Novel Dense Subgraph Discovery Primitives: Risk Aversion and Exclusion Queries

Charalampos E. Tsourakakis[1], **Tianyi Chen**[1],
Naonori Kakimura[2], Jakub Pachocki[3]
[1]Boston University, USA
[2]Keio University, Japan
[3]OpenAI, USA

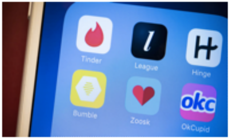
ECML PKDD

Sept, 2019

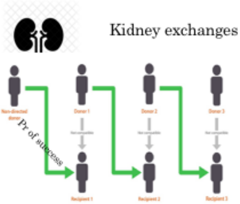
Outline of today's talk

1. Introduction
2. Related work
3. Risk-averse dense subgraphs (and a bonus extension)
4. Experiments
5. Open problems

Uncertain graphs are everywhere



Online dating



Uncertain (aka stochastic) graphs are ubiquitous!

- PPI networks [Asthana et al., 2004, Krogan et al., 2006]
- Dating apps
- Kidney exchange [Roth et al., 2004]
- Influence maximization [Kempe et al., 2003]
- Injecting privacy [Boldi et al., 2012]
- ...

Uncertain graph model

Existing work has focused on the following model (e.g., [Bonchi et al., 2014, Kollios et al., 2013])

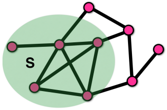
- Let $\mathcal{G} = (V, E, p)$ be an uncertain graph where $p : E \rightarrow (0, 1]$.
- Edge e exists with probability p_e independently from the rest of the edges
- We can make this model more general by replacing p with f_e, which is the probability distribution for edge e with parameters $\vec{\theta_e}$:

$$w(e) \sim f_e(x; \vec{\theta_e}) \forall e \in E.$$

- Each edge brings:
  1. Reward $\rightarrow$ expected weight
  2. Risk $\rightarrow$ variance

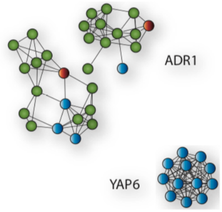
Densest subgraph problem (DSP)

Degree density: $\boxed{\rho(S) = \dfrac{e(S)}{|S|}}$. E.g.,

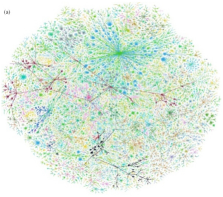


$\rho(S) = \frac{7}{5}$

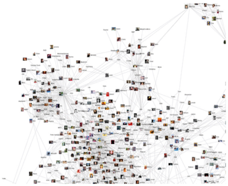
Objective: $\boxed{\max_{S \subseteq V} \rho(S)}$



DNA motif detection



Web community



Social network

Related research - Densest Subgraph Discovery

- DSD is poly-time solvable for non-negative weights! (via max flows)
  [Goldberg, 1984, Gallo et al., 1989, Khuller et al., 2009]

- 2-approximation algorithm which uses linear space $O(n+m)$ and runs in linear time $O(n+m)$ due to [Charikar, 2000] *(greedily removes the lowest degree node, and returns among the sequence of n graphs the one with the highest degree density)*

- *"The densest subgraph problem* **(DSP)** *lies at the core of large scale data mining"* [Bahmani et al., 2012]

Related research - Risk Aversion

- [Parchas et al., 2014] Proposed a heuristic to extract a good possible world to combine risk-aversion with efficiency, but lack guarantees.
- [Tsourakakis et al., 2018] Studied the problem of finding efficiently risk-averse graph and hypergraph matching algorithms.

- [Zou, 2013] DSP on uncertain graphs can be solved in polynomial time in expectation. (With limitation)
- [Miyauchi and Takeda, 2018] DSD on uncertain graphs with far different modelling assumptions and mathematical objective.

Risk-averse DSD formulation

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

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

We approach the problem as follows:

- For each edge we create two edges:
  1. A positive edge with weight equal to the expected reward, i.e., $w^+(e) = \mu_e$
  2. A negative edge with weight equal to the risk of the edge, i.e., $w^-(e) = \sigma_e^2$.

Risk-averse DSD formulation

- Our goal is to find a subgraph $S \subseteq V$ such that:
  1. large positive average degree $\frac{w^+(S)}{|S|}$ (large reward)
  2. small negative average degree $\frac{w^-(S)}{|S|}$ (small risk)

We combine the two objectives into one objective $f : 2^V \to \mathbb{R}$ that we wish to maximize:

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

Questions: But can we maximize this objective in polynomial time?

Insights

If we can answer the following query in polynomial time, then by binary search we can solve the problem:

$$\boxed{\text{Does there exist a subset of nodes } S \subseteq V \text{ such that } f(S) \geq q, \text{ where } q \text{ is a query value?}}$$

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

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

Questions: Can we solve the DSP in poly-time when the weights are negative?

Hardness

Theorem

The DSP on graphs with negative weights is NP-hard.

Reduction from MAX-CUT.

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

Any efficient algorithm?

Algorithm - DSP with Negative Weights

Algorithm 1: `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);

  $H_{i-1} \leftarrow H_i \backslash v$;

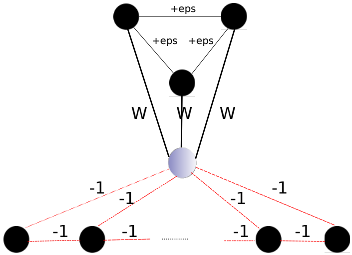
end

return H_j *that achieves maximum average degree among* H_i*s,* $i = 1, \ldots, n$.

Theorem

Let $G(V, E, w)$, $w : E \rightarrow \mathbb{R}$ *be an undirected weighted graph with possibly negative weights. If the negative degree* $deg^-(u)$ *of any node* u *is upper bounded by* Δ, *then our Algorithm outputs a set whose density is at least* $\frac{\rho^*}{2} - \frac{\Delta}{2}$.

Bad instance



Let $W = \frac{n-4}{3}$. Then, $3W - n < -3$. The degrees of the $n+4$ nodes are as follows:

$$\underbrace{3W - n}_{\text{one node}} < \underbrace{-3}_{n-2 \text{ nodes}} < \underbrace{-2}_{\text{two nodes}} < 0 < \underbrace{2\epsilon + W}_{\text{three nodes}}.$$

Heuristic

Algorithm 2: `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,* $i = 1, \ldots, n$.

Rule of thumb: Run the above heuristic for various values of C, and return the best possible subgraph!

We can use our heuristic to develop a new algorithmic primitive!

Exclusion queries

Problem

Given a multigraph $G(V, E, \ell)$, where $\ell : E \rightarrow \{1, \ldots, T\} = [T]$ is the labeling function, and T is the number of types of edges, and an input set $\mathcal{I} \subseteq [T]$ of edges, how do we find a set of nodes S that (i) induces a dense subgraph, and (ii) does not induce any edge e such that $\ell(e) \in \mathcal{I}$?

Approach: Use $-W$ weights for the excluded edge types.

Application: Given the daily Twitter interactions, find a dense subgraph in *follows* and *quotes* but with no *replies*. ($-W = -\infty$)

Uncertain graph datasets

Name	# of nodes	# of edges
■ 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

PPI datasets: Nodes represent proteins and the probability of the edge is equal to the existence probability of the interaction.

TMDB dataset: Nodes represent actors and the probability of the edge is equal to probability of two actors collaborate together.

Multilayer graph datasets

Name	# of nodes	# of edges (follows, mentions, retweets, quotes, replies)
⊙ 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)

Experimental findings – Exploring B

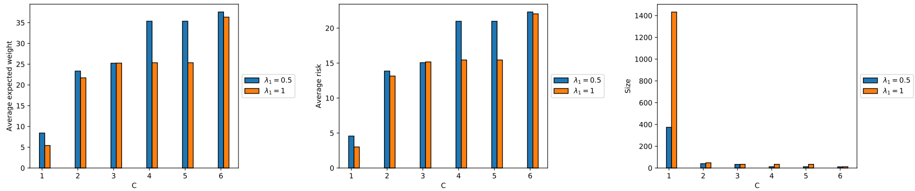
We test the trade-off between reward and risk by ranging B.

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

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

Reminder: $f(S) = \frac{w^+(S)+\lambda_1|S|}{Bw^-(S)+\lambda_2|S|}$

Experimental findings – TMDB



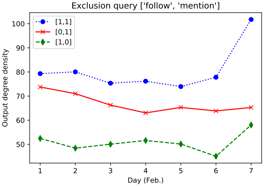
(α) (β) (γ)

Risk averse DSD results for TMDB:
(α) average expected weight, (β) average risk, (γ) output size.

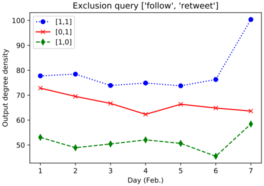
Reminder: $f(S) = \frac{w^+(S)+\lambda_1|S|}{Bw^-(S)+\lambda_2|S|}$

Experimental findings – Exclusion queries on Twitter

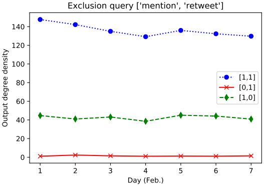
We set $C = 1, W = -\infty$:



(α)



(β)



(γ)

Degree density for three exclusion queries per each pair of interaction types over the period of the first week of February 2018. (α) Follow and mention. (β) Follow and retweet. (γ) Mention and retweet.

Experimental findings - Ranging W, C

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

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

Open problems

- **Dense subgraphs:** Study in greater depth the computational complexity of DSD with negative weights
- **General direction:** Design risk-averse algorithms that combine efficiency, and solid theoretical guarantees

Thank you! Questions?

email: ctony@bu.edu

web page: `http://c752334430.github.io`

code: `http://github.com/tsourolampis`

Slides modified from Babis's work.

references I

Asthana, S., King, O. D., Gibbons, F. D., and Roth, F. P. (2004).
Predicting protein complex membership using probabilistic network reliability.
Genome research, 14(6):1170–1175.

Bahmani, B., Kumar, R., and Vassilvitskii, S. (2012).
Densest subgraph in streaming and mapreduce.
Proc. VLDB Endow., 5(5):454–465.

Boldi, P., Bonchi, F., Gionis, A., and Tassa, T. (2012).
Injecting uncertainty in graphs for identity obfuscation.
Proceedings of the VLDB Endowment, 5(11):1376–1387.

references II

Bonchi, F., Gullo, F., Kaltenbrunner, A., and Volkovich, Y. (2014).
Core decomposition of uncertain graphs.
In *Proc. of the 20th ACM SIGKDD conference*, pages 1316–1325. ACM.

Charikar, M. (2000).
Greedy approximation algorithms for finding dense components in a graph.
In *APPROX*.

Kempe, D., Kleinberg, J., and Tardos, É. (2003).
Maximizing the spread of influence through a social network.
In *Proceedings of KDD 2003*, pages 137–146. ACM.

Kollios, G., Potamias, M., and Terzi, E. (2013).
Clustering large probabilistic graphs.
IEEE Transactions on Knowledge and Data Engineering, 25(2):325–336.

references III

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., et al. (2006).
Global landscape of protein complexes in the yeast saccharomyces cerevisiae.
Nature, 440(7084):637.

Miyauchi, A. and Takeda, A. (2018).
Robust densest subgraph discovery.
In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1188–1193. IEEE.

Parchas, P., Gullo, F., Papadias, D., and Bonchi, F. (2014).
The pursuit of a good possible world: extracting representative instances of uncertain graphs.
In *Proceedings SIGMOD 2014*, pages 967–978.

references IV

Roth, A. E., Sönmez, T., and Ünver, M. U. (2004).
Kidney exchange.
The Quarterly Journal of Economics, 119(2):457–488.

Tsourakakis, C. E., Sekar, S., Lam, J., and Yang, L. (2018).
Risk-averse matchings over uncertain graph databases.
arXiv preprint arXiv:1801.03190.

Zou, Z. (2013).
Polynomial-time algorithm for finding densest subgraphs in uncertain graphs.
In *Proceedings of MLG Workshop*.