Report of Task 1: Graph Data Mining

Name: Shimin Luo
E-mail: 12012939@mail.sustech.edu.cn
Task Duration: 5.1-5.8

I. Paper Reading and summary

A. Brief Introduction & Summary

The paper *Systematic Inequality and Hierarchy in Faculty Hiring Networks* indicates that faculty hiring success reflects social hierarchy, mainly from institutional prestige in academia. The prestige encompasses both merit-based and non-merit-based factors, such as social status or geography. Nevertheless, non-merit factors also have a significant impact, decreasing the correlation between prestige and merit and contributing to social inequality in faculty hiring.

The authors analyzed faculty hiring patterns in top-ranked American universities using a data-driven method that takes non-merit factors into account. They found **significant unequal levels and centralized structures in the hiring networks**, with highly connected individuals or institutions dominating the network. Higher institutional prestige leads to better faculty placement and more influence within a discipline. This study highlights the inequality and hierarchy in the faculty hiring network and offers important insights for promoting educational equity.

B. Innovation Point

Traditional ranking methods (e.g. U.S. News & World Report) overlook collaborations and actual outcomes, while focusing on inputs like reputation and wealth. Faculty hiring networks, in contrast, account for these factors and offer a more effective and quantitative way to measure prestige impact and identify hiring patterns, illuminating the balance between merit and status.

C. Graph Data Analysis

1) Basic Analysis: Direct analysis based on raw data

a) Faculty production is highly skewed, 25% institutions produce the majority (71-86%) of tenure-track faculty.

b) Disciplinary size is not a determining factor of hiring success rates, as demonstrated by statistical tests (KS test).

c) The Gini coefficient for faculty production is high (0.62-0.76), indicating strong inequality across disciplines.

d) The existence of strong inequality among the top faculty producers implies that non-meritocratic factors such as social status influence placement rates, and cannot be solely explained by meritocracy.

2) Aggregate Analysis:

a) Minimum Violation Ranking: To extract a consensus ranking that is most closely related to the social hierarchy.

Definitions:

1. *p:* Prestige ranking of vertices, where p_u = 1 is the highest-ranked vertex.
2. *r:* Hierarchy's strength, the fraction of edges that point downward ($p_u \leq p_v$).

Result Analysis:

1. Steep prestige hierarchies exists, only 9-14% of faculty are placed at institutions more prestigious than their doctorate.
2. Strong preference for hiring faculty with prestigious doctorates is indicated by hierarchies that are 19-33% stronger than expected from observed inequality in faculty production rates alone (Monte Carlo).

b) Changes-in-rank: $p_{hiring} - p_{doctoral}$

1. Each distribution is significantly right-skewed.
2. Only the top 18 to 36% of institutions are net producers of within-discipline faculty.
3. Faculty trained at higher-ranked institutions tend to make smaller moves down the hierarchy than those trained at lower-ranked institutions.
4. Female graduates generally place worse than male graduates from the same institution.

c) Core-periphery Pattern:

1. Academia exhibits a core-periphery pattern, where higher prestige institutions occupy a more central and influential network position.
2. This pattern has implications for the free exchange of ideas, as ideas originating from the high-prestige core spread more easily throughout the discipline.
3. The centrality of high-prestige institutions enables them to have a significant influence over research agendas, communities, and departmental norms.

D. Prediction

Using *prestige hierarchy* and *Changes-in-rank* for modeling predictions, the result shows that **the doctoral prestige alone better predicts ultimate placement than authoritative rankings** (e.g. U.S. News & World Report). This holds true in terms of both precision and universality across various academic disciplines, as measured by the AUC.

E. Discussion

The study highlights the significant role of institutional prestige in faculty hiring and the existence of gender disparities in placement quality within certain academic fields.

Similar patterns were observed across different disciplines, indicating the fundamental nature of strong prestige hierarchies. The findings call for identifying factors that distinguish exceptional institutions and individuals in faculty hiring and the adoption of data-driven methods for evaluating academic activities. The study also raises questions about the efficacy of the academic system and suggests potential applications of the methods used in studying other sectors.

II. Reproduce the data with Neo4j

A. Create Graph Database

The properties of node and edge can be observed in Fig.1. (In a Node, u is institution's rank. pi represents the score.) And the total number of nodes and Edges is equal to the description of the supplementary material.

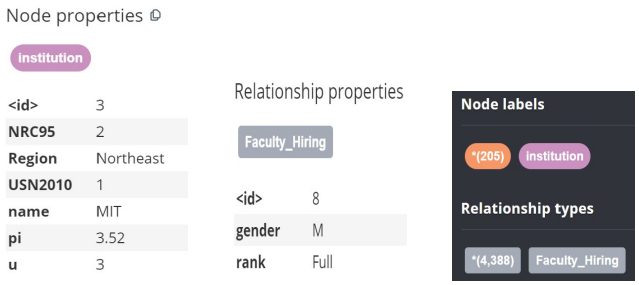

Fig. 1: Properties and Count

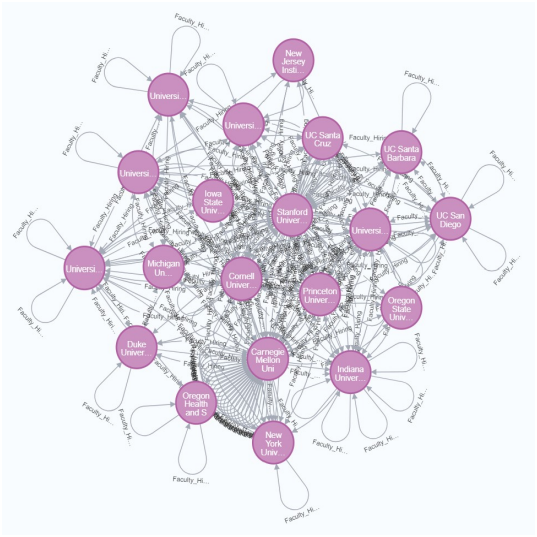

Fig. 2: Partial Graph

B. Queries

Mainly focus on verifying some data analysis results mentioned in the paper.

1) : Result is consistent with the statement: "9 to 14% of faculty are placed at institutions more prestigious than their doctorate"

```
MATCH ()-[r]→()
WITH COUNT(r) AS total_count

MATCH (n1:institution)-[r]→(n2:institution)
WHERE toInteger(n2.u) < toInteger(n1.u)
RETURN COUNT(r) * 100.0 / total_count AS percentage
```

	percentage
1	11.030082041932543

Fig. 3: Query1 & Result

2) : Result is consistent with the statement: "only 25% of institutions producing 71 to 86% of all tenure-track faculty"

```
MATCH (n:institution)
WITH COUNT(n)*0.25 AS total25p_institution

MATCH ()-[r]→()
WHERE r.rank = 'Full'
WITH COUNT(r) AS total_full, total25p_institution

MATCH (n1:institution)-[r]→(n2:institution)
WHERE toInteger(n1.u) ⩽ total25p_institution AND r.rank = 'Full'
RETURN COUNT(r) * 100.0 / total_full AS percentage
```

	percentage
1	78.71407696054554

Fig. 4: Query2 & Result

3) : Comparison of the median in the result confirms the statement: "a greater fraction of faculty trained at higher-ranked institutions make smaller moves down the hierarchy than those trained at lower-ranked institutions". (Median is chosen here as the criterion, because the mean is susceptible to extreme values)

```
MATCH (n1:institution)-[r]→(n2:institution)
WHERE toInteger(n1.u) ⩽ top_institution_30
WITH n1, toInteger(n1.u) AS rank1, n2, toInteger(n2.u) AS rank2,
  CASE
    WHEN toInteger(n1.u) ⩽ top_institution_5 THEN 'top_5%'
    WHEN toInteger(n1.u) ⩽ top_institution_15 THEN 'top_5-15%'
    WHEN toInteger(n1.u) ⩽ top_institution_30 THEN 'top_15-30%'
  END AS institution_group
RETURN institution_group, AVG(rank2 - rank1) AS avg_diff, percentileCont(rank2 -
rank1, 0.5) AS median, STDEV(rank2 - rank1) AS stdev_diff
```

	institution_group	avg_diff	median	stdev_diff
1	"top_5%"	39.65965732087225	24.0	46.43885905949497
2	"top_5-15%"	57.60238095238096	45.0	53.25422559730956
3	"top_15-30%"	53.76791120080727	50	50.84760971484621

Fig. 5: Query3 & Result

4) : Comparison of avg_diff and $median$ in the result confirms the statement: "the hierarchy is slightly steeper for elite women than for elite men".

```
MATCH (n1:institution)-[r]→(n2:institution)
WITH n1, toInteger(n1.u) AS rank1, n2, toInteger(n2.u) AS rank2, r.gender AS gender
RETURN gender, AVG(rank2 - rank1) AS avg_diff, percentileCont(rank2 - rank1, 0.5)
AS median, STDEV(rank2 - rank1) AS stdev_diff
```

	gender	avg_diff	median	stdev_diff
1	"M"	46.016345123258226	33.0	51.47385511212321
2	"F"	47.13871951219512	38.0	48.99609544381599

Fig. 6: Query4 & Result

III. Implement the MVR ranking method

A. Do you get the same results as the paper?

I coded in Python to implement the MVR algorithm, and found that the resulting institutional prestige rankings were not much different from those reported in the supplementary materials. Specifically, the mean absolute difference between paper's result and mine is 10.43, the median is 8. Moreover, the fraction of edges that violated the ranks ordering is 0.0957 finally, indicating that the algorithm performed correctly overall.

B. Do you think the "ranking" method suggested in the paper is good? Any issues with it? Do you suggest improvements?

The ranking method suggested in the paper, Minimum Violation Ranking (MVR), is a reasonable approach for studying the hiring networks of faculty. However, it also has some issues and limitations.

First, MVR method requires a sufficient amount of data to generate meaningful results. If the data is not big enough, the rankings produced by MVR may be more susceptible to noise and outliers, leading to less accurate and meaningful rankings.

Another issue is that MVR only considers the pairwise relationships between institutions and ignores larger network structures and possible indirect relationships, which may miss important information. For example, some faculty members may have trained at multiple institutions (Co-Training Program) during their PhD, creating relations that involve more than two institutions.

Improvements:

1) : Find ways to combine multiple datasets to increase the amount of data available, or develop methods to account for noise and outliers in smaller datasets.

2) : Modify MVR algorithm according to incorporating network structure information (such as redefining violation rules for multiple relationships), or develop new methods to identify and account for indirect relationships between institutions.

IV. related paper

Q: Can you find ONE related paper and read it? After reading the paper, summarize its problem (1 sentence) and key ideas (3 sentences).

A: I found an paper titled *"Human Mobility, Social Ties, and Link Prediction"* [1] by Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási, which was published in 2011.

Problem:

The paper aims to predict the likelihood of a social tie forming between individuals based on their mobility patterns and existing social network.

Key ideas:

1) : The authors use a dataset of mobile phone records to analyze the mobility patterns and social ties of a group of individuals.

2) : They develop a model that incorporates both geographical proximity and social similarity to predict the formation of social ties between individuals.

3) : The authors find that their model outperforms previous models, and the prediction accuracy can be significantly improved by combining both mobility and network measures in supervised learning.

References

[1] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 1100–1108, New York, NY, USA, 2011. Association for Computing Machinery.