Mining Large Single Networks under Subgraph Homomorphism

Mostafa H. Chehreghani
Jan Ramon
Thomas Fannes

DECLARATIVE LANGUAGES & ARTIFICIAL INTELLIGENCE
KATHOLIEKE UNIVERSITEIT LEUVEN

12/13/13

Overview

- Introduction
- Problem definition and preliminaries
- Related work and motivation
- Our contributions and the proposed algorithm
- Conclusion

KATHOLIEKE UNIVERSITEIT LEUVEN

Frequent Patterns

- Frequent patterns = pattern which occurs in a database more often than a user-defined threshold
- Two settings:
  - Transactional
  - **Single-network**
- Applications:
  - Web mining
  - Social network analysis
  - Biological & chemical interaction networks

Problem Definition

- Given:
  - a network graph *H*
  - a pattern language L_p
  - a matching operator ≤
  - a threshold *minsup*$\in R^+$
- Find (a condensed representqtion of) all patterns such that their frequency is at least *minsup*

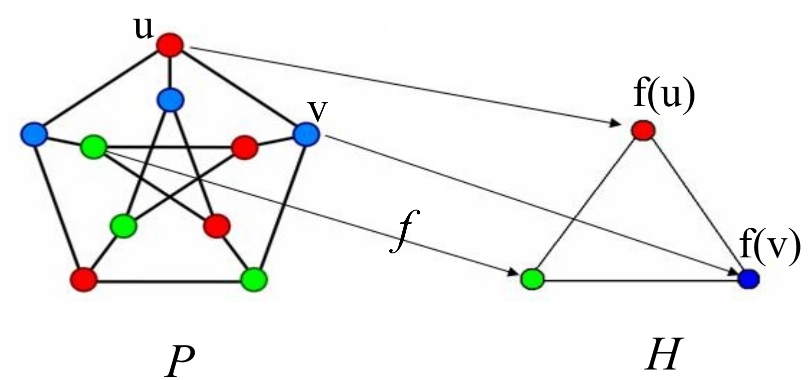
Homomorphism

- Graph homomorphism *f* from *P* to *H*:
  - Label preserving
  - If *u* and *v* of *P* are adjacent in *P*, then *f(u)* and *f(v)* are adjacent in *H*

Subgraph Homomorphism:

Homomorphism from P

to (a subgraph of) *H*



- Subgraph homomorphism is easier than subgraph isomorphism
  - Polynomial algorithms for bounded treewidth graphs

Related Work and Motivation

- Most approaches use any graph patterns
  - e.g. Kuramochi&Karypis ICDM'04
  - NP-hard under normal matching operators
    - We will limit ourselves to bounded treewidth graphs
      - This is not a strong restriction
- Most approaches use subgraph isomorphism
  - e.g. Zhu et. al., VLDB'11
  - Computationally expensive
  - A few methods use subgraph homomorphism
    - e.g. Dries&Nijssen, SDM12 (Only for trees)
    - e.g. J.Van den Bussche, (No antimonotonic pruning)

Related Work and Motivation Cont.

- Matching operator ≤
  - We use subgraph homomorphism
  - Candidate generation under homomorphism is challenging
    - Our solution: root embedding equivalent classes
- The frequency measure
  - Wang&Ramon, DMKD'13: s-measure: linear program
    - LP with one variable per embedding of pattern
    - Describes statistical power of the pattern
    - But: needs to construct overlap graph (exponential amount of embeddings)
    - We avoid overlap graph using bounded treewidth homomorphism!

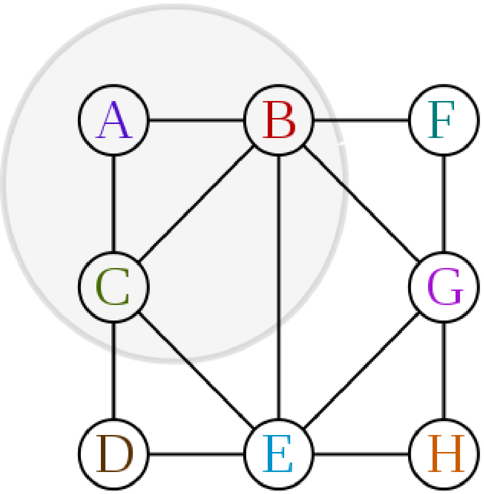
A Summary of Our Contributions

- We consider the class of rooted graphs
  - We present an efficient method to generate them from data
- We present a new notion for compactly representing all frequent patterns
  - It gives a closure operator
- Two frequency counting settings:
  - Mining patterns with frequent root embeddings

    (= embeddings of the root of the pattern)
  - Mining s-measure-frequent patterns
    - Linear program to compute s-measure

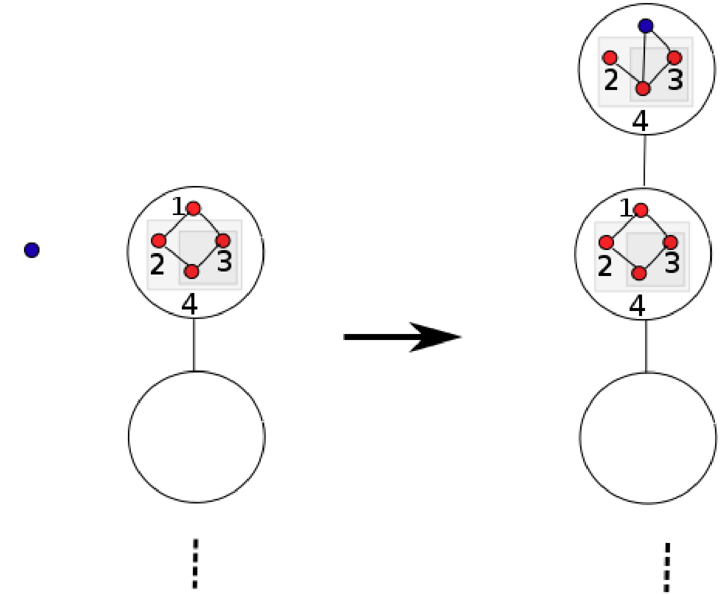
Rooted Patterns and Root Embeddings

- A rooted graph P^X, is a graph P where the set $X \subseteq V(P)$ is distinguished.
- Let H be a database graph
- Let φ be a subgraph homomorphism mapping from P to H
- $\varphi|_X$: φ restricted to the vertices in X
- $\varphi|_X$ is called a *root embedding* of P^X in H
- Two rooted graphs are equivalent under root embedding iff they have the same set of root embeddings

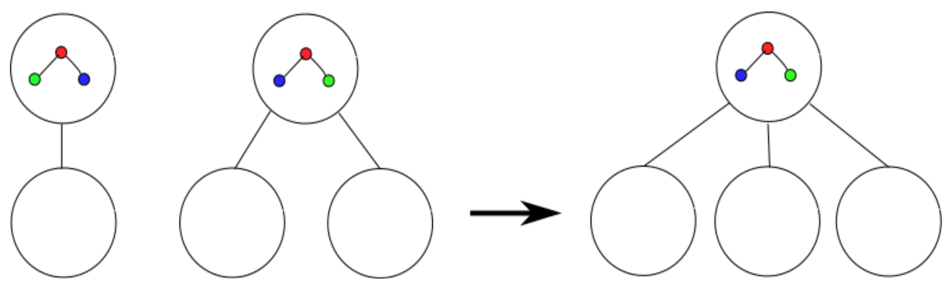


Generating Rooted Patterns

- The extension operator
  - Adds a new vertex to a pattern
- The join operator
  - Joins two existing patterns



extension



join

Closed Pattern

- f: maps a root embedding equivalence class eq to a finite set $\{P_1{}^{X_1}, \ldots, P_n{}^{X_n}\} \subseteq eq$ which contains all rooted cores of eq
-
- P^X is defined as $\rho_{join}(\ldots \rho_{join}(\rho_{join}(P_1{}^{X_1}, P_2{}^{X_2}), P_3{}^{X_3}), \ldots, P_n{}^{X_n})$
- The operator σ_f maps every member of eq to P^X
- P^X is a closed pattern

- σ_f is a closure operator
  - It is *extensive*, *increasing* and *idempotent*

KATHOLIEKE UNIVERSITEIT
LEUVEN

s-measure

- Let P^X be a rooted pattern and H be a database graph
- To every embedding φ of P^X in H a weight ω_φ is assigned
- Feasible assignment:
  - $\forall \varphi \in Emb(P^X, H) : \omega_\varphi \geq 0$
  - $\forall v \in V(H), u \in V(P^X) : \sum_{\varphi \in Emb(P^X, H) | \varphi(u) = v} w_\varphi \leq 1$
- s-measure: minimum feasible assignment
- Can be computed efficiently for rooted graphs when matching operator is subgraph homomorphism
  - Without forming overlap graph

Conclusion

- A new class of patters: rooted patterns
- Mining patterns with frequent root embeddings
- Mining patterns with minimal s-measure
- A new notion for compactly representing all frequent patterns under homomorphism
  - It gives a closure operator