

Part 2 Rubric Explanation

Weak/strong ties and LCC size after each removal

In the implementation, tie strength is taken from edge weight.

- lower weights are treated as weak ties
- higher weights are treated as strong ties
- weak/strong categories are explicitly defined using a median-weight threshold:
- weak ties: $\text{weight} \leq \text{median}$
- strong ties: $\text{weight} > \text{median}$

I run two removal orders on the LCC:

- 1 weakest to strongest
- 2 strongest to weakest

Edges are removed one by one. After every removal, the LCC is recalculated and its size is stored as node count. The x-axis is fraction of ties removed, and the y-axis is LCC size.

From the run output, the starting LCC is:

- 1662 nodes
- 26134 edges

The code also prints exact weak/strong tie statistics:

- total number of ties in the LCC: 26134
- weak-tie threshold (median weight): 0.6276
- number of weak ties ($\text{weight} \leq 0.6276$): 13067
- number of strong ties ($\text{weight} > 0.6276$): 13067

So both tie classification and total weak/strong counts are explicitly reported before the stepwise removal process.

Centrality, central papers, interpretation, correlation

Three centrality measures are computed on the LCC:

- Degree

- Closeness
- Betweenness

For each one, top-10 papers are listed in ID<TAB>Title format.

For correlation, I first convert centrality scores to ranking vectors and then compute Pearson correlation between rankings.

Results from the run:

```
Metric | Degree | Closeness | Betweenness
Degree | 1.0000 | 0.9361 | 0.8114
Closeness | 0.9361 | 1.0000 | 0.7684
Betweenness | 0.8114 | 0.7684 | 1.0000
```

- Degree vs Closeness: 0.9361
- Degree vs Betweenness: 0.8114
- Closeness vs Betweenness: 0.7684 (lowest)

The output explicitly reports the lowest-correlation pair.

Lowest pair interpretation:

- closeness measures overall proximity in the graph
- betweenness measures bridge role on shortest paths
- these are related but different structural roles, so their rankings are less aligned

Repeatedly central papers across top lists include:

- ahuja-etal-2023-mega
- ding-etal-2020-discriminatively
- shin-etal-2020-autoprompt
- weller-etal-2020-learning
- qin-etal-2023-chatgpt

The code also explicitly prints papers that appear in multiple metric top-10 lists (with metric names), which strengthens the evidence for identifying robustly central papers.

Optional Extra Credit (50%): Theme shift before and after 2023

I compare two time groups: before 2023 and 2023+.

Steps used:

- 1 split papers by year
- 2 create text from title + abstract
- 3 tokenize and clean
- 4 build one shared vocabulary
- 5 train LDA for each period
- 6 extract topic-term matrices D (before) and S (after)
- 7 compare topics with cosine similarity and rank by shift score

Run evidence:

- D shape: (5, 5000)
- S shape: (5, 5000)

Examples from output:

- emerging: After Topic 2 | shift=0.1989 | llms, large, data, tasks, knowledge, reasoning, generation, performance
- disappearing: Before Topic 4 | shift=0.1912 | question, knowledge, event, performance, questions, task, graph, can

This indicates a stronger LLM/reasoning focus in the later period.

Results (from current execution)

- Network loaded successfully; LCC size is 1662 nodes and 26134 edges.
- Weak/strong tie section reports:
- total ties: 26134
- median-weight threshold: 0.6276
- weak ties: 13067
- strong ties: 13067
- Centrality ranking correlations:
- Degree-Closeness: 0.9361
- Degree-Betweenness: 0.8114
- Closeness-Betweenness: 0.7684
- Lowest-correlation pair: Closeness vs Betweenness.
- Top-10 central papers were produced for all three metrics in ID<TAB>Title format.
- Repeated papers across multiple centrality top-10 lists are explicitly reported.

- Topic-evolution matrices were produced:
- D (before 2023): (5, 5000)
- S (2023+): (5, 5000)
- Highest-shift emerging topic: After Topic 2 (shift=0.1989) with keywords around llms, reasoning, and generation.
- Highest-shift disappearing topic: Before Topic 4 (shift=0.1912) with keywords around question, knowledge, and graph.

Findings

- The centrality rankings are strongly related overall, but not identical.
- Degree and closeness are most aligned (0.9361), indicating that papers with strong local connectivity are often globally well-positioned.
- Closeness and betweenness are least aligned (0.7684), showing that global proximity and bridge-role influence capture different node functions.
- Repeated appearance of papers such as *ahuja-etal-2023-mega*, *ding-etal-2020-discriminatively*, and *qin-etal-2023-chatgpt* across multiple lists suggests robust influence across different centrality definitions.
- Topic-shift outputs indicate post-2023 movement toward LLM-oriented and reasoning-heavy themes.
- Overall, the network remains highly connected at baseline, and the analysis pipeline covers connectivity, influence, and temporal theme evolution in a consistent way.