Dense Subgraph Discovery Primitives: Risk Aversion and Exclusion Queries

Charalampos E. Tsourakakis†, **Tianyi Chen**†, Naonori Kakimura‡, Jakub Pachocki*

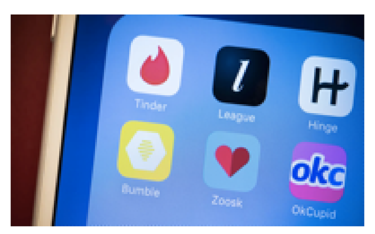
†Boston University, USA ‡Keio University, Japan *OpenAI, USA

BU Department of Computer Science

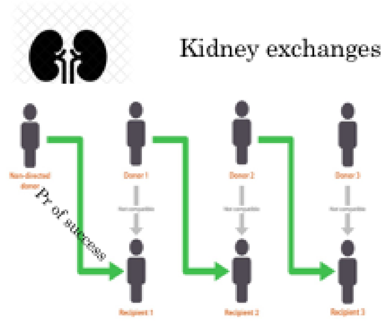
Project code: https://github.com/tsourolampis

Motivation

Uncertain graphs



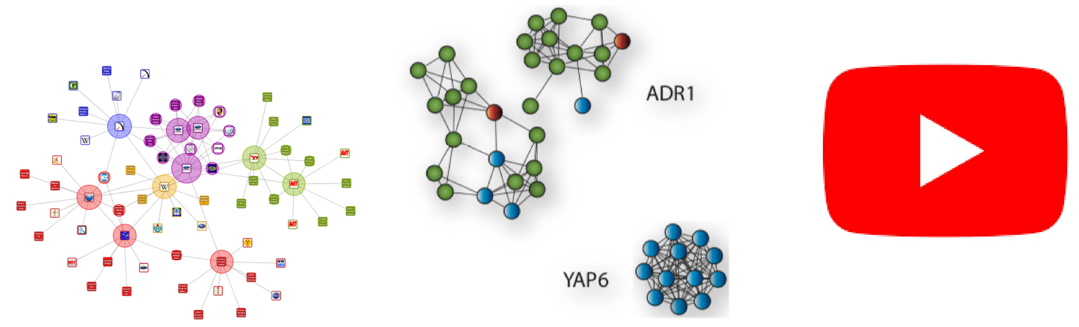
Online dating



$\mathcal{G} = (V, E, p)$, where $p : E \to (0, 1]$. Whether an edge exists or not is uncertain.

- Expected weight $\to$ reward.
- variance $\to$ risk(uncertainty).

Densest subgraph discovery



Given a graph $G = (V, E)$, the objective is to $\max_{S \subseteq V} \frac{e(S)}{|S|}$. For standard graph, this problem:

- is poly-time solvable via max flow[1].
- has $\frac{1}{2}$ approximation greedy algorithm[2].

How to do DSD on uncertain graph?

Problem1: Given an uncertain graph, how can we find a dense subgraph that has low risk?

Method

Intuitively, our goal is to find a subgraph $G[S]$ induced by $S \subseteq V$ such that:

- Its average expected reward $\frac{\sum_{e \in E(S)} w_e}{|S|}$ is large.
- The associated average risk is low $\frac{\sum_{e \in E(S)} \sigma_e^2}{|S|}$.

Transform:

- For each edge e we create two edges:
  - A positive edge, $w^+(e) = \mu_e$
  - A negative edge, $w^-(e) = \sigma_e^2$.

Objective:

$$max_{S \subseteq V} f(S) = \frac{w^+(S) + \lambda_1|S|}{w^-(S) + \lambda_2|S|}$$

$$\frac{w^+(S) + \lambda_1|S|}{w^-(S) + \lambda_2|S|} \geq q \to$$

$$\sum_{e \in E(S)} \underbrace{\left(w^+(e) - qw^-(e)\right)}_{\tilde{w}(e)} \geq |S| \underbrace{(q\lambda_2 - \lambda_1)}_{q'} \to$$

$$\sum_{e \in E(S)} \frac{\tilde{w}(e)}{|S|} \geq q' \; (\tilde{w}(e) \; can \; be \; negative)$$

Hardness: . The DSP on graphs with negative weights is NP-hard (Reduction from MAX-CUT).

Corollary: Assume that $w^+(e) \geq q_{max} w^-(e)$ for all $e \in E^+ \cup E^-$, where q_{max} is the maximum possible query value. Then, the densest subgraph problem is solvable in polynomial time.

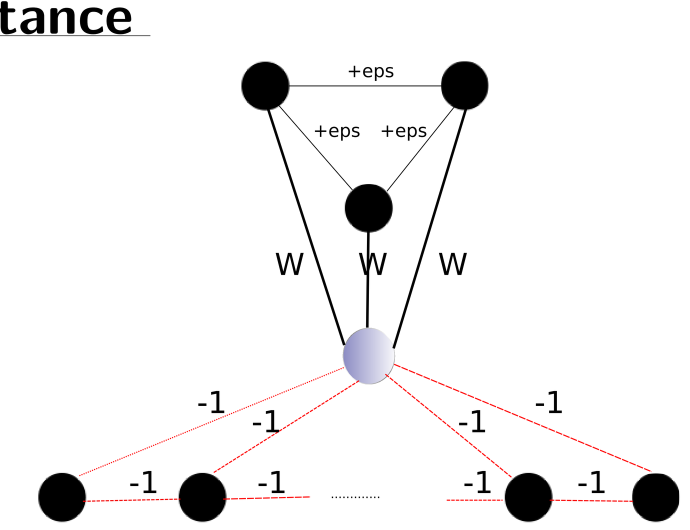
Bounding risk: $f(S) = \frac{w^+(S)+\lambda_1|S|}{Bw^-(S)+\lambda_2|S|}$ by changing parameter B.

Algorithm:

Algorithm 2: `Peeling`

Input: G
$n \leftarrow |V|, H_n \leftarrow G$;
for $i \leftarrow n$ *to* 2 **do**
  Let v be the vertex of H_i of minimum degree, i.e., $d(v) = deg^+(v) - deg^-(v)$ (break ties arbitrarily);
  $H_{i-1} \leftarrow H_i \backslash v$;
end
return H_j *that achieves maximum average degree among* $H_i s$, $i = 1, \ldots, n$.

Bad instance



- In total n nodes in the bottom row.
- Let $W = \frac{n-4}{3}$. Then, $3W - n < -3$, which leads to removing the central node at first.

Improvement

Algorithm 3: `Heuristic-Peeling`

Input: $G, C \in (0, +\infty)$
$n \leftarrow |V|, H_n \leftarrow G$ **for** $i \leftarrow n$ *to* 2 **do**
  Let v be the vertex of H_i of minimum degree, i.e., $d(v) = Cdeg^+(v) - deg^-(v)$ (break ties arbitrarily) $H_{i-1} \leftarrow H_i \backslash v$
end
return H_j *that achieves maximum average degree among* $H_i s$, $i = 1, \ldots, n$.

Extension - exclusion queries

Problem2: Given a large-scale multilayer network, how do we find a dense subgraph that excludes certain types of edges?

Approach: Use $-W$ weights for the excluded edge types. W can differ according to the scenarios.

Datasets

- Uncertain graphs

Name	n	m
Biogrid	5 640	59 748
Collins	1 622	9 074
Gavin	1 855	7 669
Krogan core	2 708	7 123
Krogan extended	3 672	14 317
TMDB	160 784	883 842

- Multilayer graphs

Name	n	m
Twitter (Feb. 1)	621 617	(902 834, 387 597, 222 253, 30 018, 63 062)
Twitter (Feb. 2)	706 104	(1 002 265, 388 669, 218 901, 29 621, 64 282)
Twitter (Feb. 3)	651 109	(1 010 002, 373 889, 218 717, 27 805, 59 503)
Twitter (Feb. 4)	528 594	(865 019, 435 536, 269 750, 32 584, 71 802)
Twitter (Feb. 5)	631 697	(999 961, 396 223, 233 464, 30 937, 66 968)
Twitter (Feb. 6)	732 852	(941 353, 407 834, 239 486, 31 853, 67 374)
Twitter (Feb. 7)	742 566	(1 129 011, 406 852, 236 121, 30 815, 68 093)

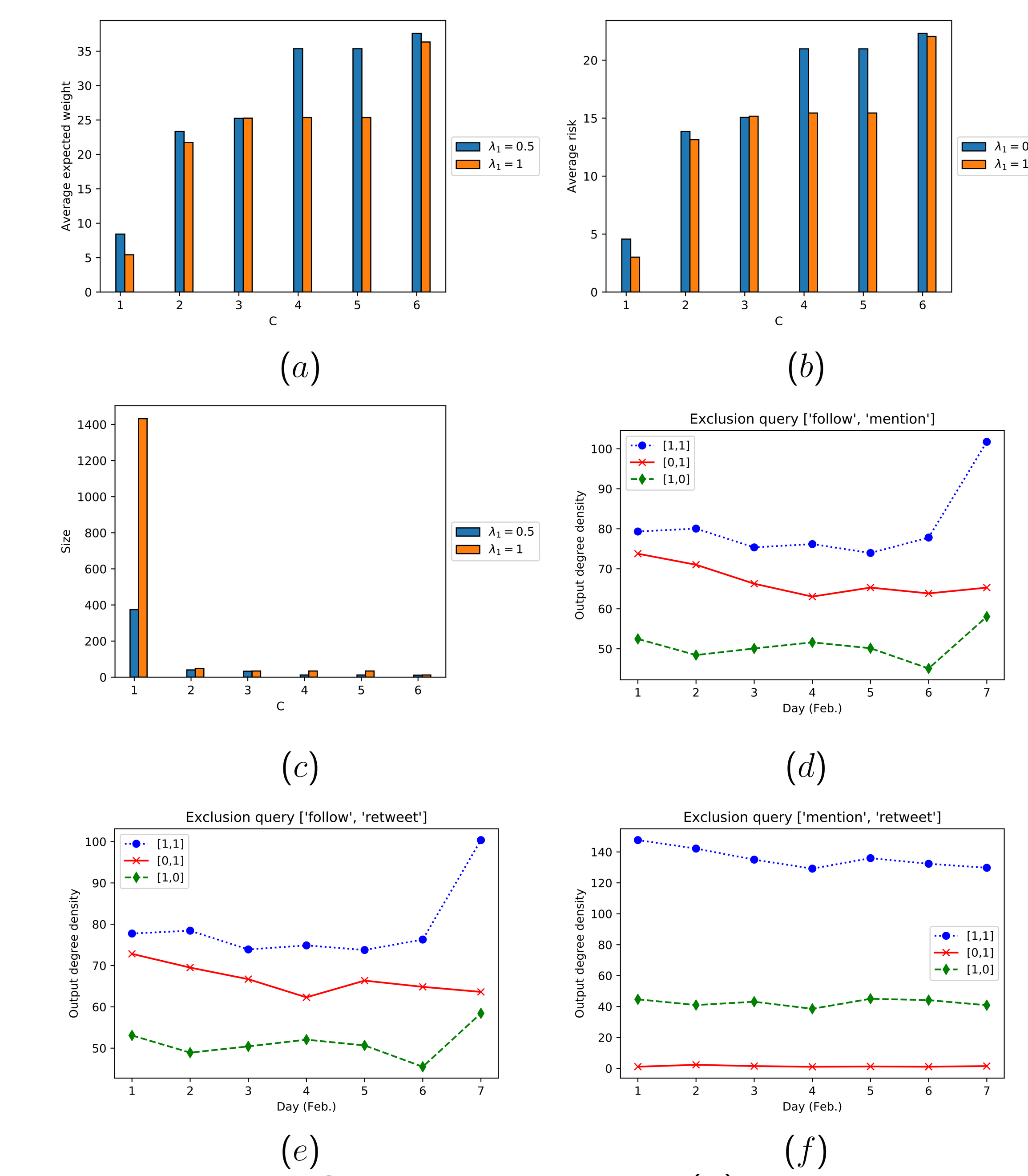
Results

- Test the trade-off between reward and risk by ranging B

B	Average exp. reward	average risk
0.25	0.18	0.09
1	0.17	0.08
2	0.13	0.06

Gavin dataset ($n = 1\,855, m = 7\,669$).

- Test the trade-off between reward, risk and size by ranging C and $\lambda 1$, see fig (a), (b) and (c).
- Test the heuristic exclusion queries on Twitter datasets, see fig (d), (e) and (f).



1. Risk averse DSD results for TMDB: (a) average expected weight, (b) average risk, (c) output size.
2. Degree density for three exclusion queries per each pair of interaction types over the period of the first week of February 2018: (d) Follow and mention. (e) Follow and retweet. (f) Mention and retweet.

- Exploring the effect of the negative weight $-W$ on the excluded edge types for various C values..

C	W	$\lvert S^*\rvert$	$\rho_{\text{retweet}}(S^*)$	$\rho_{\text{reply}}(S^*)$
0.1	1	296	63.44	-0.75
	5	99	45.67	-0.01
	200 000	200	30.37	0
1	1	346	72.70	-2.75
	5	319	68.70	-1.29
	200 000	200	30.38	0
10	1	351	73.10	-3.31
	5	351	73.10	-3.31
	200 000	200	30.37	0

References

[1] Goldberg, A. V.. Finding a Maximum Density Subgraph. *University of California at Berkeley*, 1984.

[2] Charikar, Moses. Greedy approximation algorithms for finding dense components in a graph. *APPROX*, 2000.

[3] Tsourakakis, Charalampos E and Sekar, Shreyas and Lam, Johnson and Yang, Liu. Risk-Averse Matchings over Uncertain Graph Databases, *arXiv preprint arXiv:1801.03190*, 2018.