

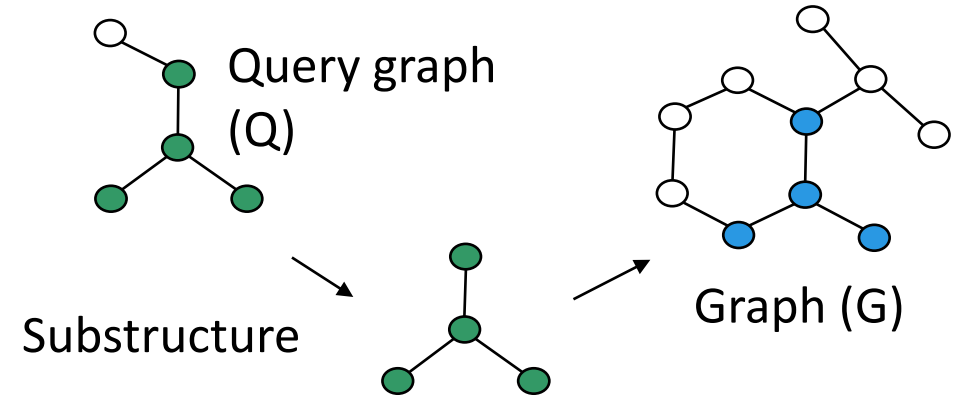
The background features a collage of various graph visualizations. At the top, there's a network with red edges and green nodes. Below it, a horizontal strip shows a sequence of small graphs with arrows indicating transitions. On the left, a vertical strip displays a graph with orange and blue nodes. The bottom half of the image shows a large, dense network with red edges and green nodes, similar to the top one. The text is overlaid on a white, angular shape in the center.

# Graph Pattern Mining

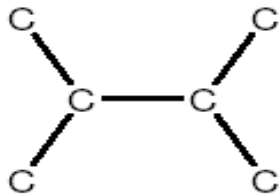
## Application I: Graph Indexing

# Application of Pattern Mining I: Graph Indexing

- ❑ Graph query: Find all the graphs in a graph DB containing a given query graph
- ❑ Index should be a powerful tool
- ❑ Path-index may not work well
- ❑ Solution: Index directly on substructures (i.e., graphs)

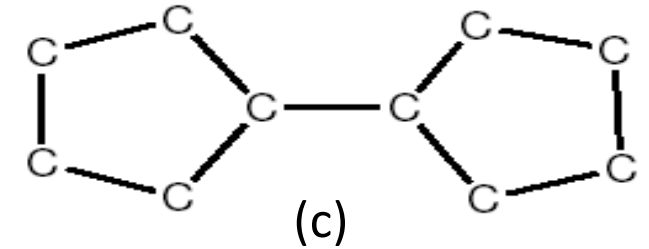
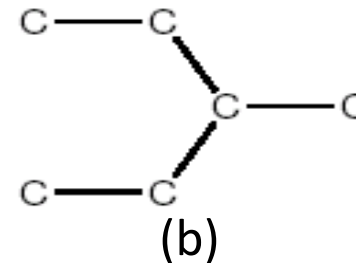
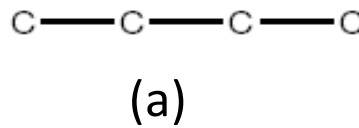


Query Q:



Only graph (c) contains Q

Graph DB:



Path-indices: C, C-C, C-C-C, C-C-C-C cannot prune (a) & (b)

# gIndex: Indexing Frequent and Discriminative Substructures

- Why index frequent substructures?
  - Too many substructures to index
  - Size-increasing support threshold
  - Large structures will likely be indexed well by their substructures
- Why discriminative substructures?
  - Reduce the index size by an order of magnitude
- Selection: Given a set of selected structures  $f_1, f_2, \dots, f_n$ , and a new structure  $x$ , the extra indexing power is measured by
$$\Pr(x|f_1, f_2, \dots, f_n), f_i \subset x$$
when  $\Pr(x|f_1, f_2, \dots, f_n)$  is small enough,  $x$  is a discriminative structure and should be included in the index
- Experiments show that gIndex is small, effective, and stable

