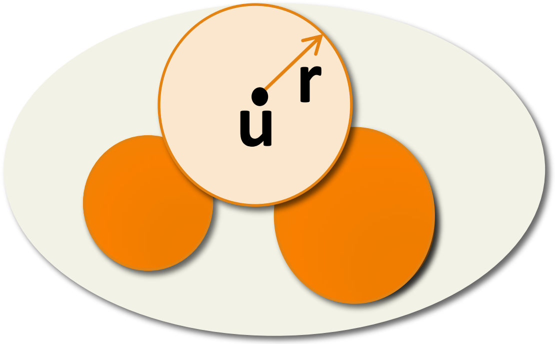
Session 6. SpiderMine: Mining Top-K Large Structural Patterns in a Single Network

SpiderMine: Mining Top-K Large Structural Patterns in a Massive Network

- Large patterns are informative to characterize a large network (e.g., social network, web, or bio-network)
- Similar to pattern fusion, mining large pattern should not aim for completeness but for representativeness of the target results
- Spider-Mine (F. Zhu, et al., VLDB'11): Mine top-*K* largest frequent substructure patterns whose diameter is bounded by D_{max} with a probability at least $1-\epsilon$
- General idea: Large patterns are composed of a number of small components ("spiders") which will eventually connect together after some rounds of pattern growth
- **r-Spider:** An r-spider is a frequent graph pattern P such that there exists a vertex u of P, and all other vertices of P are within distance r from u

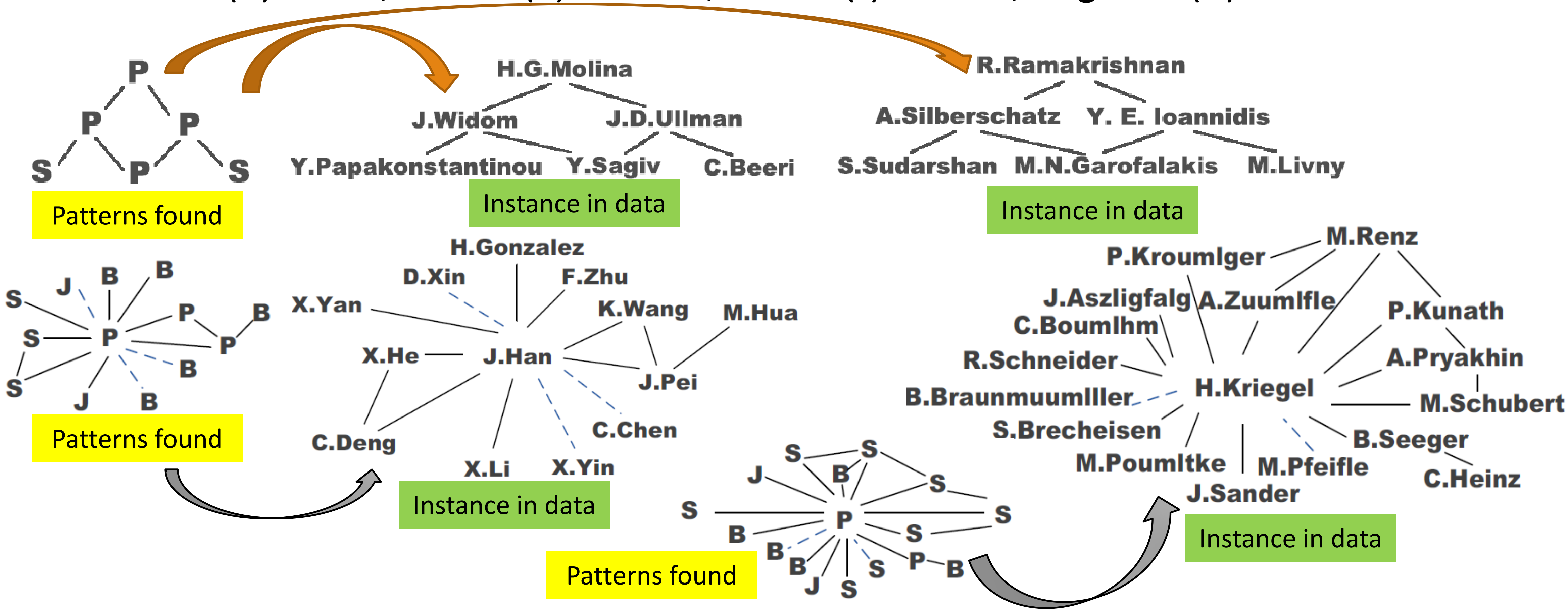


Why Is SpiderMine Good for Mining Large Patterns

- The SpiderMine Algorithm
  - Mine the set S of all the r-spiders
  - Randomly draw M r-spiders
  - Grow these M r-spiders for $t = D_{max}/2$ iterations, and merge two patterns whenever possible
  - Discard unmerged patterns
  - Continue to grow the remaining ones to maximum size
  - Return the top-K largest ones in the result
- Why is SpiderMine likely to retain large patterns and prune small ones?
  - Small patterns are much less likely to be hit in the random draw
  - Even if a small pattern is hit, it is even less likely to be hit multiple times
  - The larger the pattern, the greater the chance it is hit and saved

Mining Collaboration Patterns in DBLP Networks

- Data description: 600 top confs, 9 major CS areas, 15071 authors in DB/DM
- Author labeled by # of papers published in DB/DM
  - Prolific (P): >=50, Senior (S): 20~49, Junior (J): 10~19, Beginner(B): 5~9



Summary

- Graph pattern mining: Basic concepts
- Apriori-based graph pattern mining methods
- gSpan: A pattern-growth-based method
- CloseGraph: Mining closed graph patterns
- Graph Indexing: A graph pattern mining application example
- SpiderMine: Mining top-k large structural patterns in a large network

Recommended Readings

- C. Borgelt and M. R. Berthold, “Mining molecular fragments: Finding relevant substructures of molecules”, ICDM'02
- J. Huan, W. Wang, and J. Prins. “Efficient mining of frequent subgraph in the presence of isomorphism”, ICDM'03
- A. Inokuchi, T. Washio, and H. Motoda. “An apriori-based algorithm for mining frequent substructures from graph data”, PKDD'00
- M. Kuramochi and G. Karypis. “Frequent subgraph discovery”, ICDM'01
- S. Nijssen and J. Kok. A quickstart in frequent structure mining can make a difference. KDD'04
- N. Vanetik, E. Gudes, and S. E. Shimony. “Computing frequent graph patterns from semistructured data”, ICDM'02
- X. Yan and J. Han, “gSpan: Graph-Based Substructure Pattern Mining”, ICDM'02
- X. Yan and J. Han, “CloseGraph: Mining Closed Frequent Graph Patterns”, KDD'03
- X. Yan, P. S. Yu, and J. Han, “Graph Indexing: A Frequent Structure-based Approach”, SIGMOD'04
- F. Zhu, Q. Qu, D. Lo, X. Yan, J. Han, and P. S. Yu, "Mining Top-K Large Structural Patterns in a Massive Network", VLDB'11