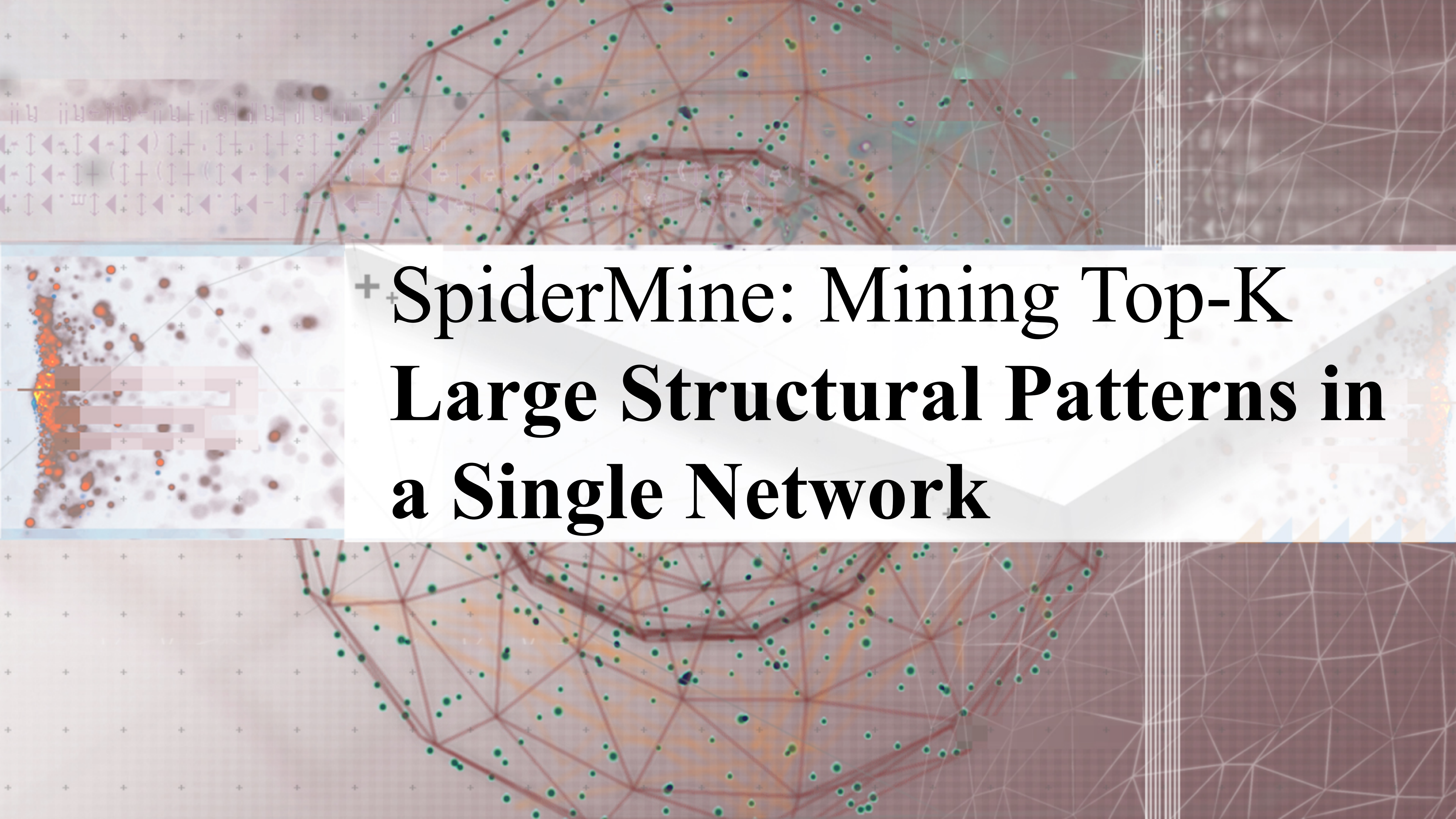
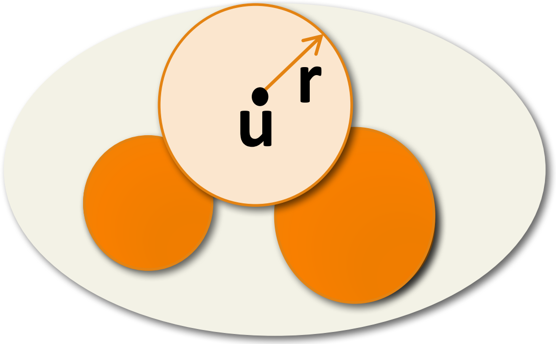


SpiderMine: Mining Top-K **Large Structural Patterns in a Single Network**

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 patterns should not aim for completeness but for representativeness of the target results
- ❑ SpiderMine (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

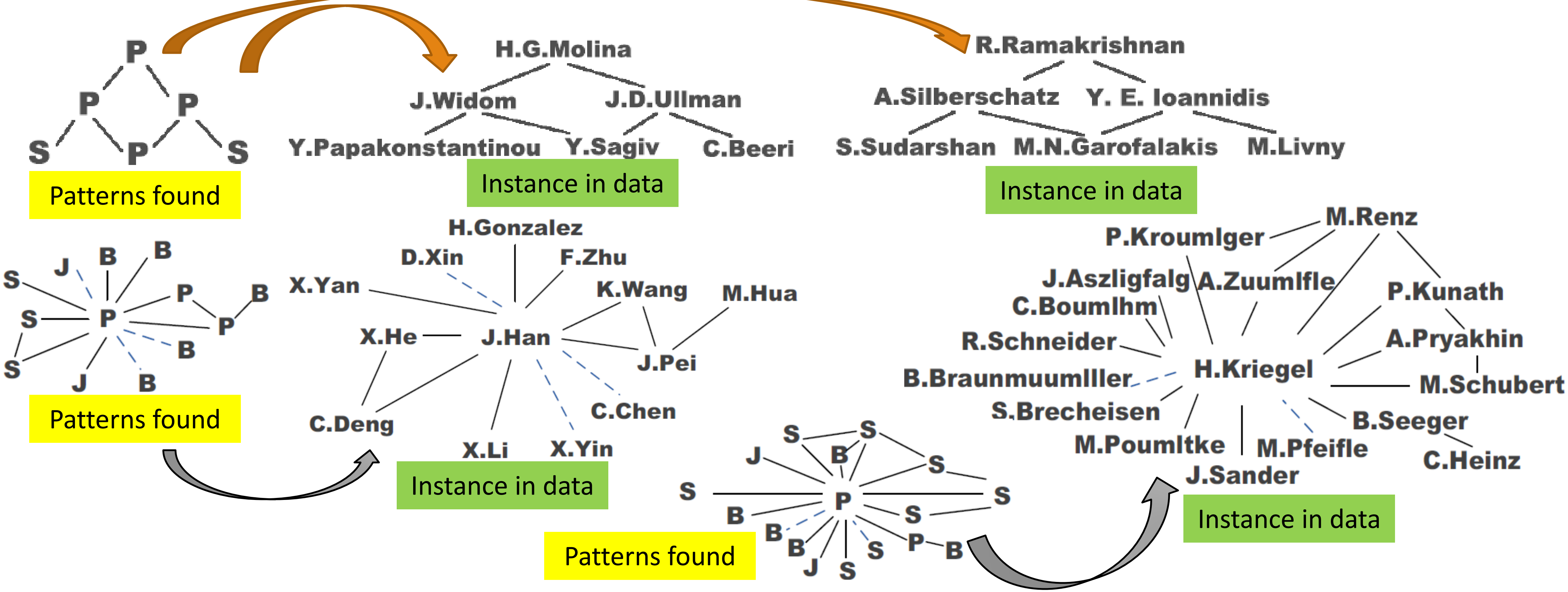


Why Is SpiderMine Good for Mining Large Patterns?

- 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 conferences, 9 major CS areas, 15,071 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



Patterns found

Instance in data

Instance in data

Patterns found

Instance in data

Patterns found

Instance in data