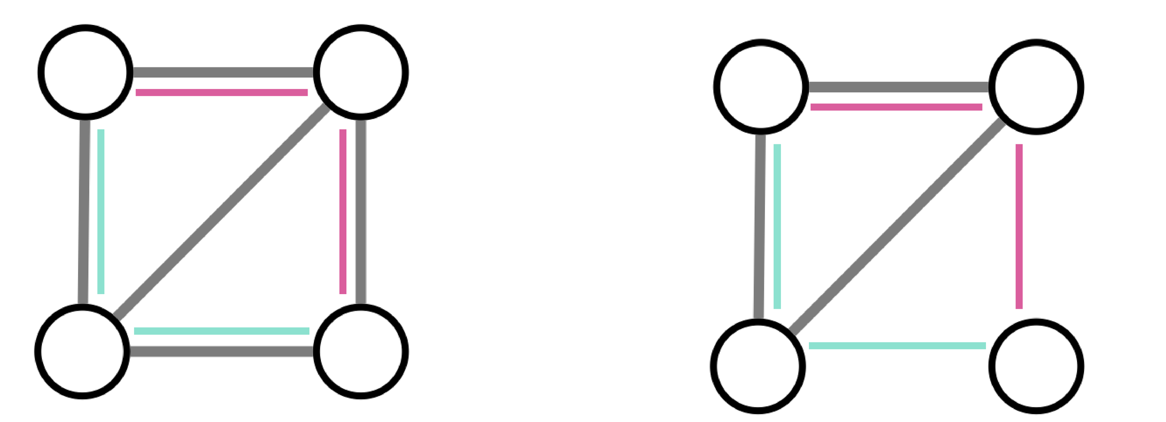
Combinatorial Approximations for Cluster Deletion: Simpler, Faster, and Better

Vicente Balmaseda[1], Ying Xu[2], Yixin Cao[2], Nate Veldt[1] [1]Texas A&M University, [2]Hong Kong Polytechnic University

Abstract. We provide improved deterministic approximation algorithms and guarantees for Cluster Deletion, and the first combinatorial algorithm for the STC relaxation

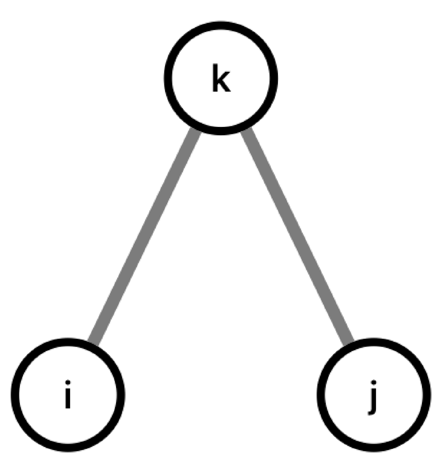
Cluster Deletion (CD)



Input: unweighted undirected graph

Goal: minimum number of edges to *delete* to obtain a disjoint set of cliques

Open wedge (i, j, k)



Principle of strong triadic closure

"At least one of these connections is weak, or else j and k would also be friends"

Min STC labeling

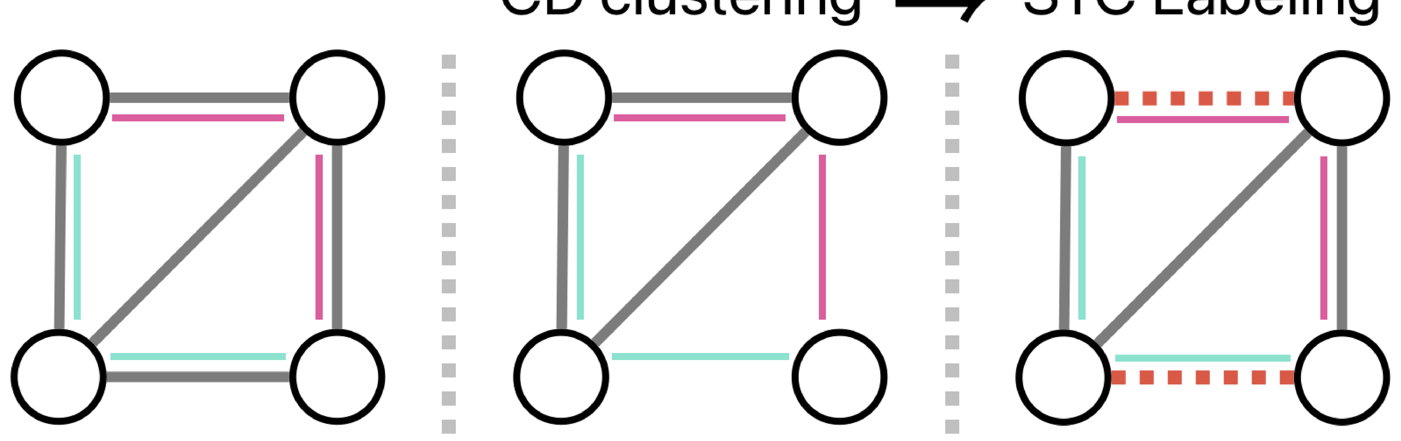
Input: unweighted undirected graph

Goal: minimum number of edges to label as *weak* to "cover" all open wedges

MinSTC ≤ CD: it is already known that MinSTC *lower bounds* CD

Note a CD clustering is a valid MinSTC labeling, but the opposite is not true

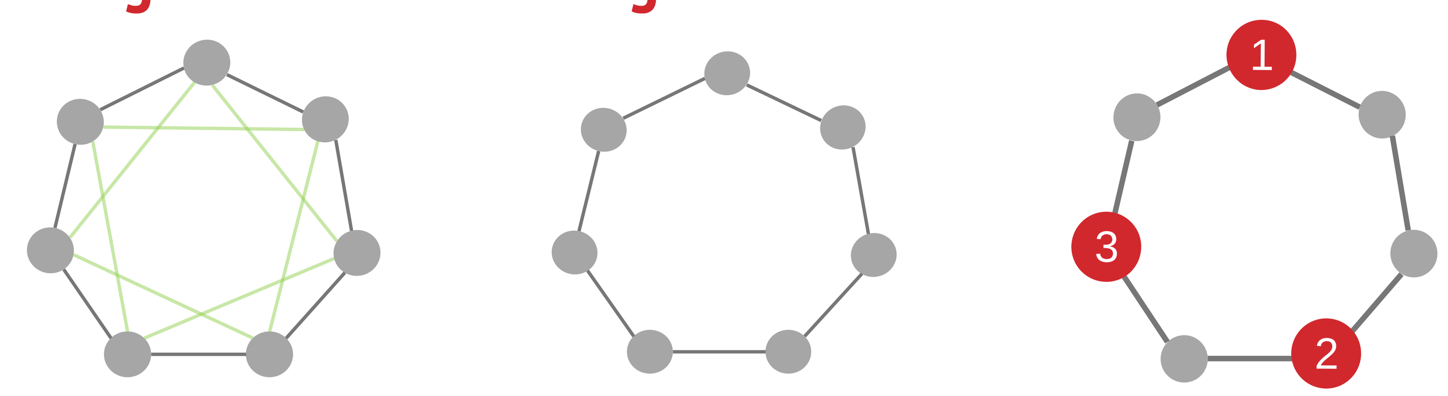
CD clustering ⟹ STC Labeling



Approximation algorithm via STC labeling

1. **STC labeling** (lower bound)
   1. Rounding STC linear program
   2. Max disjoint set of open wedges
2. **Cluster by pivoting** (i.e., cluster together the pivot node and its neighbors) after *deleting weak* edges

   Choose pivot *k* (new in red):
   1. RanMFP: Uniformly at random
   2. RatMFP: Minimize ratio between the number of boundary edges and non-edges in cluster formed by *k*
   3. **DegMFP: Maximum degree**



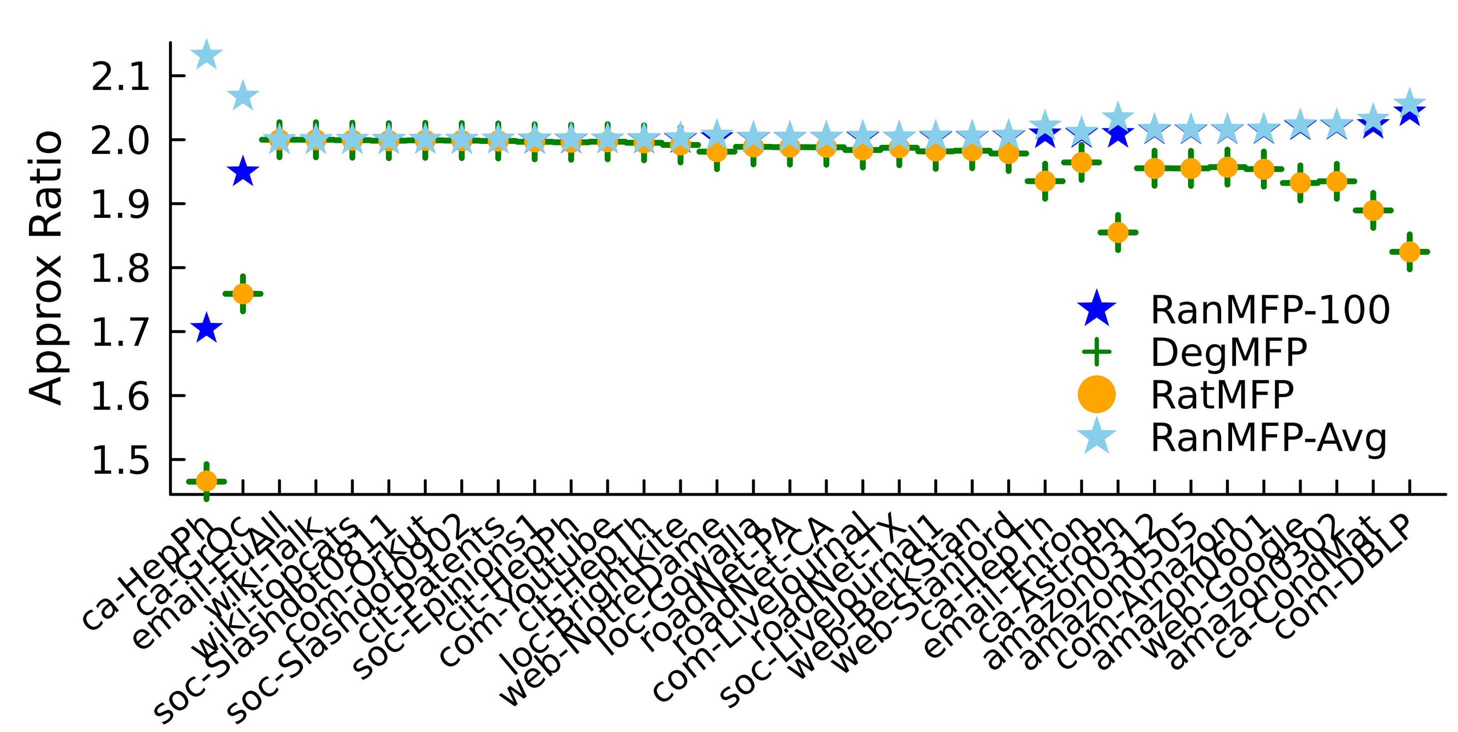
STC labeling *Delete weak edges* *Pivot*

Contributions

1. **Simpler and faster** max degree *deterministic* pivoting strategy
   - $O(m)$ time compared to previous $\Omega(m + |W|)$ of previous deterministic algorithm (W is the set of open wedges)
2. **Faster** *combinatorial* algorithm for solving the STC LP relaxation
   - Achieves the same result as black-box LP solver faster and on graphs that are an order of magnitude larger
3. **Better** approximation guarantees for framework

   Theorem. All the combinations for (1) and (2) provide a 3 approximation *(previous guarantee was 4)*

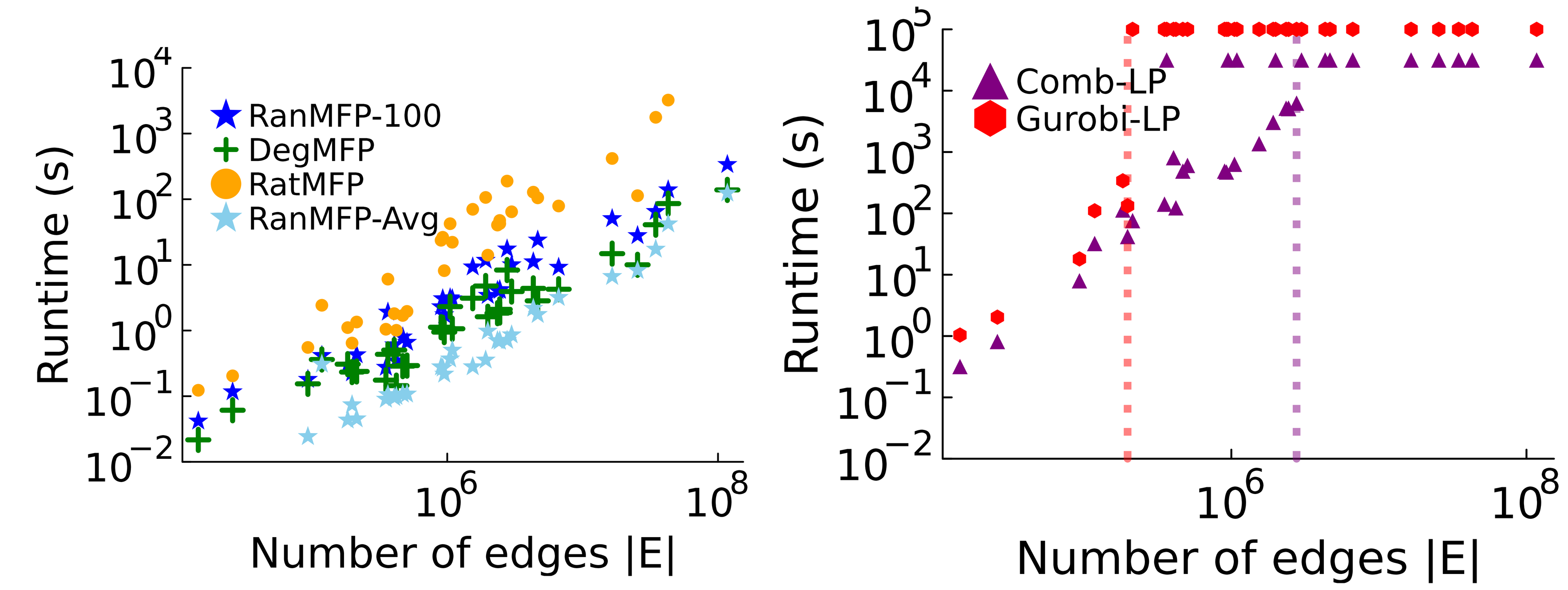
DegMFP achieves better approximation ratios



Code: *github.com/vibalcam/combinatorial-cluster-deletion*

(Left) **DegMFP is very simple and fast**

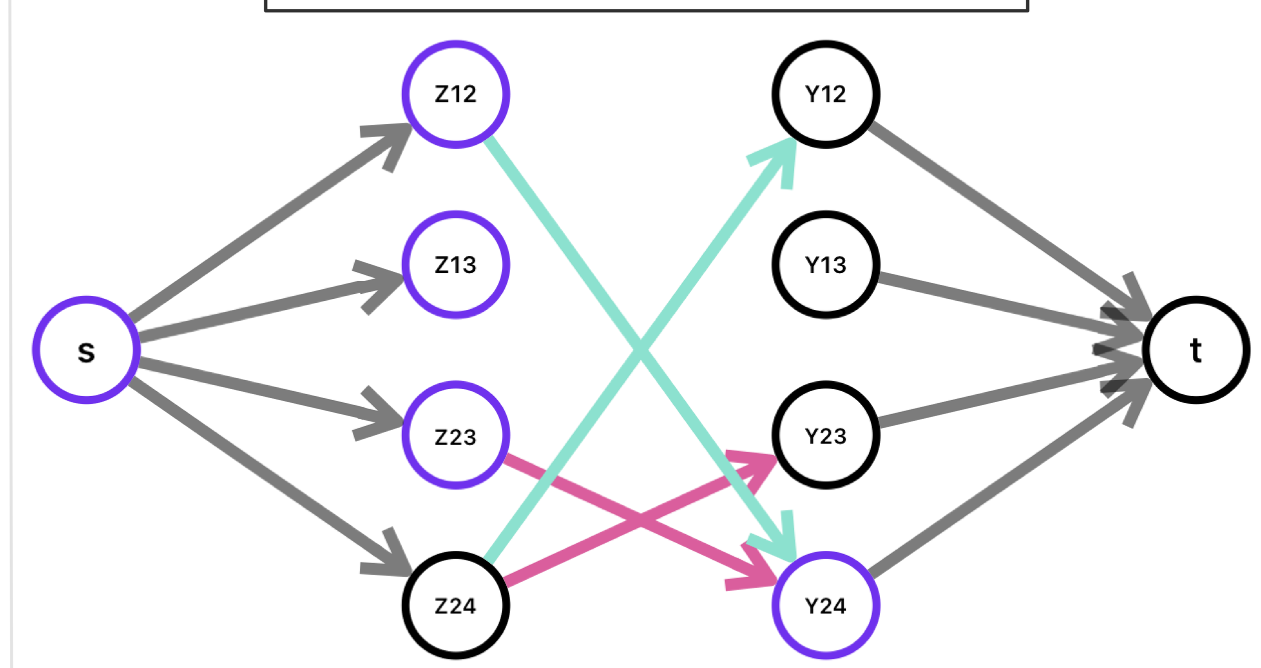
(Right) **Combinatorial STC-LP is faster and more scalable**



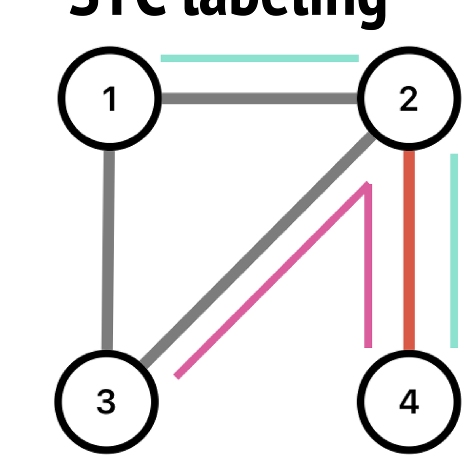
Combinatorial solver for STC LP via Min s-t cut

1. **Add source** s **and target** t and **nodes** Y and Z for each edge
3. **½-weighted edges** from Z to s and Y to t
4. **Inf-weighted edges** from Z to Y enforce STC constraints
5. **Solve min s-t cut**
6. $y_{ij} = 1, z_{ij} = 1$ if node in source set
7. $x_{ij} = \frac{1}{2}(y_{ij} - z_{ij} + 1)$
8. **Round solution** (keep edge if 0, remove otherwise)

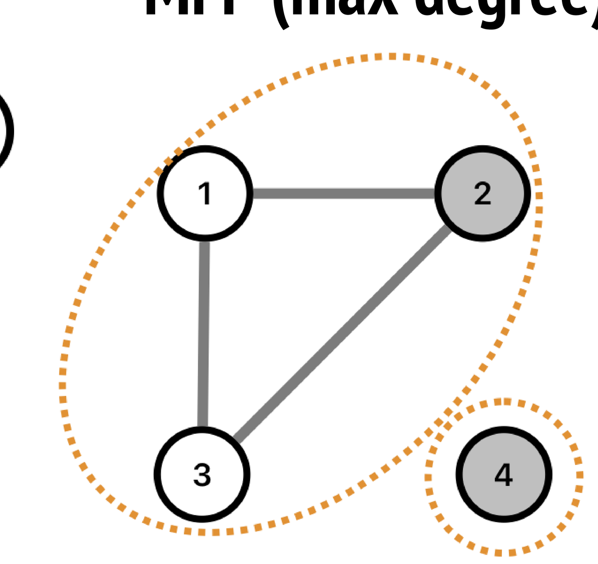
$$\begin{aligned} \min \ & \textstyle\sum_{ij \in E} x_{ij} \\ \text{s.t.} \ & x_{ij} + x_{ik} \geq 1 \text{ if } ijk \in W_i \\ & x_{ij} \geq 0 \ \forall ij \in E \end{aligned}$$



STC labeling



MFP (max degree)



Key Takeaways

1. Degree-based MFP is very fast, simple to implement, has the same theoretical guarantees, and experimentally achieves better approximations than random pivot (i.e., current **best of all worlds**)

2. Combinatorial STC-LP is faster and more scalable, allowing to solve problems on a laptop with 1.97 million nodes and 2.77 million edges (black-box solvers reached 35k nodes and 421k edges)