Mining Graph Patterns in Massive Networks (ID #70)

Nilothpal Talukder
Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY

Background

Frequent graph mining: Discover frequent graph patterns in labeled networks. Application in bio-informatics, chemistry, social network.

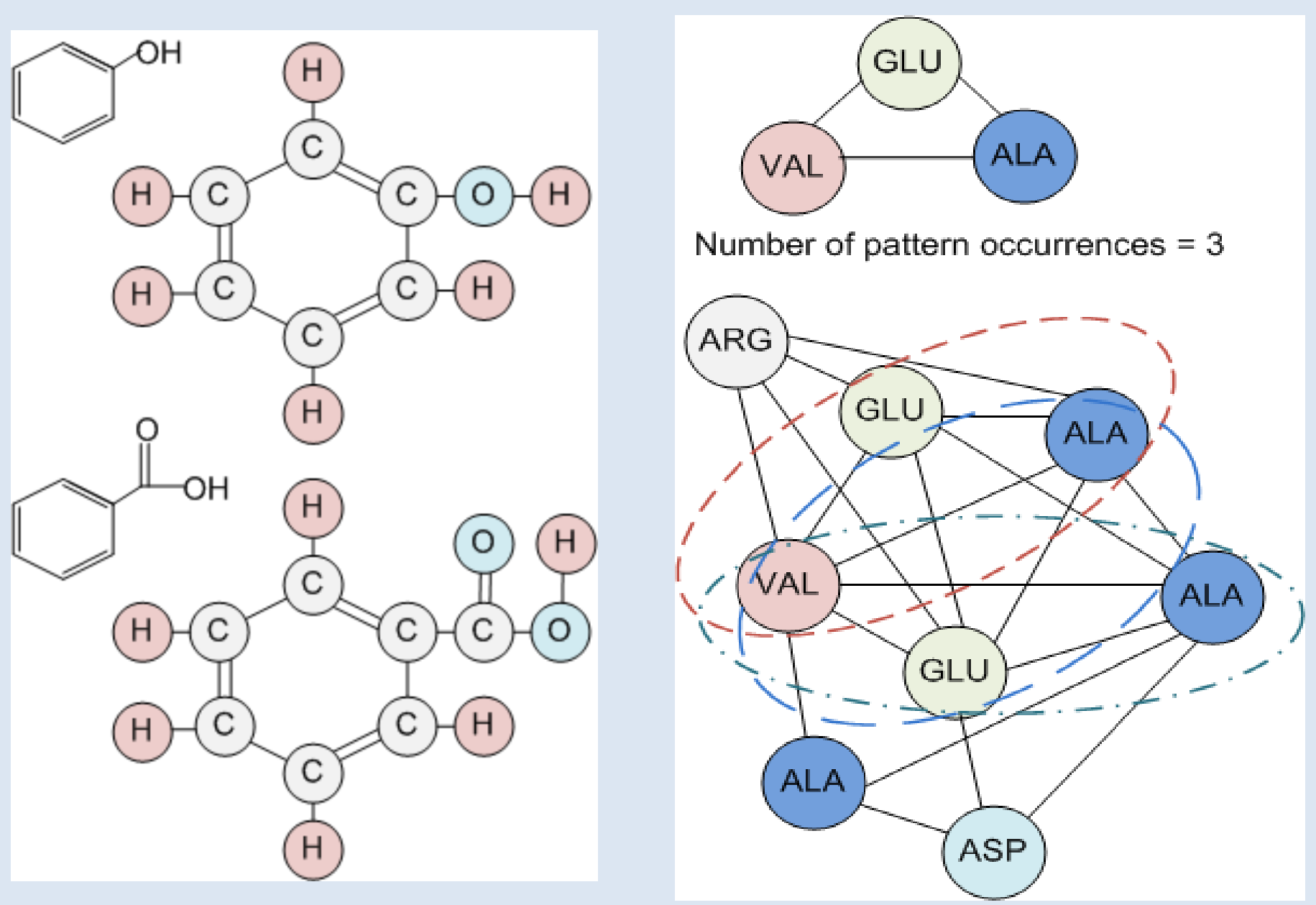


Figure: Chemical compounds and Protein Interaction graphs

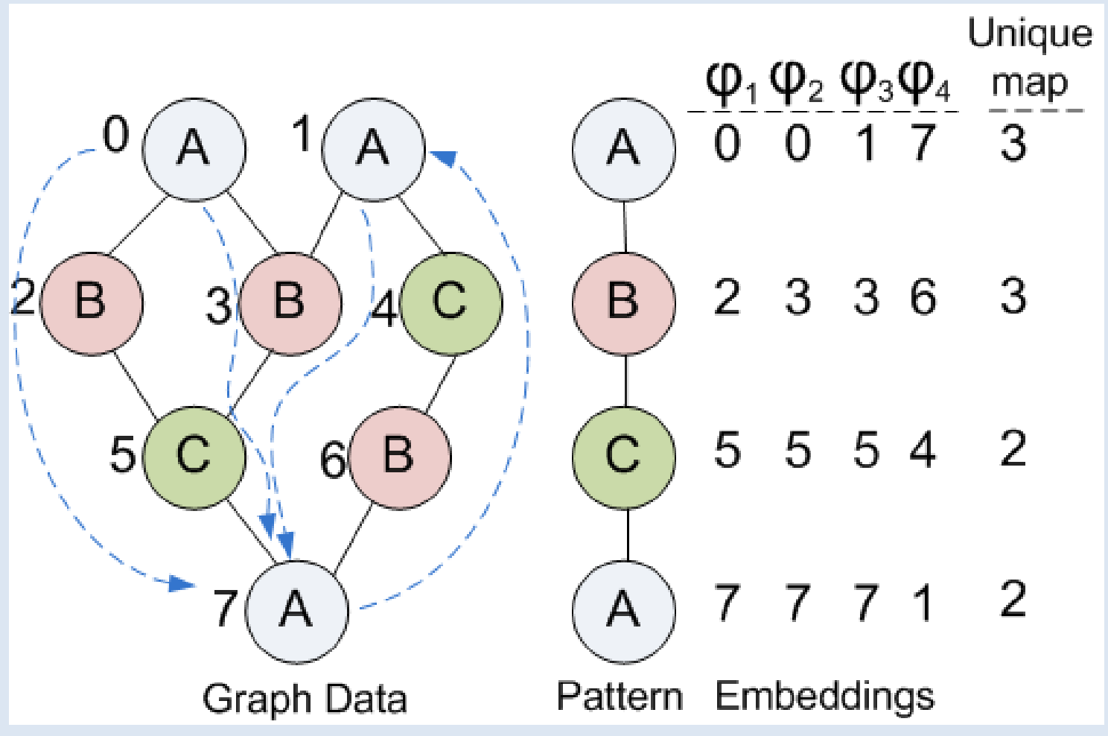


Figure: Anti-monotonic non-overlapping support =2

Transactional setting: A set of many moderate sized networks, can easily be mined in parallel.

Single graph setting: Single large sparse graph, more challenging as the input may not fit in memory of a single machine. Facebook currently has 1.4 billion active monthly users.

The following makes matters even worse:

i) Subgraph enumeration space is exponential

ii) Subgraph isomorphism is NP-complete

Existing approaches [1-3] are sequential/parallel and shared-memory based, cannot mine a very large graph. We are in need of a scalable distributed solution that can process large graph.

Contributions

We consider the following firsts:

i) Mining over partitioned large input graph

ii)Hybrid approach that leverages both MPI (message passing interface) and threads

iii)Scale over a billion vertex graph; experiments on IBM Blue Gene/Q (upto 2048 nodes)

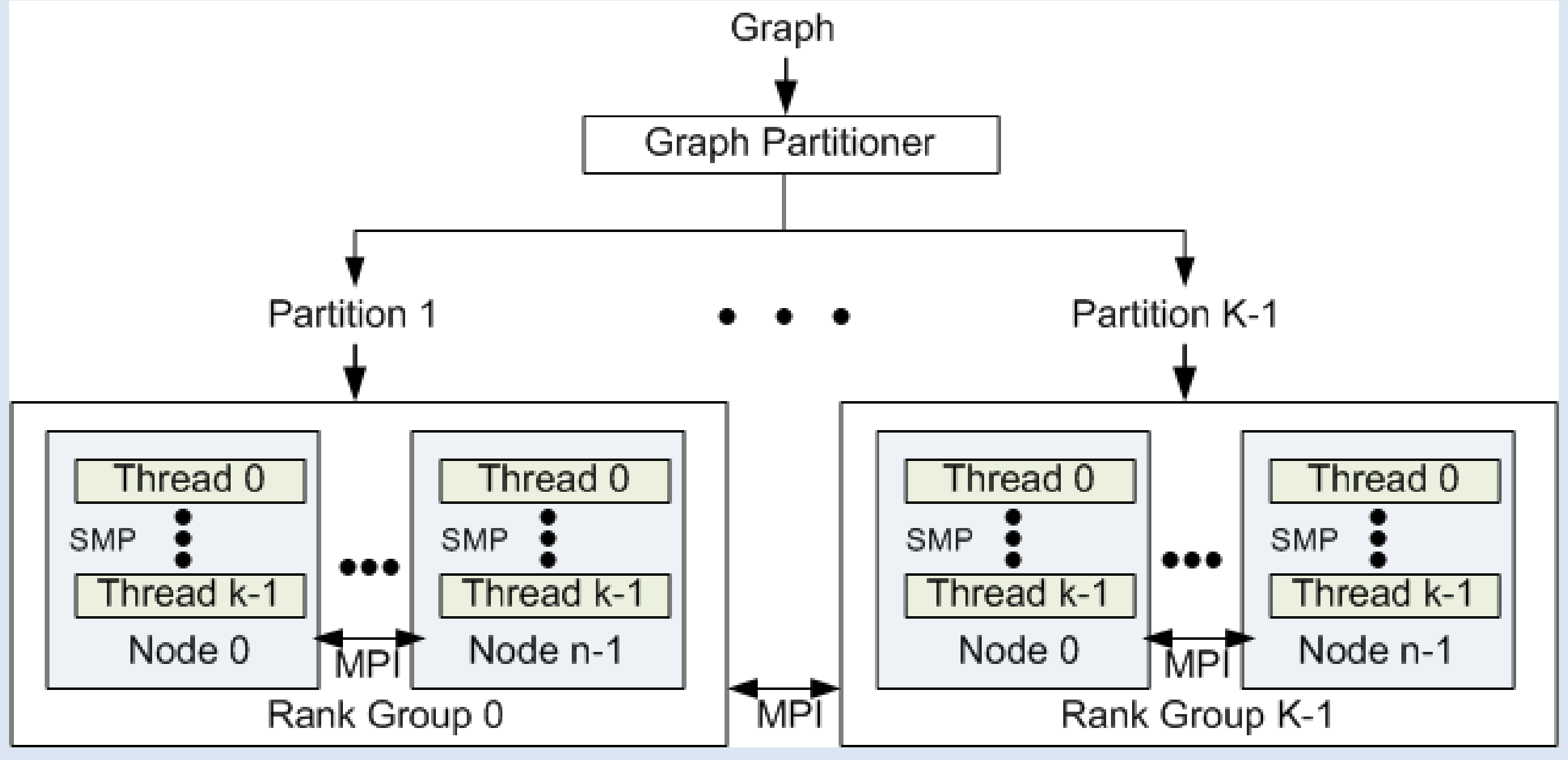


Figure: General System Architecture

Acknowledgement: Joint work with Dr. Mohammed Zaki, Professor, RPI. It was supported by NSF Award IIS-1302231.

Challenges

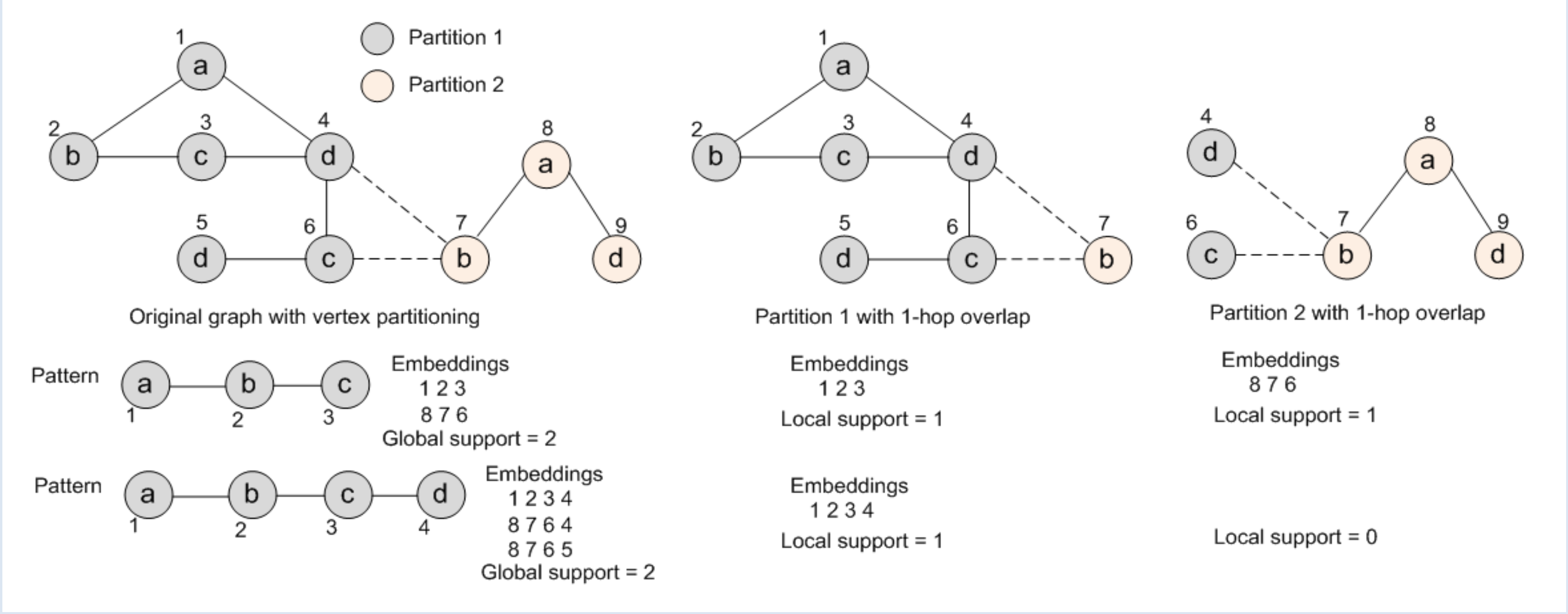
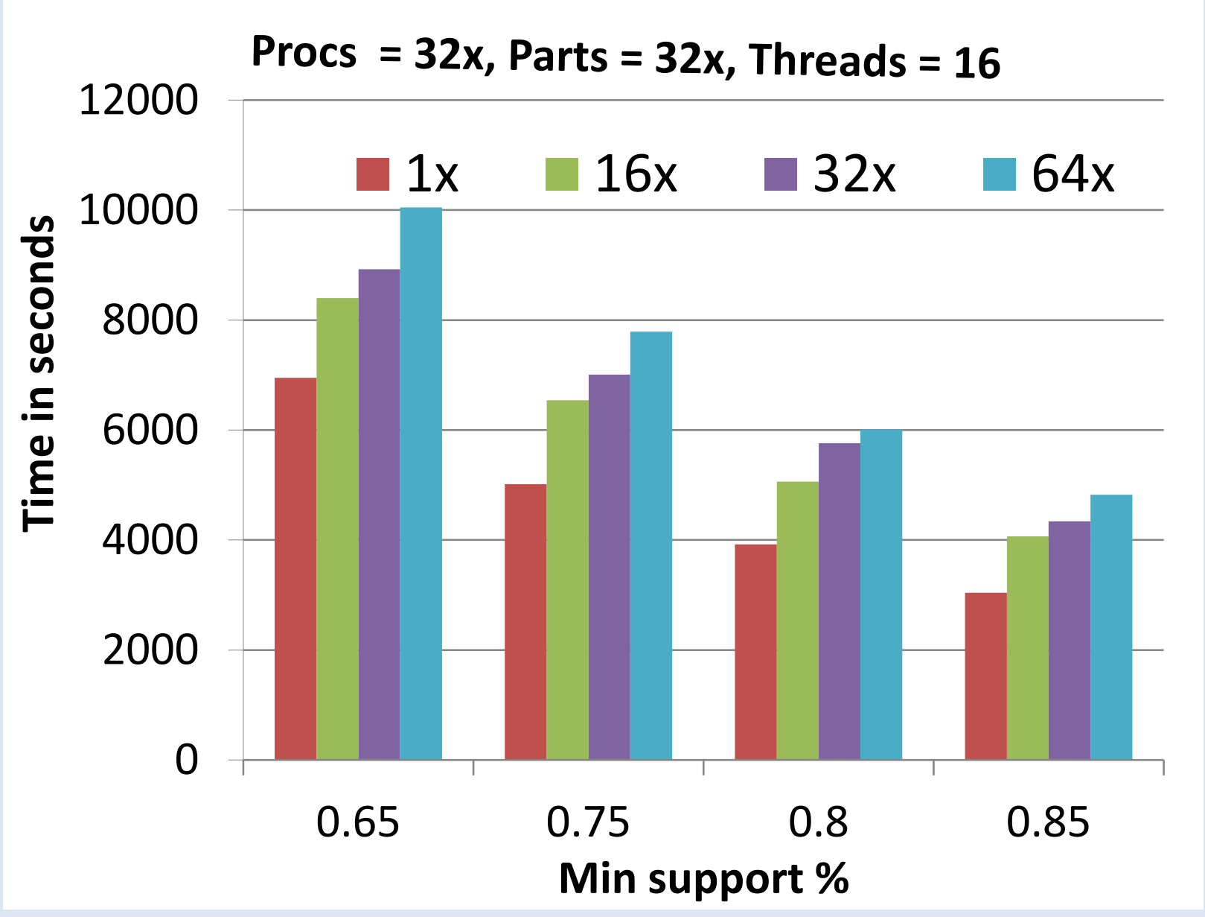


Figure: Computing support from partitioned input

1. ***Large input graph***: The input graph is split into multiple parts with 1 hop overlap. Graph partitioning is hard; existing solutions use heuristics. A balanced partition is desired.
2. ***False negatives***: Certain edges involved in the isomorphisms of a pattern can span across partitions, and thus missed.
3. ***Local support and prunning***: Need a local support measure that can effectively prune patterns that are globally infrequent.
4. ***False positives***: Vertex mappings of a pattern can be large, and are exchanged only when a pattern is estimated to be globally frequent.

Broader Impact



Our distributed mining approach made it possible to discover patterns from massive networks.

The scalability plot shows performance from scaled up Protein interaction graph (www.rcsb.org). 1x graph consists 17.4 million vertices and 68.5 million edges. Therefore, 64x consists more than 1 billion vertices.

References

1. M. Kuramochi et al., “Finding frequent patterns in a large sparse graph”, Data Mining and Knowledge Discovery, vol. 11, no. 3, 2005.
2. S. Reinhardt et al., “A multi-level parallel implementation of a program for finding frequent patterns in a large sparse graph”, Parallel and Distributed Processing Symposium, 2007.
3. M. Elseidy et al., “Grami: Frequent subgraph and pattern mining in a single large graph”, VLDB endowment, vol. 7, no. 7, 2014.