Discovering dynamic communities in interaction networks

Polina Rozenshtein · Nikolaj Tatti · Aristides Gionis

HIIT, Aalto University

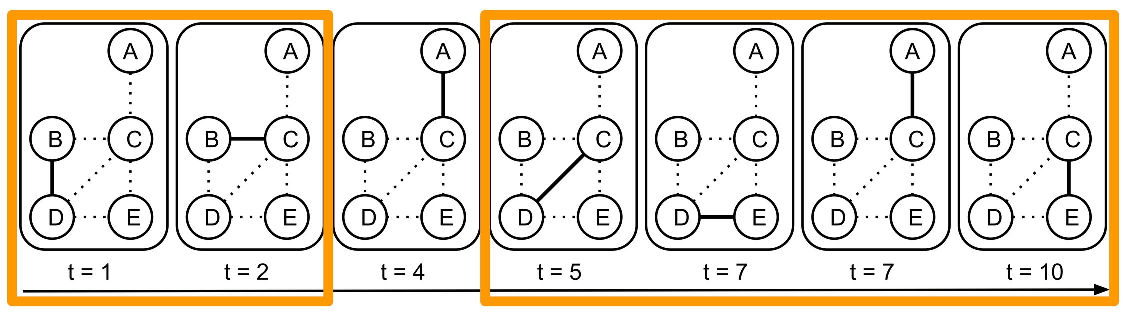
Interaction networks

Social interaction network:

- an edge between people represents some interaction: phone call, email, retweet, etc.
- multiple edges are annotated with time
- sequence of interactions

Model

Given: a graph $G = (V, E)$, m timestamps
Model: time series of edges $E = \{(u_i, v_i, t_i)\}$ with $i = 1, .., m$ and $u_i, v_i \in V$.



Consider: set of nodes $W \subseteq V$ and set of time intervals $\mathcal{T} = \{[t_{j_1}, t_{j_2}], [t_{j_3}, t_{j_4}], \ldots\}$

Measure:

- cardinality $|\mathcal{T}| = k$ and total length of time intervals $span(\mathcal{T})$
- quality of induced subgraph $G(\mathcal{T}, W)$ - average degree (density):

$$q(G(\mathcal{T}, W)) = \frac{2|E(G(\mathcal{T}, W))|}{|W|}$$

Example: $G(\{[1, 2], [5, 10]\}, \{B, C, D, E\})$

- $|\mathcal{T}| = 2, \quad span(\mathcal{T}) = 6$
- $q(G(\mathcal{T}, W)) = 2.5$

Problem

Problem 1 *Given numbers K and B. Find a set of intervals $\mathcal{T}$ and a set of nodes $W \in V$:*

$$\text{maximize} \quad q(G(\mathcal{T}, W))$$
$$\text{s.t. } |\mathcal{T}| \leq K \text{ and } span(\mathcal{T}) \leq B$$

Proposed algorithm

Iterate until convergence:

1. Solve Densest subgraph subproblem: Given fixed set of intervals $\mathcal{T}$, find optimal set of nodes W
2. Solve Optimal intervals subproblem: Given fixed set of nodes W, find optimal set of intervals $\mathcal{T}$

Densest subgraph subproblem is $O(n^3 \log n)$.
Optimal intervals subproblem is **NP**-hard

Densest subgraph

Algorithm by Charikar [1] has $\frac{1}{2}$ approximation guarantee and time complexity $O(n)$.

input : Fixed $\mathcal{T}$
start with induced graph $G(\mathcal{T}, V)$;
while *graph is not empty* **do**
  Delete node with the smallest degree;
return densest seen subgraph

Optimal intervals

- Greedy heuristics based on Maximal Coverage: edges — elements; intervals — sets
- Binary search for parameter α:
  $\max_{\mathcal{T}} q(G(\mathcal{T}, W)) - \alpha span(\mathcal{T}), \quad \text{s.t. } |\mathcal{T}| \leq K$

Optimal intervals: Greedy

Generalized Maximum Coverage:

For each set S_i item $e_j \in S_i$ has weight $w_i(e_j)$ and cost $c_i(e_j)$. Each set has a cost $c(S_i)$.

Given total cost budget B, find set of sets S, s.t.:

$$\text{maximize} \quad Q(S) = \sum_{S_i \in S} w_i(e_j)$$

Standard approach [2]: add next set R:

$$R = \arg\max \frac{Q(S \cup R) - Q(S)}{c(R)}$$

We combine two budgets.

$$R = \arg\max \frac{q(G(\mathcal{T} \cup R, W)) - q(G(\mathcal{T}, W))}{\max(x, y)}$$

where $x = \dfrac{1}{K - \mathcal{T}}$ and $y = \dfrac{span(R)}{B - span(\mathcal{T})}$

Greedy heuristics algorithm:

input : Fixed W
$\mathcal{T} \leftarrow \emptyset$;
for K times **do**
  Find R;
  $\mathcal{T} = \mathcal{T} \cup R$;
return $\mathcal{T}$

Optimal intervals: Binary

Transform the problem:

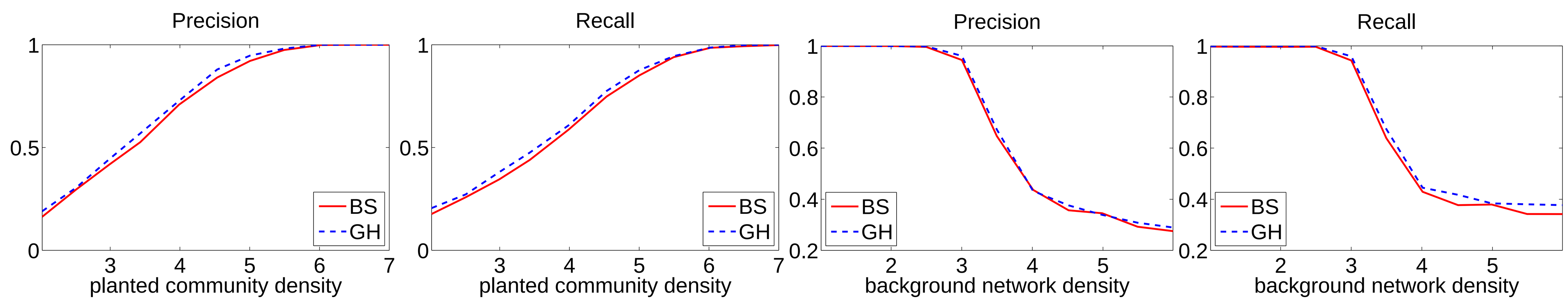
$$\max_{\mathcal{T}} q(G(\mathcal{T}, W)) - \alpha span(\mathcal{T})$$
$$\text{s.t. } |\mathcal{T}| \leq K$$

Note that for $\alpha = 0$ optimal $\mathcal{T}$ equals to whole interval.
Additionally, $span(\mathcal{T})$ decreases with α.
Thus, we can use binary search for α to fit the budget B.
However, the problem remains **NP**-hard for each fixed α.

Binary Search based algorithm:

input : Fixed W
$\mathcal{T} \leftarrow \emptyset$;
while *not converged* **do**
  Check budget B;
  Update α by binary search rule;
  With fixed α greedily find K intervals to maximize the cost function $q(G, (W\mathcal{T}))$;
return best solution that fits budget B

Experiments



					Community density			Community size		
Dataset	$d(\pi(G))$	$d(H)$	B	K	GH	BS	Base	GH	BS	Base
Facebook	2.498	5.292	1	5	3.666	3.666	2.4	6	6	5
				10	3.75	3.75	2.4	8	8	5
			7	5	3.875	4	3	16	9	6
				10	4.285	4.47	3	14	17	6
Twitter	2.608	10.119	1	5	5.111	5.333	4	9	9	6
				10	6.4	6.4	4	10	10	6
			7	5	6	6.222	4.666	14	9	9
				10	6.923	7.2	4.666	13	15	9

Twitter example

Twitter dataset: 3 months of tweets from Helsinki region.
Discovered community: 8 Twitter users with density = 6.0
Among these users: Aalto Entrepreneurship Society aaltoes and Aalto Venture Garage
Retrieved hash-tags: summerofstartups, startup, entrepreneur, slush10, aaltoes, me310, vc, churchillclub

References

[1] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. *APPROX*, 2000.
[2] R. Cohen and L. Katzir. The generalized maximum coverage problem. *Information Processing Letters*, 108, 2008.