Discovering Nested Communities

NIKOLAJ.TATTI@AALTO.FI · ARISTIDES.GIONIS@AALTO.FI
Aalto University, Helsinki Institute of Information Technology

Discovering communities

Given a graph G and a set of vertices S, find a good community around S

Graphs rarely have clear community structure

Discovering communities becomes ill-defined problem:

Should we select

- a small and tight community
- or larger and sparser community?

We could either

- introduce a score that balances between size and tightness
- or discover multiple communities

Nested communities

Given a graph $G = (V, E)$, number of communities K, and a set of nodes S, find a sequence of communities

$$S = V_0 \subsetneq V_1 \subsetneq \cdots \subsetneq V_K = V \text{ such that}$$

- V_i is more dense than V_{i+1}
- quality score $q(V_0, \ldots, V_K)$ is optimized

Density

E_i = edges of V_i

- replace each non-edge with an edge with a weight of 0

outer edges $F_i = E_i \setminus E_{i-1}$

density:

$$d(F_i) = \frac{1}{|F_i|} \sum_{e \in F_i} w(e)$$

Quality score

$$q(V_0, \ldots, V_K) = \sum_{i=1}^{K} \sum_{e \in F_i} |w(e) - \mu_i|^2,$$

where μ_i is the centroid, $\mu_i = d(F_i)$.

Fixing order

Split the problem into 2 subproblems:

1. given a graph with *ordered* nodes, find a sequence of communities respecting the order:

$$\text{if } v_j \in V_i, \text{ then } v_{j-1} \in V_i$$

2. find a good order

There are orders corresponding to the optimal solution

THEOREM: Finding optimal communities given the order is a monotonic segmentation problem

Monotonic segmentation

Input: a sequence of real-numbers $x_1, \ldots, x_N$ with weights $m_1, \ldots, m_N$

Output: partition of the sequence into K segments such that

- quality score

$$q(C_1, \ldots, C_K) = \sum_{i=1}^{K} \sum_{j \in C_i} m_j |x_j - \mu_i|^2$$

  is minimized
- C_i has a higher average than C_{i+1}.

Can be solved by a classic dynamic program

- quadratic time, linear space
- linear time approximations exist

Monotonicity can be enforced with preprocessing using PAV algorithm

Communities as segments

Express community detection as segmentation problem by setting

$$x_i = \frac{1}{i-1} \sum_{j=1}^{i-1} w((v_i, v_j)) \text{ and } m_i = i - 1$$

Then

$$q(V_0, \ldots, V_K) = q(C_1, \ldots, C_K) + \text{ const}$$

Selecting order

- degree of each node
- personalized page rank
- dense subgraph algorithm

$W \leftarrow V$;
while $W \neq S$ **do**
  $w \leftarrow \arg\min_{x \in W} \deg_W(x)$;
  delete w from W;

Weighting edges

Compute $p(v)$ = personalized pagerank.

3 options:

$$w_n(e) = \frac{p(v)}{\deg(v)} + \frac{p(w)}{\deg(w)}$$
$$w_s(e) = p(v) + p(w)$$
$$w_m(e) = \min(p(v), p(w))$$

Alternative approaches

Any subcommunity of V_i is more dense than any subcommunity of V_{i+1}

THEOREM: Let X and Y such that

$$V_{i-1} \subseteq Y \subsetneq V_i \subsetneq X \subseteq V_{i+1}$$

then

$$d(E(V_i) \setminus E(Y)) > d(E(X) \setminus E(V_i))$$

ATTEMPT 1: Add dense communities first

$W \leftarrow S$;
while $W \neq V$ **do**
  $C \leftarrow$ densest community containing W;
  add C to W;

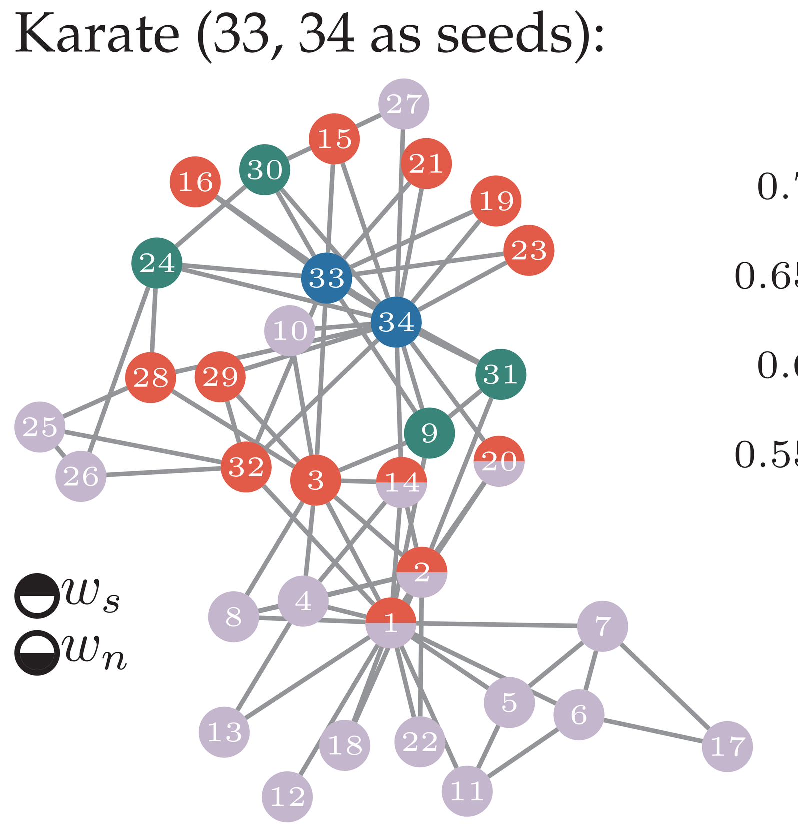
ATTEMPT 2: Delete sparse communities first

$W \leftarrow V$;
while $W \neq S$ **do**
  $C \leftarrow$ sparsest community in W s.t. $C \cap S = \emptyset$;
  remove C from W;

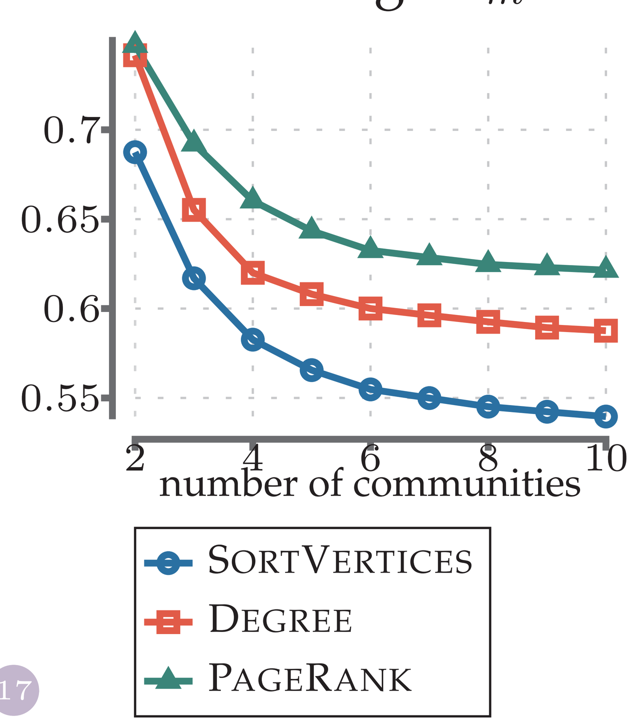
Finding C is **NP**-hard

Experiments

Karate (33, 34 as seeds):



Polblogs w_m



Name	$\|V(G)\|$	$\|E(G)\|$	Time	N	performance $q(\mathcal{V}) / q(\mathcal{H})$ w_n	w_s	w_m
Adjnoun	112	425	2ms	84	0.90/0.95	0.88/0.95	0.77/0.94
Dolphins	62	159	1ms	41	0.67/0.80	0.61/0.78	0.57/0.80
Karate	34	78	1ms	21	0.78/0.91	0.76/0.91	0.60/0.93
Lesmis	77	254	2ms	37	0.77/0.93	0.84/0.94	0.62/0.94
Polblogs	1 222	16 714	84ms	872	0.87/0.96	0.95/0.99	0.57/0.96
DBLP	703 193	2 341 362	23s	1 797	0.87/0.99	0.98/1.00	0.45/0.99