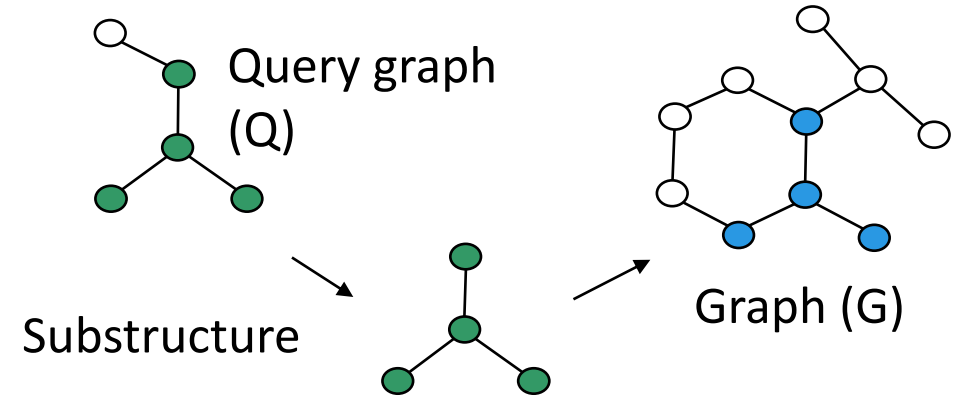


The background of the slide is a complex, abstract composition. It features a network graph with numerous green nodes and red edges, overlaid on a light blue and white geometric pattern. The text is centered in a large, bold, black font.

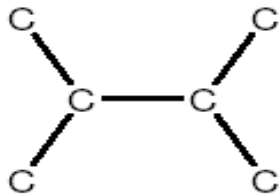
# **Session 5. glIndex: A Graph Indexing Method**

# Application of Pattern Mining: 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)

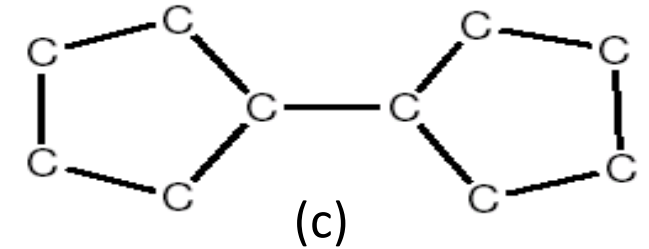
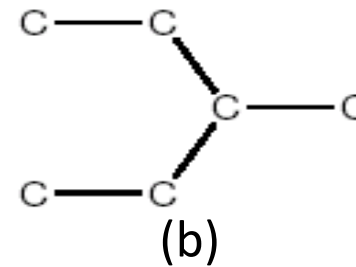
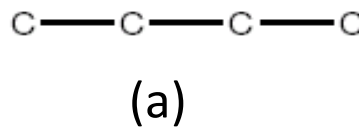


Query Q:



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 gIndex is small, effective and stable

