Outline

- Introduction
- Application domains
- Graph Mining Algorithms
Introduction

- Graph: \( G = (V, E) \)
  - \( V \) ... set of nodes,
  - \( E \subseteq V \times V \) ... set of edges
Graph: $G = (V, E)$
- $V$ ... set of nodes,
- $E \subseteq V \times V$ ... set of edges,
- $w: E \rightarrow \mathbb{R}$ ... weight function,
- $\mu: V \rightarrow L_V$ ... node labeling function,
- $\nu: E \rightarrow L_E$ ... edge labeling function,
- ...

Introduction
Application domains

- Chemical data analysis
- Computational biology
- Social networking
- Web link analysis
- Computer networks
- ...

Main Graph Mining Algorithms

- Clustering
- Classification
- Frequent pattern / substructure mining
Considerations

- **Data properties**
  - One large graph vs. set of (smaller) graphs (also *transactions*)
  - **Size**
    - Streaming of massive graphs (they are too large to fit in the main memory and random access is slow in large capacity storage devices)
  - **Static vs. dynamic**
  - ...

Node clustering
- Based on distance functions for nodes
- Related to minimum cut (polynomially solvable) and graph partitioning (NP-hard) problems
- Applications: determining dense regions (=> summarization, dimensinality reduction), ...

Graph clustering
- Based on structural behavior
- Applications: molecular biology, chemical graphs, XML data, ...
Classification Algorithms

- Node classification
- Graph classification

Diagram showing a graph with nodes labeled -1, +1, and a question mark, connected by edges, indicating different classifications.
Pattern Mining Algorithms

- Pattern ≈ subgraph
- Single-graph setting
  - Frequency: number of pattern occurrences in the single graph.
  - Examples of algorithms: SUBDUE, SEuS, GREW, SIGRAM, GBI
  - Not discussed further
- Graph-transaction setting
  - Frequency (of a pattern): number of graph transactions in which the pattern occurs
Apriori-like algorithms

- Basically two steps:
  - Generation of frequent substructure candidates
    - Based on adding nodes, edges or paths
  - Checking the frequencies of candidates
- Examples of algorithms: AGM, FSG
Pattern Mining Algorithms

Checking the frequencies of candidates:
- (Sub)graph isomorphism

- Canonical labeling
  - Unique code for the set of graphs with the same topological structure and the same labeling
  - Both problems are not known to be either in P or in NP-complete → relaxed problems
Pattern Mining Algorithms

Pattern growth algorithms

- gSpan
- Gaston
- ...

gSpan

- Without candidate generation
- Minimum DFS code as canonical label
Pattern Mining Algorithms
gSpan
- Without candidate generation
- Minimum DFS code as canonical label
- DFS lexicographic ordering on DFS codes
  → DFS code tree
Pattern Mining Algorithms

Diagram: A tree structure with labeled edges (0-edge, 1-edge, 2-edge, n-edge) and nodes labeled with $G_0$ and $G_1$. The diagram also indicates the process of pruning.
gSpan

- Without candidate generation
- Minimum DFS code as canonical label
- DFS lexicographic ordering on DFS codes
  → DFS code tree
- Searching frequent patterns: traversing DFS code tree
References

- Charu C. Aggarwal, Haixun Wang. *Managing and Mining Graph Data*. Springer, 2010