Clustering

November 24, 2017

1 Definitions

Let $G = (V, E)$ be a d-bounded degree graph and let $|V| = n$. Our algorithm is based on running lazy random walks on G. Let us first formally define the random walks that we use. Given we are currently at vertex u, in the next step we move to vertex v with probability $\frac{1}{2d}$ and we stay at u with the remaining probability. Let $\mathbf{p}_v^\ell$ denote the probability distribution of endpoints of such random walk of length ℓ starting at v.

Let A be the adjacency matrix of G and D be a diagonal matrix where D_{ii} is the degree of vertex i. Let $M = \frac{I+D^{\frac{-1}{2}}AD^{\frac{-1}{2}}}{2}$ denote the lazy random walk transition matrix and $\mathcal{L} = I - D^{\frac{-1}{2}}AD^{\frac{-1}{2}}$ be the normalized Laplacian matrix. We set $0 = \lambda_1 \leq \lambda_2 \leq \ldots \lambda_n \leq 2$ to be the eigenvalues of $\mathcal{L}$ and $D^{\frac{1}{2}}\mathbf{1}_V = v_1, , v_2, \ldots, v_n$ to be their corresponding orthonormal eigenvectors respectively. Let $1 = \eta_1 \geq \eta_2 \geq \ldots \geq \eta_n \geq 0$ denote the eigenvalues of M , then it is easy to see that for each $1 \leq i \leq n$, $\eta_i = 1 - \frac{\lambda_i}{2}$ and v_i is its corresponding eigenvector.

Throughout this paper by $x(i)$ denote the i th coordinate of vector x. Let $S \subseteq V$ be the subset of vertices. $\mathbf{1}_S$ denote the indicator vector of S such that $\mathbf{1}_S(v) = 1$ if $v \in S$ and $1_S(v) = 0$ otherwise. We let $\mathbf{1}_v = \mathbf{1}_{\{v\}}$, thus, $\mathbf{p}_v^\ell = \mathbf{1}_v W^\ell$.

Let $S \subseteq V$, the conductance of S is defined as $\phi_G(S) = \frac{e(S,V \setminus S)}{d|S|}$, where $e(S, V \setminus S)$ denotes the number of edges coming out of S. The conductance of the graph G, is defined as $\phi(G) = \min_{S \subseteq V, |S| \leq \frac{|V|}{2}} \phi_G(S)$. For any $S \subseteq V$, the inner conductance of S is defined as the conductance of the induced subgraph of G on the vertex set S. We refer to that by $\phi(G[S])$.

2 The algorithm

Our testing algorithm is given as follows.

Algorithm 1 ***k*-ClusterTest(G, ϕ, k)**

```
procedure $k$-ClusterTest$(G, \phi, k)$
    Let $S \in \mathbb{R}^{n \times (k+1)}$ be a sample matrix such that for any $1 \leq i \leq k+1$ the $i$th column of $S$ is $\mathbf{1}_v$ where $v$ is sampled independently and uniformley at random from $V$.
    Let $\mu_{k+1}$ be the $(k+1)$th largest eigenvalue of $S^T M^{2t} S$.
    if $\mu_{k+1} < n^{-20}$ then
        return Accept
    else
        return Reject
```

3 Completeness: accepting k clustreable graphs

4 Soundness: rejecting graphs ϵ-far from k clusterable