



## **Fair densest subgraph across multiple graphs**

Chamalee Wickrama Arachchi, Nikolaj Tatti

University of Helsinki, Finland

**ECML-PKDD'24**

## Dense subgraph discovery

- Finding dense subgraphs has many applications in diverse domains such as community detection, biological system analysis, and anomaly detection.
- Definition of **density** :

$$d(S) = \frac{|E(S)|}{|V(S)|}$$

- **Densest subgraph problem** : Find a subset of vertices which maximizes the density.
- Exact<sup>1</sup> and greedy<sup>2</sup> algorithms.

---

<sup>1</sup>Andrew V Goldberg. Finding a maximum density subgraph. 1984.

<sup>2</sup>Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In International workshop on approximation algorithms for combinatorial optimization, pages 84–95. Springer, 2000.

## Total densest subgraph problem (TDS)

- Extension for multiple graph snapshots.
- **TDS** : Find a **common** set of vertices which maximizes sum of the densities <sup>3</sup>.

---

<sup>3</sup>Semertzidis, K., Pitoura, E., Terzi, E. and Tsaparas, P., 2019. Finding lasting dense subgraphs. Data Mining and Knowledge Discovery, 33(5), pp.1417-1445.

## What do we do?

- We consider two problem variants of the densest subgraph problem.
  - ▶ Multiple graph snapshots are given  $(G_1, G_2, \dots, G_r)$ .
  - ▶ The goal is to find a fair densest subgraph without **over-representing** the density among the graph snapshots.

## Fair densest subgraph problem (FDS)

### Input:

- A sequence of graph snapshots.
- Input parameter:  $\alpha$ .

### Output:

- A subgraph  $S$ .

### Such that:

- The sum of the densities induced by  $S$  is maximized.
- The difference between the maximum and minimum density induced by  $S$  is at most  $\alpha$ .

## The smallest difference densest subgraph problem (SDS)

### Input:

- A sequence of graph snapshots.
- Input parameter:  $\sigma$ .

### Output:

- A subgraph  $S$ .

### Such that:

- The difference between the maximum and minimum density induced by  $S$  is minimized.
- The sum of the densities induced by  $S$  is at least  $\sigma$ .

## Extended definitions

- Density induced by  $S$  on  $i$ th graph

$$d(S, G_i) = \frac{|E(S, G_i)|}{|S|} \quad .$$

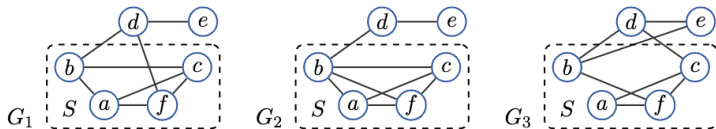
- The graph sequence :  $\mathcal{G} = (G_1, \dots, G_r)$
- Sum of the densities from all graph snapshots

$$d(S, \mathcal{G}) = \sum_{i=1}^r d(S, G_i) \quad .$$

- The difference between the maximum and minimum induced density

$$\Delta(S, \mathcal{G}) = \max_i d(S, G_i) - \min_i d(S, G_i) \quad .$$

## Example



- Let  $S = \{a, b, c, f\}$ .
- The sum of densities induced by  $S$  is  $\frac{5}{4} + \frac{6}{4} + \frac{4}{4} = 3.75$ .
- The density values are 1.25, 1.5, and 1.
- Therefore, the difference between maximum and minimum density value =  $1.5 - 1 = 0.5$ .

## Fair densest subgraph problem (FDS)

- Given a graph sequence  $\mathcal{G} = (G_1, \dots, G_r)$ , with  $G_i = (V, E_i)$  and real number  $\alpha$ ,  
find a subset of vertices  $S$ , such that  $d(S, \mathcal{G})$  is maximized and  $\Delta(S, \mathcal{G}) \leq \alpha$ .
- FDS is **NP**-hard.

## The smallest difference densest subgraph problem (SDS)

- Given a graph sequence  $\mathcal{G} = (G_1, \dots, G_r)$ , with  $G_i = (V, E_i)$  and real number  $\sigma$ ,  
find a common subset of vertices  $S$ , such that  
the density induced by  $S$  over  $\mathcal{G}$  is at least  $\sigma$  and  
 $\Delta(S, \mathcal{G})$  is minimized.
- SDS problem is **NP**-hard.

## ILP formulation of $\text{FDS}(\gamma)$

$$\begin{array}{ll}
 \text{MAXIMIZE} & \sum_{k=1}^r \sum_{ij \in E_k} x_{ij} - \gamma \sum_{i=1}^n y_i \\
 \text{SUBJECT TO} & x_{ij} \leq y_i \quad ij \in E \\
 & x_{ij} \leq y_j \quad ij \in E \\
 & x_{ij} \geq y_i + y_j - 1, \quad ij \in E \\
 & \sum_{ij \in E_k} x_{ij} - \sum_{ij \in E_\ell} x_{ij} \leq \alpha \sum_{i=1}^n y_i \quad k, \ell = 1, \dots, r \\
 & x_{ij}, y_j \in \{0, 1\} \quad .
 \end{array}$$

- $y_i$  : Whether the node  $i \in S$ .
- $x_{ij}$  : If the edge  $ij$  is in the subgraph induced by  $S$ .
- $\gamma$  : Parameter.
- $\text{FDS}(\gamma)$  yields the same solution as FDS for the **largest**  $\gamma$  with a non-empty solution.

## ILP formulation of $\text{SDS}(\gamma)$

$$\begin{array}{ll}
 \text{MINIMIZE} & u - \ell - \gamma \sum_{i=1}^n y_i \\
 \text{SUBJECT TO} & x_{ij} \leq y_i \quad ij \in E \\
 & x_{ij} \leq y_j \quad ij \in E \\
 & x_{ij} \geq y_i + y_j - 1 \quad ij \in E \\
 & \sum_{ij \in E_k} x_{ij} \geq \ell \quad k = 1, \dots, r \\
 & \sum_{ij \in E_k} x_{ij} \leq u \quad k = 1, \dots, r \\
 & \sum_{k=1}^r \sum_{ij \in E_k} x_{ij} \geq \sigma \sum_{i=1}^n y_i \\
 & x_{ij}, y_j \in \{0, 1\} \\
 & u, \ell \geq 0 \quad .
 \end{array}$$

- $y_i$  : Whether the node  $i \in S$ .
- $x_{ij}$  : If the edge  $ij$  is in the subgraph induced by  $S$ .
- $\gamma$  : Parameter.

## Greedy algorithm for SDS

---

**Algorithm 1:** SDS-GRD( $\mathcal{G}, \alpha$ ), finds greedily a subgraph  $S$  which minimizes  $\Delta(S, \mathcal{G})$  while satisfying  $d(S, \mathcal{G}) \geq \sigma$ .

---

```
1  $S \leftarrow$  The solution of TDS;  
2 while changes to  $\Delta(S)$  do  
3   Find a vertex  $v$  which minimizes  $\Delta(S)$  either by adding or removing while  
   satisfying density constraint;  
4   Add or remove  $v$  from the set  $S$ ;  
5 return  $S$ ;
```

---

Figure: Greedy algorithm

## Greedy algorithm for FDS

- **First phase:** we search for a set  $S$  such that  $\Delta(S, \mathcal{G}) \leq \alpha$ .
- **Second phase:** we start from our feasible set returned in the first phase, and at each iteration, we try to improve our density score by picking the best vertex to either add or delete until convergence.

## Experiments with Synthetic Datasets

- $d_{tds}$  : The solution of TDS.
- *constr.* : The input parameter  $\alpha$  in FDS or normalized parameter  $\sigma_{norm}$  in SDS.
- $d_{dis}$  : The discovered sum of densities.
- $d_{min}$ ,  $d_{max}$ , and  $d_a$  : The minimum, maximum, and average density induced by the discovered subgraph over the graph sequence.
- $i$  : The number of iterations of the algorithms.
- *Jacc.* : Jaccard index between the solution set and ground truth set.

<i>Algorithm</i>	<i>constr.</i>	$d_{dis}$	$d_{min}$	$d_{max}$	$d_a$	$\Delta$	$i$	$ S $	<i>Jacc.</i>	time
FDS-IP	3.9	51.12	10.84	14.74	12.78	3.9	8	100	1	127
FDS-GRD	3.9	50.53	10.73	14.58	12.63	3.85	13	98	0.96	81
SDS-IP	0.69	51.12	10.84	14.74	12.78	3.9	8	100	1	159
SDS-GRD	0.69	51.17	10.59	15.03	12.79	4.44	97	103	0.95	9

### Groundtruth vs TDS solution:

- The sum of the densities = 51.12 and difference  $\Delta = 3.9$ .
- The sum of the densities = 74.33 and difference  $\Delta = 22.84$ .

## Experiments with Real-world Datasets

- $n$  : The number of vertices.
- $m$  : The number of edges.
- $r$  : The number of snapshots.
- $d_{tds}$ : The solution of TDS.
- $\Delta_{tds}$  : The difference between the maximum and minimum density of the TDS solution.
- $d_{ind}$  : The sum of densities of individual densest subgraph from each graph snapshot.
- $d_{mds}$  and  $\Delta_{mds}$  give the total density and the difference between the maximum and minimum density of the solution.

<i>Data</i>	$n$	$m$	$r$	$d_{ind}$	$d_{tds}$	$d_{mds}$	$\Delta_{tds}$	$\Delta_{mds}$
<i>Twitter-#</i>	806	1 518	15	38.8	9.8	5.5	2	0.21
<i>Hospital</i>	75	1 885	5	41.68	30.29	17.76	6.93	2.47
<i>Airports</i>	417	3 588	37	83.75	24.54	16.34	2.98	1.26
<i>Students</i>	889	5 329	122	118.01	26.32	17.15	0.68	0.26
<i>Tumblr</i>	1 980	5 812	89	103.99	55.83	43.62	1.33	0.77
<i>Facebook</i>	4 117	8 646	104	88.65	14	5.75	0.5	0.07
<i>Twitter-user</i>	4 605	10 155	93	90.63	23	12.81	0.5	0.17

## Experiments with Real-world Datasets

- $\alpha$  : The input parameter in FDS problem.
- $d_{sum}$  : The discovered sum of densities.
- $time$  : The computational time in seconds.

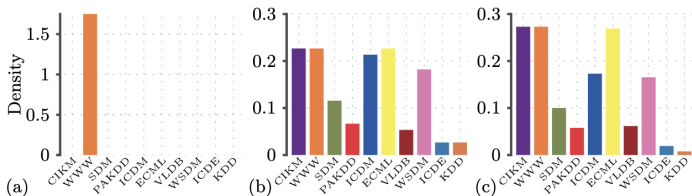
Data	$\alpha$	$d_{sum}$		$\Delta$		time		size		$i_{IP}$	$i_{GR}$
		IP	GR	IP	GR	IP	GR	IP	GR		
Twitter-#	0.3	6.45	1.75	0.3	0.25	14	2.44	20	4	18	3
	0.5	7.58	5	0.5	0.5	3	2.33	12	2	18	1
	0.7	7.92	5	0.69	0.5	4	2.19	13	2	18	1
Hospital	0.3	12.21	4.71	0.29	0.21	23	4.11	14	14	12	2
	0.5	13.86	5	0.5	0.5	23	1.48	14	16	12	5
	0.7	14.53	8.33	0.67	0.67	12	2.8	15	18	12	5
Airports	0.3	14.96	6.3	0.3	0.3	1006	25.31	70	50	17	5
	0.5	17.02	9.88	0.5	0.5	127	30.95	54	40	17	8
	0.7	18.43	12.27	0.69	0.7	106	26.32	49	33	17	1
Students	0.3	22.65	17.83	0.3	0.29	28	102.35	40	24	19	2
	0.5	25.64	25.17	0.5	0.5	5	102.45	36	18	19	2
	0.7	26.32	26.32	0.68	0.68	4	92.65	19	19	19	5
Tumblr	0.3	29.41	12.86	0.3	0.29	74	14.34	27	7	20	2
	0.5	38.56	26	0.5	0.5	14	13.37	18	8	19	1
	0.7	45.92	37	0.69	0.67	7	13.19	13	3	19	1
Facebook	0.3	11.14	7	0.29	0.25	14	26.05	7	4	22	1
	0.5	14	14	0.5	0.5	11	29.37	2	2	22	1
	0.7	14	14	0.5	0.5	12	29.61	2	2	22	1
Twitter-user	0.3	23	11.5	0.5	0.25	15	47.46	4	4	21	1
	0.5	23	23	0.5	0.5	14	42.82	4	2	21	1
	0.7	23	23	0.5	0.5	14	44.91	4	2	21	1

## Experiments with Real-world Datasets

- $\sigma$  : The input parameter in SDS where  $\sigma = \sigma_{nrm} \times d_{tds}$ .
- $d_{sum}$  : The discovered sum of densities.
- $time$  : The computational time in seconds.

Data	$\sigma_{nrm}$	$\Delta$			$d_{sum}$		time		size		$i_{IP}$	$i_{GR}$
		$\sigma$	IP	GR	IP	GR	IP	GR	IP	GR		
Twitter-#	0.3	2.94		0.33		3.33		0.12		3		5
	0.5	4.9	0.06	0.5	4.91	5	92	0.11	35	2	21	4
	0.7	6.86	0.36	0.75	7	7.25	26	0.11	14	4	18	4
Hospital	0.3	9.09	0	1.58	9.67	9.13	297	0.19	15	24	34	21
	0.5	15.14	0.93	2.73	15.2	15.17	32	0.14	15	30	13	13
	0.7	21.2	2.59	4.15	21.23	21.24	28	0.11	22	34	12	9
Airports	0.3	7.36		0.43		7.4		1.81		40		34
	0.5	12.27		0.7		12.27		1.03		33		19
	0.7	17.18	0.53	1.33	17.19	17.22	458	0.79	57	36	17	14
Students	0.3	7.89		0.13		7.93		3.72		40		30
	0.5	13.16		0.26		13.16		0.97		19		9
	0.7	18.42	0.17	0.3	18.42	18.48	556	1.02	64	23	19	9
Tumblr	0.3	16.75		0.38		17.63		0.92		8		7
	0.5	27.92	0.27	0.57	27.95	28.29	154	0.48	22	7	20	4
	0.7	39.08	0.53	0.8	39.16	39.6	61	0.27	19	5	19	2
Facebook	0.3	4.2		0.14		4.29		1.69		7		6
	0.5	7	0.07	0.25	7.02	7	962	0.84	60	4	23	3
	0.7	9.8	0.21	0.33	9.86	10	43	0.5	14	3	21	2
Twitter-user	0.3	6.9		0.14		7.14		1.67		7		6
	0.5	11.5		0.25		11.5		0.77		4		3
	0.7	16.1	0.19	0.5	16.12	23	132	0.48	26	2	22	2

## Case-study



- (a) : TDS solution with density value of 1.75.
- (b) : FDS solution with density value of 1.36 for  $\alpha = 0.2$ .
- (c) : SDS solution with density value of 1.4 for  $\sigma = 0.8d_{tds}$ .

## Conclusion

- Introduced two variants of dense subgraph discovery problem for graphs with multiple snapshots that take fairness into account.
- Given an input parameter  $\alpha$ , the goal of our first variant is to find a dense subgraph maximizing the sum of densities across snapshots such that the difference between the maximum and minimum induced density is at most  $\alpha$ .
- Considered dual problem where given an input parameter  $\sigma$ , we find a subgraph that minimizes the gap between the maximum and minimum density induced by the subgraph while inducing at least  $\sigma$  amount of total density over the graphs.
- Proved that our problems are **NP**-hard.
- proposed two exponential time, exact algorithms based on integer programming. and also proposed two polynomial time heuristics.
- Showed experimentally that the algorithms could find the ground truth using synthetic dataset.
- Showed experimentally that the number of iterations was low in iterative algorithm and computational time is reasonable.

**Thank you for your attention!!!**