Summer Research Report

Name: Shimin Luo
E-mail: shiminluo.cs@gmail.com
or: 12012939@mail.sustech.edu.cn
Summer research duration: 2023.6.17-2023.8.18

I. Paper reading Summary

A. Read Papers List

1. Systematic Inequality and Hierarchy in Faculty Hiring Networks [1]
2. Subfield Prestige and Gender Inequality among U.S. Computing Faculty [2]
3. Science of Science [3]
4. Mining Social Networks for Targeted Advertising [4]
5. Mining (Social) Network Graphs to Detect Random Link Attacks [5]
6. LightFace: A Hybrid Deep Face Recognition Framework [6]
7. HyperExtended LightFace: A Facial Attribute Analysis Framework [7]
8. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments [8]
9. Ethnea – an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database [9]
10. Gender differences in scientific productivity: a persisting phenomenon? [10]

B. Summary and Acquirement

The first three articles [1] [2] [3] for me are **introductory pieces in the field of graph data mining**. Through reading them, I have gained a fundamental understanding of the research area of graph data mining and have come to know what needs to be extracted from data for our summer research. I have also learned about the analytical perspectives and methods that aid in our data analysis and mining.

Although these three articles have different emphases – one focusing on the impact of institutional prestige on faculty recruitment through a hiring network, one exploring gender inequality among faculty in subfields, and the third investigating how scientific research can be enhanced through analyzing the staff and their publication achievements – they all analyze a particular form of network data, namely graph data. Among them, there are some innovative and pivotal points that I find worthy of learning and emulating: method like ”minimum violation ranking,” the emphasis on studying subfield dimensions, and employing a ”two-stage approach” for categorizing stages in faculty careers and so on. These perspectives and methods have greatly enriched my understanding.

The following two articles [4] [5], driven by **my curiosity about the potential applications** of graph data mining, focus on implementing specific algorithms (and introducing custom definitions as needed) on data structured as graphs to achieve a certain functionality or objective. One article involves utilizing graph data mining algorithms to create a universal user recommendation system. This research topic has been quite popular in the AI field in recent years, but what makes this article intriguing and innovative is that it achieves its goals through non-AI methods, which I found to be very interesting. The other article employs graph data mining algorithms to detect Random Link Attacks, which is also a universal approach applicable to various types of attacks, not limited to their content. Unlike NLP method, which is typically used for processing text and not suitable for analyzing speech patterns, this article’s proposed idea holds significant importance as it addresses a drawback of the AI methods that the models often require adjustments.

The following articles [6] [7] [8] [9] were read **based on the requirements of our research**. While seeking high-quality tools for gender identification based on facial images or names, I read these articles. I delved into understanding their recognition principles and thoroughly examined and analyzed their experimental results. I believe that the recognition tools they developed, HyperExtended LightFace and Ethnea, could be utilized for our research.

The last paper [10] was read when **contemplating how to advance our research**. The innovative aspect of this article is its categorization of researchers into an ”established generation” and the ”youngest generation” based on age or qualifications. It aims to investigate whether gender disparities observed in the ”established generation” persist in the ”youngest generation.” I believe that our research analysis can also follow this approach. By combining estimated age, gender, and race, we can explore whether academic performance differences exist among the older and newer generations of researchers in terms of their gender distribution, and racial diversity.

II. Research Progress Summary

A. Problem

Initially, the problem is that, according to the faculty list provided by Professor Kevin, using the names of faculties and their affiliated institutions, conduct searches on OpenAlex to extract publication information for the corresponding faculties. The whole information will be imported into

Neo4j graph database for exploration through graph data mining.

However, during the process of retrieving publications, I noticed that when using the same name and institution for searches in OpenAlex, I could obtain multiple results. Validating these results against the DBLP database revealed that, **some of these diverse returns were actually pointing to the same individual**. This indicated certain shortcomings in the OpenAlex database in terms of data accuracy. Additionally, due to the issue of non-standardized name formats, **many faculties could not be found through original name searches on OpenAlex.**

As a result, our subsequent research direction gradually shifted. We began to focus on methods like **entity-merging** or **name-matching**, coupled with the comprehensive utilization of other databases or search engines such as DBLP and Google Scholar, to gather faculties' publication information from **various sources**. I also sought ways to **predict faculties' gender, race, and age**, combining this information with existing data for further analysis.

B. Phased Goal, Approaches, Progress, and Results

1) Pre-task:

· **Goal:** Gaining an understanding of research topics in the field of graph data mining, learning about various algorithms, and acquiring a basic knowledge in using the Neo4j graph database.

· **Approach:**

1. Read the paper [1] provided by Professor Kevin and summarize it.
2. Reproduce the MVR (Minimum Violation Rank) algorithm mentioned in the paper.
3. Search another article in this field to further expand my knowledge.

· **Progress & Result:** I gained a fundamental understanding of the content within this field. Moreover, the MVR algorithm that I reproduced exhibited better performance compared to the results provided in the supplementary materials of the original article. (For specific details, refer to another submitted report "Task 1 report.pdf").

2) Week 1-2:

· **Goal:** Familiarized myself with the data format and API of OpenAlex, attempted to retrieve faculties' publications using the API, and import the whole useful information into the Neo4j graph database.

· **Approach:**

1. Access OpenAlex official documentation, review data format attributes, and understand how to use the API.
2. Meanwhile, do some faculty-name cleaning and use API to do institution-name normalization to the original data, also to find "institution_id" for later use.
3. Use the "author name" and "institution_id" to find the "author_id" through API. Then, use "author_id" to collect ones all publications.
4. Convert relevant data to the corresponding format, and load them into Neo4j. Create three nodes: "faculty", "institution" and "publication", and establish the relations between them.
5. Turn to DBLP to verify certain data problems encountered in OPenAlex, and contemplated using data from DBLP as criteria for entity-merging decisions.

· **Progress & Result:**

1. Two methods were employed to obtain the "author_id". The improved version now replaces matching the search results of institution names, to using faculty names and with "institution_id" to conduct searches. This has significantly reduced the time required and improved accuracy to a great extent.
2. Completed the **entire process**, starting with 'faculty.json' file processing, gathering all required information, and generating files (3 node files, 2 relationship files) for Neo4j graph data import. Also, wrote code to import current results into Neo4j graph database.
3. Identified issues within the Open Alex database: for instance, searching the same name within the same institution yields multiple results. After validation against the DBLP database, it became evident that **many of these results actually point to the same individual.** However, Open Alex mistakenly separates them into distinct entities.

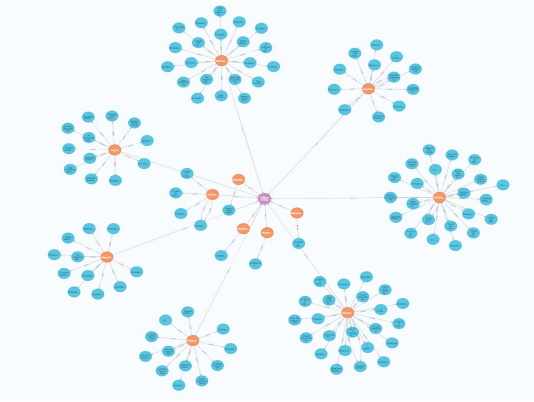

Fig. 1: Small portion graph display

Fig. 2: Nodes properties

3) Week 3-4:

· **Goal:** Evaluated the quality of OpenAlex data and investigated whether Semantic Scholar or DBLP could be used as alternative databases.

· **Approach:**

c1. Conducted code-based experiments to assess the rate of multiple results obtained when searching for individuals

with the same name and institution in OpenAlex.
2. Visited the Semantic Scholar and DBLP official website to familiarize myself with its data structure and API, conducted experiments to analyze its data quality(Eg. the missing rate of affiliation, whether a search for an individual return different records), and determined whether it could be suitable for our use.
· **Progress & Result:**
1. Using the faculty names after cleaning, search for authors in OpenAlex. Among a total of 5594 faculties, only 3477 faculties (**62.15%**) can be **matched**. Among the matched records, only **13.3%** of individuals can be **uniquely identified**, while the rest have the possibility of being assigned multiple IDs. The low rate of name matches may be due to variations in name formats. The latter reveals a significant issue with the OpenAlex data. If we intend to use it as a source for obtaining publication data, we **must do entity-merging**.
2. The same institution might be assigned into different "institution_id", which is also a problem shown in OpenAlex.
3. Through multiple experiments and averaging, **Semantic Scholar** shows a notably **higher faculty search rate at 86.36%** compared to OpenAlex. The search results are categorized by subject and individual perspectives, seemingly obviating the need for data entity-merging. However, a critical drawback is **the absence of official affiliation records** for faculty. Only 442 out of 5594 (**0.75%**) faculty members had provided affiliation information, as it relies on manual input during their Semantic Scholar account registration. Therefore, we can't find our faculties and publications through this database.
4. **DBLP** also demonstrates a **high name search rate at 91.92%**. However, **only 49.93% of faculty entries include affiliation records.** Additionally, DBLP's faculty search scope is **limited to the field** of computer science. If we intend to expand our analysis to other disciplines, DBLP cannot be used.

4) Week 5-7:
· **Goal:** Explore ways to determine the gender and race of faculty members.
· **Approach:**
1. Read numerous papers and conducted extensive research on currently reliable image-based gender and race recognition tools. Specifically, delved into [6] [7] [8] these selected articles, understood their principles, and examined the reliability based on their experimental outcomes to determine their suitability for our use.
2. Recalled the paper [2] mentioned tools for determining gender based on names. Consequently, I went on to search for and explore the sources and reliability of these methods: "gender-guesser" package in python and "Ethnea" [9] .
3. Apply the methods found above and explore their feasibility. Try to combine them with the process of searching for faculties.
4. Try to use "coauthors" information in "Scholarly" to review faculties' institutions. (However, this idea was deemed unreasonable during the discussion.)
· **Progress & Result:**
1. After addressing environmental problems, **"deepface" for image recognition, gender-guesser and Ethnea for name-based identification accurately determine gender and ethnicity**. I've also deployed "deepface" on a remote server for efficient large-scale image data recognition.
2. **Trying various approaches to obtain faculty images**. These approaches include but are not limited to, using a web crawler through the Google API (encountering Google search block) or visiting faculty homepages (hard to locate the images), utilizing libraries to bypass Google Search block (no picture information return), and employing "Scholarly"(also encountering Google Search block).
3. After Professor Kevin said that maybe we could pay for Google Search API, I used **"deepface" and Google Search API to accomplish the process** from faculty searching, finding picture-url and retrieving, to putting the pictures into "deepface" to recognize. (Need to corporate with Alicia later to implement the whole process of name-matching and identification).
4. For those **faculties who haven't pictures** on Google website, I code to find their gender and race by Using 'request' in Python to visit the tool **"Ethnea"**.

5) Week 8:
· **Goal:** Find approaches to recognize faculties' age. Reach out to Alicia to combine her name-matching code, finalizing the process of gathering faculty publications and picture-url through search and subsequently determining their gender, ethnicity, and age.
· **Approach:**
1. Conducted wide-range paper searches and extensive research to identify suitable methods for obtaining faculty ages.
2. Communicated with Alicia and focused on understanding her ideas and merging codes.
· **Progress & Result:**
1. Many articles treat age data as pre-existing information without providing details on the acquisition process. Then, I came out with an idea: faculty members usually generate their first publication during their master's or doctoral studies. **Assuming an approximate age of 25 for their first paper, we can employ this to gauge their current age.** However, this approach might need to be implemented after collaborating with Alicia to complete the acquisition of faculty publication information.
2. Understand Alicia's idea of using Google search + DBLP + OpenAlex to search faculties, and merge some of our codes.
3. After reading this week's paper and taking a quick look at the attributes and quality of the new data, I **proposed the whole idea of obtaining faculty information, identifying some properties (gender, race, age), and conducting**

analysis. (See "Whole idea & Future Work" section)

C. Whole idea & Future Work

For the faculty search and publication retrieval, follow the approach Alicia used: **Google Search + DBLP + OpenAlex**. For faculty found through **Google Search with obtainable images**, employ **"deepface"** to determine their gender and ethnicity based on their images. For faculty without facial images available, utilize the **name-based method**. Maximizing the use of facial recognition-based determination wherever possible can enhance accuracy. And **using the earliest publication's date**, assume faculty members were around 25 years old when it was published, **estimating their current age.**

Combining this data with the predicted age, gender, and ethnicity information, we analyzed **academic performance metrics such as position, the number of publications, total citations, or the ratio of citations to publications. Future work could begin by investigating potential academic performance disparities based on gender or ethnicity within the established generation. If such disparities exist, the next step would involve exploring whether these differences persist within the younger generation. In cases where disparities do not exist, research into potential differences in ethnicity and gender between young and established faculty members could be pursued.**

D. Assessment & Reflection

This is my first time engaging in the field of graph data mining and data crawling for retrieval information. Throughout this process, every encounter with a new task demands a certain amount of time for **self-directed learning** and **explorations with various approaches**. The methods of exploration are diverse, and the mode of thoughtful analysis requires **multidimensional thinking** and brainstorming. The manner in which problems are solved and analyzed is also quite **open-ended**.

This contrasts with my previous approach to researching problems in the field of database systems, such as optimizing specific query algorithms or designing a distributed framework for outlier detection in edge computing scenarios. In my earlier research endeavors, the majority of the process involved surveying existing algorithms and technologies, identifying areas for optimization or innovation, proposing personal designs, and implementing them. The process also included theoretical derivation and experimental validation of correctness and optimization. However, in this summer research project, I've realized that when it comes to graph data mining, **both what to mine and how to mine it are quite open-ended and have multiple interpretations**. Even the research ideas I had in mind could change due to issues encountered in earlier steps, such as the data acquisition process. **In other words, before getting enough information, we cannot accurately tell what can be mined and the expected outcomes for each subsequent step.**

Therefore, during this summer research, it feels like **our goals have been gradually evolving**. Initially, our goal was to retrieve publications based on faculty names and institutions from OpenAlex, importing relevant information into Neo4j for analysis and exploration. **As we encountered issues with data quality in OpenAlex and faculty name matching, our focus shifted towards seeking other reliable data sources and name-matching tasks.** We spent a significant amount of time on this, which seems to deviate from our initial research direction. However, in reality, we were solving encountered problems, ensuring a reliable data source for subsequent analysis and exploration. The later steps of obtaining faculty gender, ethnicity, and age also contribute additional information elements for data analysis.

Overall, the process of **data acquisition** turned out to be **more complex and challenging than I initially anticipated**. It feels like our team **has just about tackled the data acquisition challenge**, and yet we're approaching the end without having fully integrated and imported the data into the Neo4j graph database for research. In general, we still **have a distance to go before reaching the ultimate research outcomes.** However, during this process, I've learned how to analyze and address problems from **multiple angles**. I've gained insights into **autonomous exploration** and the **assessment of the rationality and feasibility of methods**. It has also enhanced my ability to **independently contemplate a problem, consider what can be done at present, and anticipate future analyses and explanations**. Also, I've improved my skills in paper searching and reading speed, **becoming adept at conducting broad searches in related fields and swiftly reading through papers**.

Through this summer research, **I've realized that my performance and abilities still have room for improvement. I'm grateful for Kevin's patient guidance and assistance, particularly for the understanding during times when personal reasons required a slight slowing of research progress. I also believe that the format of group study and the weekly meeting arrangement are also very effective. One aspect for improvement could be enhancing group collaboration or dividing tasks. During the week, teammates can also use Google Chat to share progress or ideas, so that they don't have to do some duplicate works and they can discuss some problem solutions. This approach can greatly enhance the efficiency of team research.**

REFERENCES

[1] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005, 2015.

[2] Nicholas Laberge, K. Hunter Wapman, Allison C. Morgan, Sam Zhang, Daniel B. Larremore, and Aaron Clauset. Subfield prestige and gender inequality among u.s. computing faculty. *Commun. ACM*, 65(12):46–55, nov 2022.

[3] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359(6379):eaao0185, 2018.

[4] Wan-Shiou Yang, Jia-Ben Dia, Hung-Chi Cheng, and Hsing-Tzu Lin. Mining social networks for targeted advertising. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 6, pages 137a–137a, 2006.

[5] Nisheeth Shrivastava, Anirban Majumder, and Rajeev Rastogi. Mining (social) network graphs to detect random link attacks. In *2008 IEEE 24th International Conference on Data Engineering*, pages 486–495, 2008.

[6] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5, 2020.

[7] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4, 2021.

[8] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, October 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.

[9] Vetle Ingvald Torvik and Sneha Agarwal. Ethnea – an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. March 2016. International Symposium on Science of Science ; Conference date: 22-03-2016 Through 23-03-2016.

[10] Pleun van Arensbergen, Inge van der Weijden, and Peter van den Besselaar. Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, 93(3):857—868, December 2012.