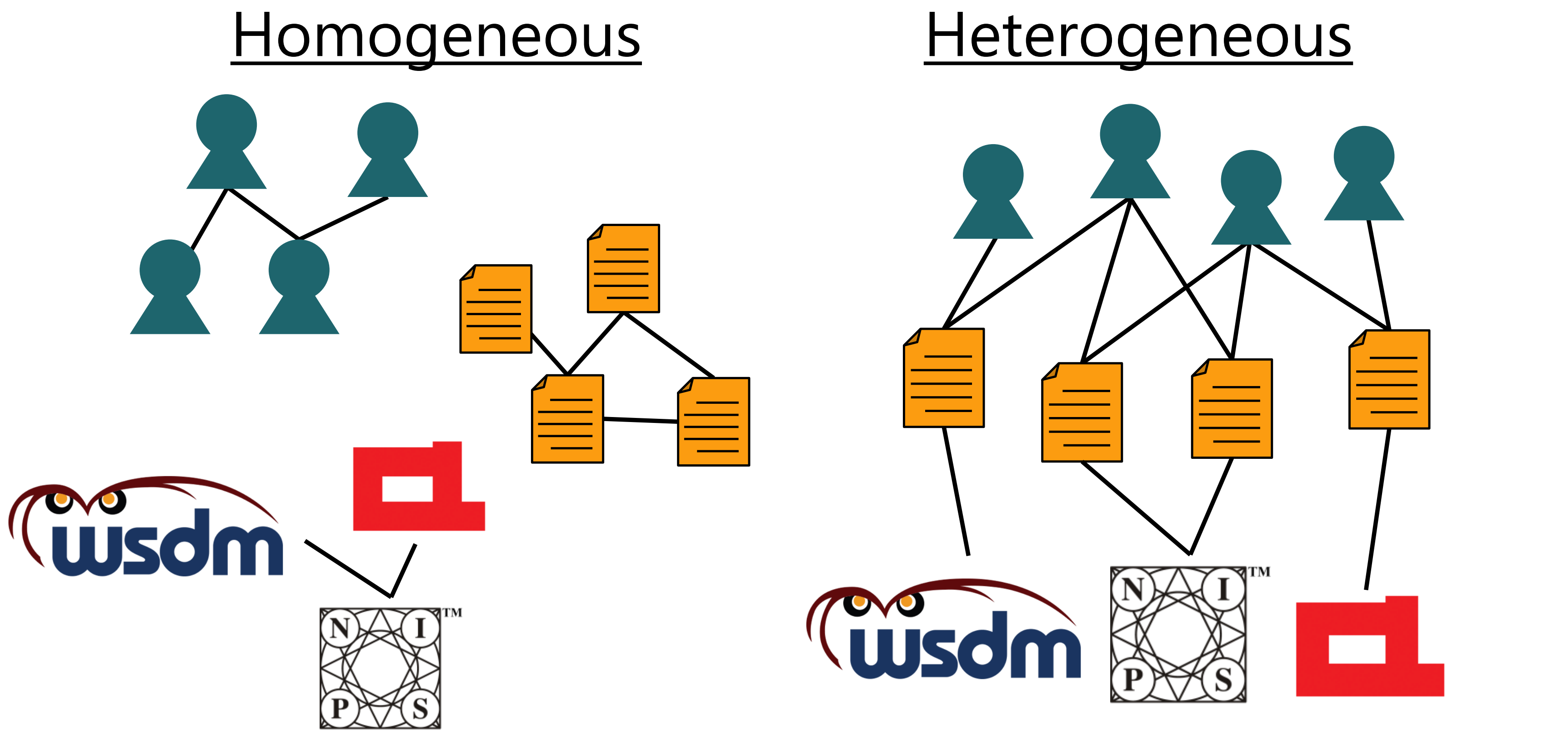
Transductive Classification on Heterogeneous Information Networks with Edge Betweenness-based Normalization

P. Bangcharoensap and T. Murata (Tokyo Tech) | H. Kobayashi and N. Shimizu (Yahoo Japan Corporation)

WSDM 2016. San Francisco, USA. Feb 22-25, 2016

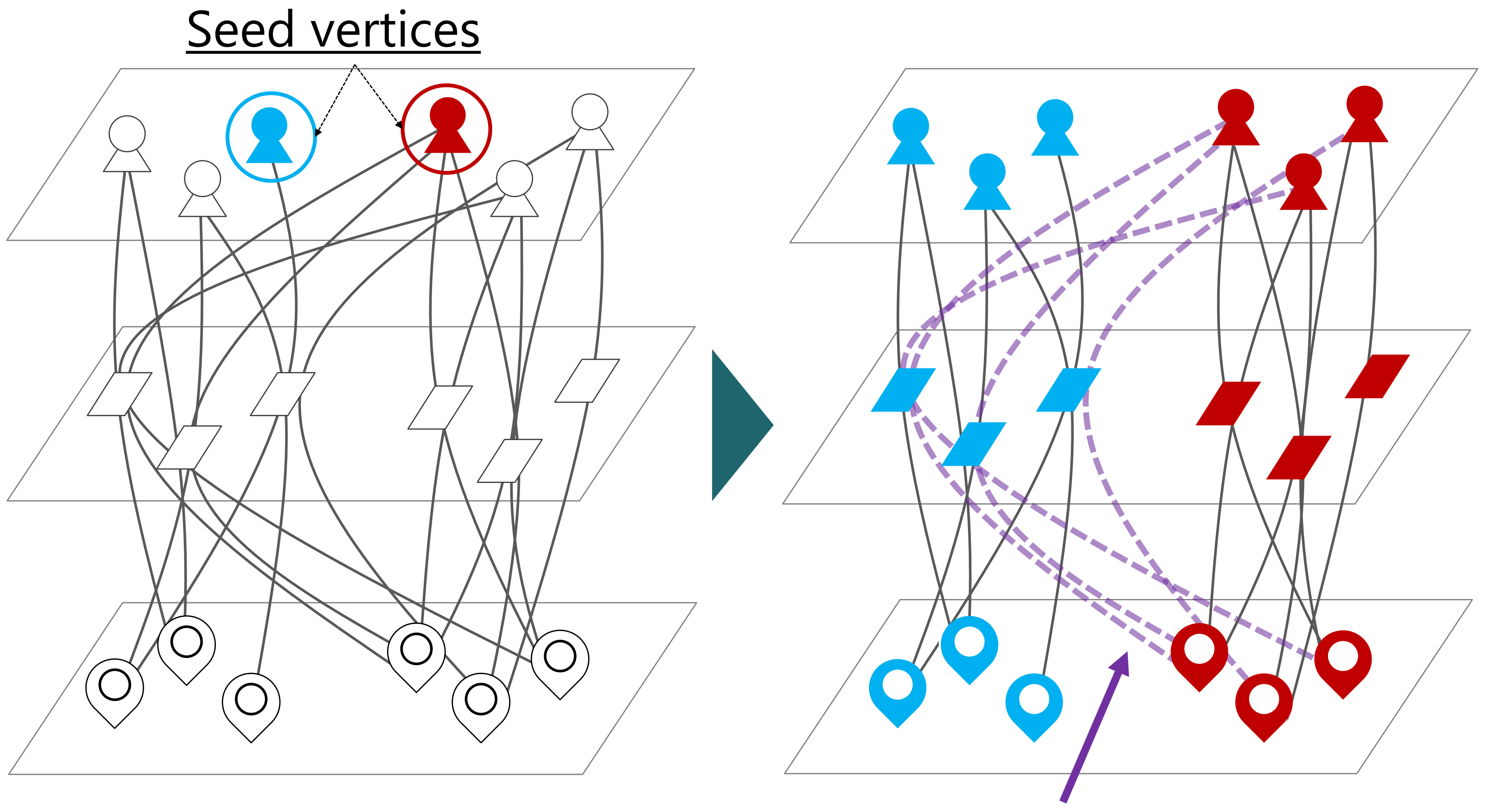
What is heterogeneous network?

Networks containing multiple types of vertices



Objective

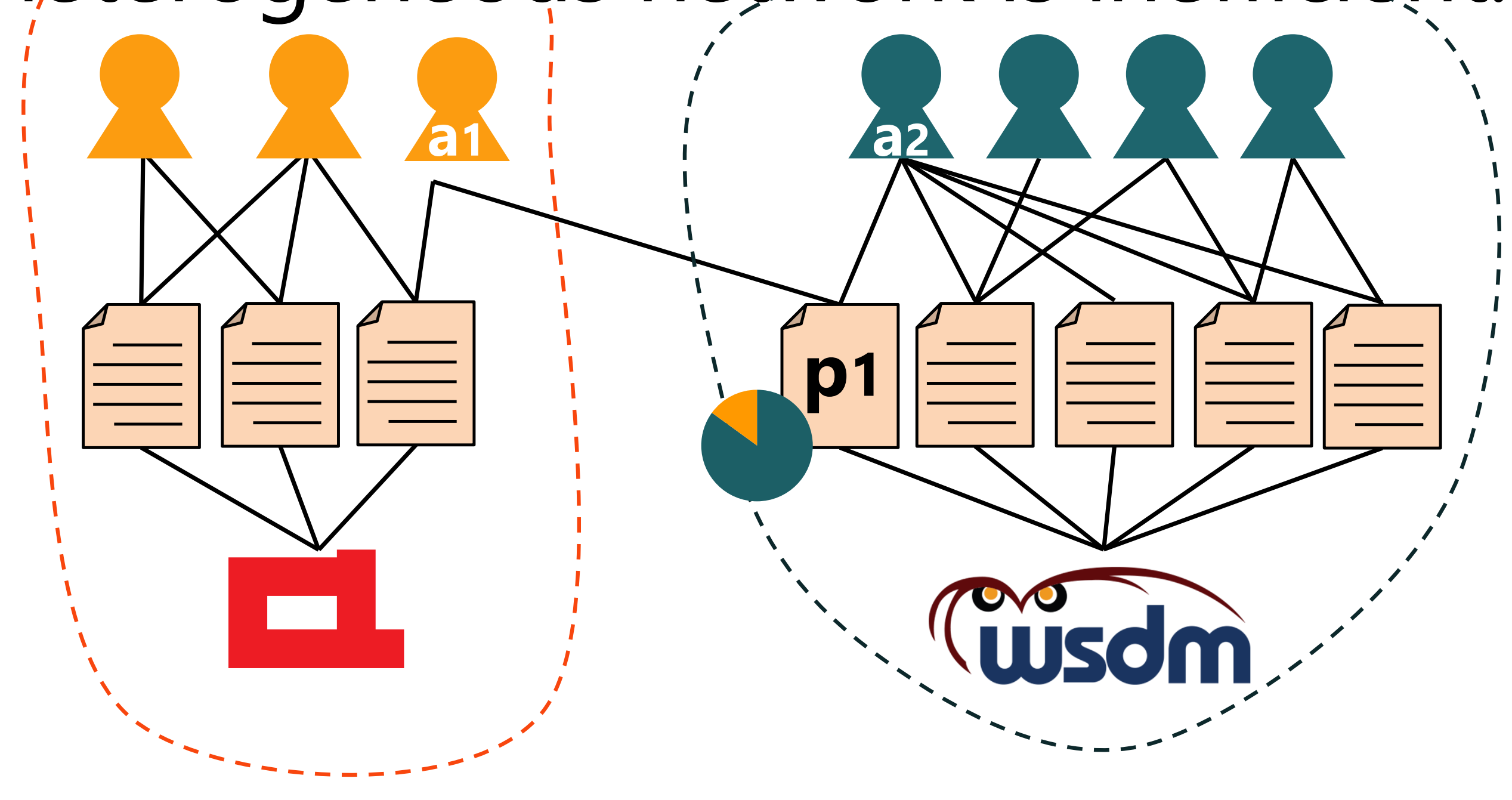
To infer labels of all vertices in a heterogeneous network where some vertices are labeled (seed vertices)



Erroneous edge: edge bridging across two classes

Key Ideas

- Penalize labels flowing through edges bridging across communities, sets of vertices densely connected, instead of edges originating from high degree vertices
- Use edge betweenness to capture the inter-community behavior
  - The number of shortest paths between all vertex pairs that pass through the edge.
- Directly applying conventional betweenness on heterogeneous network is inefficient.



The label of p1 should convey that of a2, not a1

Heterogeneous Edge Betweenness

1. Ignore flows from non-target type vertices:
   - To reduce influence of non-target vertices
2. Ignore flows originating from endpoints of considering edges:
   - To increase trustworthiness of labels flowing through dedicated edges

The betweenness of an edge$e = (a_i, p_j)$, where$a_i \in A$ and$p_i \in P$s defined as

$$\mathrm{C}(e) = 1 + \sum_{s \in A \setminus a_i} \sum_{t \in P \setminus p_i} \frac{\sigma(s,t|e)}{\sigma(s,t)}$$

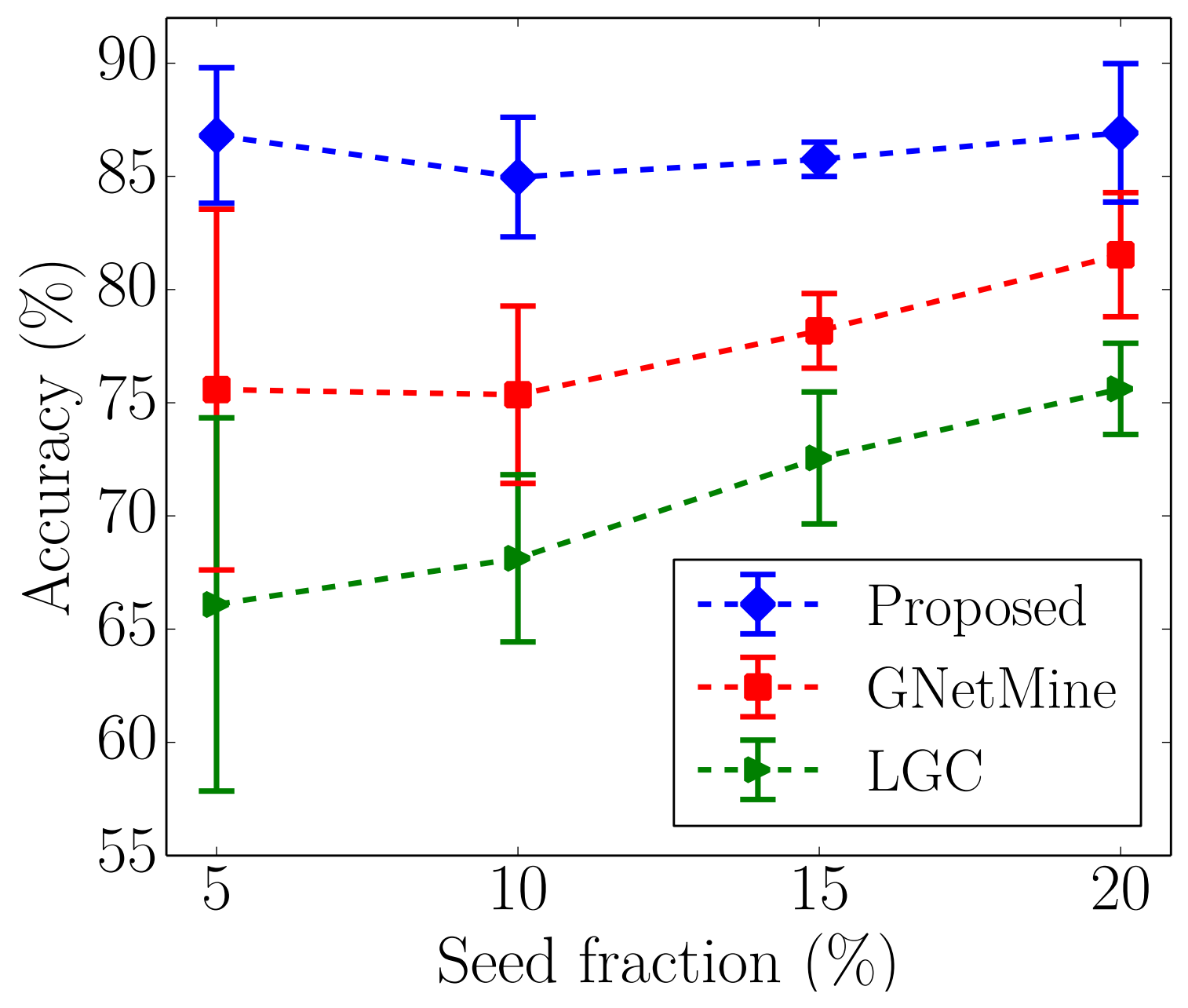
where$\sigma(s,t)$s # the shortest paths from s to t

$\sigma(s,t|e)$ is # the shortest paths from s to $t$passing through e

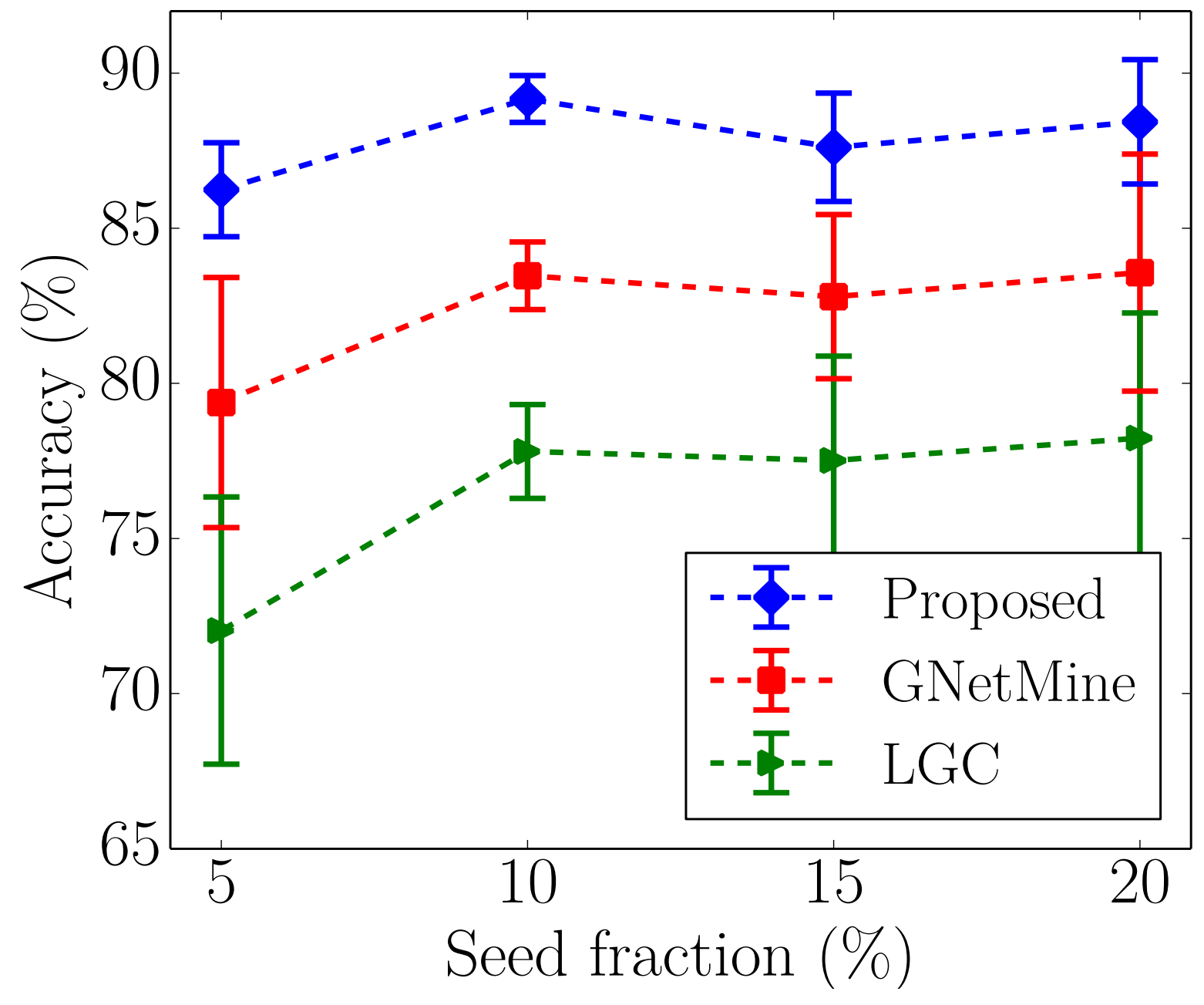
Experiment

- Dataset: DBLP network
- Goal: determine research theme of vertices
- Compare with GNetMine (Ji et al., 2010) and LGC (Zhou et al., 2004)
- Gain around 5 percentage points increase in accuracy compared with GNetMine

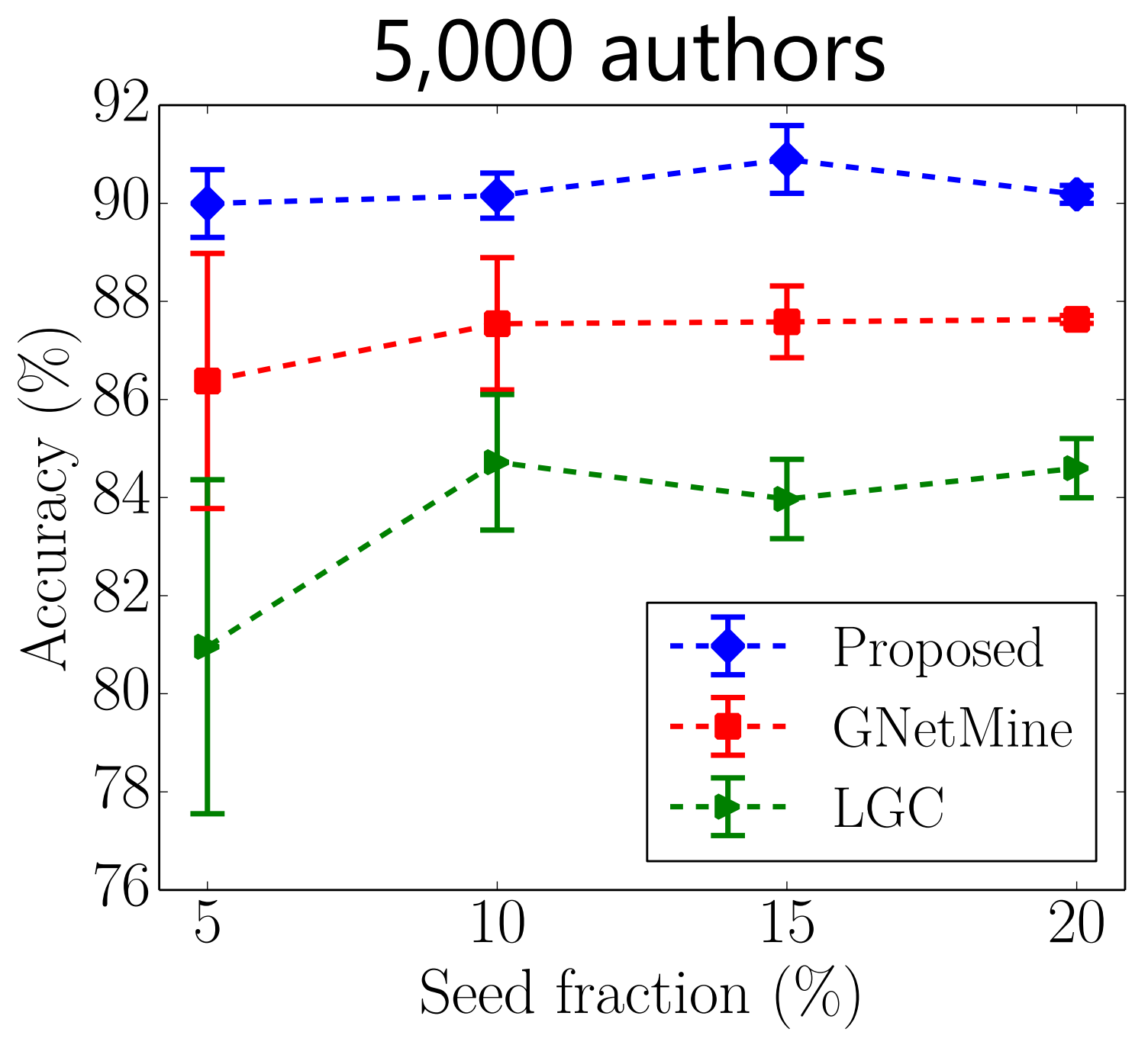
500 authors



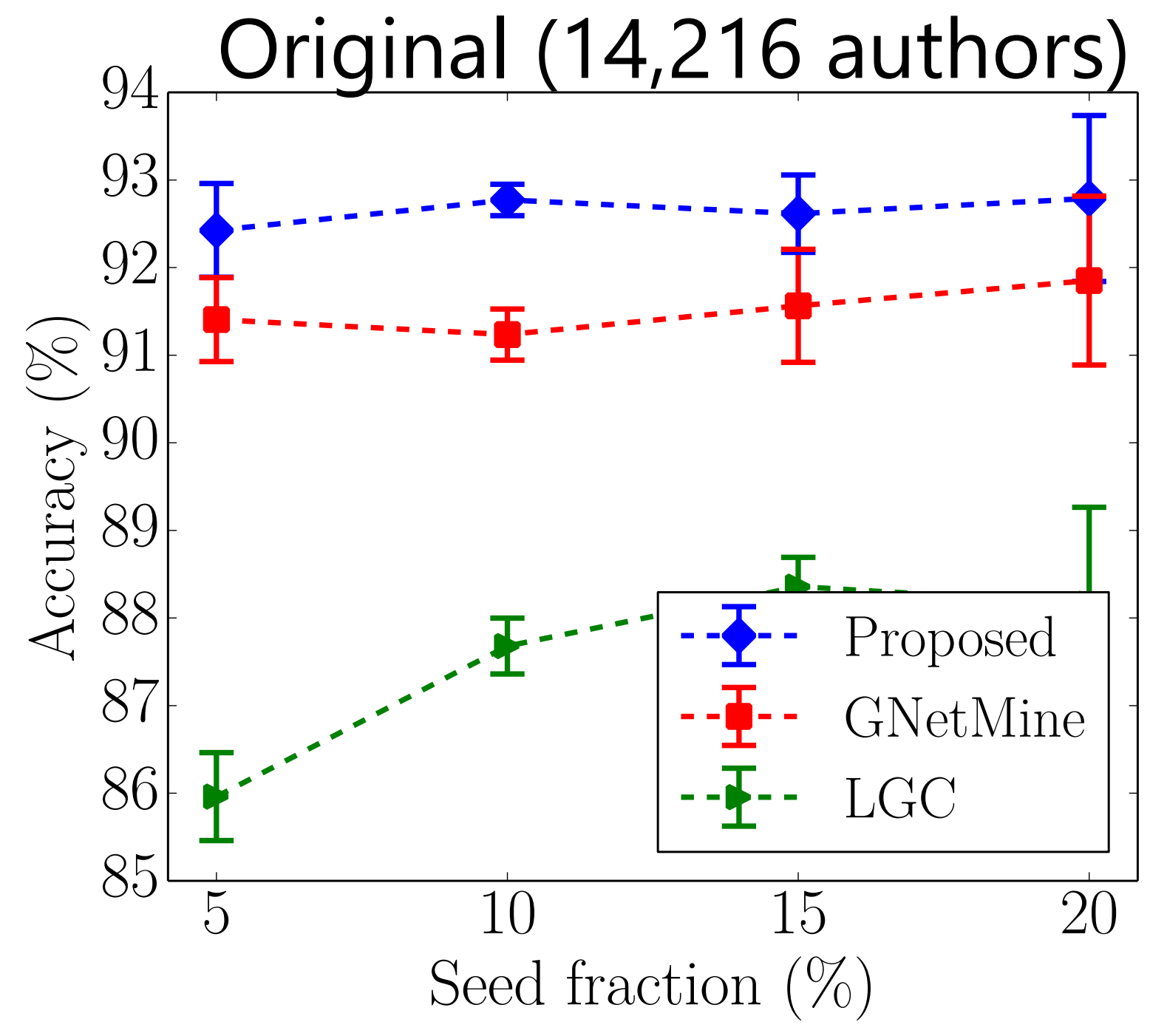
1,000 authors



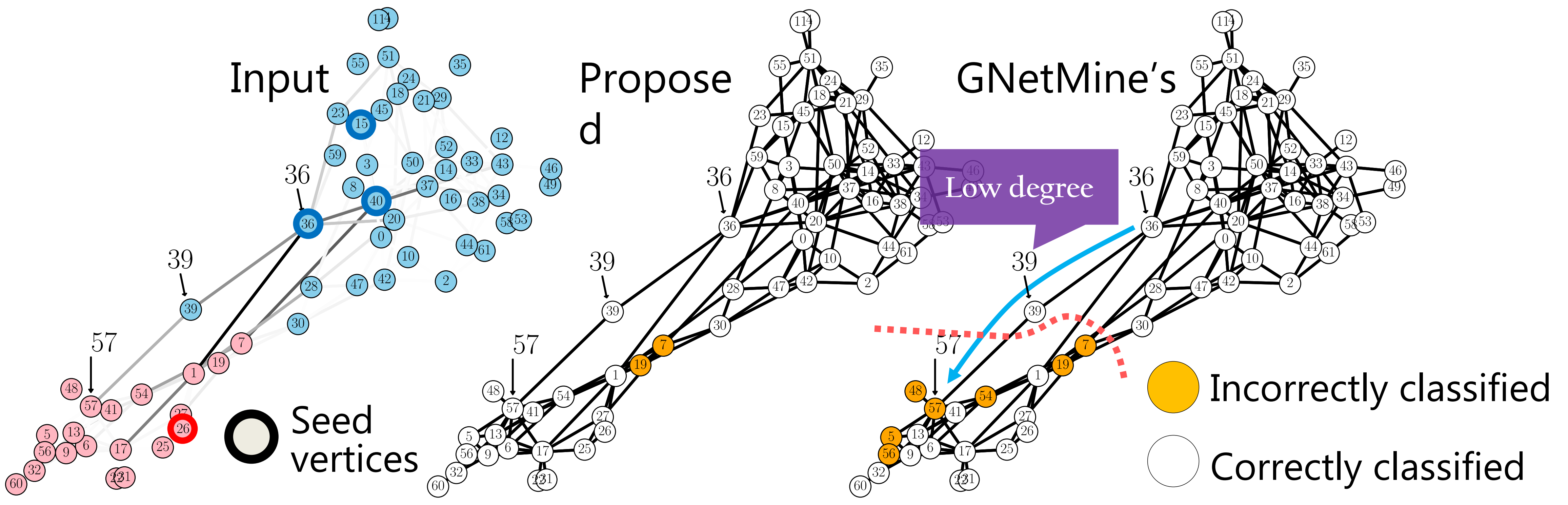
5,000 authors



Original (14,216 authors)



Case Study: Dolphin Network (Lusseau et al., 2003)


Input
Propose
d
GNetMine's
Low degree
36
39
57
Seed vertices
Incorrectly classified
Correctly classified