.

[13] Sheng Gao, Ludovic Denoyer, and Patrick Gallinari. "Temporal Link Prediction by Integrating Content and Structure Information". In: *Proceedings of the*

*20th ACM International Conference on Information and Knowledge Management*. CIKM '11. Glasgow, Scotland, UK: Association for Computing Machinery, 2011, pp. 1169–1174. ISBN: 9781450307178. DOI: 10.1145/2063576.2063744. URL: https://doi.org/10.1145/2063576.2063744.

[14] M. Jawed, M. Kaya, and R. Alhajj. "Time frame based link prediction in directed citation networks". In: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Aug. 2015, pp. 1162–1168. DOI: 10.1145/2808797.2809323.

[15] Muhan Zhang and Yixin Chen. *Link Prediction Based on Graph Neural Networks*. 2018. arXiv: 1802.09691 [cs.LG].

[16] Weiwei Gu et al. *Link Prediction via Graph Attention Network*. 2019. arXiv: 1910.04807 [cs.SI].

[17] Hanwen Liu et al. "Link prediction in paper citation network to construct paper correlation graph". In: *EURASIP Journal on Wireless Communications and Networking* 2019 (Dec. 2019). DOI: 10.1186/s13638-019-1561-7.

[18] Nataliia Pobiedina and Ryutaro Ichise. "Citation Count Prediction as a Link Prediction Problem". In: *Applied Intelligence* 44.2 (Mar. 2016), pp. 252–268. ISSN: 0924-669X. DOI: 10.1007/s10489-015-0657-y. URL: https://doi.org/10.1007/s10489-015-0657-y.

[19] E. Bütün and M. Kaya. "Predicting Citation Count of Scientists as a Link Prediction Problem". In: *IEEE Transactions on Cybernetics* (2019), pp. 1–12. ISSN: 2168-2275. DOI: 10.1109/TCYB.2019.2900495.

[20] Rui Yan et al. "Citation count prediction: Learning to estimate future citations for literature". In: Oct. 2011, pp. 1247–1252. DOI: 10.1145/2063576.2063757.

[21] Cheng-Te Li et al. "Trend-Based Citation Count Prediction for Research Articles". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Tru Cao et al. Cham: Springer International Publishing, 2015, pp. 659–671. ISBN: 978-3-319-18038-0.

[22] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. "DeepWalk". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14* (2014). DOI: 10.1145/2623330.2623732. URL: http://dx.doi.org/10.1145/2623330.2623732.

[23] Aditya Grover and Jure Leskovec. *node2vec: Scalable Feature Learning for Networks*. 2016. arXiv: 1607.00653 [cs.SI].

[24] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. "Metapath2vec: Scalable Representation Learning for Heterogeneous Networks". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '17. Halifax, NS, Canada: Association for Computing Machinery, 2017, pp. 135–144. ISBN: 9781450348874. DOI: 10.1145/3097983.3098036. URL: https://doi.org/10.1145/3097983.3098036.

[25] Palash Goyal and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey". In: *Knowledge-Based Systems* 151 (July 2018), pp. 78–94. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2018.03.022. URL: http://dx.doi.org/10.1016/j.knosys.2018.03.022.

[26] Shreya Ramaprasad. *Citation Network Analysis to Predict Citation Trends*. 2019. URL: https://github.com/ShreyaRamaprasad/CitationNetworkAnalysis/blob/master/CapstoneProjectReport_ShreyaSR.pdf.

[27] Jie Tang et al. "ArnetMiner: Extraction and Mining of Academic Social Networks". In: *KDD'08*. 2008, pp. 990–998.

[28] *Keras*. URL: https://keras.io/.