Karel Vaculík

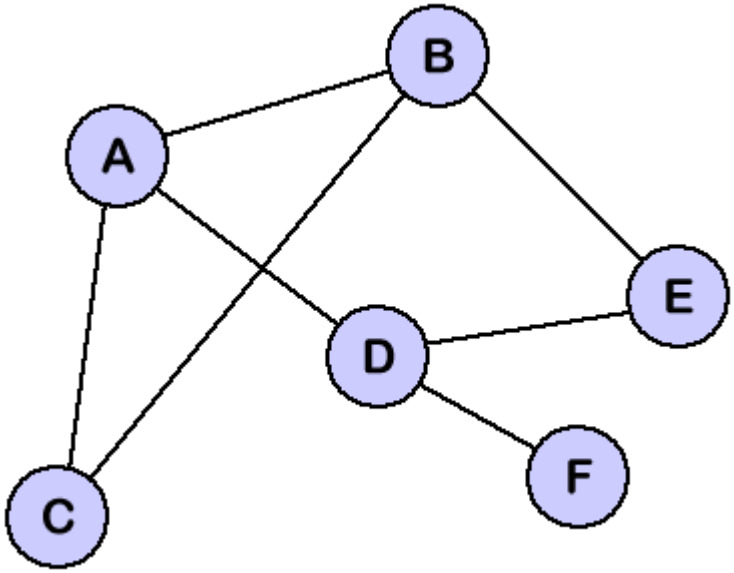
Mining Graph Data

Outline

- Introduction
- Application domains
- Graph Mining Algorithms

Introduction

- Graph: $G = (V, E)$
  - V … set of nodes,
  - $E \subseteq V \times V$… set of edges



Introduction

- 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,
  - …

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

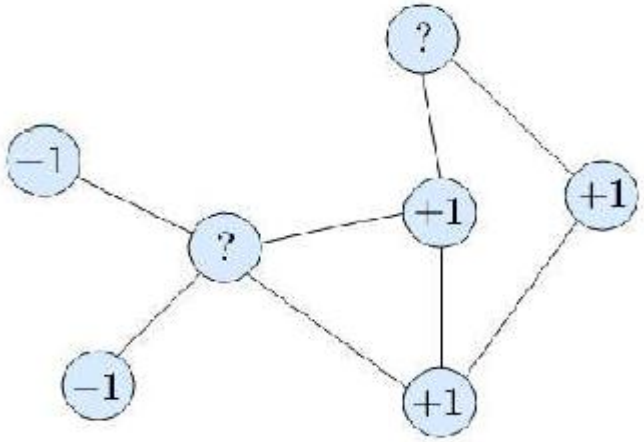
- Data properties
  - One large graph vs. set of (smaller) graphs (also *transactions*)
  - Size
    - Streaming of massive graphs (they are too large to fit in the main memory and random access is slow in large capacity storage devices)
  - Static vs. dynamic
  - …

Clustering Algorithms

- 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

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

Pattern Mining Algorithms

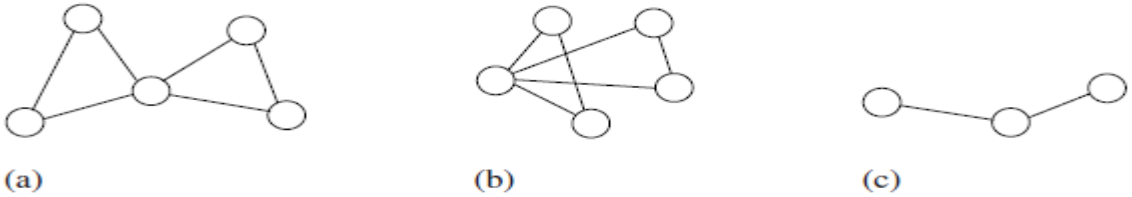
Apriori-like algorithms

- 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

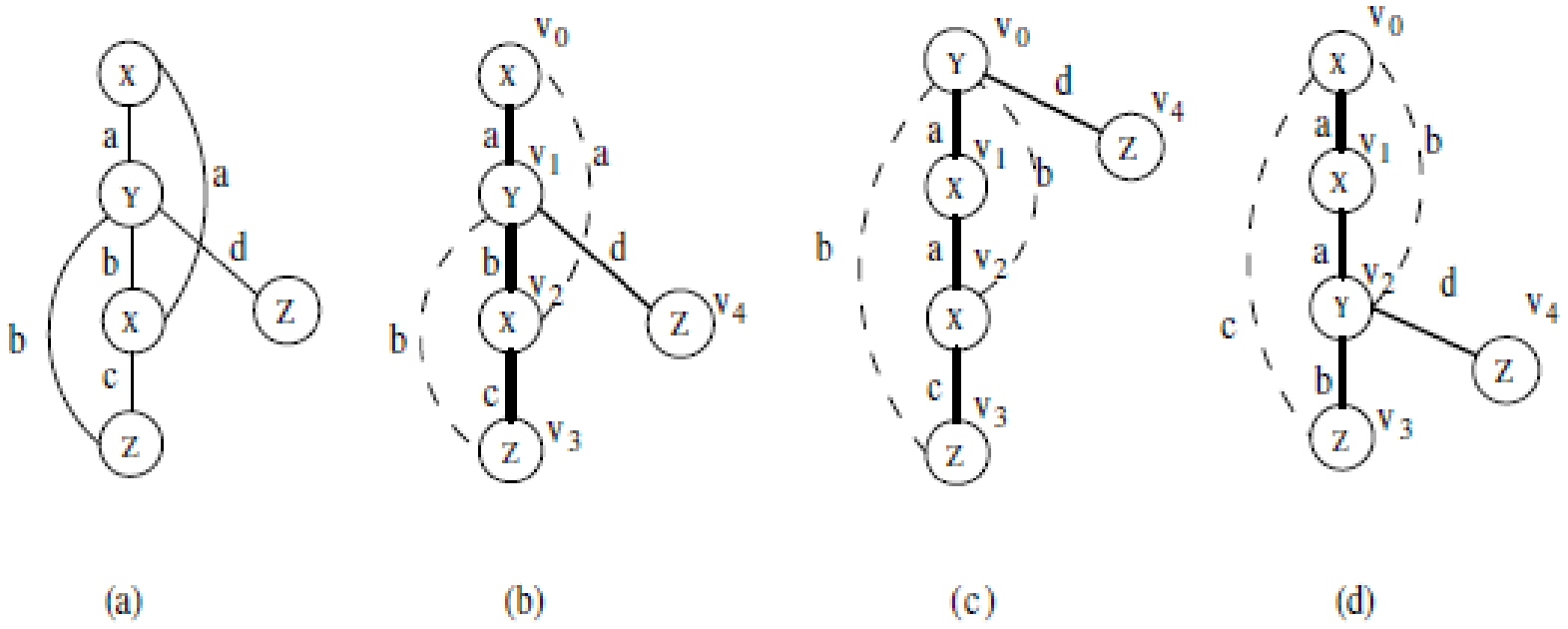
- gSpan
- Gaston
- …

Pattern Mining Algorithms

gSpan

- Without candidate generation
- Minimum DFS code as canonical label

Pattern Mining Algorithms

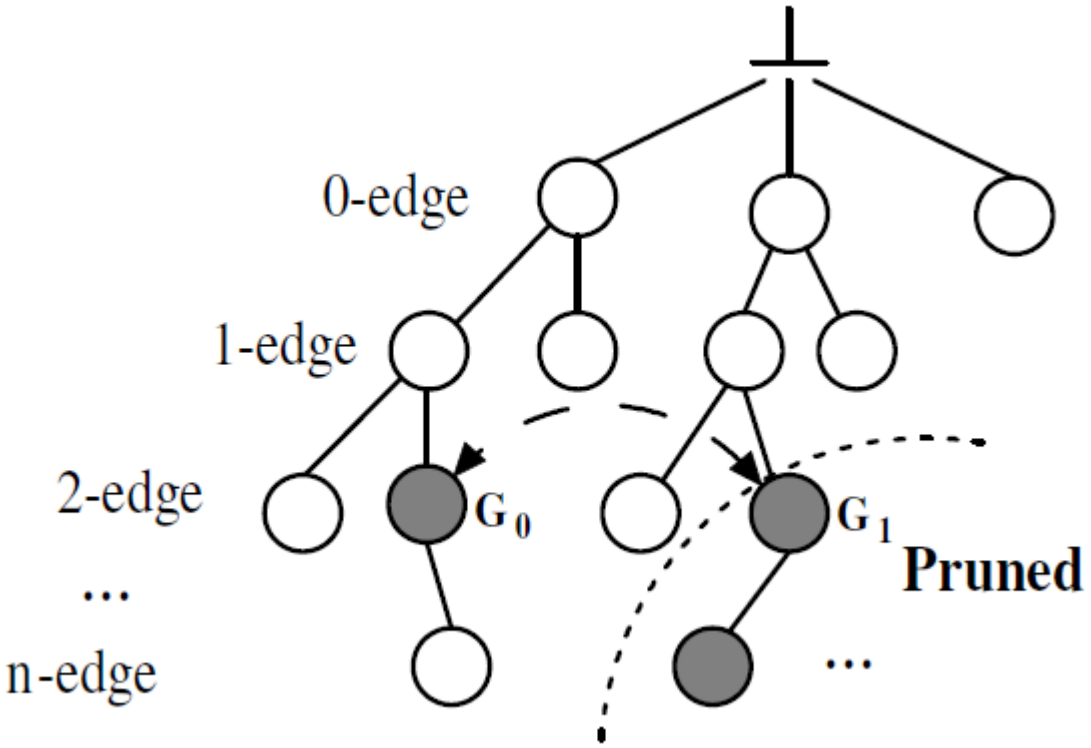


Pattern Mining Algorithms

gSpan

- Without candidate generation
- Minimum DFS code as canonical label
- DFS lexicographic orderering on DFS codes
  $\longrightarrow$ DFS code tree

Pattern Mining Algorithms



Pattern Mining Algorithms

gSpan

- Without candidate generation
- Minimum DFS code as canonical label
- DFS lexicographic orderering on DFS codes $\longrightarrow$ DFS code tree
- Searching frequent patterns: traversing DFS code tree

References

- Diane J. Cook, Lawrence B. Holder. *Mining graph data*. John Wiley and Sons, 2007.
- Charu C. Aggarwal, Haixun Wang. *Managing and Mining Graph Data*. Springer, 2010
- X. Yan and J. Han. *gSpan: Graph-based substructure pattern mining*. In Proceedings of 2002 IEEE International Conference on Data Mining (ICDM), pp. 721–724, 2002.