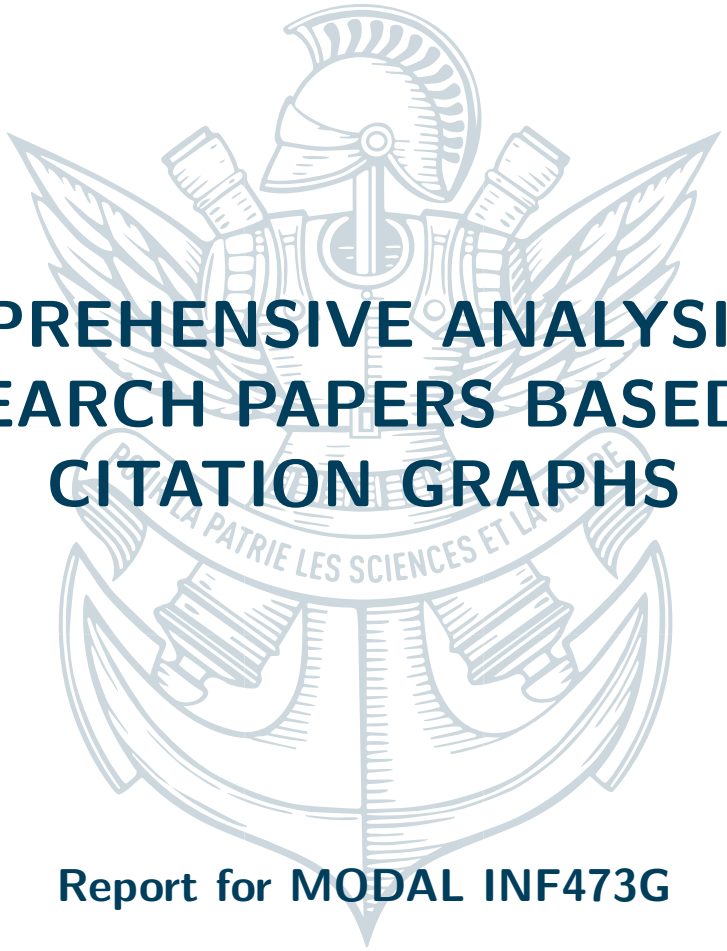


A COMPREHENSIVE ANALYSIS OF AI RESEARCH PAPERS BASED ON CITATION GRAPHS

Report for MODAL INF473G

November 23, 2023

LE QUANG Dung, LIU Zuhong

CONTENTS

1 INTRODUCTION

In recent years, the field of Artificial Intelligence (AI) has witnessed a remarkable surge in research and development, leading to significant advancements and breakthroughs across various domains. As the AI landscape expands, it becomes increasingly important to understand and review the progress that have been made within this rapidly evolving field, which will have a positive effect on the enduring stability of artificial intelligence development.

The query is related to discovering the primary research directions within the AI community in recent years. Various approaches exist for addressing this matter. In this study, we adopt a citation graph-based method. This graph is constructed using citation data from AI articles, wherein nodes represent articles and authors, while edges represent authorship or citation relationships.

In this report, Section 2 is devoted to presenting the **datasets** and elucidating the **normalization** process applied to render the data usable for the problem at hand. Utilizing this dataset, we generate logical graphs and subsequently tackle the problems outlined in Section 3. Initially, we identify **influential articles** and **authors** by employing measures such as *In-Degree Centrality* and *Betweenness Centrality*. Subsequently, we employ the *Louvain algorithm* to identify **significant communities** in the AI field. Leveraging relevant information extracted from these communities, we propose a methodology for determining the **academic Competitiveness** of institution. Our codes are published here[1].

2 CREATION OF DATASET

In this section, we would like to describe the creation of citation graph. To begin with, we briefly introduce the two basic datasets used in the project and the pipeline of data cleaning and merging.

2.1 Citation Network Dataset

The Citation Network Dataset, provided by Aminer[1], is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. Initially, the dataset can be used for clustering with network and side information, studying influence in the citation network, finding the most influential articles, topic modeling analysis. It serves as a collection of articles associated with abstract, authors, year, venue, and title. Besides, it also provides citation information for articles, including the total number of citation and references. We use its latest released version (DBLP-Citation-Network V14) which contains 5,259,858 articles and 36,630,661 citation relationships.

2.2 CS Ranking Dataset

The CS Ranking Dataset is a metrics-based ranking of top computer science institutions around the world. It includes the mean count of published articles and the number of faculty members for around 500 institutions in all subdomains of computer science, such as AI, Computer Systems, Computer Theory and other interdisciplinary areas. In our project, we only consider the ranking of institutions on artificial intelligence and its sub-domain (Computer Vision, Natural Language Processing, Web and Information Retrieval, etc). We hope to obtain more

[1] https://github.com/ZuhongLIU/INF473G/tree/main.

information of research institutions (especially universities) to enrich the correspondent entities in our citation graph.

2.3 Data Filtering and Cleaning

Due to the considerable number of articles and citation relationship that are mostly irrelevant to our subject, it is essential that we should screen the samples in the Citation Network Dataset based on when and where they are published. In practice, we select articles that are published at top artificial intelligence conferences (Conference on Neural Information Processing Systems (NIPS), Computer Vision and Pattern Recognition Conference (CVPR), etc) from 2010 till now. Then we eliminate articles whose reference information is not recorded in the dataset. After the procedure of filtering, we obtain 10,509 articles and 15,641 citation relationship in total. The following figures show the distribution of dataset based on year and conference. From the left figure, it is evident that the quantity of published articles experiences a significant decline starting from 2019. This observation suggests that due to the characteristic of the dataset, our investigation primarily concentrates on AI research during the initial to intermediate phase of the AI boom(from years 2011 to 2018).

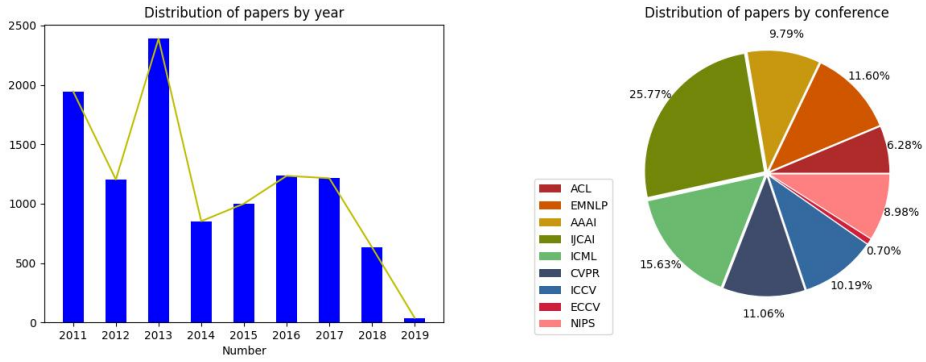


Figure 1: Distribution of article by Year and Conference

After a preliminary data filtering procedure of the original Citation Network Dataset, we aim to integrate two datasets into our citation graph. For each article, the Citation Network Dataset offers information of all of its author, including his or her id, name and affiliated organisation. However, the main difficulty lies on the incompatibility of institutions between the two datasets. The CS Ranking Dataset mainly comprises academic institutions such as universities and colleges, while omitting data pertaining to research and development departments or institutes of enterprises such as Google and Microsoft, which also appears frequently in the Citation Network Dataset. Besides, many universities are present within the dataset under an alternative designation.

We have summarized several common problems that we have encountered during the merging procedure.

- **Abbreviation**: "Univ. of California - Los Angeles" corresponds to "University of California -Los, Angeles"; "Chinese Acad Sci" corresponds to "Chinese Academy of Sciences".
- **Alias**: An alias is a different name or label assigned to an entity. Ex: "University of Tsinghua" corresponds to "Tsinghua University"; "Université Pisa" corresponds to "Univerity of Pisa"
- **Whole Address**: The Citation Network Dataset provides the full address information of research institution to which the author belongs. Ex: "Department of Computer Science and Engineering, University of California, San Diego, USA" corresponds to "Univ. of California - San Diego"; "Carnegie Mellon Univ, Pittsburgh, PA 15213 USA, Carnegie Mellon University" corresponds to "Carnegie Mellon University".

Owing to various problems mentioned above (sometimes they appear together in a single sample), it's rather difficult to correctly match the institution through simply taking the candidate with the largest cosine similarity. Therefore, we design a specific normalization algorithm for each entry of institution of two datasets.

To begin with, we change all letters to lower case and replace all the common abbreviation ("univ", "technol", "instit") in the entry with the complete word ("university", "technology", "institute"). In order to extract the name of university from the whole address, we then tokenize the entry ("Department of Computer Science and Engineering, University of California, San Diego, USA") into a list of candidate using comma as the separator (["Department of Computer Science and Engineering", "University of California", "San Diego", "USA"]). Since there are different campuses of certain universities (ex:University of California - Berkeley, Univ. of Maryland - College Park), the city information in the address is also crucial in entity matching. Hence, we append new tokens that combine the university token and city token to the list of candidate. Finally, we remove prepositions and articles ("of","the") in each candidate that may hinder the matching process.

We compute the distance between two names of institute using Levenshtein distance which calculates the minimal number of operations for transforming one string into another. Specifically, given two entity names S_1 and S_2, we have

$$d(S_1, S_2) = \min \{ld(t_1, t_2), t_1 \in Normalized(S_1), t_2 \in Normalized(S_2)\} \tag{1}$$

where ld represents the Levenshtein distance.

Then we select the target name of institution with the minimal distance. If the minimal distance is under certain threshold, we will obtain the counterpart of the author's organization in CS Ranking Dataset, otherwise we admit that this institution is not included in the CS Ranking Dataset.

In order to examine the performance of our normalization algorithm, we propose a validation set that is randomly sampled from the raw dataset. The validation set contains 200 relationship between author and his or her affiliated organization. Then we manually label the counterpart of the organization in CS Ranking Dataset as the ground-truth. Our matching algorithm achieves an accuracy of 90.3% on the validation set, which shows its robustness and reliability. Some failure cases are shown below:

Table 1: Some Failure Cases

Affliated Organization	**Ground Truth**	**Prediction**
SJTU, Shanghai, Peoples R China	Shanghai Jiao Tong University	Others
Penn State Univ, University Pk, PA 16802 USA	Pennsylvania State University	Others
Ocean Univ China, Qingdao, Peoples R China	Others	Renmin University of China

From the table, we can see that the failure cases are mainly caused by some specific abbreviations and other uncommon situations. We could indeed improve the performance of matching algorithm by exhaustively listing all alternate names for each institution in CS Ranking Dataset, which is not cost efficient for our project.

2.4 Overview of Graph Structure

Here we present the general structure of our graph 2. Our Citation Graph contains 10,509 article nodes, 13,500 author nodes, 555 institutions nodes, 15,641 article-article edges (citation relationship), 35,279 author-article edges, 8,745 institution-author edge (affiliation). We note that all relationship is considered as directed edge in the graph. In the context of research papers, it should be noted that the weight assigned to both the author-article edge and the institution-author edge is consistently set at 1. However, for the article-article edge, the weight is determined by the similarity between the titles of two articles. The purpose of this weight calculation will be elaborated upon in the following sections.

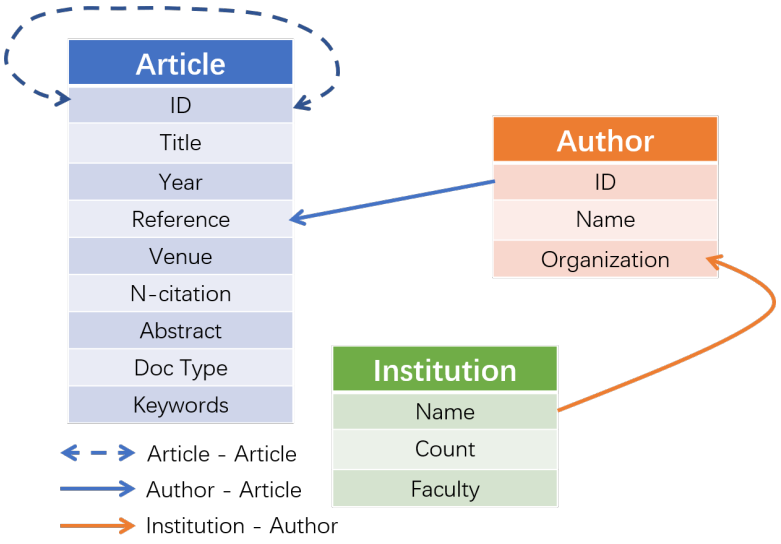


Figure 2: Construction of Citation Graph

3 ANALYSIS OF CITATION GRAPH

Following the normalization of our dataset, we conduct a comprehensive analysis encompassing influential papers, authors, as well as the identification and examination of significant communities. Furthermore, we endeavor to discern commonalities within these communities, thereby contributing to a deeper understanding of their shared characteristics.

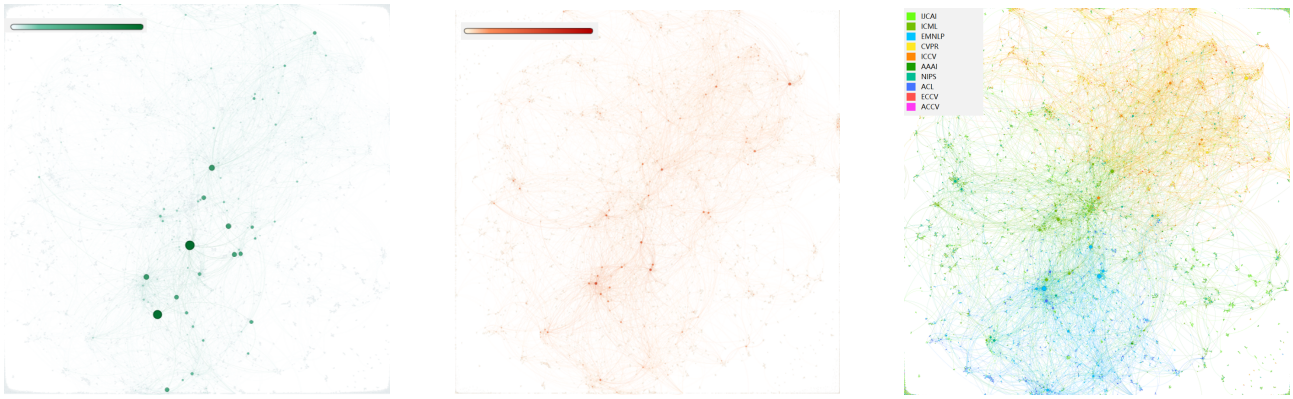
3.1 Network Analysis of Influential Articles and Authors

3.1.1 • Influential Articles

In order to figure out the most influential articles, we calculate various centrality metrics based on a subset of our citation graph that only encompasses article nodes. Several common graph centrality metrics, such as *Degree Centrality, Closeness Centrality, Betweeness Centrality* and *Eigenvector centrality* have been demonstrated to be effective in identifying important nodes. In our setting, we employ *in-Degree Centrality* and *Betweeness Centrality* as the evaluation metrics of influential articles due to the following justifications:

- In undirected graphs, nodes with a high closeness centrality value could be seen as important. However, the closeness centrality of a article node signifies its extensive referencing of prior scholarly works, without considering its influence on others.
- Eigenvector centrality works well for undirected graphs, but in a directed graph that is rather sparse (not strongly connected), such as our citation graph, the solution of the eigenvector problem is not necessarily unique and positive.

Therefore, in order to perform calculations of centrality metrics and visualization, we employed `Gephi`, a software designed for comprehensive analysis of various types of graphs and networks. The following figures 3 reflects the overall relationship between nodes in the graph. The size and the color of nodes in the first two figures respectively depend on the betweeness centrality and in-degree centrality. In the third figure, nodes are



(a) Sorted by betweeness centrality (b) Sorted by in-degree centrality (c) Classified by conference

Figure 3: Visualizations of article nodes in Citation Graph

categorized by the name of conference. The conferences within the same subdomain are categorized according to the color system, wherein Computer Vision (CV) is represented by the color red, Natural Language Processing (NLP) is represented by the color blue, and General Machine Learning (General AI) is represented by the color green. From the visualization, we notice that both "centers", characterized by articles exhibiting high citation rates (as indicated by their in-degree centrality values) and "bridges",characterized by articles that serve as vital links or intermediaries (as indicated by the value of betweeness centrality) of the graph concentrate on NLP and General AI instead of CV.

Statistically, we present a tabular featuring five articles with the highest value of centrality metrics. We observed a remarkable trend wherein the article related to natural language processing (NLP) has emerged as the predominant entry, while articles related to NLP account for less than twenty percent of the total articles. This finding suggests that the NLP domain exhibits a higher rate of technological advancement compared to other sectors during the initial phase of the AI boom. For example, the article *"Learning Phrase Representations using RNN Encoder-Decoder"*, which appears on both leaderboard, proposed the encoder-decoder and GRU structure (Gate Recurrent Unit) that have a great influence on deep learning. Subsequently, a year later, another article titled *"Effective Approaches to Attention-based Neural Machine Translation"* proposed the attention mechanism based on this encoder-decoder model.

Table 2: Top Five Most influential articles Based on In-Degree Centrality

Rank	**Title**	**Venue**	**Year**	**In-Degree Centrality**
1	Glove: Global Vectors for Word Representation.	EMNLP	2014	189
2	Effective Approaches to Attention-based Neural Machine Translation.	EMNLP	2015	80
3	Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.	EMNLP	2013	77
4	Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.	EMNLP	2014	65
5	DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.	ICML	2013	64

Moreover, it is noteworthy that research articles in the fields of CV and NLP show a significant degree of interaction with articles in the domain of General Artificial Intelligence (AI). This cross-referencing phenomenon serves to enhance the progress and advancement of both CV and NLP domains. Nevertheless, an analysis reveals

Table 3: Top Five Most influential articles based on Betweeness Centrality

Rank	Title	Venue	Year	Betweeness Centrality
1	Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.	EMNLP	2014	8630
2	Bilingual Word Embeddings for Phrase-Based Machine Translation.	EMNLP	2013	8318
3	DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.	ICML	2013	5141
4	Distributed Representations of Sentences and Documents.	EMNLP	2014	4948
5	Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books	ICCV	2015	4598

a limited occurrence of interactions between CV and NLP during this period.

3.1.2 • Influential Authors

After studying the influence of articles in different domains , our objective is to identify the influential authors who have exerted the greatest influence in that period. Due to the unavailability of explicit "influencer and follower" associations within the original citation graph, we intend to investigate this particular attribute by examining the citation relationships between the articles themselves.

We could define each author as a "hyper-node", which consists all his articles in the citation graph. In the graph of hyper-nodes, we establish the edge between hyper-nodes as the citation relationship between an article authored by one individual and an article authored by another individual and its weight as the total number of citations. The construction of graph is presented as follows :

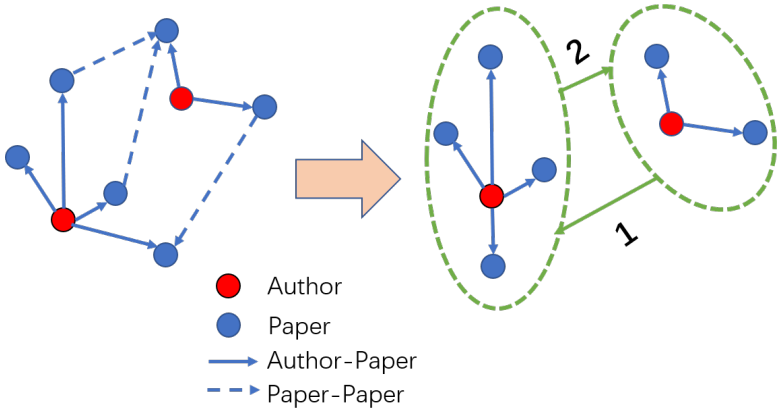
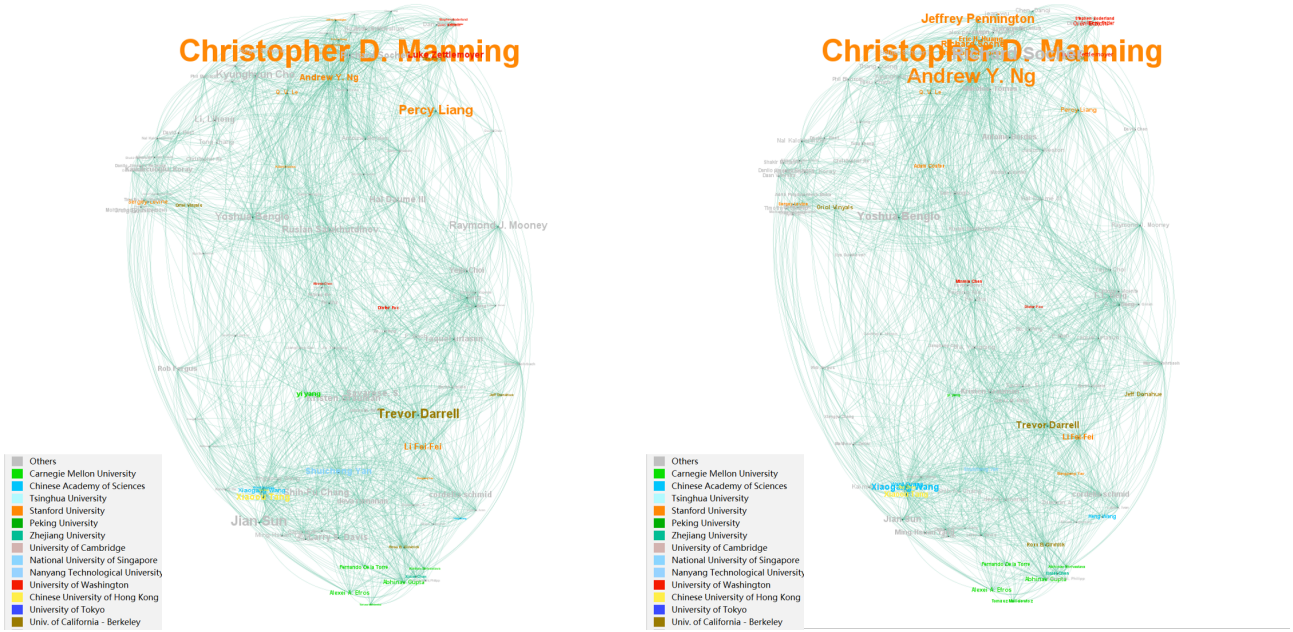


Figure 4: Construction of hyper-node graph

To determine the most influential authors, we employ a similar process as outlined in the preceding section. However, we utilize weighted in-degree centrality instead of conventional in-degree centrality due to the significance of weight information considered as influence. The following figures 5 demonstrate the results. To achieve a more concise visualization, we opted to extract a subset of nodes from our dataset with the highest in-degree centrality. This approach was employed due to the large number of authors involved in the study.

From the figures, we could easily find prominent researchers in the field of artificial intelligence, like Christopher D.Manning, Joshua Bengio, Fei-Fei Li, etc. Based on the presented figures, the dimensions of node labels



(a) Sorted by betweeness centrality

(b) Sorted by weighted in-degree centrality

Figure 5: Visualizations of author nodes in Citation Graph

show minimal variation across diverse metrics. This observation signifies the absence of contentiousness regarding influential authors.

The color of each node in the figures indicates the respective institutional affiliation of the authors. It is evident that several prestigious institutions, such as Stanford University and Carnegie-Mellon University, have already established highly effective academic exchange systems, which consequently contribute to their prominence in research endeavors. Conversely, nearly fifty percent of the researchers are affiliated with other institutions, thereby highlighting the dynamic and diverse nature of artificial intelligence research."

3.2 Community Detection

Upon completion of data cleaning, our investigation focuses on identifying influential communities through the utilization of a graph structure. Prior to conducting this analysis, it is imperative to construct a graph that accurately simulates the relationships among articles. Within our project, we limit our graph construction to articles published between 2010 and the present, with each article serving as a distinct node. For efficient data storage, we employ the graph structure provided by the `networkx` library. Whenever an article B cites article A, an edge is introduced from article A to article B. To determine the weight assigned to the edge connecting article A to article B, we explore three distinct methodologies.

- Firstly, we assign a weight of 1 to the citation from article A to article B. However, a significant challenge arises due to the fact that numerous articles cite other works but provide only minimal reference to the content of those articles, resulting in negligible correlation among these articles.
- Secondly, to determine the edge between article B and article A, we calculate their similarity based on the **abstracts** using the function `similarity` in `spacy` library. Nevertheless, this approach encounters two issues. Firstly, a considerable portion of our dataset comprises articles lacking abstracts (up to 24% of the total). Secondly, when two articles do have abstracts, the similarity values calculated are predominantly

above 0.9. This indicates that the abstracts of these articles are written in a relatively similar style, thereby rendering this metric ineffective for measuring article similarity.

- Thirdly, as an alternative, we opt to measure the similarity between **titles** rather than abstracts. We believe this approach to be more efficient, as the similarity values span a range of 0.5 to 1. Consequently, we have selected this method for subsequent stages of our project.

Once the construction of the citation graph is completed, our focus shifts towards the detection of communities within this graph. For this purpose, we employ the **Louvain algorithm**. This algorithm optimizes a quality function that evaluates the modularity of network partitioning. It achieves this by iteratively merging nodes into communities, taking into account their connectivity patterns, with the objective of maximizing modularity at each step. The iterative process continues until further improvement in modularity becomes unattainable, ultimately yielding a partition of the network into discrete communities.

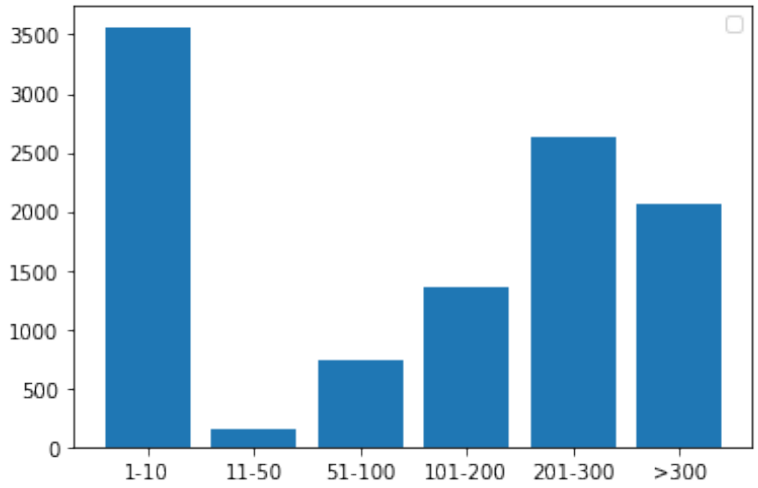


Figure 6: The variable under consideration is the count of articles within a given community. Specifically, columns 11-50 denote the number of articles attributed to communities containing 11 to 50 articles.

For the implementation of the algorithm in Python, within the `networkx` library, we simply utilize the pre-existing function `nx.community.louvain_communities`. Upon executing the algorithm, it was observed that a substantial number of exceedingly small communities emerged. Consequently, to facilitate the study of influential communities, we impose a restriction by focusing solely on communities comprising more than 50 nodes.

Upon identification of influential communities, a pertinent inquiry arises regarding the **extraction of information** from each community. This entails investigating key aspects such as the primary research direction pursued within the articles comprising each community and the research methodologies employed within these communities. For instance, our objective was to identify communities that focus on specific research domains, such as *image segmentation* problems within the field of computer vision or *machine translation* problems in natural language processing (NLP). To accomplish this in our project, we employed various methods to extract the desired information. The methods employed can be enumerated as follows.

Initially, we leveraged the **keywords** provided within the articles. These keywords were collected and compiled into a list, from which we extracted the ten most frequently occurring words. Regrettably, the obtained results were not deemed satisfactory. The frequently appearing words predominantly pertained to broad and general topics such as *computer vision* and *deep learning*, failing to shed light on the specific problems addressed within each community. Please refer to Table 4 for an illustrative example.

Subsequently, we endeavored to extract meaningful insights from the problem content, utilizing the **titles** of the articles as a key resource. Within our project, we adopted the following methods to accomplish this task:

Table 4: Example of information extraction from the community using **keyword** method. The community comprises a total of 499 articles. Analyzing the titles allows us to deduce that the primary focus of this community is *sentiment analysis, information retrieval,* etc. However, relying solely on the frequency of keywords does not provide a comprehensive understanding. In this case, the phrase *"sentiment analysis"* appears only 5 times in total.

Example of titles	High-frequency keyword
Modeling Human Reading with Neural Attention.	sentiment analysis
Joint Language and Translation Modeling	(5), novel approach
with Recurrent Neural Networks.	(4), information
A Joint Segmentation	retrieval (4),
and Classification Framework for Sentiment Analysis. .	training data (3)
Recurrent Neural Network	
for Text Classification with Multi-Task Learning. .	
Parsing Natural Scenes	
and Natural Language with Recursive Neural Networks.	

• The initial approach employed involved selecting a **sliding window** of **size 2** within the title. For instance, taking the title *"Clustering by Low-Rank Doubly Stochastic Matrix Decomposition."* the resulting terms obtained were *"Clustering by," "by low-rank," "low rank doubly," "stochastic matrix," and "matrix decomposition."*. Subsequently, by analyzing the terms present in the headings within the same community, we identified the top 10 most frequently occurring terms. The outcomes yielded satisfactory results. Through these 10 terms, we were able to deduce the prevalent concerns shared by the community. However, a limitation of this method is that numerous prepositions are attached to high-frequency terms. For example, phrases such as *"for Visual"* and *"tracking with"* frequently emerge, making it challenging to devise an automated approach to precisely identify the specific problems or topics of interest to the community.

Table 5: An example of information extraction from the community using the titles and the noun phrases, with example in Table 4. However, relying solely on the frequency of noun phrases yields somewhat better results compared to the previous method but remains less satisfactory. The term *"word embedding"* appears a total of 12 times. The term *"sentiment analysis"* does not appear.

Example of titles	High-frequency word
Modeling Human Reading with Neural Attention.	word embeddings
Recurrent Neural Network	(12), word (6),
for Text Classification with Multi-Task Learning.	neural networks (6),
A Joint Segmentation	machine translation
and Classification Framework for Sentiment Analysis. .	(6), text
Joint Language and Translation Modeling	classification (6)
with Recurrent Neural Networks.	
Parsing Natural Scenes	
and Natural Language with Recursive Neural Networks.	

• The second approach employed in our study involves the utilization of the `noun_chunks` function from the `spacy` library to extract **noun phrases**. For instance, considering the title *"Query Adaptive Similarity for Large Scale Object Retrieval"*, the extracted noun phrases include *"Query Adaptive Similarity"* and *"Large Scale Object Retrieval"*. Subsequently, for each title, after extracting the noun phrases, we identified the top 10 most frequently occurring noun phrases within that sentence. However, it was observed that most of these noun phrases appeared a maximum of three times. This observation can be attributed to the fact that although the problems discussed in different articles are quite similar, they are approached and resolved differently. For example, when comparing the titles *"Recognizing Named Entities in Tweets"* and *"Named entity recognition in tweets: an experimental study,"* we observe that both address the issue of named entity recognition in tweets,

but the wording and word positioning differ significantly. Additionally, in several instances, the `noun_chunk` function fails to recognize proper nouns, particularly when they are utilized. For instance, considering the title *"Deep Learning for Chinese Word Segmentation and POS Tagging,"* the extracted noun phrases are *"Deep Learning"*, *"Chinese Word Segmentation"*, and *"POS Tagging"*. However, if we modify the capitalization to *"Deep learning for Chinese word segmentation and POS tagging"*, we only obtain *"Chinese word segmentation"* and *"POS tagging"*. To address these limitations, we introduce a modification wherein all extracted noun phrases are converted to **lowercase**, and then all **two-word combinations** (preserving the word order) within the noun phrase are selected. For instance, for the phrase *"Chinese Word Segmentation"*, the resulting combinations are *"chinese word" "word segmentation"* and *"chinese segmentation"*. After obtaining these terms, we identify the top 10 terms with the highest frequency. This revised method yields more satisfactory outcomes and higher levels of accuracy compared to the first method.

Table 6: The example presented in Table 4 demonstrates the utilization of **noun phrases** and word splitting method. This approach proves to be superior to the previously discussed methods as it yields a higher frequency count of 25 occurrences for the term *"word embedding"* and 17 occurences for the term *"sentiment analysis"*.

Example of titles	High-frequency word
Modeling Human Reading with Neural Attention. Recurrent Neural Network for Text Classification with Multi-Task Learning. A Joint Segmentation and Classification Framework for Sentiment Analysis. . Joint Language and Translation Modeling with Recurrent Neural Networks. Parsing Natural Scenes and Natural Language with Recursive Neural Networks.	neural networks (30), word embeddings (25), sentiment analysis (17), word representations (12), machine translation (11)

The third approach employed in our study is based on the observation of common patterns within the titles of articles. Most titles can be divided into sections representing the **problem** and the **method**, which are typically separated by conjunctions such as *"with," "for," "using," "in," "by,"* etc. Therefore, our approach involves preparing **templates** corresponding to these conjunctions and the positions of the problem and method sections mentioned earlier. Using these templates, we extract the problem and method sections from the titles within a community, creating two lists: one for strings of methods and another for string of problems. Within each list, we convert the words to lowercase. For each string, represented as a sequence of numbered words in the order $w_0, w_1, ..., w_{n-1}$, we extract terms of the form $w_i w_j$ where $0 < j - i \leq 3$, aggregating them to form a new list. From these two new lists, we select the top 10 phrases with the highest occurrences.

Table 7: The illustrative instance presented in Table 4 showcases the implementation of problem-method analysis. This approach exhibits superiority over previously examined methods as it effectively distinguishes between the problem and method within an article. Furthermore, it is noteworthy that this method exhibits a notable frequency of occurrence for crucial words, further reinforcing its efficacy.

Example of titles	Problem	Method
Modeling Human Reading with Neural Attention. Recurrent Neural Network for Text Classification with Multi-Task Learning. A Joint Segmentation and Classification Framework for Sentiment Analysis. . Joint Language and Translation Modeling with Recurrent Neural Networks. Parsing Natural Scenes and Natural Language with Recursive Neural Networks.	word embeddings (18), sentiment analysis (15), machine translation (11), dependency parsing (11), word representations (11)	neural networks (20), language models (5), word embeddings (5), deep learning (5), recursive neural network (4)

This method allows us to capture important information words that are separated by other words. For example, with the phrase *"Word Recognition and Segmentation,"* we can extract two significant phrases: *"Word Recognition"* and *"Word Segmentation."* By separating the methods and problems, we can identify the important problems studied within the community, as well as the methods employed. Through experimentation, we observed that words in the method section have a lower repetition frequency (approximately 4.5 times for the most frequent words), while words in the problem section exhibit higher repetition frequency (10-30 times for the most repeated words). This suggests that within a community, problems are approached and studied in various ways.

3.3 Analysis on Academic Competitiveness of Institutions

For individuals who aspire to pursue AI research, the selection of an institution holds significant importance, as it has a direct influence on their potential for future research excellence and the establishment of a robust collaborative network. While the original CS Ranking Dataset has provided ranking information as well as faculty numbers and counts, these details prove insufficient for conducting a comprehensive evaluation of academic power.

Therefore, we aim to have a comprehensive analysis on academic performance of institutions based on the community information in Section 3.2. To be more precise, we undertake the classification of each research articles based on the affiliations of its authors and the specific scholarly community to which it belongs. The following graph7 visualizes the academic power of several universities on several domains.

Based on the bar chart, the relative strengths of universities within specific domains can be readily discerned, thereby providing a more comprehensive understanding than relying solely on a singular ranking. Furthermore, it is observed that domains associated with NLP, such as Sentiment Analysis, Machine Translation, and General Artificial Intelligence (specifically Variational Inference), exhibit a higher global output of research papers compared to domains related to CV, namely Action Recognition and Object/Saliency Detection. This phenomenon aligns with the preceding analysis conducted in the previous section.

Moreover, We aim to use linear regression to study how each community (research domain), as well as the faculty number and count, contributes to the final CS ranking. In practice, To achieve this, we combine the aforementioned count vector with the original faculty number and count data from the CS Ranking Dataset, thereby creating a feature vector for each institution. The dependent variable in the linear regression model is the ranking of each institution.

However, the obtained results from linear regression analysis are deemed unsatisfactory. The coefficient of determination, denoted as R^2, which serves as an indicator of the performance of the linear regression model, falls below the threshold of 0.5 and it varies by the selection of test samples. Furthermore, the presence of numerous negative coefficients contradicts the underlying assumption that an increase in the number of articles would correspondingly result in higher academic influence and ranking.

Based on our analysis, a potential factor contributing to the failure is the lack of sufficient accuracy in community detection, as approximately forty percent of papers are categorized as singleton communities, classified as papers in `Other`research domains.

4 FUTURE WORK

Our project has achieved global success in conducting an in-depth analysis of the influence of papers, authors, institutions, and communities in the field of AI research based on our citation graphs. However, there remain several areas that need further investigation and improvement in future studies:

- The filtering procedure employed in the present study relies on the selection of a specific library of venues. In this project, our primary focus is on incorporating articles sourced from leading AI conferences.

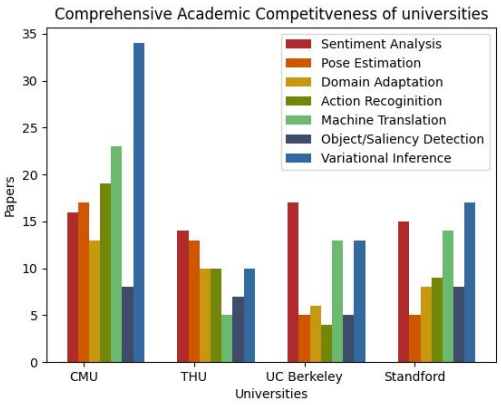


Figure 7: The academic power of several universities on several domains

Nevertheless, it is worth noting that a substantial quantity of articles exists within other journals as well. The omission of this particular subset of data introduces a potential source of bias to our dataset. Consequently, it is imperative to establish a more nuanced criterion in order to establish our dataset more objectively.

- In the context of dataset integration, certain information has been omitted for the sake of simplicity. In the CS Ranking Dataset, researcher names affiliated with each institution have been excluded. Similarly, in the Citation Network Dataset, additional details such as the field of study of each paper and its weight have been discarded. It is our aspiration to incorporate this omitted information in future studies, as it holds the potential to yield further insights and enhance the depth of analysis.

- In our pursuit of identifying significant information within a community, we have employed numerous methods, resulting in improved outcomes. However, one persisting challenge lies in the absence of an automated approach to characterize articles, necessitating manual determination of their relevance within important communities. To address this issue, we propose the utilization of a comprehensive and well-labeled dataset for training purposes. Furthermore, our current algorithms primarily rely on statistical analysis of phrase occurrences, without accounting for the semantic understanding of words. As a consequence, distinct meanings of closely related terms, such as *"3D reconstruction"* and *"3D restitution,"* are treated as separate occurrences. To mitigate this, a model capable of capturing semantic information within sentences is required.

5 CONCLUSION

In this project, we present an approach utilizing a citation graph to investigate advancements in AI research over recent years. Our methodology involves the integration of two distinct datasets, the Citation Network Dataset and the CS Ranking Dataset, to construct the citation graph. Multiple techniques are employed to carefully select and merge the desired data, with the aim of preserving a maximum amount of information throughout the combination.

Through the analysis on most influential articles and authors, it has been determined that the majority of influential research papers originate from the field of natural language processing (NLP). This observation suggests that NLP demonstrates a higher pace of technological progress in comparison to other domains. Furthermore, we have observed substantial technological interplay between NLP and General Artificial Intelligence

(AI), as well as between Computer Vision (CV) and General AI. It is noteworthy that unlike the present scenario where the transformer model, initially devised for NLP tasks, has found widespread adoption in Computer Vision, there is a lack of substantial technological exchanges between these two domains.

Through the identification of communities, it becomes evident that these groups frequently exhibit a concentrated focus on a limited number of challenges within specific domains, such as computer vision (CV), natural language processing (NLP), and optimization. For instance, within the realm of CV, the problem of image segmentation garners significant attention, while NLP encompasses machine translation and entity recognition, among others. Moreover, an observation can be made regarding the word distribution in the method section of article titles, wherein words tend to exhibit repetition with relatively lower frequencies (typically ranging from 3 to 5 occurrences). Conversely, the problem component of titles demonstrates a higher frequency of repetition (typically appearing around 20 to 30 times). This phenomenon suggests that within a given community, individuals frequently employ multiple methods to address a particular problem.

Finally, we present a thorough analysis of the academic competitiveness of each institution, drawing upon the findings from our preceding investigations. This examination offers valuable insights that can facilitate prospective AI researchers in making informed decisions regarding their career paths.

REFERENCES

[1] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.