Output Space Sampling for Graph Patterns

Mohammad Al Hasan, Mohammed Zaki

Computer Science Department, Rensselaer Polytechnic Institute, Troy, New York

Motivation and Problem Statement

1. Subgraph Mining (SGM) algorithms follow an *enumeration and test* paradigm, where every candidate pattern is generated and tested for support.
2. But in many domains, graphs are large and dense, for which complete *enumeration and test* is infeasible
3. Furthermore, with reasonable *support*, the output-set of an SGM algorithm is generally too large for an analyst to interpret
4. So, sampling frequent subgraphs from the output space of Subgraph Mining is a very useful research direction.
   - Scalable, since it does not obtain the entire output-set
   - Return good quality pattern by adopting desired sampling criteria

Sampling based graph Pattern Mining

Existing Algorithm

- Depth-first or Breadth-first walk on the subgraph partial order graph (POG)
- Rightmost extension
- Complete Algorithm

Our Approach

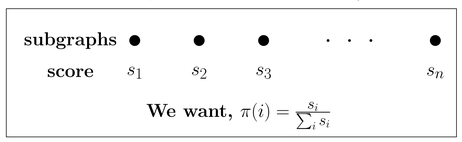
- Random walk on the subgraph POG
- Arbitrary extension
- Sampling Algorithm

Features:

- **Quality: Sampling quality guaranty**
- **Scalability: Visits only a small part of the search space**
- **Non-redundancy: Finds very dissimilar patterns by virtue of randomness**
- **Genericity: In terms of pattern type and sampling objective**

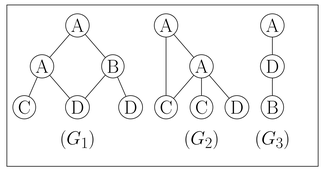
MCMC Framework for subgraph sampling

- MCMC is a family of algorithms for sampling from probability distribution
- Construct an ergodic Markov chain
- If π is the stationary distribution, and P is the transition matrix, in equilibrium, we have $\pi = \pi P$
- Main task is to choose P, so that a desired stationary distribution is achieved

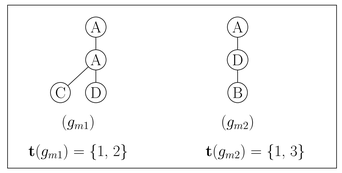


- How to Choose score?
  - Depends on application need
- For exploratory data analysis (EDA), every frequent pattern can have uniform score
- For subgraph summarization task, user may be interested to obtain uniform sample of maximal graphs (Hasan & Zaki, SDM 2009)
- For graph classification, discriminatory subgraph should be preferred
- Score can be changed dynamically based on the already sampled patterns

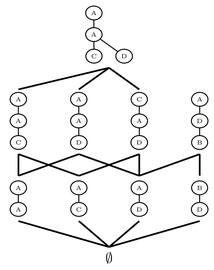
Uniform Frequent subgraph Sampling



A graph database



Maximal Frequent Patterns ($minsup = 2$)

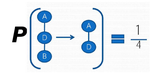


Partial Order Graph

- Compute Neighbors locally
- Compute appropriate transition probability
- For uniform random walk, choose a symmetric matrix for P

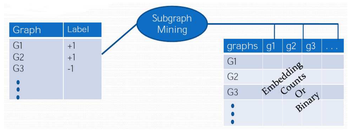
$$P(u,v) = \begin{cases} \frac{1}{\max(d_u, d_v)} & \text{if } u \neq v \text{ and } v \in \text{adj}(u) \\ 1 - \sum_{x \in \text{adj}(u)} P(u,x) & \text{if } u = v \\ 0 & \text{otherwise} \end{cases}$$

- For example,



Sampling Discriminatory Subgraphs

- Database graphs are labeled
- Mine subgraphs to use as
  - Features for supervised classification
  - To make graph kernels
- Interestingness score (feature quality)
  - Entropy
  - Delta Score = abs (positive support - negative support)
- Direct Mining is difficult, as the scores are neither monotone nor anti-monotone



Discriminatory subgraphs as features

Metropolis-Hastings based Subgraph Sampling

- Objective is to sample a pattern from a target distribution π, with

$$\pi(i) = \frac{s_i}{C}, \quad i \in \mathscr{X} \tag{1}$$

Where, C is the normalizing constant $C = \sum_{i=1}^{m} b_i$ (difficult to calculate)

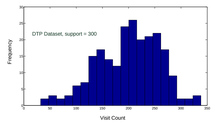
- First, constructs a Markov chain $X_t, t = 0, 1, \ldots$ on $\mathscr{X}$ with a proposal distribution, Q
- Choose a transition $i \rightarrow j$ using Q
- Accept it with probability α_{ij} Where,

$$\alpha_{ij} = \min\left\{\frac{\pi(j)\ q_{ji}}{\pi(i)\ q_{ij}}, 1\right\} = \min\left\{\frac{s_j\ q_{ji}}{s_i\ q_{ij}}, 1\right\}$$

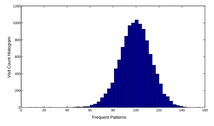
- Continue the walk until Markov Chain converges (to target distribution)

Uniform Sampling Results

- Run the sampling algorithm for sufficient number of iterations and observe the visit count distribution
- In ideal case, we expect a Gaussian-like distribution



Distribution on DTP Chemical Datasets



Distribution on Itemset Patterns

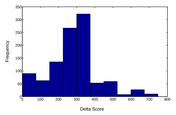
	Our Algorithm				Ideal	
Dataset	Max	Min	Median	Std	median	Std
DTP	338	32	209	59.02	200	14.11
Itemset	156	6	100	13.64	100	10

Uniform sampling performance

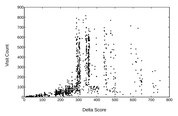
- Sampling quality depends on the convergence rate of MCMC
- If the POG is well-connected, convergence is good
- For clique dataset, sampling quality is almost perfect

Result on Discriminatory Pattern Sampling

Sample No	Delta Score	Rank	% of POG-Explored
1	404	132	5.7
2	644	21	11.0
3	707	10	10.8
4	282	593	2.4
5	646	17	5.5
6	280	595	2.8
7	627	27	3.3
8	709	9	7.7
9	721	5	9.1
10	725	4	8.9



Delta-score dist. in entire FP-set



Delta-score dist. in sampling

Future works and Conclusion

1. Sampling based approach is an effective paradigm to cope with *lack of scalability* and *information overload* problem in graph mining
2. Biased random walk finds discriminatory patterns that have high quality score. Extension can be made by dynamically adapting the proposal distribution so that non-correlated patterns can be sampled more often.
3. For large graph, support counting is still a bottleneck, need to find a way to sample good patterns without explicit *subgraph isomorphism* test.

References

M. Al Hasan, M. Zaki: **Output Space Sampling for Graph Patterns**, *submitted for publication*

M. Al Hasan, M. Zaki: **MUSK: Uniform Sampling of k Maximal Patterns**, *SIAM Data Mining, 2009*

M. Al Hasan, V. Chaoji, S. Salem, J. Besson, M. Zaki: **ORIGAMI: Mining Representative Orthogonal Graph Patterns**, *In Proceedings of of IEEE International Conference on Data Mining, NE, USA, 2007, pp. 153*