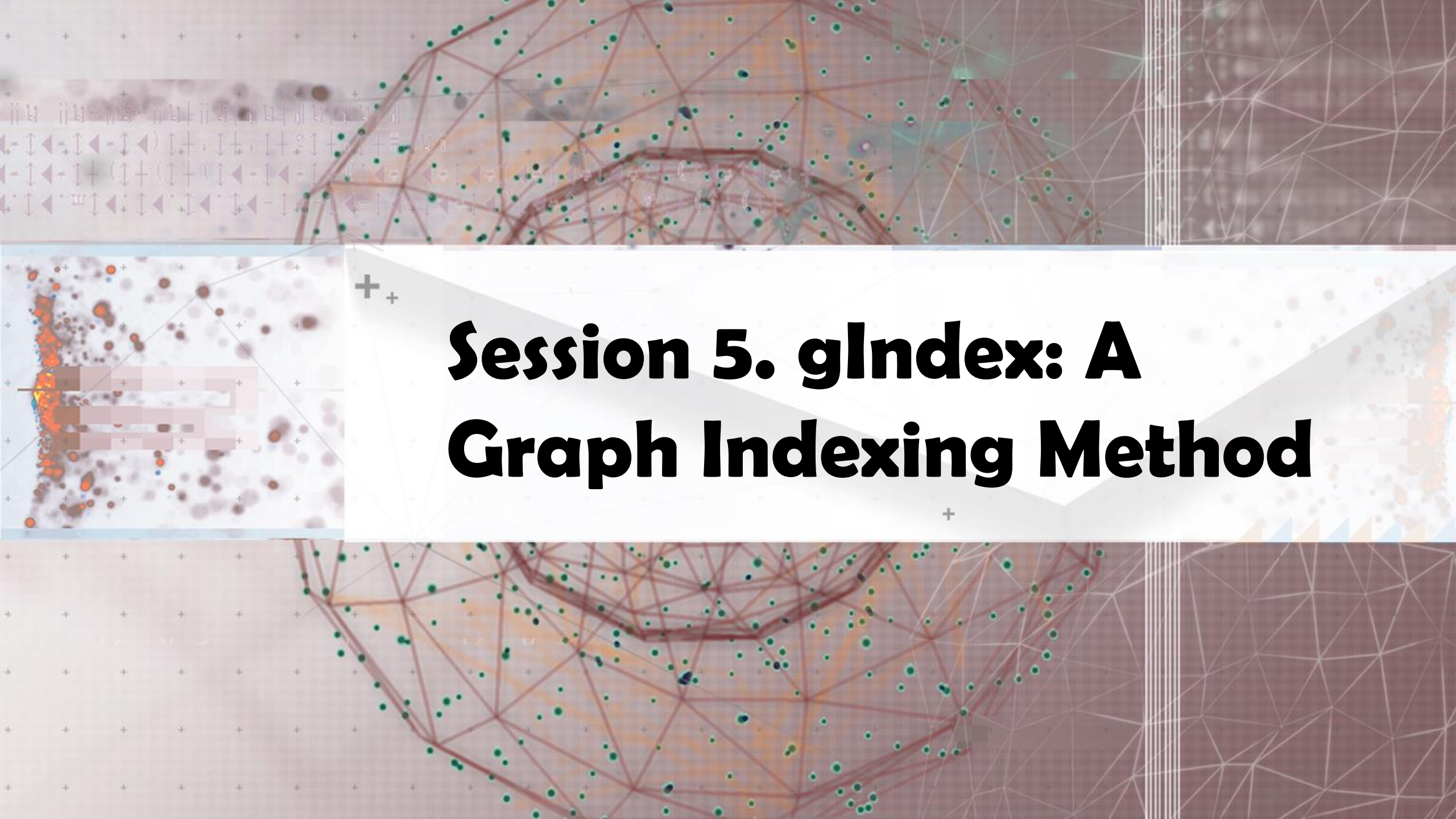
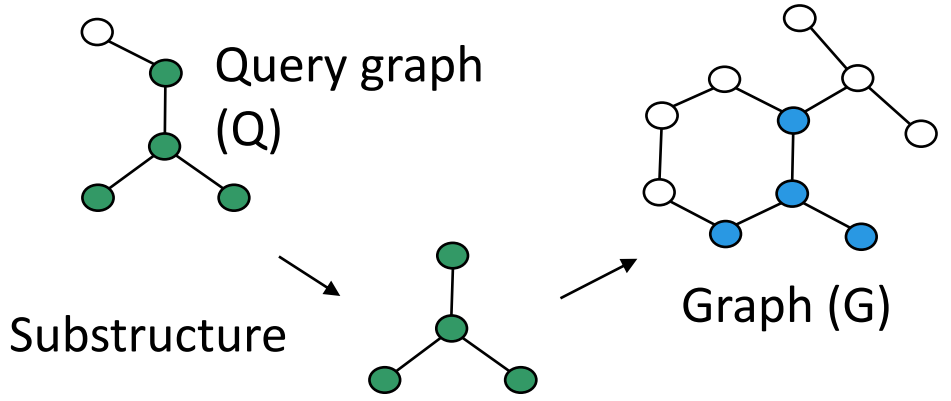


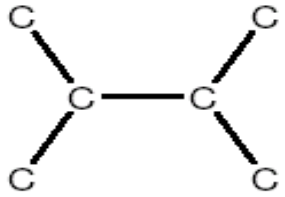
Session 5. gIndex: A Graph Indexing Method

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)

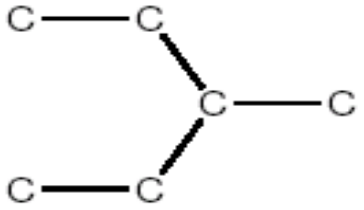


Query Q:

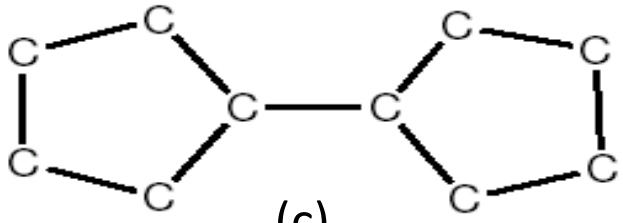


Graph DB:

(a)



(b)



(c)

Only graph (c) contains Q

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 $f_1, f_2, .. f_n$, and a new structure x, the extra indexing power is measured by

$$\Pr(x | f_1, f_2, \ldots f_n), f_i \subset x$$

when $\Pr(x | f_1, f_2, \ldots, f_n)$ is small enough, x is a discriminative structure and should be included in the index

- Experiments show gIndex is small, effective and stable