SpecGreedy: Unified Dense Subgraph Detection

Wenjie Feng[1,2](✉), Shenghua Liu[1,2](✉), Danai Koutra[3],
Huawei Shen[1,2], and Xueqi Cheng[1,2]

[1] CAS Key Laboratory of Network Data Science & Technology,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] University of Michigan, Ann Arbor, MI, USA
wenchiehfeng.us@gmail.com, liushenghua@ict.ac.cn, dkoutra@umich.edu,
shenhuawei, cxq@ict.ac.cn

Abstract. In this supplementary document, we provide additional proofs of theorems, datasets information, and experimental results, all of which supplement the main paper [1].

A Proof of Lemma 9.

Here we give proof for the Lemma 9 in Section 3, that is, the densest subgraph detection in a bipartite graph $\hat{\mathcal{G}}$ can be reduced to the GenDS framework by converting $\hat{\mathcal{G}}$ to be a mono-partite graph.

Lemma 9. *Given a bipartite graph $\hat{\mathcal{G}} = (L \cup R, E)$ with $|L| + |R| = n$, the densest bipartite subgraph detection problem over $\hat{\mathcal{G}}$ corresponds to the setting that $\boldsymbol{x} = [\boldsymbol{y}, \boldsymbol{z}]$, where $\boldsymbol{y} \in \{0,1\}^{|L|}, \boldsymbol{z} \in \{0,1\}^{|R|}$, and $\mathbf{P}$, $\mathbf{Q} \in \mathbb{R}^{n \times n}$,*

$$\mathbf{P} = \begin{bmatrix} \mathbf{D}_{\boldsymbol{c}_L} & \frac{\mathbf{A}}{2} \\ \frac{\mathbf{A}^T}{2} & \mathbf{D}_{\boldsymbol{c}_R} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{D}_{\boldsymbol{c}_L} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{\boldsymbol{c}_R} \end{bmatrix}, \ \mathbf{Q} = \begin{bmatrix} \mathbf{I}_{|L|} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{|R|} \end{bmatrix} \tag{1}$$

where $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$ are the node weight vectors for the nodesets L and R respectively, $\mathbf{I}_{|L|}$ is the identity matrix of size $|L| \times |L|$, and $\mathbf{I}_{|R|}$ is similar.

Proof. Let $\delta(\boldsymbol{x}) = \{i : \boldsymbol{x}_i = 1, i \in \lceil n \rceil\}$ for the indicator vector $\boldsymbol{x}$, the selected node subset as $S = \delta(\boldsymbol{x}) = \delta(\boldsymbol{y}) \cup \delta(\boldsymbol{z})$, $L_S = \delta(\boldsymbol{y})$, and $R_S = \delta(\boldsymbol{z})$. Then,

$$\begin{aligned} \boldsymbol{x}^T \mathbf{P} \boldsymbol{x} &= \boldsymbol{y}^T \mathbf{D}_{\boldsymbol{c}_L} \boldsymbol{y} + \boldsymbol{z}^T \mathbf{D}_{\boldsymbol{c}_R} \boldsymbol{z} + \boldsymbol{y}^T \frac{\mathbf{A}}{2} \boldsymbol{z} + \boldsymbol{z}^T \frac{\mathbf{A}}{2}^T \boldsymbol{y} \\ &= \boldsymbol{y}^T \mathbf{D}_{\boldsymbol{c}_L} \boldsymbol{y} + \boldsymbol{z}^T \mathbf{D}_{\boldsymbol{c}_R} \boldsymbol{z} + \boldsymbol{y}^T \mathbf{A} \boldsymbol{z} \\ &= \sum_{i \in S} c_i + \sum_{(i,j) \in E \wedge i \in L_S \wedge j \in R_S} a_{ij}, \end{aligned}$$

and $\boldsymbol{x}^T \mathbf{Q} \boldsymbol{x} = \boldsymbol{y}^T \mathbf{I}_{|L|} \boldsymbol{y} + \boldsymbol{z}^T \mathbf{I}_{|R|} \boldsymbol{z} = |L_S| + |R_S| = |S|$. Moreover, if $\mathbf{D}_{\boldsymbol{c}_L} = \mathbf{0}$ and $\mathbf{D}_{\boldsymbol{c}_R} = \mathbf{0}$, $\boldsymbol{x}^T \mathbf{P} \boldsymbol{x} = \boldsymbol{y}^T \mathbf{A} \boldsymbol{z} = |E(S)|$ for the unweighted graph $\hat{\mathcal{G}}$. ■

B Proof of Theorem 12 (Bigraph Spectral).

Here we provide the details of proof for the Theorem 12 in Section 4.

Proof. Given an asymmetric matrix $\mathbf{A}_r \in \mathbb{R}^{m\times n}$, its singular value decomposition is denoted as $\mathbf{A}_r = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{i=1}^{K} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ where $\boldsymbol{u}_i^T \boldsymbol{u}_j = \boldsymbol{v}_i^T \boldsymbol{v}_j = 0$ if $i \neq j$ otherwise 1 and $\mathbf{\Sigma} = diag(\sigma_1, \ldots, \sigma_K)$ for singular values $\sigma_1 \geq \cdots \geq \sigma_K > 0$; K is the rank of $\mathbf{A}_r$ and $K \leq \min\{m, n\}$.

Given the non-zero vectors $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{z} \in \mathbb{R}^n$, we only consider the case with $\|\boldsymbol{y}\| = \|\boldsymbol{z}\| = 1$ for the quadratic optimization problem related to $\mathbf{A}_r$, thus, $R(\mathbf{A}_r; \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{2}\boldsymbol{y}^T \mathbf{A}_r \boldsymbol{z}$. As for the case where $\|\boldsymbol{y}\| \neq 1$ or/and $\|\boldsymbol{z}\| \neq 1$, we have following result by using the normalized vectors $\hat{\boldsymbol{y}} = \frac{\boldsymbol{y}}{\sqrt{\boldsymbol{y}^T\boldsymbol{y}}}$ and $\hat{\boldsymbol{z}} = \frac{\boldsymbol{z}}{\sqrt{\boldsymbol{z}^T\boldsymbol{z}}}$, i.e.,

$$R(\mathbf{A}_r; \boldsymbol{y}, \boldsymbol{z}) = \frac{\boldsymbol{y}^T \mathbf{A}_r \boldsymbol{z}}{\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{z}^T\boldsymbol{z}} = \frac{\sqrt{\boldsymbol{y}^T\boldsymbol{y}}\sqrt{\boldsymbol{z}^T\boldsymbol{z}}}{\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{z}^T\boldsymbol{z}} \hat{\boldsymbol{y}}^T \mathbf{A}_r \hat{\boldsymbol{z}} \leq \frac{\hat{\boldsymbol{y}}^T \mathbf{A}_r \hat{\boldsymbol{z}}}{2} = R(\mathbf{A}_r; \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}),$$

where the inequality follows from the *Cauchy–Schwarz Inequality.*

Let $\mathcal{U} = \{\boldsymbol{u}_i; i \in \lceil K \rceil\}$ and $\mathcal{V} = \{\boldsymbol{v}_i; i \in \lceil K \rceil\}$, we discuss the following three cases for the Theorem 12.

Case 1. If $\boldsymbol{y} = \boldsymbol{u}_i \in \mathcal{U}$ and $\boldsymbol{z} = \boldsymbol{v}_j \in \mathcal{V}$, then

$$R(\mathbf{A}_r; \boldsymbol{y}, \boldsymbol{z}) = \begin{cases} \frac{\sigma_i}{2} & i = j, \\ 0 & i \neq j. \end{cases}$$

So, $R(\mathbf{A}_r; \boldsymbol{y}, \boldsymbol{z}) \leq \frac{\sigma_1}{2}$, the equation holds only when $i = j = 1$.

Case 2. If $\boldsymbol{y} = \boldsymbol{u}_i \in \mathcal{U}$ and $\boldsymbol{z} \notin \mathcal{V}$, then

$$R(\mathbf{A}_r; \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{2}\sigma_i \boldsymbol{v}_i^T \boldsymbol{z} < \frac{1}{2}\sigma_i \boldsymbol{v}_i^T \boldsymbol{v}_i = \frac{\sigma_i}{2} \leq \frac{\sigma_1}{2}.$$

For the other case where $\boldsymbol{y} \notin \mathcal{U}$ and $\boldsymbol{z} = \boldsymbol{v}_j \in \mathcal{V}$, we also have the similar result.

Case 3. If $\boldsymbol{y} \notin \mathcal{U}$ and $\boldsymbol{z} \notin \mathcal{V}$.

We define the basis for any vector belonging to $\mathbb{R}^m$ as $B_m = \mathcal{U} \cup \tilde{\mathcal{U}}$, where $\tilde{\mathcal{U}} = \{\boldsymbol{u}_{K+1}, \ldots, \boldsymbol{u}_m\}$ is the `NULL` space of the matrix $\mathbf{A}_r^T$ and $\forall i, j \in \lceil m \rceil$, $\boldsymbol{u}_i^T \boldsymbol{u}_j = 0$ if $i \neq j$ otherwise 1. Similarly, the basis $B_n = \mathcal{V} \cup \tilde{\mathcal{V}}$ for $\mathbb{R}^n$ where $\tilde{\mathcal{V}} = \{\boldsymbol{v}_{K+1}, \ldots, \boldsymbol{v}_n\}$ is the `NULL` space of $\mathbf{A}_r$ and $\forall i, j \in \lceil n \rceil$, $\boldsymbol{v}_i^T \boldsymbol{v}_j = 0$ if $i \neq j$ otherwise 1. The singular values for `NULL` space are zeros, i.e. $\sigma_i = 0$ for $i > K$.

Using the basis B_m and B_n, the vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ then can be represented as $\boldsymbol{y} = \sum_{j=1}^{m} s_j \boldsymbol{u}_j$ and $\boldsymbol{z} = \sum_{l=1}^{n} t_l \boldsymbol{v}_l$ where $\sum_{j=1}^{m} s_j^2 = \sum_{l=1}^{n} t_l^2 = 1$. Then

$$
\begin{aligned}
R(\mathbf{A}_r; \boldsymbol{y}, \boldsymbol{z}) &= \frac{1}{2}\boldsymbol{y}^T \mathbf{A}_r \boldsymbol{z} \\
&= \frac{1}{2}\sum_{i=0}^{K} \sigma_i \boldsymbol{y}^T \boldsymbol{u}_i \boldsymbol{v}_i{}^T \boldsymbol{z} = \frac{1}{2}\sum_{i=1}^{K} \sigma_i (\boldsymbol{y}^T \boldsymbol{u}_i)(\boldsymbol{z}^T \boldsymbol{v}_i)^T \\
&= \frac{1}{2}\sum_{i=1}^{K} \sigma_i \left(\left(\sum_{j=1}^{m} s_j \boldsymbol{u}_j^T \right) \boldsymbol{u}_i \right) \left(\left(\sum_{l=1}^{n} t_l \boldsymbol{v}_l^T \right) \boldsymbol{v}_i \right)^T \\
&= \frac{1}{2}\sum_{i=1}^{K} \sigma_i \left(s_i + \sum_{j=K+1}^{m} s_j \boldsymbol{u}_j^T \boldsymbol{u}_i \right) \left(t_i + \sum_{l=K+1}^{n} t_l \boldsymbol{v}_l^T \boldsymbol{v}_i \right)^T \\
&= \frac{1}{2}\sum_{i=1}^{K} \sigma_i s_i t_i \leq \frac{1}{2}\sum_{i=1}^{K} \sigma_i |s_i||t_i| \\
&\leq \frac{1}{2} \max_{i \in \lceil K \rceil} \sigma_i = \frac{\sigma_1}{2}.
\end{aligned}
$$

where the last inequality can be achieved for the following property

$$
\sum_{i=1}^{K} |s_i||t_i| \leq \frac{1}{2}(\sum_{j=1}^{K} s_j^2 + \sum_{l=1}^{K} t_l^2) \leq 1.
$$

Therefore, we can conclude that for any non-zero vectors $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{z} \in \mathbb{R}^n$, $R(\mathbf{A}_r, \boldsymbol{y}, \boldsymbol{z}) \leq \frac{\sigma_1}{2}$, the equality holds if and only if $\boldsymbol{y} = \boldsymbol{u}_1$ and $\boldsymbol{z} = \boldsymbol{v}_1$.

Moreover, it is also easy to extend the above result to the general case for σ_k where $k > 1$. ■

C Datasets information.

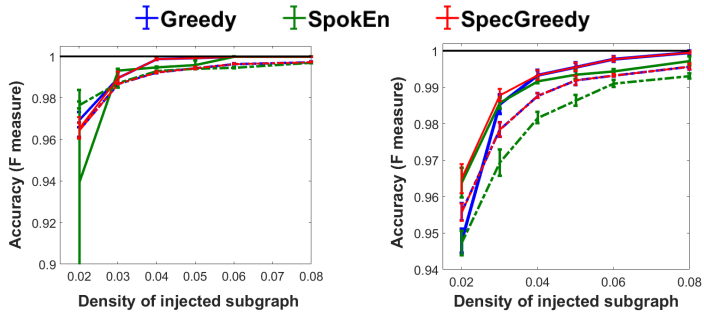
In our experiments in Section 6 of the main paper, we used a variety of datasets (40 in total) obtained from some publicly available sources, including Stanford's SNAP database [3], ASU's Social Computing Data Repository [7], Network Repository [5], AMiner scholar datasets [6], and from Koblenz Network Collection [2]. Table 1 lists the detailed information about the real-world networks, where the first cluster contains 32 monopartite graphs, and the second is some bipartite graphs; we also consider the edge weights for some marked datasets in Table 1.

D Additional Experiments

As in Section 6 of the main paper, we measure the performance of SpecGreedy, including the injection detection in Section 6.2 and the its scalability in Section 6.3. Here, we give the detailed experimental results as follows.

Fig.1 shows the detection accuracy of each methods for detecting injected dense subgraphs with different densities. As it shows, we can see that SpecGreedy achieves equally high accuracy as Greedy and is better than SpokEn.

Fig.2 illustrates linear-scalability of SpecGreedy with respect to the different size of nodes and edges for the `twitter-ASU` graph.



(a) Injection without camouflage (b) Injection with random camouflages

Fig. 1. Performance comparison for injection detection in the synthetic graphs. Some dense subgraphs with different density are injected into graph; the solid and dash lines correspond to two different subsets of amazon-Art data. SpecGreedy achieves similar accuracy as the Greedy algorithm and outperforms SpokEn.

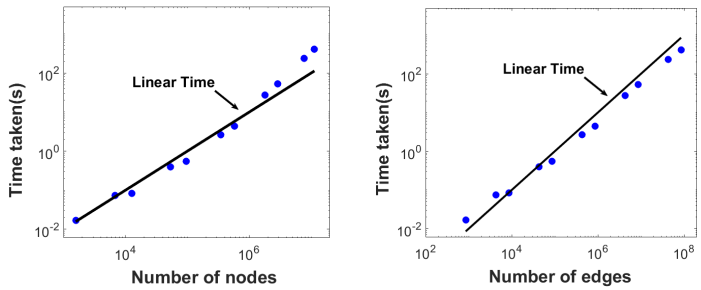


Fig. 2. The linear scalability of SpecGreedy. The time taken of SpecGreedy grows linearly with # of nodes and # of edges in graph.

References

1. W. Feng, S. Liu, D. Koutra, H. Shen, and X. Cheng. Specgreedy: Unified dense subgraph detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2020.

2. J. Kunegis. Konect: the koblenz network collection. In *WWW*, pages 1343–1350, 2013.
3. J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.
4. A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *IMC'07*, October 2007.
5. R. A. Rossi and N. K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.
6. H. Wan, Y. Zhang, J. Zhang, and J. Tang. Aminer: Search and mining of academic social networks. *Data Intelligence*, 1(1):58–76, 2019.
7. R. Zafarani and H. Liu. Social computing data repository at ASU, 2009.

Table 1. Summary of real world datasets used in experiment.

Name	$\|V\|$	$\|E\|$	**Content**
soc-twitter [2]	41.7M	1.47B	Social network
soc-Sinaweibo [5]	58.7M	261M	Social network
com-Orkut [3]	3.07M	117M	Social network
twitter-ASU [7]	11.3M	85.3M	Social network
livejournal-MPI [4]	5.28M	76.9M	Social network
ca-DBLP-NET [5]	1.31M	19.0M	Co-authorship
ego-gplus [3]	108K	12.2M	Social network
as-Skitter [3]	1.7M	11.1M	Internet topology
web-BerkStan [3]	685K	6.65M	Web
soc-Flickr [7]	80.5K	5.90M	Social
road-CA [3]	1.97M	5.53M	Road Net
com-WikiTalk [3]	2.39M	5.02M	Communication
web-Google [3]	876K	4.32M	Web graph
ca-Aminer [6]	1.56M	4.26M	Collaboration
road-TX [3]	1.38M	3.84M	Road Net
road-PA [3]	1.97M	3.08M	Road Net
soc-Youtube [3]	1.13M	2.99M	Social network
web-Stanford [3]	282K	2.31M	Web graph
ca-DBLP2012 [5]	317K	1.05M	Collboration
com-Amazon [3]	548K	926K	Community
twitter-ICWSM [2]	820K	835K	Social network
soc-Slashdot0902 [3]	82.2K	504K	Social network
soc-Slashdot0811 [3]	77.4K	469K	Social network
soc-Epinions [3]	75.9K	406K	Social network
blogcatalog [7]	10.3K	334K	Blog
ca-AstroPh [3]	18.7K	198K	Collaboration
email-Enron [3]	36.7K	183K	Communication
ca-HepPh [3]	12K	118K	Collaboration
soc-Hamsterster [5]	2.4K	16.6K	Social network
ca-GrQc [3]	5.2K	14.5K	Collaboration
*ca-Patents-AM [6]	2.08M	11.5M	Co-authorship
*ego-twitter [3]	81.3K	2.42M	Social network
livejournal-group [4]	10.7M	112M	Social network
cit-Patents-AM [6]	6.84M	54.0M	Citation
cit-Patents [3]	3.77M	16.5M	Citation
yelp-business [3]	86.4K	3.22M	Rating
beerAdvocate [3]	33.4K	65.9K	Review
*weibo-retweet	10.8M	50.1M	Social network
*amazon-Good [3]	3.38M	5.84M	Rating
*amazon-Art [3]	28.3K	28.0K	Rating

*: We also consider the edge weights for these marked datasets.