Information Processing Letters 114 (2014) 349–352

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

On a connection between small set expansions and modularity clustering

Bhaskar DasGupta [a,*,1], Devendra Desai [b]

[a] Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, United States
[b] Department of Computer Science, Rutgers University, Piscataway, NJ 08854, United States

ARTICLE INFO

Article history:
Received 30 May 2013
Received in revised form 22 October 2013
Accepted 11 February 2014
Available online 18 February 2014
Communicated by Tsan-sheng Hsu

Keywords:
Theory of computation
Small-set expansion
Modularity clustering
Social network

ABSTRACT

In this paper we explore a connection between two seemingly different problems from two different domains: the *small-set expansion* problem studied in unique games conjecture, and a popular community finding approach for social networks known as the *modularity clustering* approach. We show that a sub-exponential time algorithm for the small-set expansion problem leads to a sub-exponential time constant factor approximation for some hard input instances of the modularity clustering problem.

1. Introduction and definitions

All graphs considered in this note are *undirected* and *unweighted*.[2] Let $G=(V,E)$ denote the given input graph with $n=|V|$ nodes and $m=|E|$ edges, let d_v denote the degree of a node $v \in V$, and let $A(G)=[a_{u,v}(G)]$ denote the adjacency matrix of G, i.e.,

$$a_{u,v}(G)=\begin{cases}1, & \text{if } \{u,v\}\in E\\ 0, & \text{otherwise.}\end{cases}$$

Since our result spans over two distinct research areas, we summarize the relevant definitions from both research fields [1,6] below for convenience.

(a) By a "set of (k) *communities*" we mean a partition of the set of nodes V into (k) non-empty parts.
(b) If G is d-regular for some given d, then its symmetric *stochastic walk* matrix is denoted by $\widehat{A}(G)$, and is defined as the $n\times n$ real symmetric matrix $\widehat{A}(G)=[\frac{a_{u,v}(G)}{d}]$.
(c) For a real number $\tau\in[0,1)$, the τ-*threshold rank* of G, denoted by $\mathrm{rank}_\tau(G)$, is the number of eigenvalues λ of $\widehat{A}(G)$ satisfying $|\lambda|>\tau$.
(d) For a subset $\emptyset\subset S\subset V$ of nodes, the following quantities are defined:
   - The (normalized) *measure* of S is $\mu(S)=\frac{|S|}{n}$.
   - The (normalized) *expansion* of S is
   $$\Phi(S)=\frac{|\{\{u,v\}\mid u\in S, v\notin S, \{u,v\}\in E\}|}{\sum_{v\in S} d_v}$$
   - The (normalized) *density* of S is $\mathsf{D}(S)=1-\Phi(S)$.
   - The *modularity* value of S is
   $$\mathsf{M}(S)=\frac{1}{2m}\left(\sum_{u,v\in S}\left(a_{u,v}-\frac{d_u d_v}{2m}\right)\right)$$
(e) The modularity of a set of communities $\mathbf{S}$ is $\mathsf{M}(\mathbf{S})=\sum_{S\in\mathbf{S}}\mathsf{M}(S)$.

* Corresponding author.
E-mail addresses: dasgupta@cs.uic.edu (B. DasGupta), devdesai@cs.rutgers.edu (D. Desai).
URLs: http://www.cs.uic.edu/~dasgupta (B. DasGupta), http://paul.rutgers.edu/~devdesai (D. Desai).
1 Partially supported by NSF grant IIS-1160995.
2 Our result can be extended for the more general case of directed weighted graphs by using the correspondence of these versions with unweighted undirected graphs as outlined in [4, Section 5.1].

http://dx.doi.org/10.1016/j.ipl.2014.02.004
0020-0190/© 2014 Elsevier B.V. All rights reserved.

(f) The goal of the *modularity k-clustering* problem on an input graph G is to find a set of at most k communities $\mathbf{S}$ that *maximizes* $\mathsf{M}(\mathbf{S})$. Let $\mathsf{OPT}_k(G) = \max_{\mathbf{S} \text{ is a set of at most } k \text{ communities}}\{\mathsf{M}(\mathbf{S})\}$ denote the optimal modularity value for a modularity k-clustering; it is easy to verify that $0 \leqslant \mathsf{OPT}_k(G) < 1$.
(g) The goal of the *modularity clustering* problem on G is to find a set of (unspecified number of) communities $\mathbf{S}$ that *maximizes* $\mathsf{M}(\mathbf{S})$. Let $\mathsf{OPT}(G)$ denote the optimal modularity value for a modularity clustering; obviously, $\mathsf{OPT}(G) = \mathsf{OPT}_n(G)$.
(h) $\exp(\xi)$ denotes $2^{c\xi}$ for some constant $c > 0$ that is independent of ξ.

The modularity clustering problems as described above is *extremely popular* in practice in their applications to biological networks [8,9] as well as to social networks [5–7]. For relevant computational complexity results for modularity maximization, see [2,4]. The following results from [4] demonstrate the computational hardness of $\mathsf{OPT}_2(G)$ and $\mathsf{OPT}(G)$ even if G is a regular graph.

Theorem 1.1. *(See [4].)*

(a) *For every constant $d \geqslant 9$, there exists a collection of d-regular graphs G of n nodes such that it is* NP*-hard to decide if $\mathsf{OPT}_2(G) \geqslant \frac{1}{2} - \frac{2c}{dn}$ or if $\mathsf{OPT}_2(G) \leqslant \frac{1}{2} - \frac{2c+2}{dn}$ for some positive $c = O(\sqrt{n})$.*
(b) *There exists a collection of $(n-3)$-regular graphs G of n nodes such that it is* NP*-hard to decide if $\mathsf{OPT}(G) > \frac{0.9388}{n-4}$ or if $\mathsf{OPT}(G) < \frac{0.9382}{n-4}$.*

2. Our result

Theorem 2.1. *Let G be a d-regular graph. Then, for some constant $0 < \varepsilon < \frac{1}{2}$, there is an algorithm $\mathcal{A}_\varepsilon$ with the following properties:*

- *$\mathcal{A}_\varepsilon$ runs in sub-exponential time, i.e., in time $\exp(\delta n)$ for some constant $0 < \delta = \delta(\varepsilon) < 1$ that depends on ε only.*
- *$\mathcal{A}_\varepsilon$ correctly distinguishes instances G of modularity clustering with $\mathsf{OPT}(G) \geqslant 1 - \varepsilon$ from instances G with $\mathsf{OPT}(G) \leqslant \varepsilon$.*

(*Note that we make no claim if $\varepsilon < \mathsf{OPT}(G) < 1 - \varepsilon$.*)

Remark 2.2 *(Usability of the approximation algorithm in Theorem 2.1).* We prove Theorem 2.1 for $\varepsilon = 10^{-6}$. It is natural to ask if there are in fact infinite families of d-regular graphs G that satisfy $\mathsf{OPT}(G) \geqslant 1 - 10^{-6}$ or $\mathsf{OPT}(G) \leqslant 10^{-6}$. The answer is affirmative, and we provide below examples of infinite families of such graphs.

$\mathsf{OPT}(G) \geqslant 1 - 10^{-6}$: Consider, for example, the following explicit bound was demonstrated in [2, Corollary 6.4]:

if G is a union of k disjoint cliques each with $\frac{n}{k} > 3$ nodes then $\mathsf{OPT}(G) = 1 - \frac{1}{k}$.

Based on this and other known results on modularity clustering, examples of families of regular graphs G for which $\mathsf{OPT}(G) \geqslant 1 - 10^{-6}$ include:

(1) G is a union of k disjoint cliques each with $\frac{n}{k} > 3$ nodes for any $k > 10^6$.
(2) G is obtained by a local modification from the graph in (1) such as:
    - Start with a union of k disjoint cliques $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ each with $\frac{n}{k} > 3$ nodes for any k sufficiently large with respect to 10^6 ($k \geqslant 10^7$ suffices).
    - Remove an arbitrary edge $\{u_i, v_i\}$ from each clique $\mathcal{C}_i$. Let $U = \bigcup_{i=1}^{k}\{u_i\}$ and $V = \bigcup_{i=1}^{k}\{v_i\}$.
    - Add to G the edges corresponding to any perfect matching in the complete bipartite graph with node sets U and V.

$\mathsf{OPT}(G) \leqslant 10^{-6}$: Theorem 1.1 [4] involves infinitely many graphs of $n > 4 + 0.9388 \times 10^6$ nodes satisfying $\mathsf{OPT}(G) < \frac{0.9388}{n-4} < 10^{-6}$ (these graphs are edge complements of appropriate families of 3-regular graphs used in [3]).

Proof of Theorem 2.1. [3] Set $\varepsilon = 10^{-6}$. We assume that G is d-regular, and either $\mathsf{OPT}(G) \geqslant 1 - 10^{-6}$ or $\mathsf{OPT}(G) \leqslant 10^{-6}$.

Preliminary algebraic simplification

Let $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ be a set of communities of G. The objective function $\mathsf{M}(\mathbf{S})$ can be equivalently expressed as follows via simple algebraic manipulation [2,5–7]. Let m_i denote the number of edges whose both endpoints are in S_i, m_{ij} denote the number of edges one of whose endpoints is in S_i and the other in S_j and $D_i = \sum_{v \in S_i} d_v$ denote the sum of degrees of nodes in S_i. Then, $\mathsf{M}(\mathbf{S}) = \sum_{S_i \in \mathbf{S}} (\frac{m_i}{m} - (\frac{D_i}{2m})^2)$.

We will provide an approximation for $\mathsf{OPT}_2(G)$ and then use the result that $\mathsf{OPT}_2(G) \geqslant \frac{\mathsf{OPT}(G)}{2}$ proved in [4]. Note that if $\mathsf{OPT}(G) \leqslant 10^{-6}$ then obviously $\mathsf{OPT}_2(G) \leqslant 10^{-6}$, whereas if $\mathsf{OPT}(G) \geqslant 1 - 10^{-6}$ then $\mathsf{OPT}_2(G) \geqslant \frac{1}{2} - \frac{10^{-6}}{2}$. Consider a partition $\mathbf{S}$ of V into exactly two sets, say S and $\overline{S} = V \setminus S$ with $0 < \mu(S) \leqslant \frac{1}{2}$. By Lemma 2.2 of [4], $\mathsf{M}(S) = \mathsf{M}(\overline{S})$ and thus

$$\begin{aligned}
\mathsf{M}(\mathbf{S}) &= 2 \times \left(\frac{m_1}{m} - \left(\frac{|S|}{n}\right)^2\right) \\
&= 2 \times \left(\frac{\frac{1}{2}\mathsf{D}(S)d|S|}{\frac{1}{2}dn} - \mu(S)^2\right) \\
&= 2 \times \left(\mathsf{D}(S)\mu(S) - \mu(S)^2\right)
\end{aligned}$$

Thus, letting $\mathsf{D} = \mathsf{D}(S)$, $\mu = \mu(S)$ and $\Phi = \Phi(S)$, we have $\Phi = 1 - \mathsf{D}$ as per our notations used in page 349 and the goal of modularity 2-clustering is to maximize the following function f over all possible valid choices of D and μ:

$$f(\mu, \mathsf{D}) = 2 \times \left(\mu\mathsf{D} - \mu^2\right) = 2 \times \left(\mu(1 - \Phi) - \mu^2\right)$$

[3] We have made no significant attempts to optimize the constants in Theorem 2.1.