

Graph Mining and Its Applications

Jian Pei

Simon Fraser University, Canada

www.cs.sfu.ca/~jpei

jpei@cs.sfu.ca

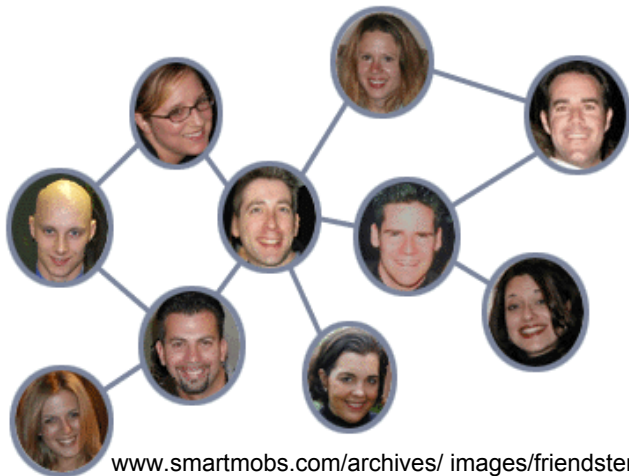
Joint work with Daxin Jiang at NTU, Singapore, and Aidong Zhang at SUNY Buffalo

Graphs as Data Models

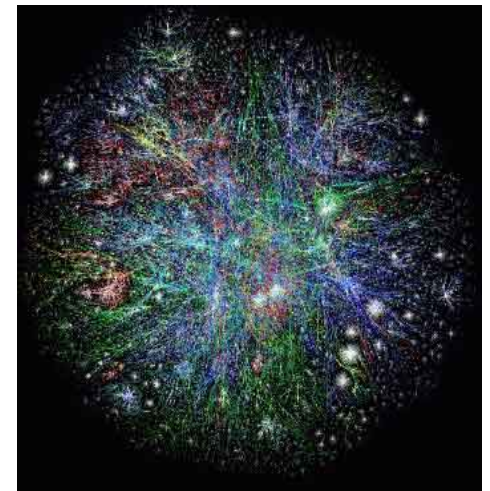
- Have been used extensively in many applications
 - Undirected graphs versus directed graphs
 - Weighted graphs, edge-labeled graphs, ...
- Structured and unstructured models
 - Well structured in terms of nodes and edges
 - Random graphs
- Understanding graphs – How?
 - Micro view: nodes and edges
 - Macro view: global structures and properties
 - How to build the bridge?

WWW and Social Networks

- A huge (random) graph – millions or billions of nodes
- “Small world”
 - On average, sparse, small in-/out-degree per node
 - Short diameter in expectation



www.smartmobs.com/archives/images/friendster.gif



www.anotherthink.com/my_graphics/Internet-map.jpg

Biological Networks

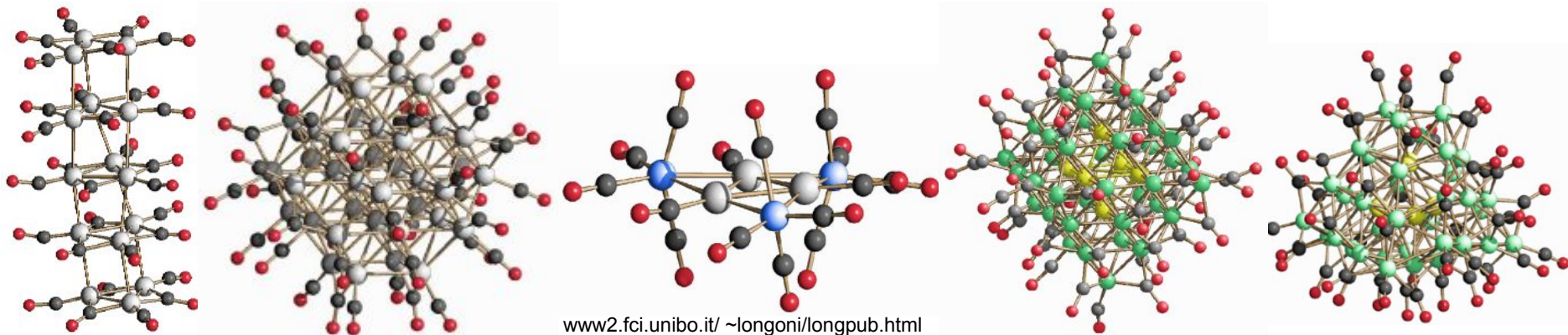
- Example: Protein-protein interaction network
 - A few proteins interact with a large number of other proteins, while most proteins have only one or two links
- Thousands of nodes
- Somehow predictable structure
- A few networks exist



www.nd.edu/~networks/linked/newfile17.htm

Chemical Compound Structures

- Up to several thousand nodes
- Highly regular structures
- Millions of compounds

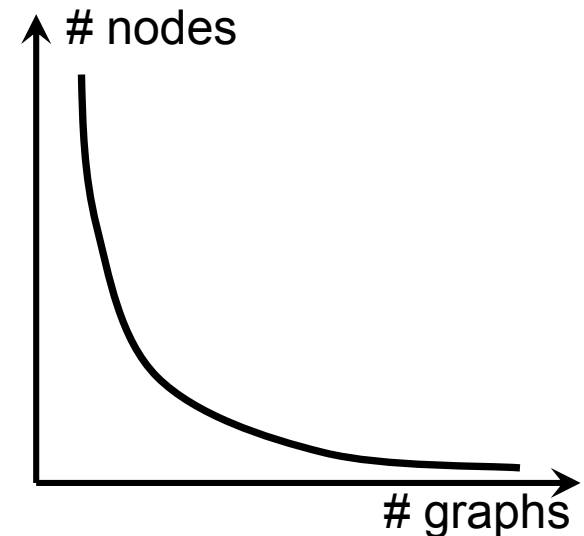


Graph Databases

- Data entries: graphs
 - Nodes and edges are highly structured, storage seems not a big problem
 - But, how to index?
- Queries
 - Conceptually, search for entries
 - But, what is the query language?
 - How to execute queries?

A Spectrum of Graph Databases

- Two major dimensions
 - Number of nodes per graph
 - Number of graphs in the database
- Another instance of power law?
 - Zipf distribution
- One dimension missing
 - The complexity of graphs!

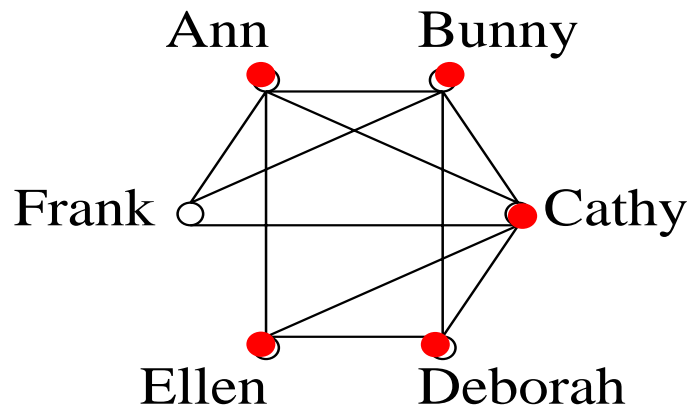


Graph Mining

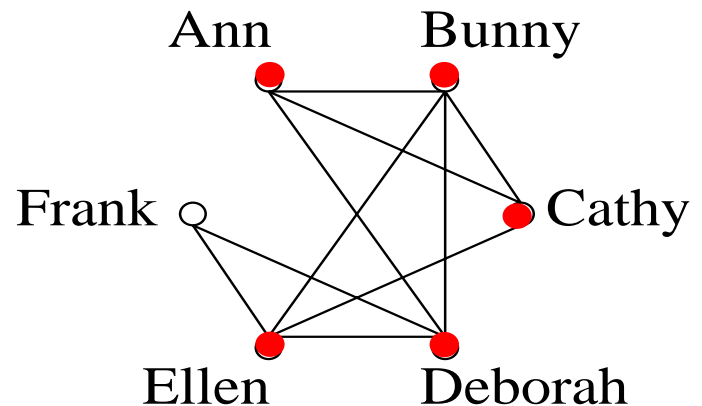
- Finding novel and potentially useful graph structures and properties from graph databases
- Graph structure mining
 - Frequent subgraph mining
 - Community mining
- Graph property mining
 - Changes of graph properties over time

X-Market Customer Segmentation

- $\{A, B, C, D, E\}$ is interesting – in each market, each customer is similar to at least three of the other four customers
 - Both the customers and the connectivity matter



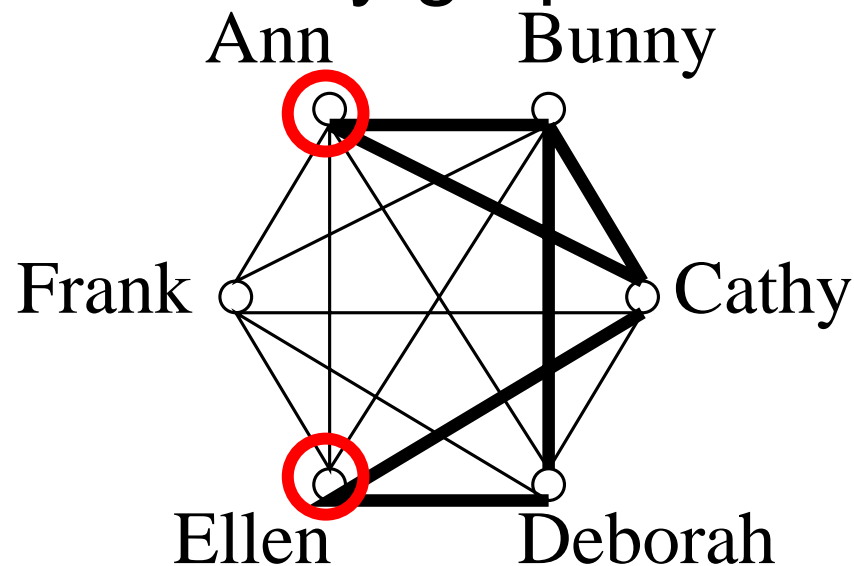
G_f : similarity graph in financial product market



G_c : similarity graph in consumer product market

Weighted Similarity Graph?

- Cluster {A, B, C, D, E} and the connectivity information cannot be found from the weighted similarity graph!

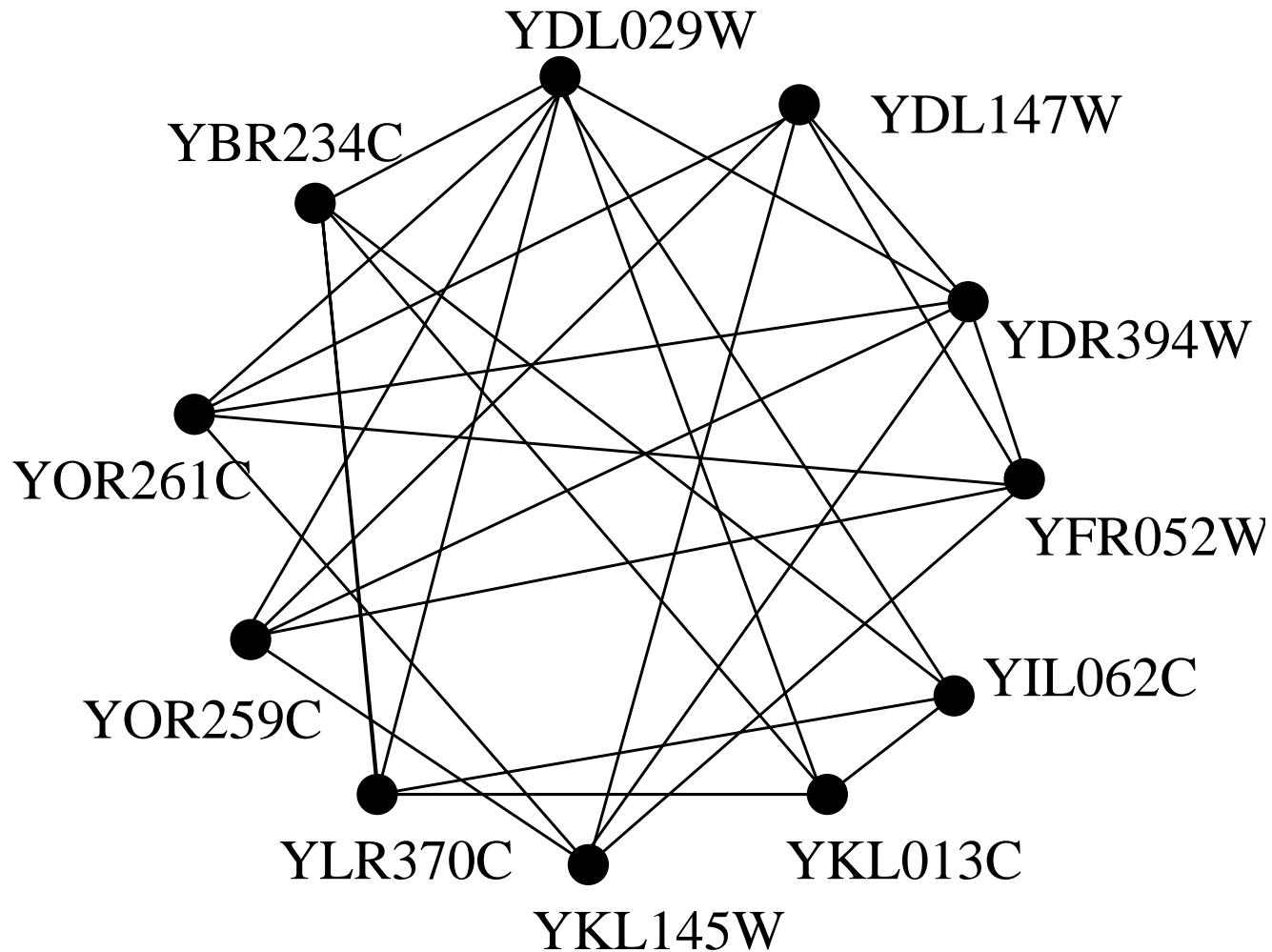


G_w : weighted similarity graph

Joint Mining in Bioinformatics

- Co-expressed genes from microarray data
- Interacting proteins from protein interaction data
- Connection: a protein is a product of a gene
- Joint mining of microarray data and protein interaction data
 - Both microarray data and protein interaction data are typically noisy
 - Can we validate clusters from mining the two types of data?
 - For many pathways, their genes exhibit a similar gene expression profile, and the protein products of the genes often interact

A Pattern from Yeast Data Sets



Intuition and Challenges

- Given a set of vertices and multiple graphs on the vertices, find the maximal subsets of vertices whose induced subgraphs in each graph are almost complete
- A weighted graph approach cannot capture the same clusters and information

Is It Frequent Graph Pattern Mining?

- Frequent graph pattern mining: given a set of graphs and a support threshold min_sup , find the complete set of graphs each of which is an embedded subgraph in at least min_sup graphs in the database
- Cross-graph quasi-cliques do not confine edges, only the connectivity (in degree) are checked
 - More realistic in some applications, especially for large graphs
 - Frequent graph patterns constrain both vertices and edges

Is It Mining Dense Areas?

- Intuitively, yes
- Technically, if only a density measure is used, a very dense area may give free rides to some poorly connected vertices
 - Example: a large clique + an isolated vertex
- Quasi-cliques set a threshold for minimal contribution from vertices to the quasi-cliques (“dense areas”)

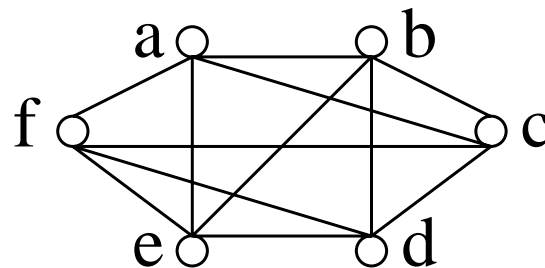
Quasi-Complete Graphs/Cliques

- A connected graph is a γ -complete graph ($0 < \gamma \leq 1$) if every vertex in the graph has a degree at least $\gamma(|V(G)| - 1)$
 - A 1-complete graph is a conventional complete graph
- In a graph G , a subset of vertices $S \subseteq V(G)$ is a γ -quasi-clique if $G(S)$ is a γ -complete graph, and no proper superset of S has this property

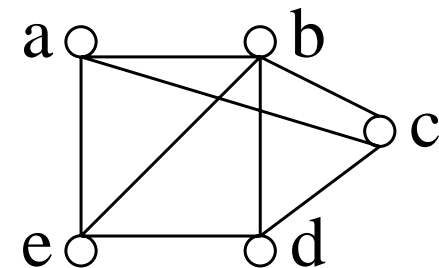
Quasi-Complete Graphs Monotonic?

- If $G(S)$ is a complete graph, then for any non-empty $S' \subseteq S$, $G(S')$ is also a complete graph
- The monotonicity does not hold for γ -complete graph if $\gamma < 1$

$\gamma=0.8$



G



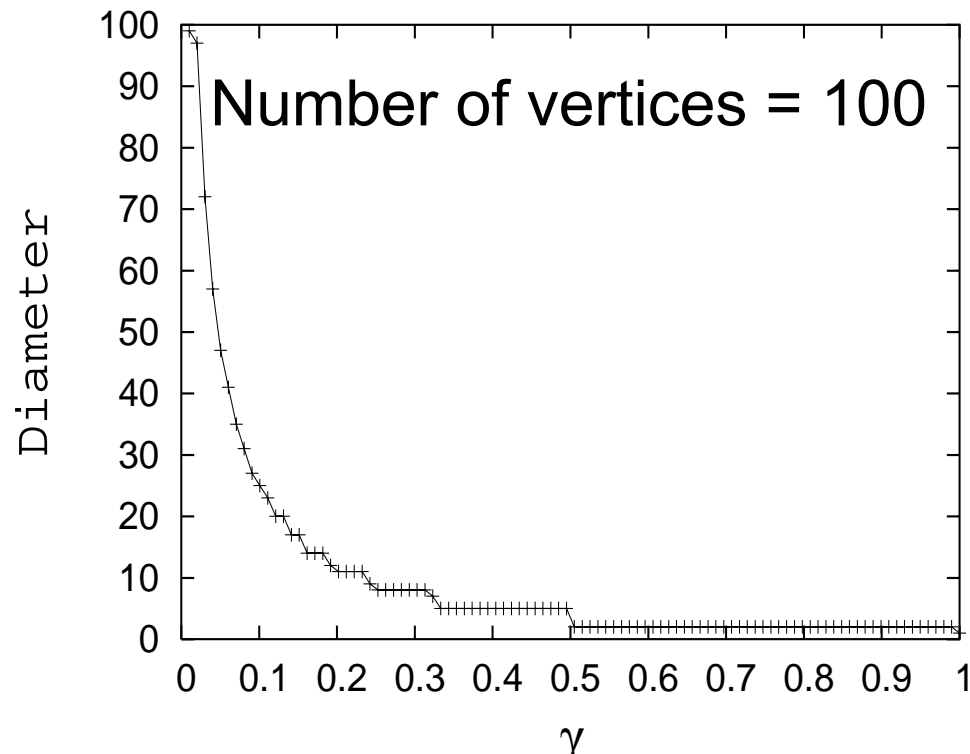
$G(\{a, b, c, d, e\})$

A Property of Quasi-Cliques

$$\text{diam}(G) \begin{cases} = 1 & \text{if } 1 \geq \gamma > \frac{n-2}{n-1} \\ \leq 2 & \text{if } \frac{n-2}{n-1} \geq \gamma \geq \frac{1}{2} \\ \leq 3 \lfloor \frac{n}{\gamma(n-1)+1} \rfloor - 3 & \text{if } \frac{1}{2} > \gamma \geq \frac{2}{n-1} \text{ and} \\ & n \bmod (\gamma(n-1)+1) = 0 \\ \leq 3 \lfloor \frac{n}{\gamma(n-1)+1} \rfloor - 2 & \text{if } \frac{1}{2} > \gamma \geq \frac{2}{n-1} \text{ and} \\ & n \bmod (\gamma(n-1)+1) = 1 \\ \leq 3 \lfloor \frac{n}{\gamma(n-1)+1} \rfloor - 1 & \text{if } \frac{1}{2} > \gamma \geq \frac{2}{n-1} \text{ and} \\ & n \bmod (\gamma(n-1)+1) \geq 2 \\ \leq n-1 & \text{if } \gamma = \frac{1}{n-1} \end{cases}$$

How to Tune Parameter γ ?

- Good News: the diameter of a γ -complete graph is relatively insensitive to γ

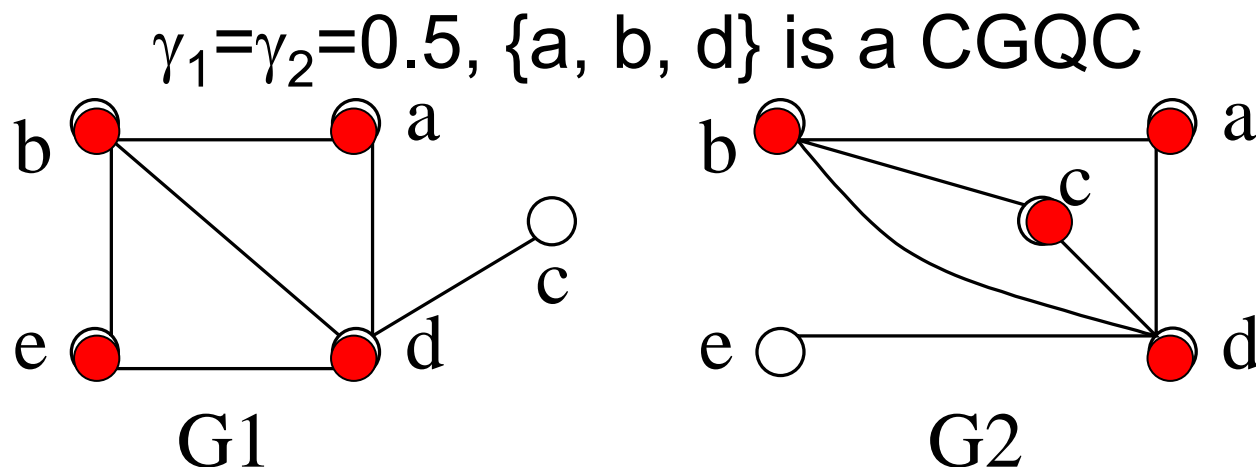


Cross-Graph Quasi-Cliques

- A set of graphs G_1, \dots, G_n and parameters $\gamma_1, \dots, \gamma_n$ such that $V(G_1) = \dots = V(G_n) = U$ and $0 < \gamma_1, \dots, \gamma_n \leq 1$
- A subset $S \subseteq U$ is a cross-graph quasi-clique (CGQC) if
 - (Quasi-complete) $G_i(S)$ is a γ_i -complete graph for $1 \leq i \leq n$
 - (Maximal) No any proper superset of S has the property
 - (Significant) S has at least \min_s vertices

CGQC and Quasi-Cliques

- If $n=1$ (only one member graph), a CGQC is a quasi-clique
- Generally, a CGQC may not be a quasi-clique in member graphs – may not be local maximal



How to Mine CGQCs?

- Generally, we cannot mine from an integrated graph
 - The integrated graph method works if $\gamma_1 = \dots = \gamma_n = 1$
 - Generally, we have to take a joint-mining approach
- The problem of counting the number of cross-graph quasi-cliques is in #P-Complete
- A difficult problem!

The First Try

- Observation: If S is a CGQC, then $G_1(S)$ must be a γ_1 -complete graph
- Step 1: mine the complete set of γ_1 -complete subgraphs in G_1 that have at least \min_s vertices
 - Do not need to find all γ_i -complete subgraphs
- Step 2: for the set of vertices in each γ_1 -complete subgraph found in step 1, check whether it is a CGQC

The Rudimentary Method Good?

- May not be efficient
 - Compute all γ_1 -complete subgraphs in G_1
 - All vertices and edges are taken into account in the mining, while some of them may not lead to CGQCs
- The rudimentary method conducts the joint-mining late!
 - A method more “aggressive” on joint-mining?

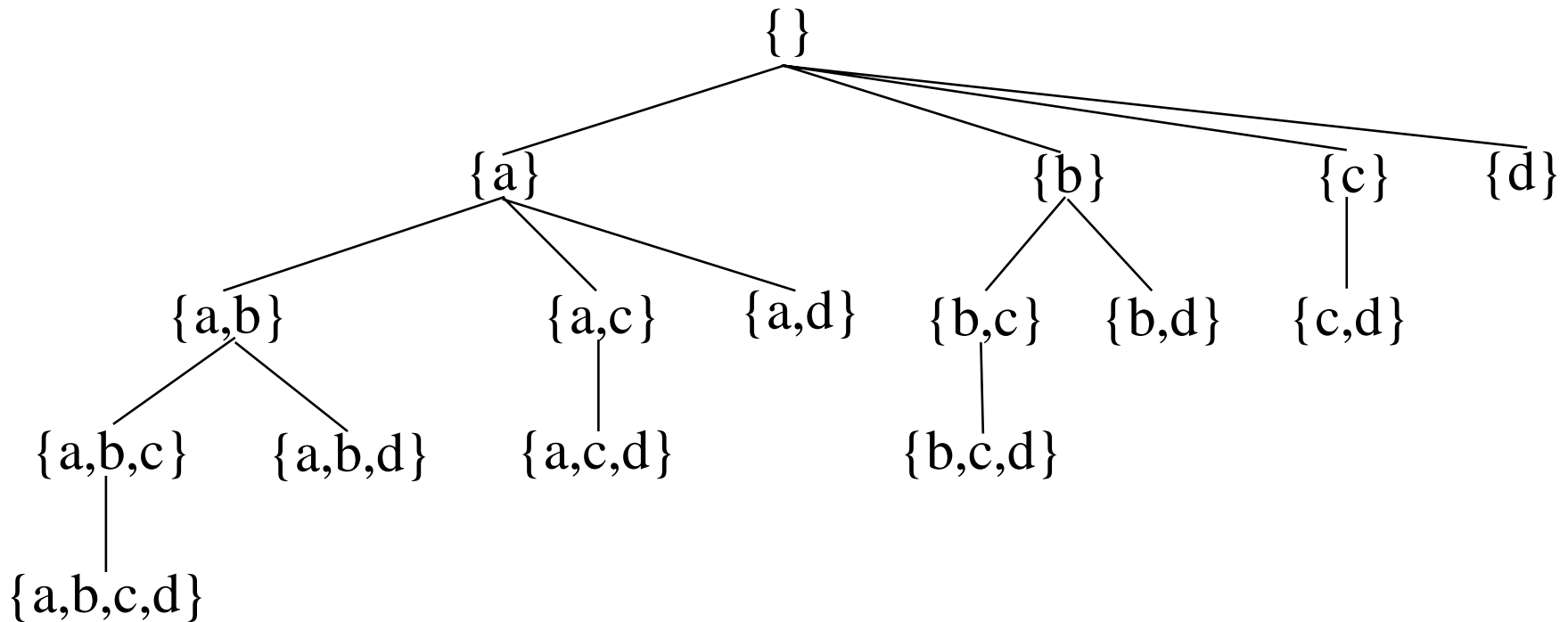
Algorithm Crochet

- Depth-first enumeration of vertex subsets
 - Guarantee the completeness of answers
- At each step, be efficient!
 - Aggressively reduce graphs
 - Dynamically choose the order to search children
 - Sharply prune futile subtrees



Depth-First Set Enumeration

- Use the popularly used set enumeration tree



Reducing Vertices and Edges

- Reducing vertices
 - For vertex v , if $\text{deg}(v)$ in one member graph is insufficient to form a quasi-complete subgraph, then v can be pruned
 - The pruning can be applied repeatedly since removing a vertex may reduce the degrees of some other vertices
- Reducing edges
 - If there exists a member graph G_i with $\gamma_i=1$, then for any vertices u and v that are not connected in G_i , (u, v) can be removed from the other member graphs
- Reducing vertices and edges iteratively

Combining Graphs

- If $\gamma_1 = \gamma_2 = 1$, S is a cross-graph quasi-clique if and only if $G_1(S)$ and $G_2(S)$ are both complete graphs
 - S must be a clique in $G = (E_1 \cap E_2, V_1 \cap V_2)$
- All the member graphs with $\gamma_i = 1$ can be combined into one graph
- Reducing graphs can make the mining faster

Pruning Set Enumeration Trees

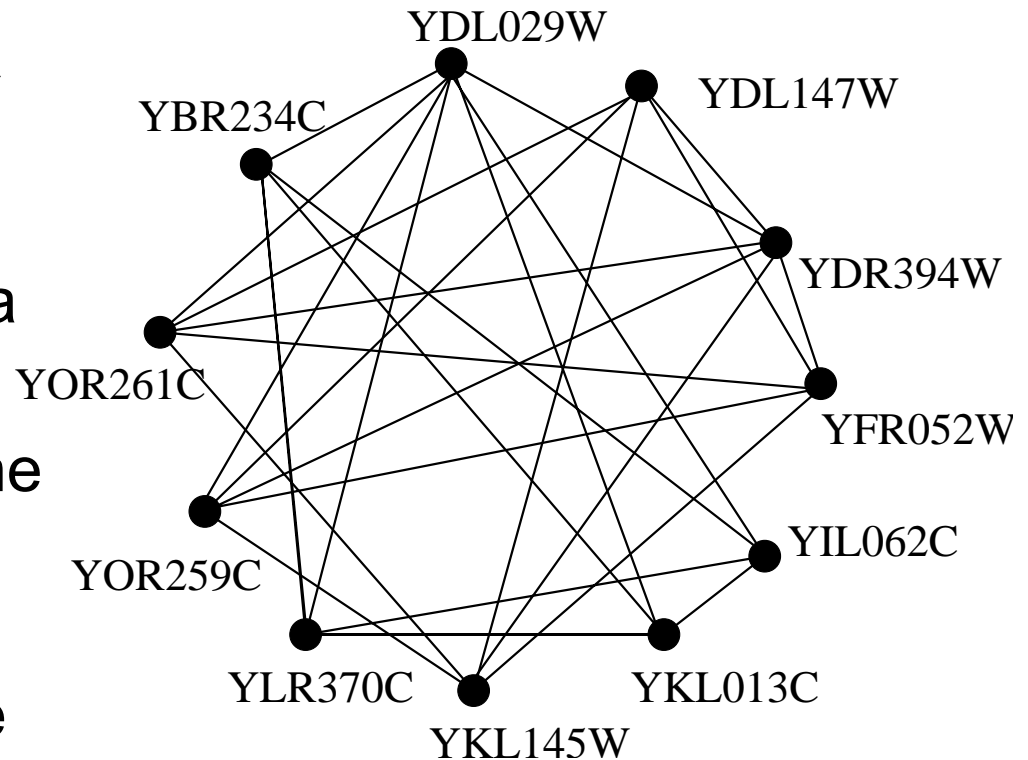
- At node S , only the vertices that are “well connected” to S should be used to expand S
 - Using the diameter of quasi-cliques as bounds, details in the paper
- At node S , let S' be the set of all vertices that are “well connected” to S , if $S \cup S'$ does not contain any quasi-clique, then the subtree of S can be pruned
- Search the children nodes in the well-connected-ness descending order

Generalization and Extensions

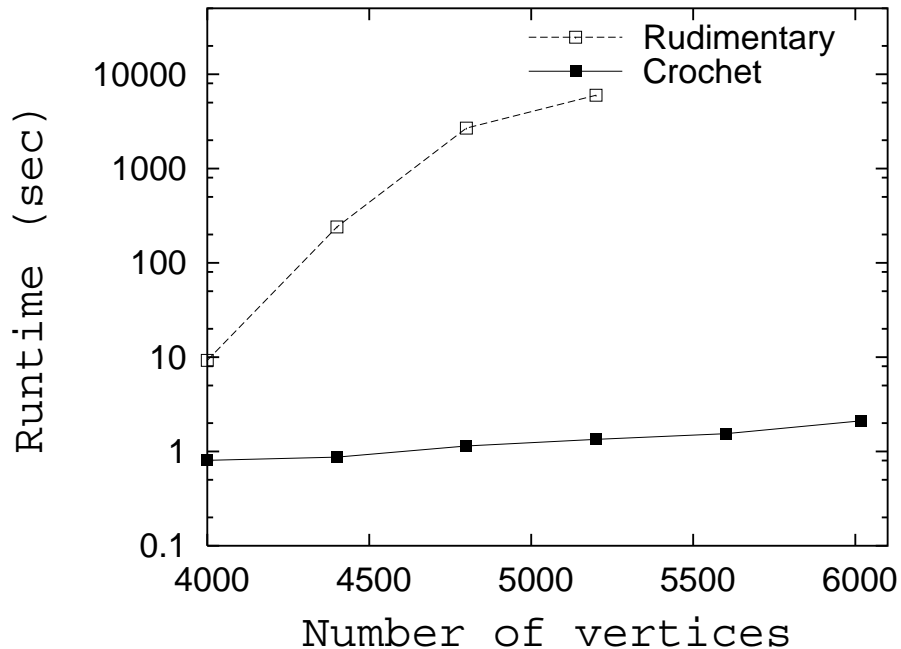
- Mining more than two graphs
- Mining with non-bijective mapping functions
- Details in the paper
- Algorithm Crochet is designed for mining a few (e.g., tens of) large graphs (i.e., each graph contains thousands of vertices and tens of thousands of edges)

Experimental Results – Yeast Data

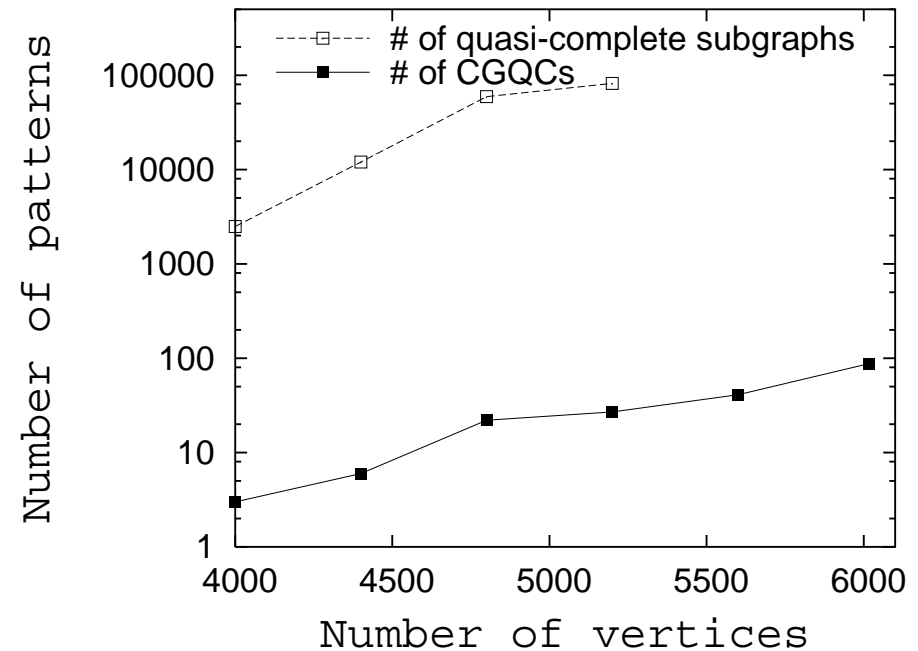
- CDC28 microarray data set and protein-protein interaction data DIP
- 4,668 vertices after data cleaning
- 865,080 edges in G_E , the microarray data graph, $\gamma_E=1$
- 15,115 edges in G_P , the protein-protein data graph, $\gamma_P=0.4$



Crochet vs. the Rudimentary Method



$$\gamma_1=1, \gamma_2=0.5, \min_s=5$$



Effectiveness of the Techniques

Technique	Runtime w/o the tech.	Runtime w/ the tech.	Speedup
Graph reduction & projection	27.819	16.392	1.697
Heuristic 1	44.274	16.392	2.701
Rule (1), Lemma 3.6	16.765	16.392	1.023
Rule (2), Lemma 3.6	130.460	16.392	7.959
Rule (3), Lemma 3.6	10.838	16.392	0.661
Rule (4), Lemma 3.6	338.916	16.392	20.676

Heuristic 1: dynamic ordering vertices in the vertex enumeration tree

Rule 1: pruning subtrees containing insufficient vertices

Rule 2: pruning by vertex and edge reduction

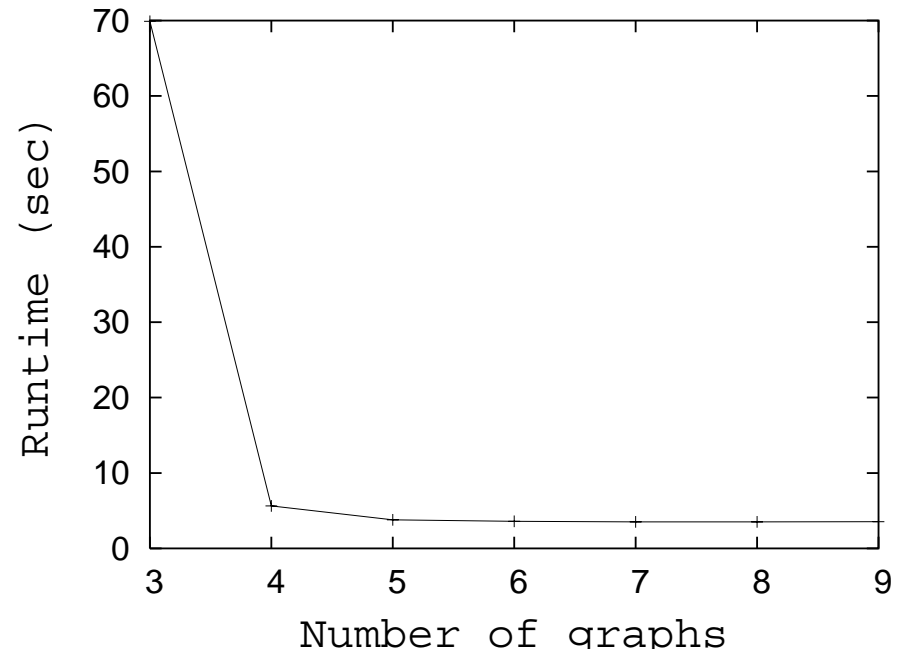
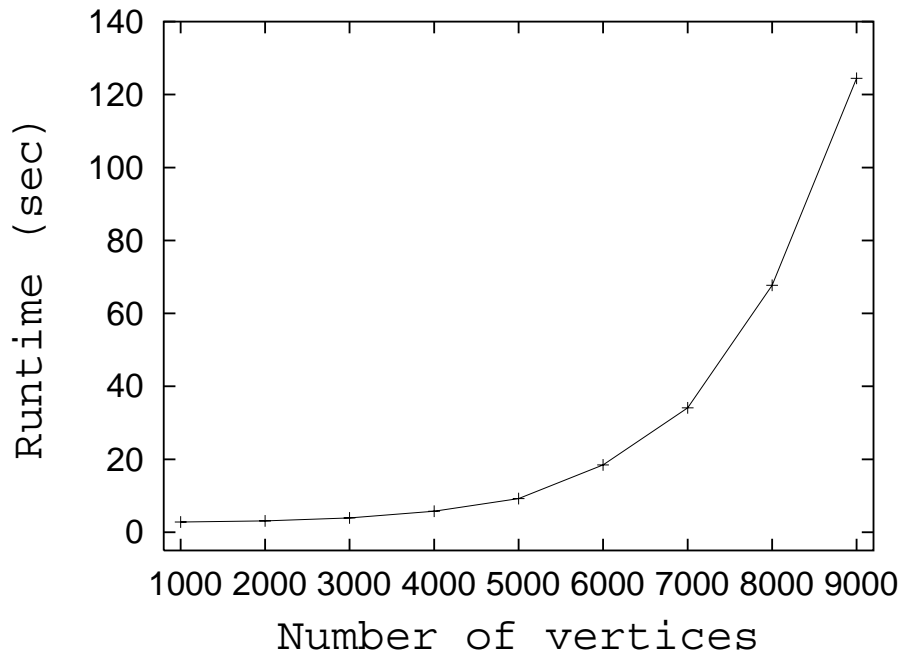
Rule 3: outputting only maximal patterns (required by finding quasi-cliques)

Rule 4: pruning using parameter γ_i

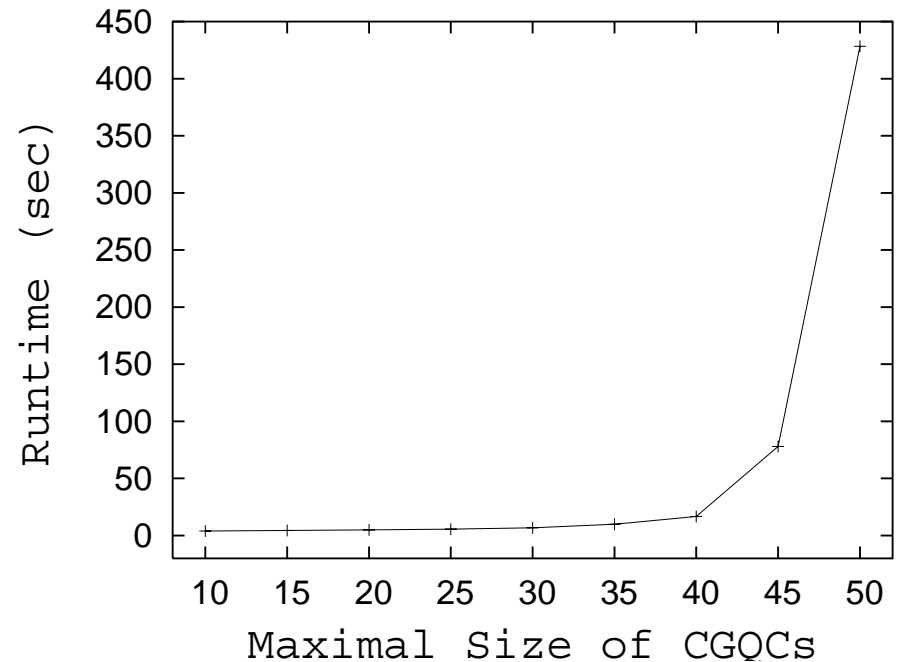
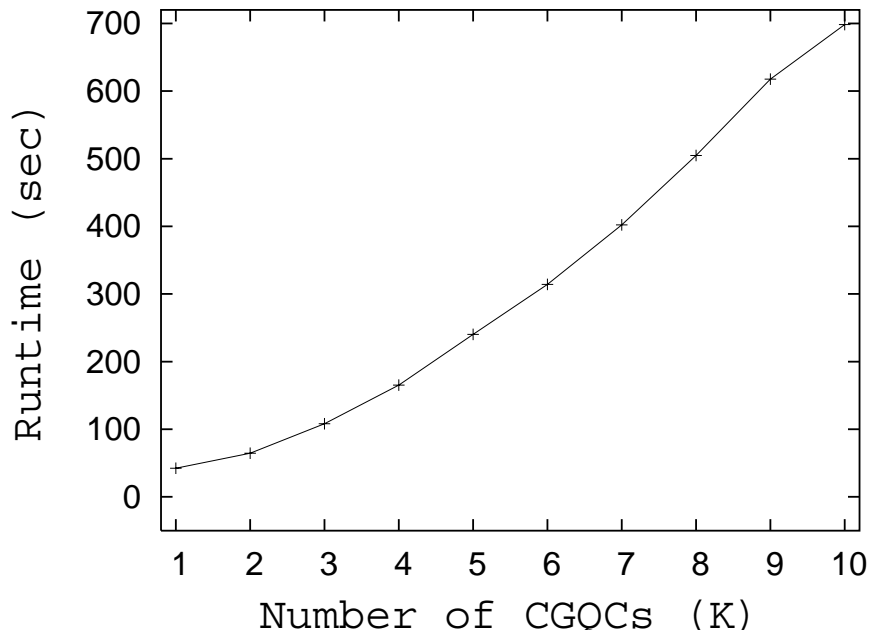
Synthetic Data Sets

- Generate graphs with various distributions
- Parameters
 - Number of graphs
 - Number of vertices
 - $\gamma_1, \dots, \gamma_n$
 - Expected number of cross-graph quasi-cliques
 - Density of the graphs (the number of edges versus the number of vertex pairs)

Mining Large/Many Graphs



Mining Large/Numerous Patterns



Conclusions

- Graph mining has important applications
- Graph mining is challenging
 - Effectiveness: what to mine?
 - Efficiency: how to mine?
- Research on graph mining
 - Graph databases and indexes
 - Efficient mining methods
 - Graph mining visualization

References

- J. Pei, D. Jiang, and A. Zhang. "On Mining Cross-Graph Quasi-Cliques". In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'05), Chicago, IL, USA, August 21-24, 2005.
- C. Wang, W. Wang, J. Pei, Y. Zhu and B. Shi. "Scalable Mining of Large Disk-based Graph Databases". In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), Seattle, WA, USA, August 22 - 25, 2004.