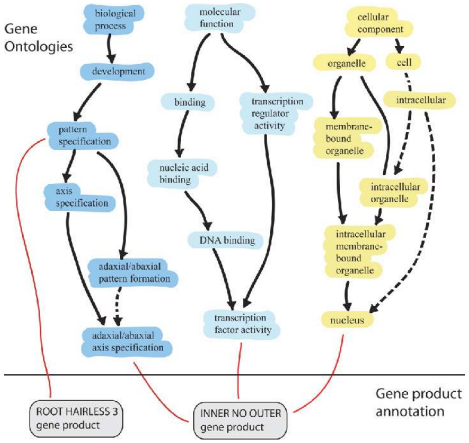
GraphBlast: multi-feature graphs database searching

Alfredo Ferro, Rosalba Giugno, Misael Mongioví,
Alfredo Pulvirenti, Dmitry Skripin, Dennis Shasha

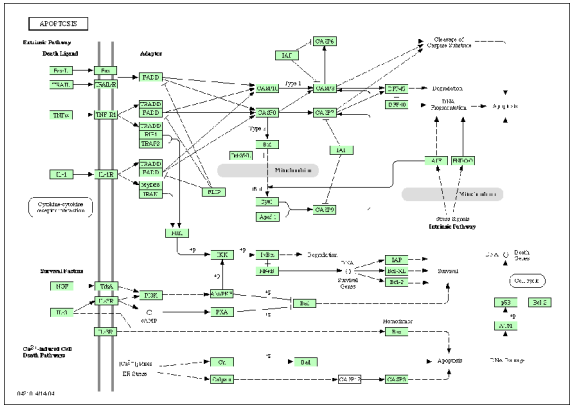
University of Catania,
New York University

June 12-15, 2007, University of
Pisa, Italy

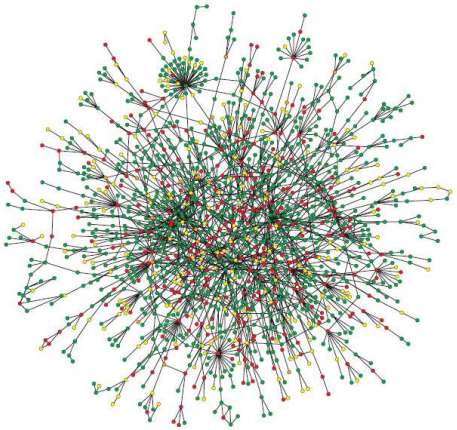
Graphs in Bio-Chemistry



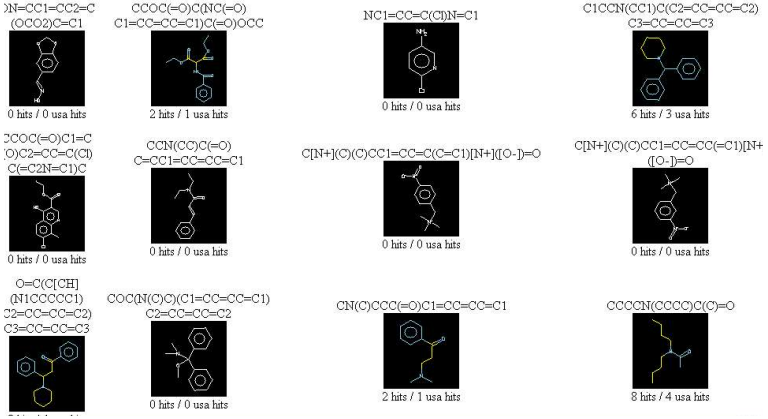
Gene Ontologies



Pathways



Biological Networks



Collection of Molecules

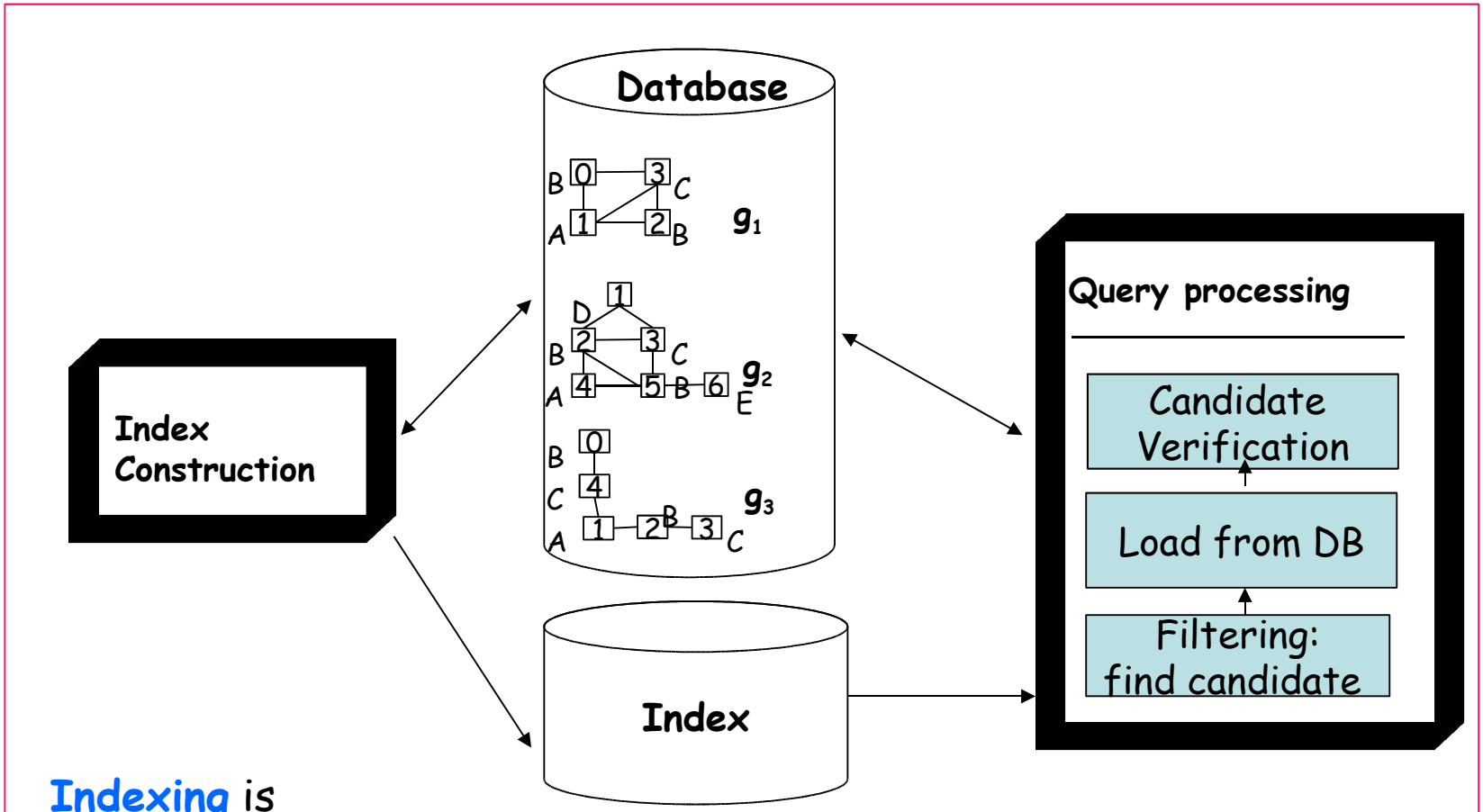
Motivation for Searching in Graphs (molecules, networks)

- Prediction of the functionality of new natural or synthesized compounds
- Make a compound Q more active

- Find fragment with the same function among different species
- Predict protein function
- Predict protein interaction

Graphs Searching Steps

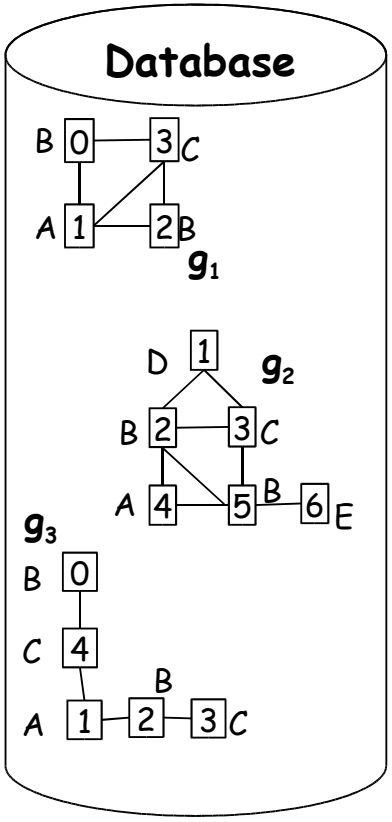
Graphs searching is an NP-problem



Indexing is crucial to **reduce the search** space and make the problem affordable!

GraphBlast Index

- For each graph in DB
  - Find all paths of length from 1 to L (4,10)
  - Count how many occurrences of each path in each graph



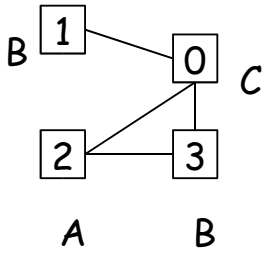
Key	g_1	g_2	g_3
h(CB)	2	2	2
….	…	…	….
h(ABCA)	2	0	0
….	….	….	….

GraphBlast Filtering

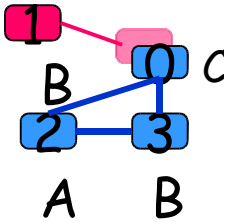
Query Processing

Query Fingerprint

key	Query
h(CB)	2
….	…
h(ABCA)	1
….	….



Query Decomposition in Patterns

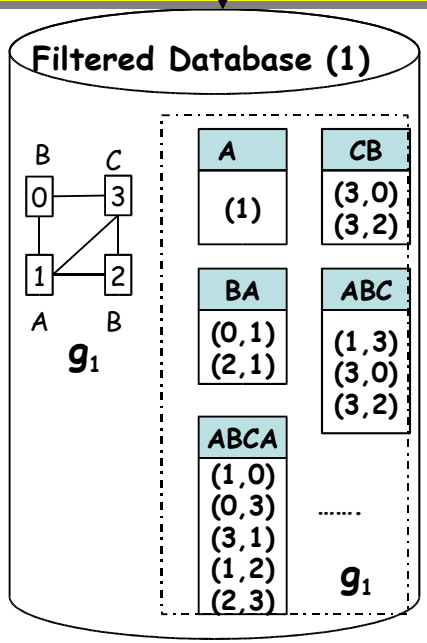


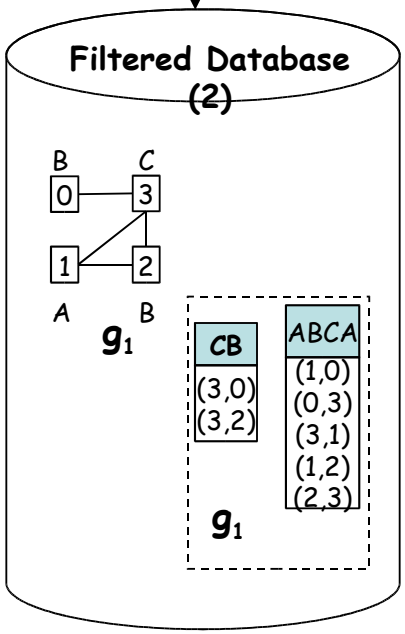
Key	g_1	g_2	g_3
h(CB)	2	2	2
….	…	…	….
h(ABCA)	2	0	0
….	….	….	….

Database Fingerprint

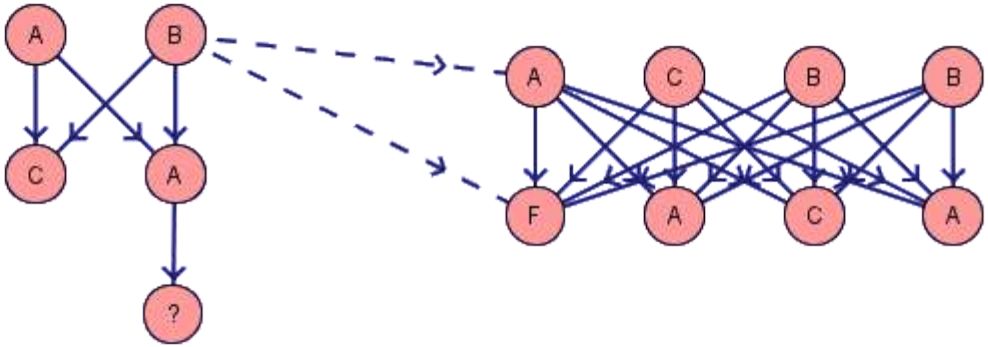
Filtering Step1: Select all Candidate Graphs

Filtering Step2: Select all Candidate SubGraphs

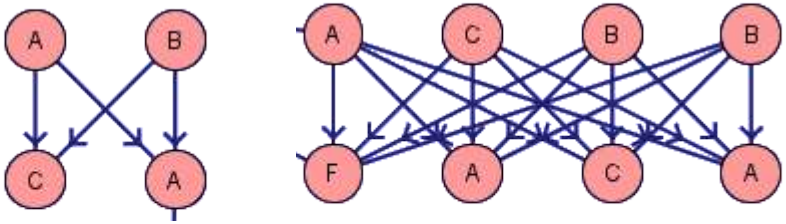




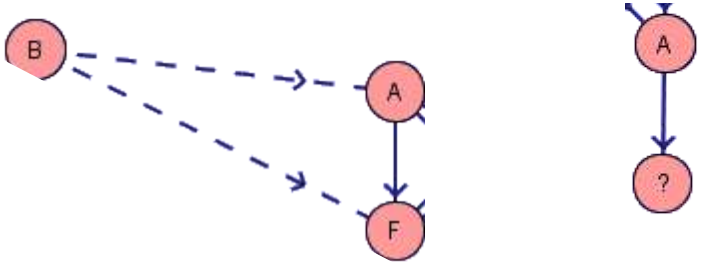
Approximate Searches



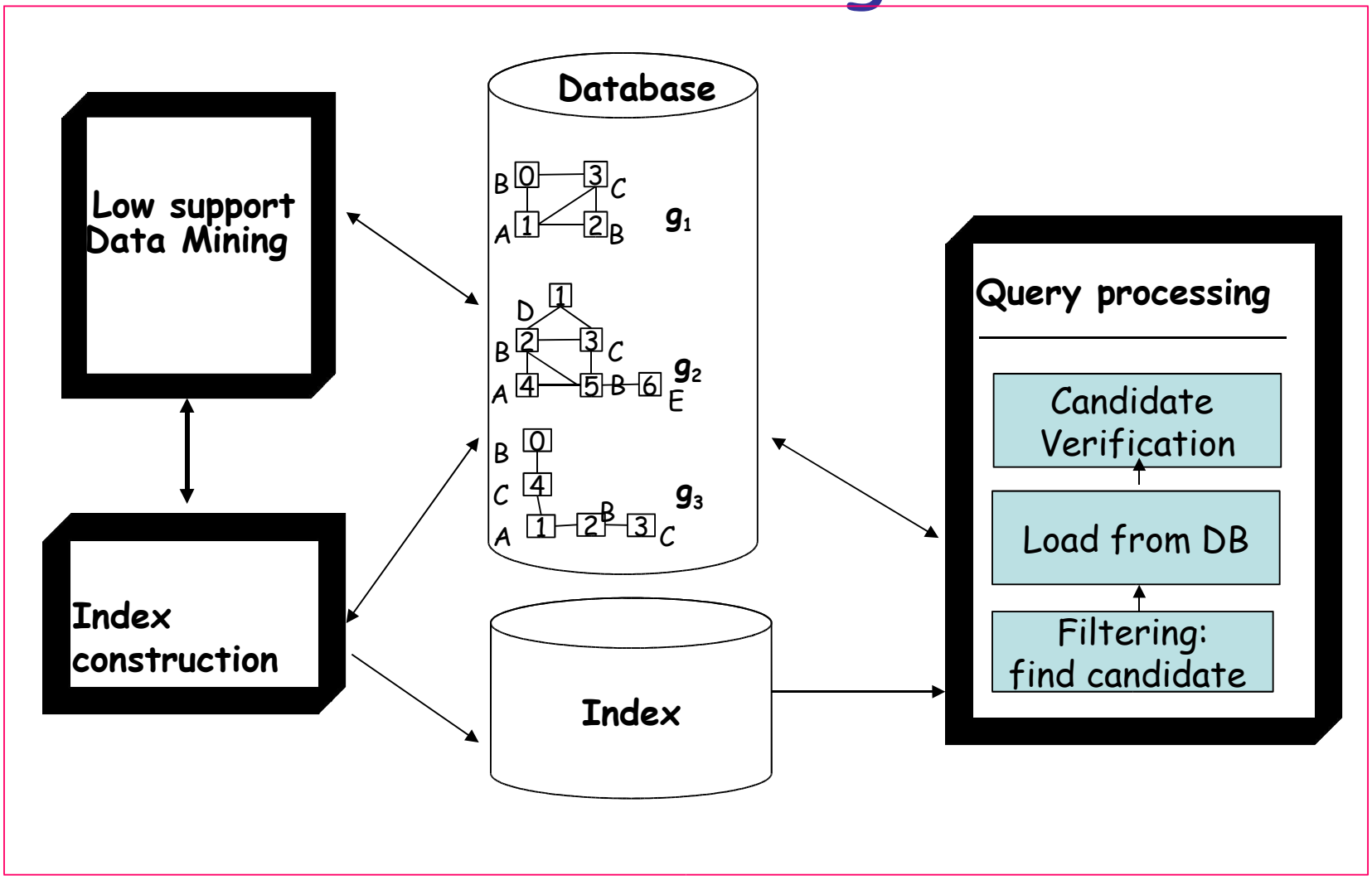
Run GraphBlast to Find occurrences for each full specified query subgraph



Check (DFS) if the approximate connections exist



Improving Indexing using Graph Data Mining



Min hashing: Low Support Data Mining Technique for Indexing Size Reduction

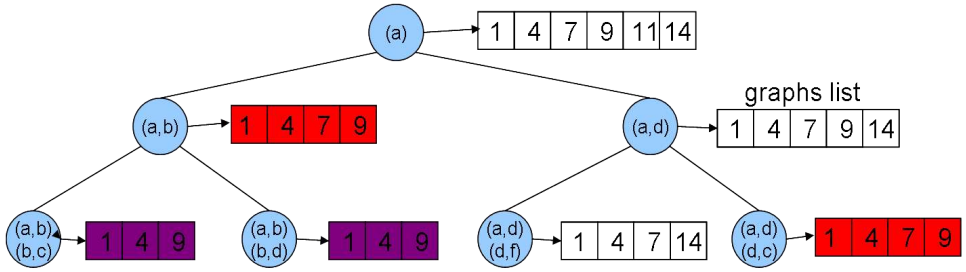
GraphBlast

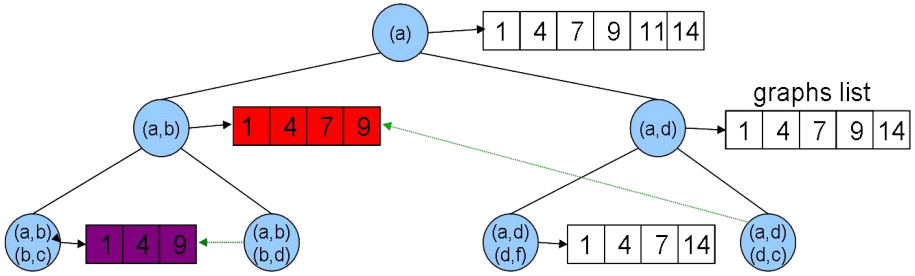
Key	g_1	g_2	g_3
h(CB)	4	8	2
h(AB)	2	2	2
h(ABC)	3	5	9
…..	…..	…..	….
h(ABCA)	4	8	2
h(ACB)	4	0	0
h(BACA)	2	2	2
…..	…..	…..	….

Min-Hashing

Key	g_1	g_2	g_3
h(CB); h(ABCA)	4	8	2
h(AB); h(BACA)	2	2	2
h(ABC)	3	5	9
…..	…..	…..	….
h(ACB)	4	0	0
…..	…..	…..	….

gIndex (TODS 2005)





Performance

Compression ratio % (Min-Hashing)

Database	GraphBlas	gIndex
Molecular	62.5	30.4
Regular 2D	56.4	76.6
Irregular 2D	56.9	52.9
Valence	48.0	9.4
Irr. Valence	48.0	7.9
Random	43.4	70.5

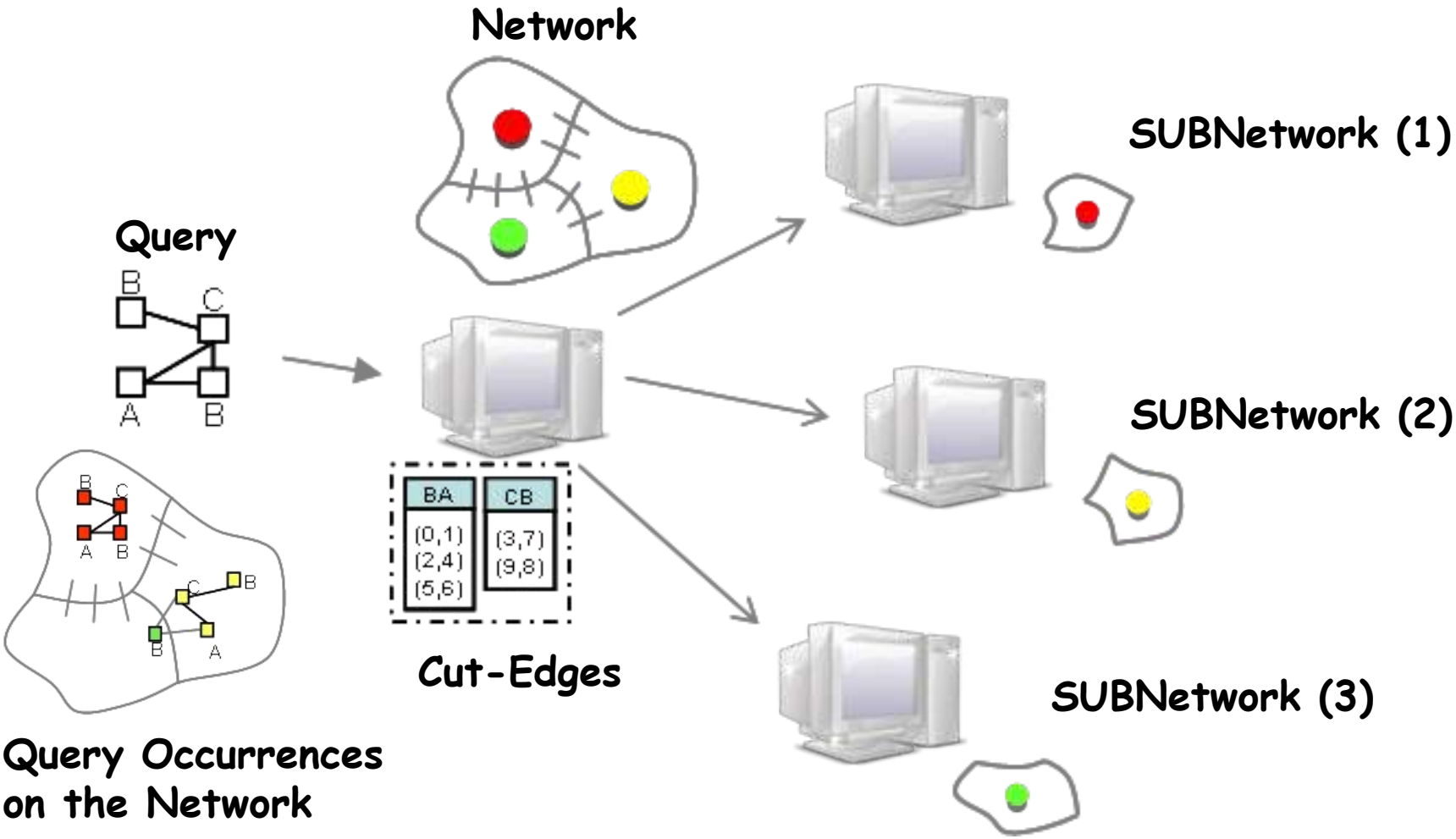
Preprocessing Time

Database	GraphBlast	gIndex
Molecular	1314.0	13750.0
Regular 2D	6.4	746.0
Irregular 2D	11.6	4587.3
Valence	6.2	7.2
Irr. Valence	6.3	7.6
Random	3511.0	> 3 days

Query Time-Molecular DB

Query Size	GraphBlast	gIndex
11	0.00	0.04
19	0.01	0.08
43	0.00	0.05
58	0.01	0.42
148	0.01	0.31
239	0.00	0.60

Distributed GraphBlast for searching in a Large Network



Reference

- D. Shasha, J.T-L Wang, and R. Giugno. Algorithmics and applications of tree and graph searching. Proceeding of the ACM Symposium on Principles of Database Systems (PODS), pages 39–52, 2002.
- Luigi P. Cordella, Pasquale Foggia, Carlo Sansone, Mario Vento: A (Sub) Graph Isomorphism Algorithm for Matching Large Graphs. IEEE Trans. Pattern Anal. Mach. Intell. 26(10): 1367-1372 (2004)
- R. Giugno, D. Shasha, GraphGrep: A Fast and Universal Method for Querying Graphs. Proceeding of the IEEE International Conference in Pattern recognition (ICPR), Quebec, Canada, August 2002.
- Yan X, Yu PS, Han J: Graph Indexing Based on Discriminative Frequent Structure Analysis. ACM Transactions on Database Systems 2005, 30 (4):960-993
- Cohen E, Datar M, Fujiwara S, Gionis A, Indyk P, Motwani R, Ullman JD, Yang C: Finding interesting associations without support pruning. IEEE Transactions on Knowledge and Data Engineering 2001, 13:64-78.
- Michelle Girvan, M. E. J. Newman, Community structure in social and biological networks, PNAS, June 11, 2002 vol. 99 no. 12 7821–7826.