TOP-k OVERLAPPING DENSEST SUBGRAPHS

ESTHER GALBRUN, ARISTIDES GIONIS, NIKOLAJ TATTI

Inria informatics mathematics — Aalto University — Helsinki Institute for Information Technology

PROBLEM DEFINITION

Dense subgraphs:
The density of a subgraph $G = (V, E)$ is

$$dens(G) = \frac{|E|}{|V|} \quad .$$

Distance between subgraphs:
The distance of two subgraphs $G = (V, E)$ and $H = (W, A)$ is

$$D(G, H) = 2 - \frac{|V \cap W|^2}{|V|\,|W|},$$

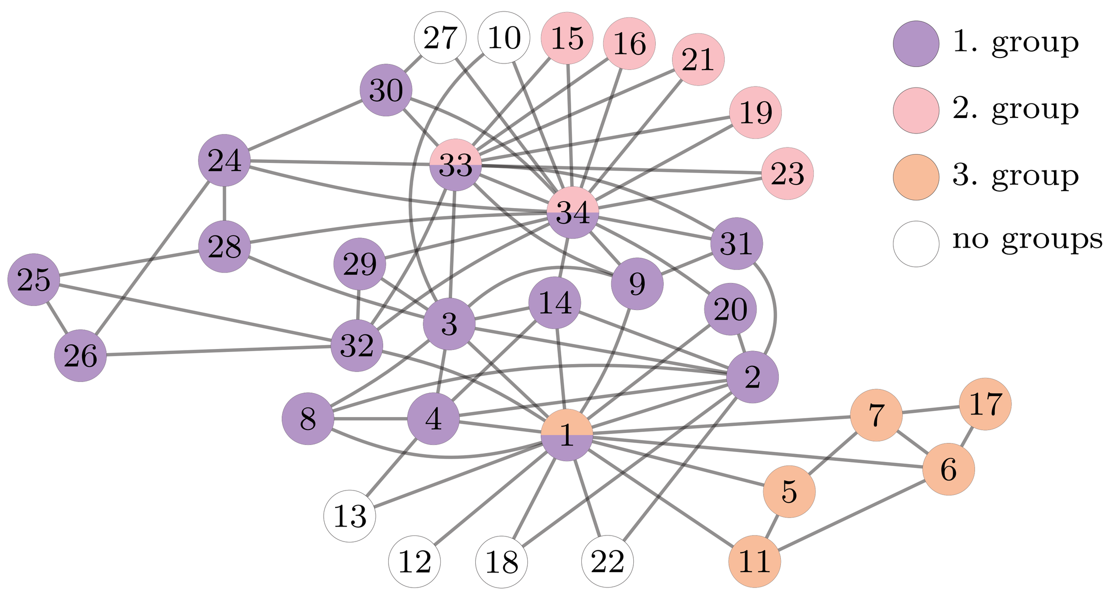
and 0 if $V = W$.

Objective: Find k subgraphs such that

$$\sum_{i=1} dens(G_i) + \lambda \sum_{i<j} D(G_i, G_j) \quad .$$

- first term: graphs should be dense
- second term: graphs should be diverse
- λ controls the balance of the terms

EXAMPLE



MAX-SUM DIVERSIFICATION

Find k elements S maximizing

$$f(S) + \lambda \sum_{x,y \in S} d(x, y),$$

where

- f is a submodular function
- d is a metric.

Greedy yields $1/2$ approximation [1]:

$S \leftarrow \emptyset$;
foreach $i = 1, \ldots, k$ **do**
  add x to S maximizing the gain

$$\frac{1}{2} f(S \cup \{x\}) + \lambda \sum_{y \in S} d(x, y)$$

If the maximization step is approximative with c, then greedy approximates with $c/2$.

GAIN PROBLEM

Given a set $\mathcal{S}$ of subgraphs, find a subgraph G maximizing

$$\frac{1}{2} dens(G) + \lambda \sum_{H \in \mathcal{S}} D(G, H) \quad .$$

- exponential number of subgraphs
- Gain problem is **NP**-hard

GREEDY FOR GAIN

Warm-up, $\lambda = 0$:

Known greedy algorithm [2] for finding dense subgraphs

$W \leftarrow V$;
while $W \neq \emptyset$ **do**
  $v \leftarrow$ vertex with the smallest degree;
  delete v from W;
return the best observed subgraph;

This yields $1/2$ approximation.

General case, $\lambda \geq 0$:

Given a set $\mathcal{S}$ of subgraphs and a set of vertices W, define

$$p(v; W) = \sum_{H=(U,E) \in \mathcal{S} \mid v \in U} \frac{|U \cap W|}{|U|} \quad .$$

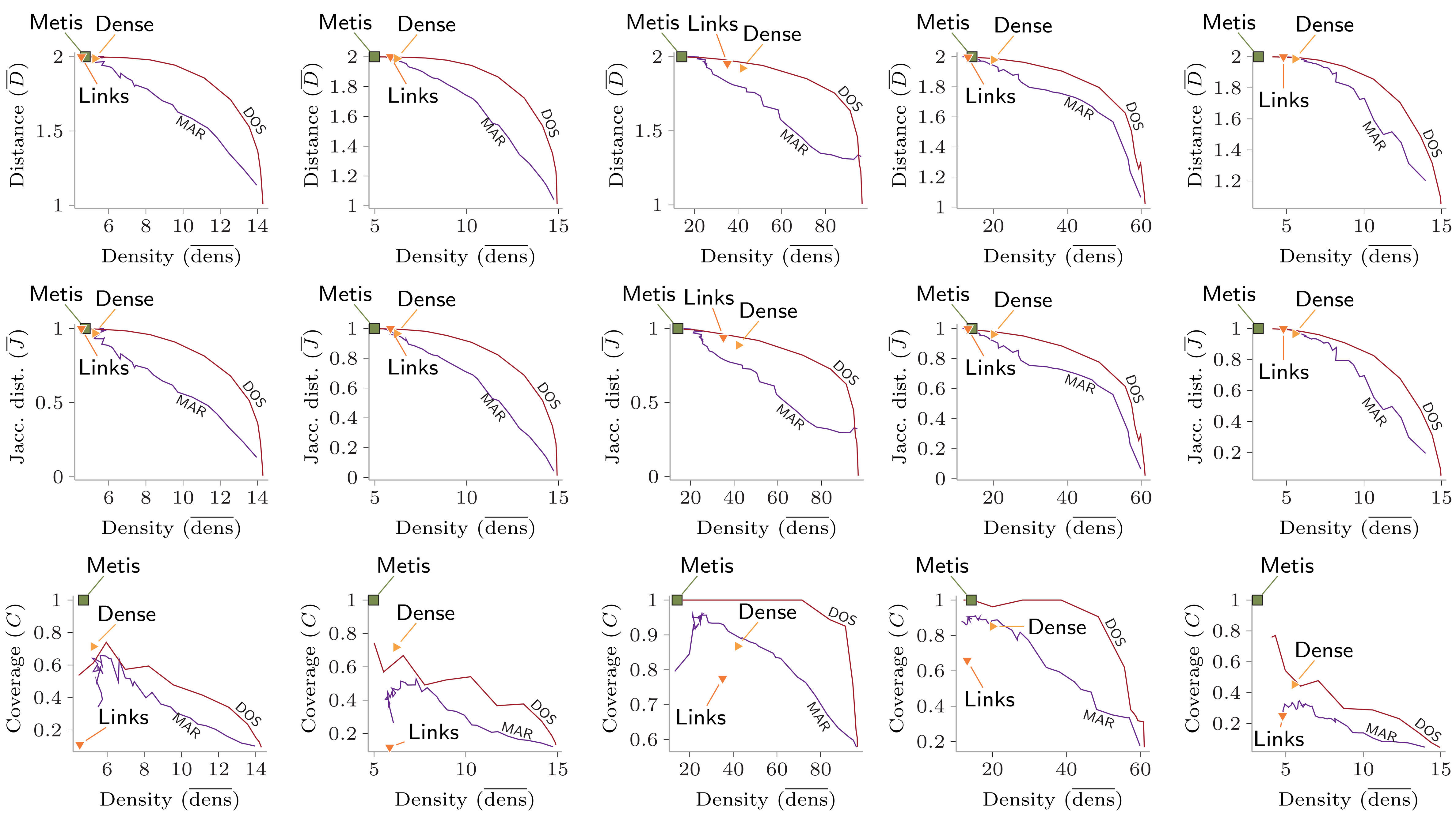
Large $p(v; W)$ = node is shared with many previous communities.

$W \leftarrow V$;
while $W \neq \emptyset$ **do**
  $v \leftarrow$ vertex minimizing $\deg(v) - 4\lambda p(v; W)$;
  delete v from W;
return the best observed subgraph;

This yields $c = 1/5$ approximation.

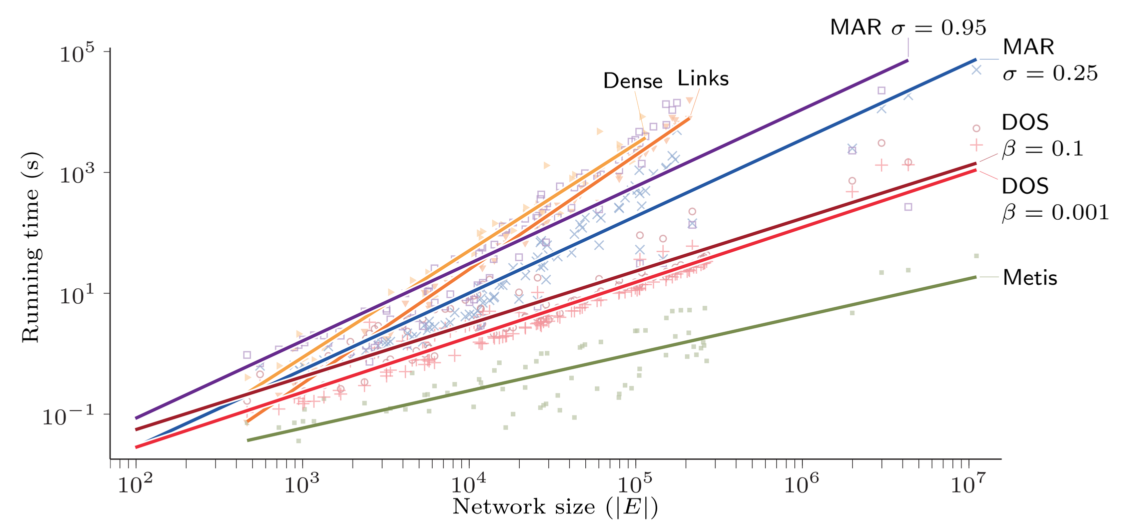
$1/10$ approximation for the general problem.

EXPERIMENTAL EVALUATION



DBLP.E2 Papadimitriou, $|E| = 2616$, $dens(G) = 3.62$ — DBLP.C KDD, $|E| = 2891$, $dens(G) = 3.88$ — G+.S 1183...6467, $|E| = 694$, $dens(G) = 40.24$ — FB 1684, $|E| = 786$, $dens(G) = 17.84$ — BKGW BK Latin-America, $|E| = 1216$, $dens(G) = 2.22$

RUNNING TIME



REFERENCES

[1] Allan Borodin, Hyun Chul Lee, and Yuli Ye. Max-sum diversification, monotone submodular functions and dynamic updates. PODS, pages 155–166, 2012.

[2] Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. APPROX, pages 84–95, 2000.

[3] Oana Denisa Balalau, Francesco Bonchi, TH Chan, Francesco Gullo, and Mauro Sozio. Finding subgraphs with maximum total density and limited overlap. In WSDM, pages 379–388, 2015.