Mining Frequent Closed Graphs on Evolving Data Streams

A. Bifet, G. Holmes, B. Pfahringer and R. Gavaldà

University of Waikato
Hamilton, New Zealand

Laboratory for Relational Algorithmics, Complexity and Learning, **LARCA**
UPC-Barcelona Tech

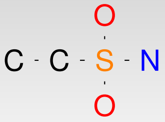
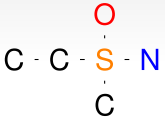
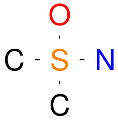
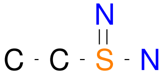
1st Graph-TA Workshop, Feb. 19th, 2013
Originally presented at KDD'11

Mining Evolving Graph Data Streams

Problem

Given a stream $\mathscr{D}$ of graphs,
maintain the set of frequent closed subgraphs

Graph Dataset

Transaction Id	Graph	Weight
1		1
2		1
3		1
4		1

Frequent Closed Pattern Mining

- Universe U of patterns
- Subpattern partial order, denoted $P \preceq P'$
- Support of a pattern P in a multiset $\mathscr{D}$ =

  = fraction of $\mathscr{D}$ elements that are have P as subpattern
- Pattern P is closed in $\mathscr{D}$ if every superpattern of P has smaller support

The frequent closed pattern mining problem

Given $\mathscr{D}$, find the set of closed patterns with support $\geq \varepsilon$

The Data Stream Computation Model

Five constraints:

1. Input is sequence of items; t-th item available at time t
2. Answers must be anytime, may be approximate
3. Low processing time per item
4. Sublinear memory; keep only summaries or sketches
5. Data distribution evolves over time; forget, react, adapt

Previous work

- *CloseGraph* [Yan-Han 03]
  - depth-first search, based on gSpan ICDM'02
- *MoSS* [Borgelt-Berthold 05]
  - breadth-first search, based on MoFa ICDM'02

Non-streaming: Non-incremental, multipass, linear memory

Graph Coresets

Coreset of a set P with respect to some problem

Small subset that approximates the original set P

- Solving the problem for the coreset provides an approximate solution for the problem on P

δ-tolerance Closed Graph

A graph g is *δ-tolerance closed* if none of its proper frequent supergraphs has a weighted support $\geq (1-\delta) \cdot \mathit{support}(g)$

- Maximal graph: 1-tolerance closed graph
- Closed graph: 0-tolerance closed graph

Graph Coresets

Relative support of a closed graph

Support of a graph minus the relative support of its closed supergraphs

- The sum of the closed supergraphs' relative supports of a graph and its relative support is equal to its own support

(s, δ)-coreset for computing closed graphs

Weighted multiset of frequent δ-tolerance closed graphs with minimum support s using their relative support as a weight

Dealing with evolution over time

- Keep a window on recent stream elements
  - Actually, just its lattice of closed elements!
- Keep track of number of closed trees in lattice, N
- Use some change detector on N
- When change is detected:
  - Drop stale part of the window
  - Update lattice to reflect this deletion, using deletion rule

Alternatively, sliding window of some fixed size

WinGraphMiner

WinGraphMiner(D, W, min_sup)

Input: A graph dataset D, a size window W and min_sup.
Output: The frequent graph set G.

$G \leftarrow \emptyset$
for every batch b_t of graphs in D
    do C $\leftarrow$ Coreset(b_t, min_sup)
        Store C in sliding window
        if sliding window is full
          then $\overline{R} \leftarrow$ Oldest C stored in sliding window, negate all support values
          else $\overline{R} \leftarrow \emptyset$
        $G \leftarrow$ Coreset$(G \cup C \cup \overline{R}, min_sup)$
return G

Experimental Evaluation

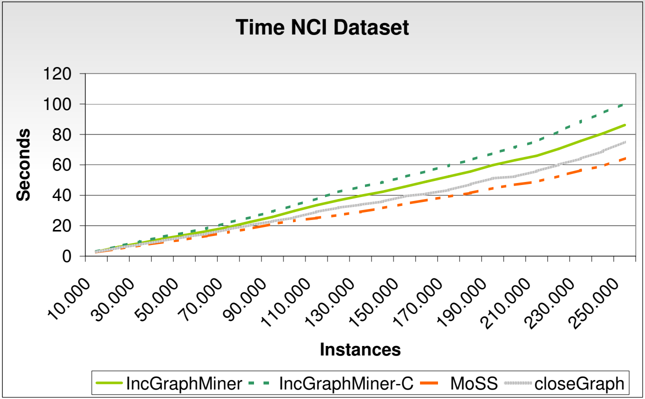
ChemDB dataset

- Public dataset
- 4 million molecules
- Institute for Genomics and Bioinformatics at the University of California, Irvine

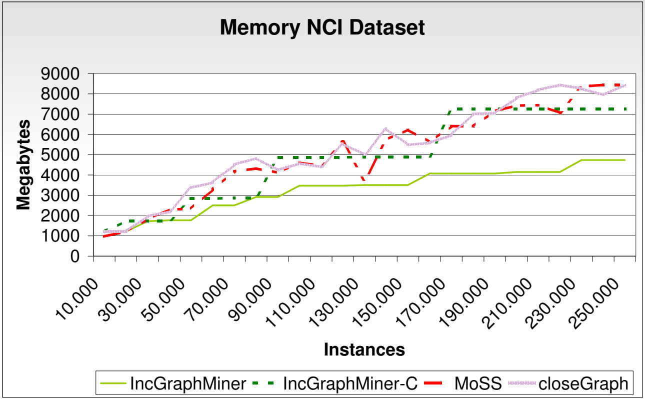
Open NCI Database

- Public domain
- 250,000 structures
- National Cancer Institute

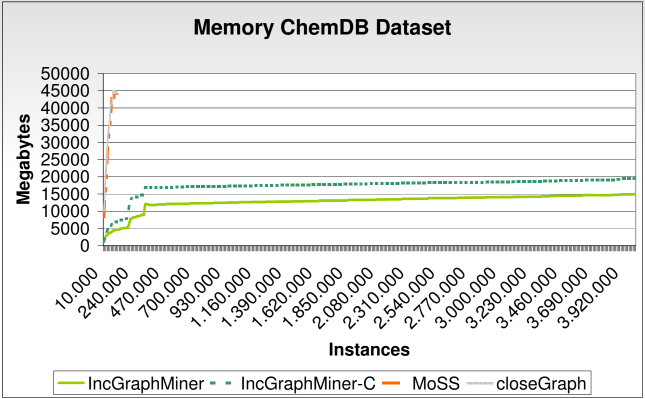
Open NCI dataset



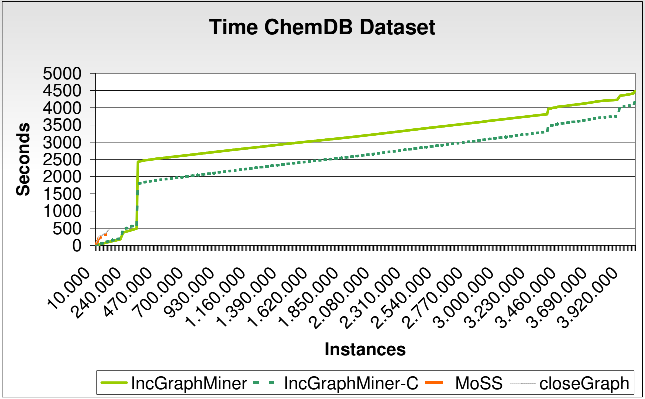
Open NCI dataset



ChemDB dataset



ChemDB dataset



ADAGRAPHMINER

Summary

We provide three algorithms of increasing power:

- Incremental
- Sliding Window
- Adaptive

To our knowledge, first algorithms for mining frequent (closed) subgraphs from evolving data streams