COL761: Data Mining

Assignment 1: Frequent Subgraph Mining and Indexing

Question 3: Discriminative Subgraph Identification

Team Members:

Anant Gupta (2025AIB2556)
Vedant Gupta (2025AIB2559)
Devanshu Patel (2025AIB2567)

Indian Institute of Technology Delhi
February 3, 2026

1 Methodology: Discriminative Subgraph Identification

Our approach utilizes a **Multi-Pass Feature Mining Strategy** to balance the two competing requirements of graph indexing: **Coverage** (ensuring no query results in a full-database scan) and **Precision** (minimizing the candidate set size $|C_q|$).

1.1 The Hybrid Feature Pool (Optimization)

Originally, a single mining pass led to the "Empty Vector Bug" for topologically simple queries. We transitioned to a dual-pass approach:

- **Coverage Pass (The Safety Net):** We mine 30 features with high support (10%–50%) and short path lengths (max 5).
- **Precision Pass (The Spear):** We mine 20 features with very low support (2%–10%) and longer path lengths (up to 10).

1.2 Custom FSM Algorithm: Path-Based Mining

Instead of general subgraph mining, our algorithm specifically mines **Canonical Paths and Rings**.

- **Efficiency:** While general subgraph isomorphism is NP-Complete, verifying a specific path is significantly faster ($O(V + E)$).
- **Canonical Representation:** Paths are converted into a canonical string format (e.g., `C-1-N-1-C`), allowing for high-performance hash map support counting.

2 Justification: Why not gSpan or Gaston?

While gSpan and Gaston are state-of-the-art for frequent subgraph mining, they were not optimal for this indexing task:

1. **Computational/Memory Overhead:** For 40,000 graphs at low support, the pattern count explodes. gSpan requires significant RAM for the DFS Lexicographic Tree.
2. **Redundancy:** Most subgraphs found by gSpan are redundant for indexing (high correlation). Our custom miner allows for length and support constraints.

3 Database and Domain Analysis

Analysis of **NCI-H23** and **Mutagenicity** shows that these datasets are ring-dominated (Mutagenicity: ∼82.6%; NCI-H23: ∼96.9%). Atom labels are highly skewed (NCI-H23 top label ∼770k occurrences), making atom-only features ineffective.

4 Experimental Results and Evaluation

To evaluate the effectiveness, we implemented an assessment tool based on the provided grading formula: $S_q = |R_q|/|C_q|$.

Dataset	Average $\lvert C_q \rvert$	Average S_q Score	Space Pruned
Mutagenicity	1424.64	0.21685	∼99%
NCI-H23	22061.76	0.19123	∼45%

Table 1: Overall Performance Metrics

4.1 Comparative Performance Summary

The table below summarizes our filtering effectiveness across the two datasets:

4.2 Execution Efficiency

By using a multiprocessing Pool across 24 cores, the total verification time for 50 queries was **12.66s** for Mutagenicity and **108.62s** for NCI-H23.

5 References and Attributions

- **Methodology & Code Implementation:** Developed by the team using a custom path-based mining algorithm.
- **Strategic Refinement:** Technical logic for the dual-pass optimization and feature pruning was refined in collaboration with **Gemini**.
- **Report Compilation:** The structure, analysis synthesis, and LaTeX formatting of this report were prepared with assistance from **Gemini**.