A local algorithm for finding dense bipartite-like subgraphs [⋆]

Pan Peng[1,2]

[1]State Key Laboratory of Computer Science, Institute of Software,
Chinese Academy of Sciences
[2]School of Information Science and Engineering,
Graduate University of China Academy of Sciences, P.R.China.
pengpan@ios.ac.cn

Abstract. We give a *local algorithm* to extract dense bipartite-like subgraphs which characterize cyber-communities in the Web [13]. We use the *bipartiteness ratio* of a set as the quality measure that was introduced by Trevisan [20]. Our algorithm, denoted as `FindDenseBipartite`(v, s, θ), takes as input a starting vertex v, a volume target s and a bipartiteness ratio parameter θ and outputs an induced subgraph of G. It is guaranteed to have the following approximation performance: for any subgraph S with bipartiteness ratio θ, there exists a subset $S_\theta \subseteq S$ such that $\mathrm{vol}(S_\theta) \geq \mathrm{vol}(S)/9$ and that if the starting vertex $v \in S_\theta$ and $s \geq \mathrm{vol}(S)$, the algorithm `FindDenseBipartite`(v, s, θ) outputs a subgraph (X, Y) with bipartiteness ratio $O(\sqrt{\theta})$. The running time of the algorithm is $O(s^2(\Delta + \log s))$, where Δ is the maximum degree of G, independent of the size of G.

1 Introduction

A local algorithm for massive graphs is one that explores only portion of the given graph and finds a solution with good approximation guarantee. Given a graph G as an oracle, from which the algorithm can request the degree of a vertex or the adjacency list of a vertex, and a numerical property P of subgraphs (such as diameter, conductance), a local algorithm is supposed to have the following form: it takes as input a starting vertex v (or a small set of vertices), only traverses the vertices that near v and outputs a subgraph S such that $P(S)$ is close to $P(S^*)$, where S^* is the subgraph containing v that has the optimal value for P. However, in the design of local algorithms, an approximation guarantee result as above is too strong, if possible, to obtain and it is usually relaxed as follows: if S is a subgraph, then there exists a large subset $S' \subseteq S$ such that for any starting vertex $v \in S'$, the algorithm will output a subgraph for which the P value is close to $P(S)$. This algorithmic paradigm was introduced by Spielman and Teng, who gave a local algorithm for finding subgraph with small conductance [18]. Building

[⋆] The author is partially supported by the Grand Project "Network Algorithms and Digital Information" of the Institute of Software, Chinese Academy of Sciences.

on their work, other local clustering algorithms with better approximation ratio and running time have been proposed by Anderson, Chung and Lang [3] and Anderson and Yepes [4]. Local algorithm for finding dense subgraphs have been studied by Anderson [2]. These algorithms have important applications in graph sparsificasion, solving linear equations [17], Laplacian algorithmic paradigm [19] and have also been used to handle real networks data (e.g., [14, 15]). However, to our knowledge, only a few number of problems are shown to have such local algorithms.

In this paper, we add one more problem to this list, and give a local algorithm for extracting *dense bipartite-like* subgraphs. Such a subgraph serves as a good channel for us to understand the link structures of the Web graph (that is, the nodes are the Web pages and a directed edge (i, j) represents a hyperlink from i to j). We are interested in extracting useful information from this huge graph, one of particular interest are the cyber-communities, which gives insights into the intellectual evolution of the Web and facilitates adverting at a more precise level [13]. As found by Kumar et al [13], the cyber-communities are characterized by dense bipartite subgraphs.

To measure the property of a group of Web pages being cyber-community like, that is, whether the group is close to a dense bipartite subgraph or not, we will adopt a concept *bipartiteness ratio* introduced by Trevisan [20]. Given a graph $G = (V, E)$, a subgraph S and one of its partitions $S = (L, R)$, the bipartiteness ratio $\beta(S_{L,R})$ of S under partition (L, R) is defined to be

$$\beta(S_{L,R}) = \frac{2e(L, L) + 2e(R, R) + e(S, V\backslash\bar{S})}{\mathrm{vol}(S)}.$$

The bipartiteness ratio $\beta(S)$ of the subgraph S is the minimum value of $\beta(S_{L,R})$ over all its possible partitions (L, R). Intuitively speaking, if $\beta(S)$ is small, then there must exist a partition (L, R) such that the number of edges from S to the outside as well as the number of edges that lie entirely in L or R is relatively small compared with all the edges involved with S. Thus, (L, R) can be seen close to a dense bipartite subgraph and S can be seen as a good Web community. Using the bipartiteness ratio as the measure of a set being dense and bipartite-like has the advantage that it unifies both properties in a natural way and admits theoretical analysis, which is difficult for many other measures.

We give a local algorithm for finding a subgraph with small bipartiteness ratio around a starting vertex v. In particular, we show that for any subgraph S with bipartiteness ratio θ and volume at most s, there exists a subset $S_\theta \subseteq S$ of large volume such that for any $v \in S_\theta$, our algorithm finds a subgraph of bipartiteness ratio $O(\sqrt{\theta})$ and runs in time $O(s^2(\log s + \varDelta))$, where $\varDelta$ is the maximum degree of the graph.

Our algorithm is composed of two parts. In the first part, the algorithm simulates the power method for the largest eigenvalue by a truncated process, which has already been used in previous local algorithms (eg.,[18, 2]). Such a process iteratively multiply a vector by some matrix of the graph (for example, the normalized Laplacian matrix in this paper), and during each iterative step, it only keeps a small fraction of non-zero elements of the vector by truncating elements of relatively small values. We will see that this process allows our algorithm to be "local" in that it will only traverse a small portion of the graph.

In the second part, the algorithm will sweep over all the vectors produced in the truncated process. Such a sweep operation is implied in a spectral algorithm for the bipartiteness ratio and is similar to the sweep operation that are widely used to find small conductance from the second smallest eigenvalue of the normalized Laplacian $\mathcal{L}$. The approximation guarantee of our algorithm is derived from this part and relies on the relation between the bipartiteness ratio and the largest eigenvalue of $\mathcal{L}$ given by Trevisan [20].

Other related works: Previous work on extracting dense bipartite subgraphs from the Web graph have used different measures and mainly focused on giving heuristic methods (e.g., [13, 1, 7, 6]). All of them did not give theoretical analysis on the performance of the corresponding algorithms on general graphs.

The definition of bipartiteness ratio is closely related to the notion of conductance and dense subgraphs. The conductance of a vertex subset S is defined as $\frac{e(S,V\setminus S)}{\min\{\mathrm{vol}(S),\mathrm{vol}(V\setminus S)\}}$. A set of small conductance can be thought of a good community as the connections crossing the set are relatively smaller than the total number of edges involved with the set. In particular, Kannan, Vempala and Veta gave a bicriteria measure of the quality of clustering based on the concept of conductance and analyzed a corresponding spectral algorithm [12]. For the literature on dense subgraphs, Kannan and Vinay defined $d(S,T) = \frac{e(S,T)}{\sqrt{|S||T|}}$ as the measure of the density of a subgraph induced on $S \cup T$ in a directed graph and gave a spectral algorithm for finding subgraphs with large density [11]. Other density measures are also extensively studied. For example, Goldberg [8] introduced the average degree as the density measure of a set S, that is, $d(S) = \frac{e(S,S)}{|S|}$. Though both conductance and the density provide us good measures to study the communities of the networks, they do not give us any information on the bipartiteness of these subgraphs, which is the main motivation of the paper.

As mentioned above, the measure we are using here was introduced by Trevisan, who found its deep connections with the Max Cut problem, the Cheeger inequality and the Geomans-Williamson Relaxation [20]. Soto [16] and Kale and Seshadhri [10] gave further analysis on the quantity that is related to the bipartiteness ratio. Both of their work are motivated by designing approximation algorithms for Max Cut.

In section 2, we give the basic definitions of the problem and some processes that will be used in our algorithm. Then we give our local algorithm and main theorem in Section 3. In section 4, we give the proof of our main theorem.

2 Preliminaries

Let $G = (V, E)$ be an undirected weighted graph. Let A denote the adjacency matrix of the graph such that $A_{u,v}$ is the weight of edge (u, v). We let d_v denote the (weighted) degree of vertex v. Let D denote the diagonal degree matrix of

G such that $D_{u,u} = d_u$ and $D_{u,v} = 0$ for $u \neq v$. Let $\mathcal{L} = I - D^{-1/2}AD^{-1/2}$ be the *normalized Laplacian* (or just Laplacian) of the graph. It is well known that $\mathcal{L}$ is a positive semi-definite (PSD, for short) matrix; that is, $\mathcal{L}$ is a real symmetric matrix and all its eigenvalues are non-negative. Let $\varDelta$ denote the maximum degree of G. We define the volume $\mathrm{vol}(S)$ of a subset S to be the sum of degrees of the vertices in S, that is, $\mathrm{vol}(S) = \sum_{v \in S} d_v$. For any two vertex sets L and R, let $e(L, R)$ denote the number of edges between L and R. We define $\mathbf{1}_v$ to be the indicator vector of vertex v. In the following, we will let S denote subgraphs induced on the vertex set S and also let $S = (L, R)$ denote the subgraphs induces on the $S = L \cup R$. For a vector x, we let $\|x\|$ denote its Euclid norm and let $\mathrm{supp}(x)$ denote the support of it (the set of vertices on which x is non-zero).

Definition 1. *For any subgraph S and a partition (L, R) of S, that is, $L \cup R = S$ and $L \cap R = \emptyset$, we define the bipartiteness ratio $\beta(S_{L,R})$ of S under partition (L, R) by*

$$\beta(S_{L,R}) = \frac{2e(L,L) + 2e(R,R) + e(S, V\backslash S)}{vol(S)}.$$

We define the bipartiteness ratio $\beta(S)$ of the subgraph S to be the minimum value of $\beta(S_{L,R})$ over all its possible partitions (L, R), that is

$$\beta(S) = \min_{(L,R)\, partion\ of\ S} \beta(S_{L,R});$$

and define the bipartiteness ratio of the graph $\beta(G)$ to be the minimum value of $\beta(S)$ over all induced subgraphs in G, that is

$$\beta(G) = \min_{S} \beta(S).$$

When it is clear, we will use $\beta(L, R)$ to denote $\beta(S_{L,R})$. Our algorithm for finding subgraphs with small bipartiteness ratio is based on the power method for the largest eigenvector of a matrix. This method is also the base of many other local algorithms. We start from a vector x and iteratively multiply the Laplacian $\mathcal{L}$. We will then make use of these vectors to find the subgraphs with good properties. To guarantee that our algorithm is local, instead of doing the dense matrix vector multiplication, in each step, we will only keep track of the set of vertices u whose value is larger then a certain threshold. We will use the following truncated process as defined in [2].

Definition 2.
1. *Given a vector x and a nonnegative real number ϵ, the truncated vector is*
$$[x]_\epsilon(u) = \begin{cases} x(u) \text{ } if\ |x(u)| \geq \epsilon\|x\| \\ 0 \quad\quad otherwise \end{cases}$$
2. *Given a vector $x = x_0$ and a set of real numbers $\epsilon_t \in [0, 1]$ for $t \leq T$, the truncated process with staring vector x_0 and parameters $\{\epsilon_t\}$ is defined to be the process that generates a sequence of vectors $x_0, \cdots, x_T$ such that $x_{t+1} = [x_t\mathcal{L}]_{\epsilon_{t+1}}$.*

Note that for a given vector x, since the absolute value of $[x]_\epsilon$ is at least $\epsilon\|x\|$ whenever it is nonzero, and $\|x\|^2 \geq \|[x]_\epsilon\|^2$, we have that the number of nonzero entries in $[x]_\epsilon$ is at most $1/\epsilon^2$. That is, $|\text{supp}([x]_\epsilon)| \leq 1/\epsilon^2$.

After we get a set of vectors $x_0, \cdots, x_T$ of the truncated process, we will perform the following *sweep* process to produce subgraphs for each x_t.

Definition 3. *Given a vector $x \in \mathbb{R}^V$ such that $|supp(x)| = s$, the sweep process over vector x is defined to be the following process:*

1. *Order the vertices so that $\frac{|x(v_1)|}{\sqrt{d_{v_1}}} \geq \frac{|x(v_2)|}{\sqrt{d_{v_2}}} \geq \cdots \geq \frac{|x(v_s)|}{\sqrt{d_{v_s}}}$.*
2. *For each $i \leq s$, define $L_i = \{v_j : x(v_j) > 0 \, and \, j \leq i\}$ and $R_i = \{v_j : x(v_j) \leq 0 \, and \, j \leq i\}$ and compute the bipartiteness ratio of the subgraph $S_i = (L_i, R_i)$.*
3. *Output the subgraph $S_m = (L_m, R_m)$ that achieves the minimum bipartiteness ratio among all the s subgraphs. Let $\beta(x) = \beta(L_m, R_m)$.*

3 Description of the algorithm and the main theorem

Now we describe our algorithm as follows.

```
FindDenseBipartite($v, s, \theta$)
Input: A vertex $v$, a target volume $s$ and a target bipartiteness ratio $\theta < 1/4$.
Output: A subgraph $(X, Y)$.

 1. Let $x_0 = \frac{\mathbf{1}_v}{\sqrt{d_v}}$, $T = \log_{2-4\theta}(8s)$, and $\epsilon_t = (2-4\theta)^{t/2}/\sqrt{8}s$.
 2. Compute $x_1, \cdots, x_T$ of the truncated process with starting vector $x_0$ and
    parameters $\epsilon_1, \cdots, \epsilon_T$.
 3. For each time $t \leq T$, sweep over $x_t$ and find the subgraph $(X_t, Y_t)$ such that
    $\beta(X_t, Y_t) = \beta(x_t)$. Output the subgraph with the smallest bipartiteness ratio
    among all such pairs.
```

Our main theorem about the algorithm is the following.

Theorem 1. *If $S = (L, U)$ is a subgraph with bipartiteness ratio $\beta(S_{L,U}) \leq \theta$, then there exists a subset $S_\theta \subseteq S$ such that*

1. *$vol(S_\theta) \geq vol(S)/9$,*
2. *for any $v \in S_\theta$, and $s \geq vol(S)$, the algorithm* `FindDenseBipartite`*(v, s, θ) outputs a subset (X, Y) satisfying that $\beta(X, Y) \leq 2\sqrt{2\theta}$.*

Remark: we can make the bound condition on the bipartiteness ratio $\theta < 1/4$ be $\theta < 1-\delta$, for any constant δ smaller than 1, just with a different (constant fraction) bound on $\text{vol}(S_\theta)$.

The proof of Theorem 1 is given in Section 4. Roughly speaking, we will first give the spectral algorithm of Trevisan [20] for the bipartiteness ratio of graph G. We give an alternative proof of the approximation performance of this algorithm and show that under certain conditions, a vector can be used to find subgraphs with small bipartiteness ratio. Then we will show that there exists a large subset of "good" starting vertices so that the truncated process from a scaled indicator vector of such a vertex will produce a vector that satisfies the conditions of the former spectral algorithm, and thus finish the proof.

In the remaining of this section, we bound the running time of our algorithm.

Theorem 2. *The running time of* `FindDenseBipartite`(v, s, θ) *is* $O(s^2(\Delta + \log s))$.

Proof. We note that in each step t of the truncated process, the number of vertices in the support $\mathrm{supp}(x_t)$ of x_t is at most $1/\epsilon_t^2$. The running time of computing $x_t\mathcal{L}$ is bounded by the volume of the degrees of the vertices in $\mathrm{supp}(x_t)$, which is at most $O(\Delta/\epsilon_t^2) = O(\Delta s^2(2-4\theta)^{-t})$.

The running time of the whole truncated process is thus $\sum_{t=0}^{T} O(\Delta s^2(2-4\theta)^{-t}) = O(\Delta s^2)$.

Finally, the computation of the sweep process might require sorting the vectors in x_t, which could take time $O(|\mathrm{supp}(x_t)| \log |\mathrm{supp}(x_t)|) = O(s^2 \log s(2-4\theta)^{-t})$. Thus, the running time of the whole sweep process is $\sum_{t=0}^{T} O(s^2 \log s(2-4\theta)^{-t}) = O(s^2 \log s)$.

Thus, the running time of `FindDenseBipartite` is bounded by $O(s^2(\Delta + \log s))$.

4 Analysis of the local algorithm

4.1 A spectral algorithm for finding subgraphs with small bipartiteness ratio

In this section, we show that under certain conditions on a vector x, the sweep over x will produce a good subgraph with low bipartiteness ratio, which is proved by Trevisan [20], and further analyzed by Soto [16] and Kale and Seshadhri [10]. The result is given in the following Lemma 1. Here, we give a self-contained proof that is somewhat different from the previous proofs. In fact, former proofs of the lemma all proceed by designing and analyzing a probabilistic algorithm. Instead, we prove the lemma by directly analyzing the deterministic version of the algorithm, which provides us more insight on the combinatorial property of the bipartiteness ratio and may be of independent interest.

Lemma 1. *For any graph G and $\theta < 1/4$, if there exists a vector $x \in \mathbb{R}^V$ such that $x\mathcal{L}x^T \geq (2-4\theta)\|x\|^2$, then the sweep over x produces a subgraph (X, Y) with bipartiteness ratio $\beta(X, Y) \leq 2\sqrt{2\theta}$.*

Proof. Let $z = xD^{-1/2}$. Let $u \sim v$ denote that $(u, v) \in E$ and let $\bar{S}$ denote $V\backslash S$. By the condition of the lemma, we have that $x(2I - \mathcal{L})x^T \leq 4\theta\|x\|^2$ and thus that

$$\begin{aligned}
4\theta &\ge \frac{x(I+D^{-1/2}AD^{-1/2})x^T}{\|x\|^2}\\
&= \frac{z(D+A)z^T}{\langle z, zD\rangle}\\
&= \frac{\sum_{u\sim v}(z(u)+z(v))^2}{\sum_{v\in V} z^2(v)d_v}\\
&= \frac{\sum_{u\sim v}(z(u)+z(v))^2 \sum_{u\sim v}(|z(u)|+|z(v)|)^2}{\sum_{v\in V} z^2(v)d_v \sum_{u\sim v}(|z(u)|+|z(v)|)^2}\\
&\ge \frac{(\sum_{u\sim v}|z(u)+z(v)|(|z(u)|+|z(v)|))^2}{2(\sum_{v\in V} z^2(v)d_v)^2}, \qquad (1)
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality.

Assume the support of x has size s. We perform a sweep over x so that $\frac{|x(v_1)|}{\sqrt{d_{v_1}}} \ge \frac{|x(v_2)|}{\sqrt{d_{v_2}}} \ge \cdots \ge \frac{|x(v_s)|}{\sqrt{d_{v_s}}}$. Equivalently, we have $|z(v_1)| \ge |z(v_2)| \ge \cdots \ge |z(v_s)|$.

Let $L_i = \{v_j : j \le i, z(v_j) > 0\}, R_i = \{v_j : j \le i, z(v_j) \le 0\}$ and $S_i = L_i \cup R_i$. Recall that $\beta(x) = \min_i \beta(L_i, R_i)$. Then we have for any i, $\beta(x)\text{vol}(S_i) \le 2e(L_i, L_i) + 2e(R_i, R_i) + e(S_i, \bar{S}_i)$.

Now we consider the square root of the numerator of (1) to obtain

$$\begin{aligned}
&\sum_{u\sim v}|z(u)+z(v)|(|z(u)|+|z(v)|)\\
&\ge \sum_{u\sim v, z(u)z(v)<0}|z^2(u)-z^2(v)| + \sum_{u\sim v, z(u)z(v)\ge 0}(z(u)+z(v))^2\\
&\ge \sum_{\substack{i<j, v_i\sim v_j,\\ z(v_i)z(v_j)<0}}(z^2(v_i)-z^2(v_j)) + \sum_{\substack{i<j, v_i\sim v_j,\\ z(v_i)z(v_j)\ge 0}}(z^2(v_i)+z^2(v_j)) \qquad (2)\\
&= \sum_{i=1}^{s}(z^2(v_i)-z^2(v_{i+1}))(2e(L_i,L_i)+2e(R_i,R_i)+e(S_i,\bar{S}_i)) \qquad (3)\\
&\ge \beta(x)\sum_{i=1}^{s}(z^2(v_i)-z^2(v_{i+1}))\text{vol}(S_i)\\
&= \beta(x)\sum_{i=1}^{s} z^2(v_i)d_{v_i},
\end{aligned}$$

where we define $z(v_{n+1})$ to be 0 if $s = n$. The main difficulty lies in the third equation, which can be obtained by comparing the coefficient of $z^2(v_k)$ on both sides for every $k \le n$ and we defer the proof of it at the end. Now from the above calculations, we have that $4\theta \ge \frac{\beta(x)^2(\sum_{v\in V} z^2(v)d_v)^2}{2(\sum_{v\in V} z^2(v)d_v)^2} = \frac{\beta(x)^2}{2}$, and the lemma follows if we set $(X, Y) = (L_m, R_m)$ for which the bipartiteness ratio achieves $\beta(x)$.

Now we show that formula (2) is equivalent to formula (3). Let $\mathrm{coef}_1(k)$ and $\mathrm{coef}_2(k)$ be the coefficient of $z^2(v_k)$ in (2) and (3), respectively. We only need to show that for each $k \leq n$, $\mathrm{coef}_1(k) = \mathrm{coef}_2(k)$. Assume that $z(v_k) \leq 0$. The case when $z(v_k) > 0$ is similar.

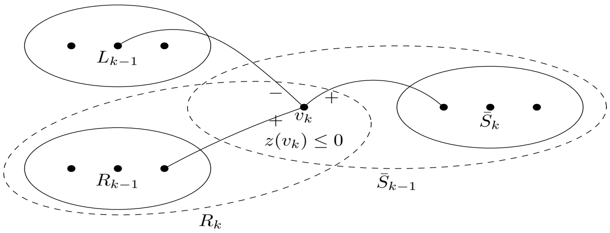

Fig. 1. The case when $z(v_k) \leq 0$. The sign on an edge denotes whether it contributes 1 or -1 to the coefficient $\mathrm{coef}_1(k)$ of $z^2(v_k)$ in (2)

By definition and our assumption that $z(v_k) \leq 0$, we know that $L_{k-1} = L_k$ and $R_k = R_{k-1} \cup \{v_k\}$ (see Figure 1). It is easy to see that only edges incident to vertex v_k can contribute to $\mathrm{coef}_1(k)$. More specifically, for each edge $u \sim v_k$, if $u \in R_{k-1} \cup \bar{S}_k$, it contributes 1 to $\mathrm{coef}_1(k)$ and if $u \in L_{k-1}$, it contributes -1 to $\mathrm{coef}_1(k)$. Totally, we have $\mathrm{coef}_1(k) = e(\{v_k\}, R_{k-1}) + e(\{v_k\}, \bar{S}_k) - e(\{v_k\}, L_{k-1})$.

On the other hand, from (3), we can get that

$$\begin{aligned}
\mathrm{coef}_2(k) &= (2e(L_k, L_k) + 2e(R_k, R_k) + e(S_k, \bar{S}_k)) \\
&\quad -(2e(L_{k-1}, L_{k-1}) + 2e(R_{k-1}, R_{k-1}) + e(S_{k-1}, \bar{S}_{k-1})) \\
&= 2e(\{v_k\}, R_{k-1}) + e(S_{k-1}, \bar{S}_k) + e(\{v_k\}, \bar{S}_k) \\
&\quad -e(S_{k-1}, \{v_k\}) - e(S_{k-1}, \bar{S}_k) \\
&= 2e(\{v_k\}, R_{k-1}) + e(\{v_k\}, \bar{S}_k) - e(S_{k-1}, \{v_k\}) \\
&= e(\{v_k\}, R_{k-1}) + e(\{v_k\}, \bar{S}_k) - e(\{v_k\}, L_{k-1}) \\
&= \mathrm{coef}_1(k).
\end{aligned}$$

This completes the proof.

4.2 The abundance of good starting vertices

We now show that for any given subgraph $S = (L, R)$ with small bipartiteness ratio θ, there exists a large subset $S_\theta \subseteq S$ of "good" vertices, such that for any $v \in S_\theta$, the truncated process with starting vector $\mathbf{1}_v/\sqrt{d_v}$ produces a vector $x \in \mathbb{R}^V$ that satisfies the condition of Lemma 1.

We will consider the normalized Laplacian $\mathcal{L}_S$ of the subgraph S. Here, we extend the dimension of $\mathcal{L}_S$ to $|V|$ by adding the corresponding zero entries. Note that $\mathcal{L}_S$ is a submatrix of $\mathcal{L} = \mathcal{L}_G$ restricted on the vertex set S. In particular, $\mathcal{L} - \mathcal{L}_S$ is still positive semidefinite. So $x\mathcal{L}x^T \geq x\mathcal{L}_S x^T$ holds for all $x \in \mathbb{R}^V$. Let

$l = |S|$ and $2 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_l = \lambda_{l+1} = \cdots = \lambda_n = 0$ be the eigenvalues of $\mathcal{L}_S$ and let $\mu_1, \mu_2, \cdots, \mu_l, \mu_{l+1}, \cdots, \mu_n$ be the corresponding orthonormal left eigenvectors. By the definition of $\mathcal{L}_S$, we can assume that for all i such that $i \leq l$, the support of μ_i's are contained in S, and for all i such that $l < i \leq n$, the support of μ_i's are contained in $V - S$. It is easy to see that for $i \leq l$, the vectors obtained by restricting μ_i's on S form an orthonormal basis of $\mathbb{R}^S$. Any vector $x \in \mathbb{R}^V$ can be expressed in terms of the eigenvectors of $\mathcal{L}_S$ so that $x = \sum_{k=1}^{n} \alpha_k \mu_k$. For an integer m such that $1 \leq m \leq n$, define the m-norm $\|x\|_m$ of x to be the length of the projection of x onto the subspace spanned by the first m eigenvectors; that is, $\|x\|_m = \sqrt{\sum_{k \leq m} \alpha_k^2}$. It is well known that $\|x\|_m$ is a norm [5]. (Also note that $\|x\| = \|x\|_n$.) For any nonnegative number $\epsilon < 1$, let $h_\epsilon = \max\{k : \lambda_k \geq 2 - \epsilon\}$. Note that $h_\epsilon \leq l$. So, $\|x\|_{h_\epsilon} \leq \|x\|_l$.

To show there are many "good" vertices, we show in the following lemma that there is a large subset S_θ of vertices whose h_ϵ-norm is large, for any $\epsilon \geq \theta$. We will see in Section 4 that the algorithm starting from a vertex in S_θ is guaranteed to produce a vector satisfying the condition of Lemma 1 in $T = O(\log s)$ steps. In the following, when the value ϵ is clear, we will abbreviate h_ϵ as h.

Lemma 2. *Let $\epsilon = 4\theta$. If $S = (L, R)$ is a subgraph with bipartiteness ratio $\beta(L, R) \leq \theta$, then there exists a subset $S_\theta \subseteq S$ satisfying that*

1. *vol$(S_\theta) \geq$ vol$(S)/9$, and*
2. *for any $v \in S_\theta$, $\|\mathbf{1}_v/\sqrt{d_v}\|_h \geq 1/\sqrt{8vol(S)}$.*

Proof. Define a vector ψ as follows: $\psi(v) = \sqrt{d_v}/\mathrm{vol}(S)$ if $v \in L$, $\psi(v) = -\sqrt{d_v}/\mathrm{vol}(S)$ if $v \in R$ and $\psi(v) = 0$ otherwise. Then, we have

$$\begin{aligned}\psi(2I - \mathcal{L})\psi^T = \psi(I + D^{-1/2}AD^{-1/2})\psi^T &= \sum_{u \sim v} (\psi(u)/\sqrt{d_u} + \psi(v)/\sqrt{d_v})^2 \\ &= \frac{4e(L, L) + 4e(R, R) + e(S, V \backslash S)}{\mathrm{vol}(S)^2} \\ &\leq 2\theta/\mathrm{vol}(S)\end{aligned}$$

Now let $\psi = \sum_{k=1}^{n} \alpha_i \mu_i$, then $\|\psi\|^2 = 1/\mathrm{vol}(S) = \sum_{k=1}^{n} \alpha_i^2$, and

$$\begin{aligned}\psi(2I - \mathcal{L})\psi^T = \frac{2}{\mathrm{vol}(S)} - \sum_{k=1}^{n} \lambda_k \alpha_k^2 &\geq \frac{2}{\mathrm{vol}(S)} - 2\sum_{i \leq h} \alpha_k^2 - (2 - \epsilon)\sum_{k > h} \alpha_k^2 \\ &= \frac{2}{\mathrm{vol}(S)} - 2\|\psi\|_h^2 - (2 - \epsilon)(\frac{1}{\mathrm{vol}(S)} - \|\psi\|_h^2).\end{aligned}$$

From the above bounds, we can get $\|\psi\|_h^2 \geq \frac{\epsilon - 2\theta}{\epsilon \mathrm{vol}(S)} = \frac{1}{2\mathrm{vol}(S)}$.

We now define $T = \{v \in S : \|\mathbf{1}_v/\sqrt{d_v}\|^2 < \frac{1}{8\mathrm{vol}(S)}\}$. Assume that $\mathrm{vol}(T) \geq 8\mathrm{vol}(S)/9$, we will derive a contradiction.

Define a vector η as follows: $\eta(v) = \sqrt{d_v}/\mathrm{vol}(T)$ if $v \in L \cap T$, $\eta(v) = -\sqrt{d_v}/\mathrm{vol}(T)$ if $v \in R \cap T$ and $\eta(v) = 0$ otherwise. Then, we have

$$\|\eta\|_h^2 = \|\sum_{v \in T} \frac{d_v}{\mathrm{vol}(T)} \cdot \frac{\mathbf{1}_v}{\sqrt{d_v}}\|_h^2 \leq \sum_{v \in T} \frac{d_v}{\mathrm{vol}(T)} \cdot \|\frac{\mathbf{1}_v}{\sqrt{d_v}}\|_h^2 < \frac{1}{8\mathrm{vol}(S)},$$

where the second inequality follows by the Jensen's inequality.

To get a lower bound of $\|\eta\|_h$, we first get an upper bound of $\|\eta - \psi\|_h$.

$$\|\psi - \eta\|_h^2 \leq \|\psi - \eta\|^2 = \sum_{v\in T}(\frac{\sqrt{d_v}}{\mathrm{vol}(T)} - \frac{\sqrt{d_v}}{\mathrm{vol}(S)})^2 + \sum_{v\in S\setminus T}(\frac{\sqrt{d_v}}{\mathrm{vol}(S)})^2 = \frac{1}{\mathrm{vol}(T)} - \frac{1}{\mathrm{vol}(S)}$$
$$\leq \frac{1}{8\mathrm{vol}(S)},$$

where the last inequality follows from our assumption on $\mathrm{vol}(T)$. By the triangle inequality, we have

$$\|\eta\|_h \geq \|\psi\|_h - \|\psi - \eta\|_h \geq \sqrt{\frac{1}{2\mathrm{vol}(S)}} - \sqrt{\frac{1}{8\mathrm{vol}(S)}} = \sqrt{\frac{1}{8\mathrm{vol}(S)}},$$

which contradicts the upper bound we obtained for $\|\eta\|_h$. Therefore, we must have $\mathrm{vol}(T) \leq 8\mathrm{vol}(S)/9$. Let $S_\theta = S\setminus T$. We have $\mathrm{vol}(S_\theta) \geq \mathrm{vol}(S)/9$ and for any $v \in S_\theta$, $\|\mathbf{1}_v/\sqrt{d_v}\|_h^2 \geq \frac{1}{8\mathrm{vol}(S)}$.

This complete the proof of the lemma.

4.3 Proof of Theorem 1

Proof. For any $S = (L, R)$ with bipartiteness ratio $\beta(L, R) \leq \theta < 1/4$, we will consider its extended Laplacian matrix $\mathcal{L}_S$ and the corresponding h-norm of vectors determined by $\mathcal{L}_S$, where h is as defined in Subsection 4.2. Let S_θ be the subset as described in Lemma 2. We show that for any s and a vertex $v \in S_\theta$, the algorithm `FindDenseBipartite`(v, s, θ) produces a subgraph (X, Y) such that $\beta(X, Y) \leq 2\sqrt{2\theta}$, where we have assumed that $s \geq \mathrm{vol}(S)$.

Let $x_1, \cdots, x_T$ be the vectors produced by the pruned process with starting vector $\mathbf{1}_v/\sqrt{d_v}$ and parameters $\epsilon_1, \cdots, \epsilon_T$. If there exists a vector x_t such that $x_t\mathcal{L}x_t^T \geq (2 - 4\theta)\|x_t\|^2$, then by Lemma 1, we are done. Now assume that there is no such vector, that is, for all $t \leq T$, $x_t\mathcal{L}x_t^T < (2 - 4\theta)\|x_t\|^2$. We will derive a contradiction.

First, note that by the definition of Laplacian matrix $\mathcal{L}$, we have that for any $k \geq 0$, $\mathcal{L}^k - \mathcal{L}^{k+1}$ is PSD. This follows from the fact that $\mathcal{L}^k - \mathcal{L}^{k+1} = \mathcal{L}^k \cdot D^{-1/2}AD^{-1/2}$, that $D^{-1/2}AD^{-1/2}$ is a PSD matrix, and that the multiplication of two commutable PSD matrices is also PSD [9]. Therefore, by our assumption, for any $t \leq T$, we have

$$\begin{aligned}\|x_t\|^2 \leq \|x_{t-1}\mathcal{L}\|^2 = x_{t-1}\mathcal{L}^2x_{t-1}^T \leq x_{t-1}\mathcal{L}x_{t-1}^T &\leq (2 - 4\theta)\|x_{t-1}\|^2\\ &\leq (2 - 4\theta)^t\|x_0\|^2\\ &\leq (2 - 4\theta)^t. \qquad (4)\end{aligned}$$

Now we show that $\|x_t\|_h$ increase exponentially with t such that

$$\|x_t\|_h \geq \frac{1}{\sqrt{8\mathrm{vol}(S)}}(2 - 4\theta)^t. \qquad (5)$$

By $\|x_t\| \geq \|x_t\|_h$, this contradicts equation (4) when $t = \log_{2-4\theta}(8\mathrm{vol}(S)) \leq T$, and this completes the proof.

We will prove equation (5) by induction. When $t = 0$, it is true by the choice of v. Now let $r_t = x_{t-1}\mathcal{L} - x_t = x_{t-1}\mathcal{L} - [x_{t-1}\mathcal{L}]_{\epsilon_t}$ be the vector that is removed from the pruned process for any t and let r'_t be the vector that is equal to r_t on S, and zero elsewhere. Now recall that μ_i's are the eigenvectors of $\mathcal{L}_S$, that for all $i \leq l = |S|$, the support of μ_i is contained in S and the vectors obtained by restricting μ_i's on S form an orthonormal basis of $\mathbb{R}^S$. We have,

$$\begin{aligned}\|r_t\|_h \leq \|r_t\|_l = \sqrt{\sum_{i\leq l} |\langle r_t, \mu_i\rangle|^2} = \sqrt{\sum_{i\leq l} |\langle r'_t, \mu_i\rangle|^2} &= \|r'_t\| \\ &\leq \epsilon_t \|x_{t-1}\mathcal{L}\| \sqrt{|S|} \\ &\leq \epsilon_t \sqrt{\mathrm{vol}(S)}(2-4\theta)^{t/2},\end{aligned}$$

where the last inequality follows from inequality (4) and that $|S| \leq \mathrm{vol}(S)$.

Also note that for any $x = \sum_{k=1}^n \alpha_k \mu_k$, we have

$$\|x\mathcal{L}\|_h = \|x\mathcal{L}_S\|_h = \sqrt{\sum_{k\leq h} \lambda_k^2 \alpha_k^2} \geq (2-4\theta)\sqrt{\sum_{k\leq h} \alpha_k^2} = (2-4\theta)\|x\|_h,$$

where the first equation follows from the fact that $\|x\mathcal{L}\|_h = \sqrt{\sum_{i\leq h} |\langle x\mathcal{L}, \mu_i\rangle|^2}$; and that $\langle x\mathcal{L}, \mu_i\rangle = \langle x(\mathcal{L} - \mathcal{L}_S) + x\mathcal{L}_S, \mu_i\rangle = \langle x\mathcal{L}_S, \mu_i\rangle$, since the support of μ_i is contained in S and the corresponding elements in the matrix $\mathcal{L} - \mathcal{L}_S$ are all zero.

Now assume that the induction hypothesis holds for $t-1$, that is, $\|x_{t-1}\|_h \geq \frac{1}{\sqrt{8\mathrm{vol}(S)}}(2-4\theta)^{t-1}$. Then

$$\begin{aligned}\|x_t\|_h &= \|x_{t-1}\mathcal{L} - r_t\|_h \\ &\geq \|x_{t-1}\mathcal{L}\|_h - \|r_t\|_h \\ &\geq (2-4\theta)\|x_{t-1}\|_h - \epsilon_t\sqrt{\mathrm{vol}(S)}(2-4\theta)^{t/2} \\ &\geq (2-4\theta)\frac{1}{\sqrt{8\mathrm{vol}(S)}}(2-4\theta)^{t-1} - \epsilon_t\sqrt{\mathrm{vol}(S)}(2-4\theta)^{t/2} \\ &\geq \frac{1}{\sqrt{8\mathrm{vol}(S)}}(2-4\theta)^t,\end{aligned}$$

where the last inequality follows because $\epsilon_t = (2-4\theta)^{t/2}/\sqrt{8}s$ and that $s \geq \mathrm{vol}(S)$. This completes the proof.

References

[1] Abello, J., Resende, M., Sudarsky, S.: Massive quasi-clique detection. LATIN 2002: Theoretical Informatics pp. 598–612 (2002)

[2] Andersen, R.: A local algorithm for finding dense subgraphs. ACM Trans. Algorithms 6, 60:1–60:12 (2010)

[3] Andersen, R., Chung, F., Lang, K.: Local graph partitioning using pagerank vectors. In: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science. pp. 475–486. IEEE Computer Society, Washington, DC, USA (2006)

[4] Andersen, R., Peres, Y.: Finding sparse cuts locally using evolving sets. In: Proceedings of the 41st annual ACM symposium on Theory of computing. pp. 235–244. STOC '09, ACM, New York, NY, USA (2009)

[5] Bhatia, R.: Matrix Analysis. Springer-Verlag, New York (1997)

[6] Dourisboure, Y., Geraci, F., Pellegrini, M.: Extraction and classification of dense implicit communities in the web graph. ACM Transactions on the Web (TWEB) 3(2), 7 (2009)

[7] Gibson, D., Kumar, R., Tomkins, A.: Discovering large dense subgraphs in massive graphs. In: Proceedings of the 31st international conference on Very large data bases. pp. 721–732. VLDB Endowment (2005)

[8] Goldberg, A.V.: Finding a maximum density subgraph. Tech. rep., Berkeley, CA, USA (1984)

[9] Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press (1985)

[10] Kale, S., Seshadhri, C.: Combinatorial approximation algorithms for maxcut using random walks. In 2nd Symposium on Innovations in Computer Science (ICS 2011) (2011)

[11] Kannan, R., Vinay, V.: Analyzing the structure of large graphs. Unpublished manuscript (1999), http://research.microsoft.com/en-us/um/people/kannan/papers/webgraph.pdf

[12] Kannan, R., Vempala, S., Vetta, A.: On clusterings: Good, bad and spectral. J. ACM 51(3), 497–515 (2004)

[13] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the web for emerging cyber-communities. In: Proceedings of the eighth international conference on World Wide Web. pp. 1481–1493. WWW '99, Elsevier North-Holland, Inc., New York, NY, USA (1999)

[14] Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics, 6(1), 29-123 (2009)

[15] Liao, C.S., Lu, K., Baym, M., Singh, R., Berger, B.: IsoRankN: spectral methods for global alignment of multiple protein networks. Bioinformatics 25(12), p253–p258 (2009)

[16] Soto, J.: Improved analysis of a max cut algorithm based on spectral partitioning. CoRR abs/0910.0504 (2009)

[17] Spielman, D.A.: Algorithms, graph theory, and linear equations. Proceedings of the International Congress of Mathematicians 2010, 2698–2722 (2010)

[18] Spielman, D.A., Teng, S.H.: Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing. pp. 81–90. STOC '04, ACM, New York, NY, USA (2004)

[19] Teng, S.H.: The laplacian paradigm: Emerging algorithms for massive graphs. In: Theory and Applications of Models of Computation, Lecture Notes in Computer Science, vol. 6108, pp. 2–14. Springer Berlin / Heidelberg (2010)

[20] Trevisan, L.: Max cut and the smallest eigenvalue. In: Proceedings of the 41st annual ACM symposium on Theory of computing. pp. 263–272. STOC '09, ACM, New York, NY, USA (2009)