Detecting Communities Is Hard (And Counting Them Is Even Harder)*

Aviad Rubinstein

University of California at Berkeley, Berkeley, USA
aviad@eecs.berkeley.edu

Abstract

We consider the algorithmic problem of community detection in networks. Given an undirected friendship graph $G = (V, E)$, a subset $S \subseteq V$ is an (α, β)-community if:

- Every member of the community is friends with an α-fraction of the community;
- Every non-member is friends with at most a β-fraction of the community.

Arora et al [3] gave a quasi-polynomial time algorithm for enumerating all the (α, β)-communities for any constants $\alpha > \beta$.

Here, we prove that, assuming the Exponential Time Hypothesis (ETH), quasi-polynomial time is in fact necessary - and even for a much weaker approximation desideratum. Namely, distinguishing between:

- G contains an $(1, o(1))$-community; and
- G does not contain a $(\beta + o(1), \beta)$-community for any $\beta \in [0, 1]$.

We also prove that counting the number of $(1, o(1))$-communities requires quasi-polynomial time assuming the weaker #ETH.

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity

Keywords and phrases Community detection, stable communities, quasipolynomial time

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.42

1 Introduction

Identifying communities is a central graph-theoretic problem with important applications to sociology and marketing (when applied to social networks), biology and bioinformatics (when applied to protein interaction networks), and more (see e.g. Fortunato's classic survey [21]). Defining what exactly is a *community* remains an interesting problem on its own (see Arora et al [3] and Borgs et al [11] for excellent treatment from a theoretical perspective). Ultimately, there is no single "right" definition, and the precise meaning of community should be different for social networks and protein interaction networks.

In this paper we focus on the algorithmic questions arising from one of the simplest and most canonical definitions, which has been considered by several theoretical computer scientists [30, 3, 5, 12] (see Subsection 1.1 for further discussion):

▶ **Definition 1** ((α, β)-Community)**.** Given an undirected graph $G = (V, E)$ an (α, β)-community is a subset $S \subseteq V$ that satisfies:

Strong ties inside the community For every $v \in S$, $|\{v\} \times S| \cap E \geq \alpha \cdot |S|$; and

Weak ties to nodes outside the community For every $u \notin S$, $|\{u\} \times S| \cap E \leq \beta \cdot |S|$.

* This research was supported by a Microsoft Research PhD Fellowship, as well as NSF grant CCF1408635 and Templeton Foundation grant 3966. This work was done in part at the Simons Institute for the Theory of Computing.

Arora et al [3, Theorem 3.1] gave a simple quasi-polynomial ($n^{O(\log n)}$) time for detecting (α, β)-communities whenever $\alpha - \beta$ is at least some positive constant. The algorithm enumerates over $O(\log n)$-tuples of vertices. For each tuple, consider the set of vertices that are neighbors of an $(\alpha + \beta)/2$-fraction of the tuple; test whether this candidate set is indeed a community.

Arora et al's algorithm and analysis are very similar to related algorithms for approximate Nash equilibrium [27], Densest k-Subgraph [8] and Dughmi's Zero-Sum Signaling problem [17]. Recently, matching quasi-polynomial hardness results have been proved for approximate Nash equilibrium [13, 4, 35, 18], Densest k-Subgraph [12, 29], and Zero-Sum Signaling [34, 10] using or inspired by the technique of "birthday repetition" [1]. A natural question, made explicit in [12], is whether similar techniques can be shown to prove quasi-polynomial time hardness, assuming the Exponential Time Hypothesis (ETH)[1], for (α, β)-community detection, for any constants $\alpha > \beta \in [0, 1]$.

Here we show that, for *every* constants $\alpha > \beta \in (0, 1]$, community detection requires quasi-polynomial time (assuming ETH). For example, when $\alpha = 1$ and $\beta = 0.01$, this means that we can hide a clique C, such that every single vertex not in C is connected to at most 1% of C. Our main result is actually a much stronger inapproximability: even in the presence of a $(1, o(1))$-community, finding any $(\beta + o(1), \beta)$-community is hard.

▶ **Theorem 2.** *For every n there exists an $\epsilon = \epsilon(n) = o(1)$ such that, assuming ETH, distinguishing between the following requires time $n^{\tilde{\Omega}(\log n)}$:*
Completeness *G contains an $(1, \epsilon)$-community; and*
Soundness *G does not contain an $(\beta + \epsilon, \beta)$-community for any $\beta \in [0, 1]$.*

Unlike all quasi-polynomial approximation schemes mentioned above, Arora et al's algorithm has the unique property that it can also *exactly count* all the (α, β)-communities. Our second result is that counting even the number of $(1, o(1))$-communities requires quasi-polynomial time. A nice feature of this result is that we can base it on the much weaker #ETH assumption, which asserts that counting the satisfying assignment for a 3SAT instance requires time $2^{\Omega(n)}$. (Note, for example, that #ETH is likely to be true even if $\mathsf{P} = \mathsf{NP}$.)

▶ **Theorem 3.** *For every n there exists an $\epsilon = \epsilon(n) = o(1)$ such that, assuming #ETH, counting $(1, \epsilon)$-communities requires time $n^{\log^{1-o(1)} n}$.*

1.1 Related works

The most closely related work is a reduction by Balcan, Borgs, Braverman, Chayes, and Teng [5, Theorem 5.3] from Planted Clique to finding $(1, 1 - \gamma)$-communities, for some small (unspecified) constant $\gamma > 0$. Note that our inapproximability in Theorem 2 is much stronger in all parameters; furthermore, although formally incomparable, our ETH assumption is preferable over the average-case hardness assumption of Planted Clique.

Algorithms for special cases

Mishra, Schreiber, Stanton, and Tarjan [30] gave a polynomial-time algorithm for finding (α, β)-communities that contain a vertex with very few neighbors outside the community.

[1] The Exponential Time Hypothesis (ETH) [25] asserts that solving 3SAT requires time $2^{\Omega(n)}$. Note that (given our current understanding of complexity) this assumption is essentially necessary - an NP-hardness result is very unlikely given [3]'s quasi-polynomial algorithm. Recall also that ETH is a significantly weaker assumption than the related SETH [24, 14] and NSETH [15],

Balcan et al [5] give a polynomial-time algorithm for enumerating (α, β)-communities in the special case where the degree of every node is $\Omega(n)$.

Arora, Ge, Sachdeva, and Schoenebeck [3] consider several semi-random models where the edges inside the community are generated at random, according to the expected degree model. (In fact, their quasi-polynomial time algorithm is also stated in this setting, but only their "Gap Assumption", which is equivalent to $\alpha - \beta = \Omega(1)$, is used in the analysis.)

Stochastic Block Model

Variants of the community detection problem on graphs generated by different stochastic models are extremely popular (see e.g. [6, 7, 16, 20, 22, 28, 31, 33, 37] for papers in conference proceedings from June 2016). Perhaps the most influential is the *Stochastic Block Model* [23]: The graph is partitioned into two disjoint communities; the edges within each community are present with probability α, independently, whereas edges between communities are present with probability β. Hence this model can also be seen as a special case of the (α, β)-Community Detection problem.

Stochastic models are extremely helpful in physics, for example, because atoms' interactions obey simple mathematical formulas with high precision. Unfortunately, for applications such as social networks, existing models do not describe human behavior with atomic precision, hence casting a shadow over the applicability of algorithms that work on ideal stochastic models. Recent works [31, 28] attempted to bridge the gap from ideal model to practice by showing that certain SDP-based algorithms continue to work in a particular semi-random model where a restricted adversary is allowed to modify the random input graph. These success stories beg the question of how strong can one make the adversary? The current paper illuminates some of the computational barriers.

Alternative approaches to modeling communities

As we mentioned above, there are many different definitions of "communities" in networks. For in-depth discussion of different definitions see Arora et al [3] or Borgs et al [11]. As pointed out by the latter, for some definitions even verifying that a candidate subset is a community is intractable.

There is also an important literature on axiomatic approaches to the related problem of clustering (e.g. [26, 9, 38]); note that while clustering typically aims to partition a set of nodes, our main focus is on detecting just a single community; in particular, different communities may intersect.

1.2 Overview of proofs

A good starting point for the technical discussion is a recent subexponential reduction from 3SAT to the related problem of Densest-k-Subgraph [12]. In Densest-k-Subgraph, we seek a subgraph of size k of maximal density. The two ingredients in [12]'s reduction are "birthday repetition" [1] and the "FGLSS graph" [19]:

"Birthday repetition" Starting with an instance of Label Cover (see definition in Section 2), the reduction considers a mega-variable for every ρ-tuple of variables, for $\rho \approx \sqrt{n}$. By the birthday paradox, almost every pair of ρ-tuples of variables intersect, inducing a consistency constraint on the two mega-assignments. Similarly, we expect to see some Label Cover edges in the union of the two ρ-tuples, inducing an additional Label Cover constraint between the two mega assignments. Notice that we have $\binom{n}{\rho} \approx 2^{\sqrt{n}}$

mega variables, and the alphabet size is also approximately $N = 2^{\sqrt{n}}$. Therefore, assuming ETH, finding an approximately satisfying assignment for the mega-variables requires time $2^{\Omega(n)} \approx N^{\log N}$.

FGLSS Similarly to the classic reduction by Feige et al. [19] for the Clique problem, [12] construct a vertex for each mega assignment to each mega variable, and draw an edge between two vertices if the induced assignments do not violate any consistency or Label Cover constraints. Notice that if the Label Cover instance has a satisfying assignment, then the graph contains a clique of size $\binom{n}{\rho}$ where each mega variable receives the mega assignment induced by the globally satisfying assignment. On the other hand, any subgraph that corresponds to a consistent assignment which violates many constraints must be missing most of its edges.

Now this simple reduction is still far from working for the Community Detection problem, and indeed the latter was listed as an open problem in [13]. Below we describe some of the obstacles and outline how we overcome them.

Completeness

Surprisingly, the main problem with using the same reduction for Community Detection is the completeness: even if the Label Cover instance has a satisfying assignment, the resulting graph has no (α, β)-communities, for any constants $\alpha > \beta$! Observe, in particular, that the clique that corresponds to the satisfying assignment does not satisfy the weak ties condition. For any vertex v in that clique, consider any vertex v' that corresponds to changing the assignment to just one variable x_i in v's assignment. If v agrees with the assignments of all other vertices in the clique, v' agrees with almost all of them - except for the negligible fraction that cover x_i or its neighbors in the Label Cover graph.

To overcome this problem of vertices that are "just outside the community", we use error correcting codes. Namely, we encode each assignment as a low-degree bivariate polynomial over finite field $\mathcal{G}$ of size $|\mathcal{G}| \approx \sqrt{n}$. Now vertices correspond to low-degree assignments to rows/columns of the polynomial. This guarantees that the assignments induced by every two vertices are far. If v agrees with all other vertices in the community, then almost all of those vertices disagree with v'.

Soundness

The main challenge for soundness is ruling out communities that do not correspond to a single, globally consistent assignment to the Label Cover instance. The key idea is to introduce auxiliary vertices that punish such communities by violating the weak ties desideratum.

Let us begin with the reduction to the counting variant (Theorem 3), which is easier, mostly because we are not concerned with approximation (i.e. we only have to show that subsets that are exactly $(1, \epsilon)$-communities correspond to satisfying assignments). Here we further simplify matters by sketching a construction with weighted edges. The full reduction (Section 3) uses unweighted edges and is only slightly more involved. Consider, for every $g \in \mathcal{G}$, an auxiliary vertex that is ϵ-connected to all proper vertices that do not correspond to assignments to the g-th row/column. Now if a $(1, \epsilon)$-community C does not contain a vertex with assignment to the g-th row/column, the auxiliary vertex must simultaneously: (i) belong to C so as not to violate the weak ties desideratum; yet (ii) it cannot belong to C because all its edges have weight ϵ (this would violate the strong ties desideratum). Therefore every $(1, \epsilon)$-community assigns values to every row/column in $\mathcal{G}^2$.

The reduction we described above suffices to show that (assuming ETH) deciding whether the graph contains a $(1, \epsilon)$-community also requires quasi-polynomial time. To get the stronger statement of Theorem 2 we must rule out even $(\beta, \beta + \epsilon)$-communities in case the Label Cover instance is far from satisfiable. In particular, we need to show that subsets that do not correspond to unique, consistent assignments are never $(\beta, \beta + \epsilon)$-communities. Instead of a single column/row, we let each proper vertex correspond to a subset of $t \approx \log n$ columns/rows. Instead of a single $g \in \mathcal{G}$, each auxiliary vertex corresponds to subset $H \subset \mathcal{G}$ of size $|H| = |\mathcal{G}|/2$. We draw an edge between an auxiliary vertex and a proper vertex if the indices of all t columns/rows are contained in H; if they are picked randomly this only happens with polynomially small probability. If, however, a β-fraction of the community is restricted to a small subset $R \subset \mathcal{G}$, then there are auxiliary vertices for $H \supseteq R$ that connect to all those nodes and violate the weak ties desideratum. Roughly, we show that at least a $(1 - \beta)$-fraction of the vertices have assignments that are "well spread" over $\mathcal{G}^2$, and among those assignments there are many violations of the Label Cover constraints.

2 Preliminaries

2.1 Label Cover

▶ **Definition 4** (Label Cover)**.** Label Cover is a maximization problem. The input is a bipartite graph $G = (A, B, E)$, alphabets Σ_A, Σ_B, and a projection $\pi_e : \Sigma_A \to \Sigma_B$ for every $e \in E$.

The output is a labeling $\varphi_A : A \to \Sigma_A$, $\varphi_B : B \to \Sigma_B$. Given a labeling, we say that a constraint (or edge) $(a, b) \in E$ is *satisfied* if $\pi_{(a,b)}(\varphi_A(a)) = \varphi_B(b)$. The *value of a labeling* is the fraction of $e \in E$ that are satisfied by the labeling. The value of the instance is the maximum fraction of constraints satisfied by any assignment.

▶ **Theorem 5** (Moshkovitz-Raz PCP [32, Theorem 11])**.** *For every n and every $\epsilon > 0$ (in particular, ϵ may be a function of n), solving* 3SAT *on inputs of size n can be reduced to distinguishing between the case that a (d_A, d_B)-bi-regular instance of* Label Cover*, with parameters $|A| + |B| = n^{1+o(1)} \cdot \mathsf{poly}(1/\epsilon)$, $|\Sigma_A| = 2^{\mathsf{poly}(1/\epsilon)}$, and $d_A, d_B, |\Sigma_B| = \mathsf{poly}(1/\epsilon)$, is completely satisfiable, versus the case that it has value at most ϵ.*

Counting the number of satisfying assignments is even harder. The following hardness is well-known, and we sketch its proof only for completeness:

▶ **Fact 1.** *There is a linear-time reduction from #3SAT to counting the number of satisfying assignments of a* Label Cover *instance.*

Proof. Construct a vertex in A for each variable and a vertex in B for each clause. Set $\Sigma_A \triangleq \{0, 1\}$ and let $\Sigma_B \triangleq \{0, 1\}^3 \setminus (000)$ (i.e. Σ_B is the set of satisfying assignments for a 3SAT clause, after applying negations). Now if variable x appears in clause C, add a constraint that the assignments to x and C are consistent (taking into account the sign of x in C). Notice that any assignment to A: (i) corresponds to a unique assignment to the 3SAT formula; and (ii) if the 3SAT formula is satisfied, this assignment uniquely defines a satisfying assignment to B. Therefore there is a one-to-one correspondence between satisfying assignments to the 3SAT formula and to the instance of Label Cover. ◀

2.2 Finding a good partition

▶ **Theorem 6** (k-wise independence Chernoff bound [36, Theorem 5.I]). *Let $x_1 \dots x_n \in [0,1]$ be k-wise independent random variables, and let $\mu \triangleq \mathsf{E}\left[\sum_{i=1}^{n} x_i\right]$ and $\delta \leq 1$. Then*

$$\Pr\left[\left|\sum_{i=1}^{n} x_i - \mu\right| > \delta\mu\right] \leq \mathrm{e}^{-\Omega\left(\min\left\{k,\delta^2\mu\right\}\right)}.$$

We use Chernoff bound with $\Theta(\log n)$-wise independent variables to deterministically partition variables into subsets of cardinality $\approx \sqrt{n}$. Our (somewhat naive) deterministic algorithm for finding a good partition takes quasi-polynomial time ($n^{O(\log n)}$), which is negligible with respect to the sub-exponential size ($N = 2^{\tilde{O}(\sqrt{n})}$) of our reduction[2].

▶ **Lemma 7.** *Let $G = (A, B, E)$ be a bipartite (d_A, d_B)-bi-regular graph, and let $n_A \triangleq |A|$, $n_B \triangleq |B|$; set also $n \triangleq n_B + n_A$ and $\rho \triangleq \sqrt{n}\log n$. Let $T_1, \dots, T_{n_B/\rho}$ be an arbitrary partition of B into disjoint subsets of size ρ. There is a quasi-polynomial deterministic algorithm (alternatively, linear-time randomized algorithm) that finds a partition of A into $S_1, \dots, S_{n_A/\rho}$, such that:*

$$\forall i \quad \left||S_i| - \rho\right| < \rho/2, \tag{1}$$

and

$$\forall i,j \quad \left|\left|(S_i \times T_j) \cap E\right| - \frac{d_A\rho^2}{n_B}\right| < \frac{d_A\rho^2}{2n_B}. \tag{2}$$

Proof. Suppose that we place each $a \in A$ into a uniformly random S_i. By Chernoff bound and union bound, (1) and (2) hold with high probability. Now, by Chernoff Bound for k-wise independent variables (Theorem 6), it suffices to partition A using a $\Theta(\log n)$-wise independent distribution. Such distribution can be generated with a sample space of $n^{O(\log n)}$ (e.g. [2]). Therefore, we can enumerate over all possibilities in quasi-polynomial time. By the probabilistic argument, we will find at least one partition that satisfies (1) and (2). ◀

3 Hardness of Counting Communities

▶ **Theorem 8.** *There exists an $\epsilon(n) = o(1)$ such that, assuming #ETH, counting $(1,\epsilon)$-communities requires time $n^{\log^{1-o(1)} n}$.*

Construction

Begin with an instance (A, B, E, π) of Label Cover of size $n = n_A + n_B$ where $n_A \triangleq |A|$ and $n_B \triangleq |B|$. Let $\mathcal{G}$ be a finite field of size $\sqrt{n}/\epsilon^3$, and let $\mathcal{F} \subset \mathcal{G}$ be an arbitrary subset of size $|\mathcal{F}| = \sqrt{n}$. We identify between $A \cup B$ and points in $\mathcal{F}^2$; we also identify between a subset of $\mathcal{G}$ and $\Sigma_A \cup \Sigma_B$. Thus there is a one-to-one correspondence between a subset of assignments to $P_{\mathcal{F}} \colon \mathcal{F}^2 \to \mathcal{G}$ and assignments to the Label Cover instance. We can extend any such $P_{\mathcal{F}}$ to an individual-degree-$(|\mathcal{F}|-1)$ polynomial $P : \mathcal{G}^2 \to \mathcal{G}$. In the other

[2] Do not confuse this with the quasi-polynomial lower bound ($N^{\tilde{O}(\log N)}$) we obtain for the running time of the community detection problem.

direction, we think of each low individual degrees polynomial $P : \mathcal{G}^2 \to \mathcal{G}$ as a (possibly invalid) assignment to the Label Cover instance.

For every $g \in \mathcal{G}$, and degree-$(|\mathcal{F}|-1)$ polynomials $p_1, p_2 : \mathcal{G} \to \mathcal{G}$ such that $p_1(g) = p_2(g)$, we construct $1/\epsilon$ vertices $\{v_{g,p_1,p_2,i}\}_{i=1}^{1/\epsilon} \subset V$ in the communities graph. Each vertex naturally induces an assignment (p_1, p_2) on $(\mathcal{G} \times \{g\}) \cup (\{g\} \times \mathcal{G})$. We draw an edge between two vertices in V if they agree on the intersection of their lines, and if their induced assignments satisfy all the Label Cover constraints.

For every $g \in \mathcal{G}$ and $i \in [1/\epsilon]$, we also add two identical auxiliary vertices $u_{g,i}$ which are connected to every $v_{g',p_1,p_2,i}$ for $g' \neq g$ (but not to each other).

Completeness

For each assignment to the Label Cover instance, we construct a $(1,\epsilon)$-community by taking the induced assignment $P_{\mathcal{F}} \colon \mathcal{F}^2 \to \mathcal{G}$ and extending it to an individual-degree-$(|\mathcal{F}|-1)$ polynomial $P : \mathcal{G}^2 \to \mathcal{G}$. Let C be all the vertices $v_{g,p_1,p_2,i}$ such that p_1, p_2 are the restrictions of P to $(\mathcal{G} \times \{g\}), (\{g\} \times \mathcal{G})$. This correspondence is one-to-one and we need to show that the resulting C is actually a $(1,\epsilon)$-community.

Because all the vertices correspond to a consistent satisfying assignment, C is a clique. Let $v_{g,q_1,q_2,i} \notin C$; wlog q_1 disagrees with the restriction of P to $(\mathcal{G} \times \{g\})$. Since both q_1 and the restriction of P are degree-$(|\mathcal{F}|-1)$ polynomials, they must disagree on all but at most $(|\mathcal{F}|-1)$ elements of $\mathcal{G}$. For all other $h \in \mathcal{G}$, the vertex $v_{g,q_1,q_2,i}$ does not share edges with any $v_{h,p_1,p_2,j} \in C$. Therefore, $v_{g,q_1,q_2,i}$ has edges to less than an $(|\mathcal{F}| / |\mathcal{G}|)$-fraction of vertices in C. Finally, every auxiliary vertex $u_{g,i}$ has edges to a $\frac{|\mathcal{G}|-1}{|\mathcal{G}|} \cdot \epsilon < \epsilon$-fraction of the vertices in ϵ. Therefore, C is a $(1,\epsilon)$-community.

3.1 Soundness

Structure of $(1,\epsilon)$-communities

▶ **Claim 1.** *Every $(1,\epsilon)$-community C contains exactly $1/\epsilon$ vertices $\{v_{g,p_1,p_2,i}\}_{i=1}^{1/\epsilon}$ for each g.*

Proof. First, observe that C cannot contain any auxiliary vertices: if C contains one copy of $u_{g,i}$, it must also contain the other; but they don't have an edge between them, so they cannot both belong to a $(1,\epsilon)$-community.

Now, assume by contradiction that for some $g \in \mathcal{G}$, C does not contain any vertices with assignments for $(\mathcal{G} \times \{g\}) \cup (\{g\} \times \mathcal{G})$. Then every vertex in C is connected to (both copies of) $u_{g,i}$, for some $i \in [1/\epsilon]$. Therefore there is at least one $i \in [1/\epsilon]$ such that $u_{g,i}$ is connected to an ϵ-fraction of the vertices in C. But this is a contradiction since $u_{g,i} \notin C$.

If we ignore the auxiliary vertices (which, as we argued, C does not contain), the different vertices $v_{g,p_1,p_2,i}$ that correspond to the same assignment to the same lines (i.e. if we only change i) are indistinguishable. Therefore if C contains one of them, it must contain all of them (hence, at least $1/\epsilon$ vertices for each g).

Finally, since C is a clique, it cannot contain vertices that disagree on any assignments. (In particular, it cannot contain more than $1/\epsilon$ vertices for each g.) ◀

Completing the proof

Proof of Soundness. By Claim 1, every $(1,\epsilon)$-community C contains exactly $1/\epsilon$ vertices $\{v_{g,p_1,p_2,i}\}_{i=1}^{1/\epsilon}$ for each g. Furthermore, since C is a clique, all the induced assignments agree on all the intersections. So every $(1,\epsilon)$-community corresponds to a unique consistent

assignment to the Label Cover instance. Finally, appealing again to the fact that C is a clique, this assignment must also satisfy all the Label Cover constraints. ◀

4 Hardness of Detecting Communities

▶ **Theorem 9.** *There exists an $\epsilon(n) = o(1)$ such that, assuming ETH, distinguishing between the following requires time $n^{\tilde{\Omega}(\log n)}$:*

Completeness *G contains an $(1, \epsilon)$-community; and*

Soundness *G does not contain an $(\beta + \epsilon, \beta)$-community for any $\beta \in [0, 1]$.*

The rest of this section is devoted to the proof of Theorem 9. Our starting point is the Label Cover of Moshkovitz-Raz (Theorem 5). We compose the birthday repetition technique of [1] with a bi-variate low-degree encoding. We then encode this as a graph a-la FGLSS [19]. We add auxiliary vertices to ensure that any $(\beta + \epsilon, \beta)$-community corresponds, approximately, to a uniform distribution over the variables.

Construction

Begin with a (d_A, d_B)-bi-regular instance (A, B, E, π) of Label Cover of size $n = n_A + n_B$ where $n_A \triangleq |A|$ and $n_B \triangleq |B|$. Let $\rho \triangleq \sqrt{n} \log n$; let $\mathcal{G}$ be a finite field of size $\rho/\epsilon^3 = \tilde{O}(\rho)$, and let $\mathcal{F} \subset \mathcal{G}$ be an arbitrary subset of size $|\mathcal{F}| = 2\rho$. Let $\mathcal{F}_A, \mathcal{F}_B \subset \mathcal{F}$ be disjoint subsets of size n_A/ρ, n_B/ρ, respectively. By Lemma 7, we can partition A and B into subsets $X_1, \ldots, X_{|\mathcal{F}_A|}$ and $Y_1, \ldots, Y_{|\mathcal{F}_B|}$ of size at most $|\mathcal{F}|$ such that between every two subsets there are approximately $\frac{d_A \rho^2}{n_B} = \frac{d_B \rho^2}{n_A}$ constraints. For $i \in \mathcal{F}_A$, we think of the points $\{i\} \times \mathcal{F} \subset \mathcal{G}^2$ as representing assignments to variables in X_i; for $j \in \mathcal{F}_B$, we think of $\mathcal{F} \times \{j\} \subset \mathcal{G}^2$ as representing assignments to variables in Y_j. Notice that each point in $\mathcal{F}^2$ may represent an assignment to both a vertex from A and a vertex from B, to one of them, or to neither. In particular, any assignment $P \colon \mathcal{G}^2 \to \mathcal{G}$ induces an assignment for the Label Cover instance; note that since $|\mathcal{G}| > |\Sigma_A| \, |\Sigma_B|$, one value $P(f_1, f_2) \in \mathcal{G}$ suffices to describe assignments to both $a \in A$ and $b \in B$.

Let $t \triangleq \log n \cdot \left(\frac{|\mathcal{G}|}{|\mathcal{F}_A|} + \frac{|\mathcal{G}|}{|\mathcal{F}_B|} \right) = \mathsf{polylog}(n)$. We say that a subset $S \in \binom{\mathcal{G}}{t}$ is *balanced* if: $|S \cap \mathcal{F}_A| = \frac{|\mathcal{F}_A|}{|\mathcal{G}|} \cdot t$ and $|S \cap \mathcal{F}_B| = \frac{|\mathcal{F}_B|}{|\mathcal{G}|} \cdot t$. For every balanced subset S, consider $2t$ polynomials $q_\ell \colon \mathcal{G} \to \mathcal{G}$ of degree at most $|\mathcal{F}| - 1$, representing an assignment[3] $Q \colon (S \times \mathcal{G}) \cup (\mathcal{G} \times S) \to \mathcal{G}$. For balanced S and $2t$-tuple of polynomials (q_ℓ), we construct a corresponding vertex $v_{S,(q_\ell)}$ in the communities graph. Let V denote the set of vertices defined so far. For $g \in \mathcal{G}$ we abuse notation and say that $g \in v_{S,(q_\ell)}$ if $g \in S$. We construct an edge in the communities graph between two vertices in V if their assignments agree on the variables in their intersection, and their induced assignments to $A \cup B$ satisfy all the Label Cover constraints.

Additionally, for every $H \subset \mathcal{G}$ of size $|H| = |\mathcal{G}|/2$, define $|V|^2$ identical auxiliary vertices u_H in the communities graph. We draw an edge between auxiliary vertex u_H and vertex $v_{S,(q_\ell)}$ if $S \subset H$. Similarly, for every $H_A \subset \mathcal{F}_A$ of size $|H_A| = |\mathcal{F}_A|/2$, we define $|V|^2$ identical auxiliary vertices u_{H_A} with edges to every vertex $v_{S,(q_\ell)}$ such that $(S \cap \mathcal{F}_A) \subset H_A$. For $H_B \subset \mathcal{F}_B$ of size $|H_B| = |\mathcal{F}_B|/2$, we draw edges between u_{H_B} and $v_{S,(q_\ell)}$ such that $(S \cap \mathcal{F}_B) \subset H_B$.

[3] We will only consider polynomials that correspond to a consistent assignment Q; i.e. for each point in $S \times S$ we expect the two corresponding polynomials to agree with each other.

Completeness

Suppose that the Label Cover instance has a satisfying assignment. Let $\mathcal{Z} \subseteq \mathcal{G}^2$ denote the subset of points that correspond to at least one variable in A or B. Let $P_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{G}$ be the induced function on $\mathcal{Z}$ that corresponds to the satisfying assignment, and let $P : \mathcal{G}^2 \to \mathcal{G}$ be the extension of $P_{\mathcal{Z}}$ by setting $P(f_1, f_2) = 0$ for $(f_1, f_2) \in \mathcal{F}^2 \setminus \mathcal{Z}$ (this choice is arbitrary), and then extending to an $(|\mathcal{F}| - 1)$-individual-degree polynomial over all of $\mathcal{G}^2$.

Let C be the set of vertices that correspond to restrictions of P to balanced sets, i.e.

$$C = \left\{ v_{S,(P|_S)} : S \text{ is balanced} \right\},$$

where $P\ |_S$ denotes the restriction of P to $(S \times \mathcal{G}) \cup (\mathcal{G} \times S)$. Since all those vertices correspond to a consistent satisfying assignment, C is a clique.

For any vertex $v_{S,(q_\ell)} \notin C$, at least one of the polynomials, q_{ℓ^*} disagrees with the restriction of P to the corresponding line. Since both q_{ℓ^*} and the restriction of P to that line are degree-$(|\mathcal{F}| - 1)$ polynomials, they must disagree on at least $\left(1 - \frac{|\mathcal{F}|}{|\mathcal{G}|}\right)$-fraction of the coordinates. The probability that a random balanced set S' is contained in the $O\left(\epsilon^3\right)$-fraction of coordinates where they do agree is smaller than ϵ (and in fact polynomially small in n). Therefore $v_{S,(q_\ell)}$ has inconsistency violations with all but (less than) an ϵ-fraction of the vertices in C.

For any auxiliary vertex u_{H_A}, the probability that a random vertex $v_{S,(P|_S)} \in C$ is connected to u_{H_A} is $2^{-|S \cap \mathcal{F}_A|} < 1/n$, and similarly for u_{H_B} and u_H. Therefore, every auxiliary vertex is connected to less than a $(1/n)$-fraction of the vertices in C.

4.1 Soundness

▶ **Lemma 10.** *If the* Label Cover *instance has value at most ϵ^3, then there are no $(\beta + \epsilon, \beta)$-communities.*

4.1.0.1 Auxiliary vertices

▶ **Claim 2.** *Every $(\beta + \epsilon, \beta)$-community does not contain any auxiliary vertices.*

Proof. There are $|V|^2$ identical copies of each auxiliary vertex. Since they are identical, any community must either contain all of them, or none of them. If the community contains all $|V|^2$ copies, then it has a vast majority of auxiliary vertices, so none of them can have edges to an ϵ-fraction of the community. ◀

4.1.0.2 List decoding

▶ **Claim 3.** *The vertices in any $(\beta + \epsilon, \beta)$-community C induce at most $4/\epsilon$ different assignments for each variable.*

Proof. Suppose by contradiction that this is not the case. Then, wlog, there is a line $\{g_1\} \times \mathcal{G}$ that receives at least $2/\epsilon$ different assignments from vertices in C. Every two assignments agree on at most $|\mathcal{F}|$ points (g_1, g') on the line, so in total there are at most $2|\mathcal{F}|/\epsilon^2$ points where at least two assignments agree. Let $R \subseteq \mathcal{G}$ denote the set of g' such that no two assignments agree on (g_1, g'); we have that $|R| \geq |\mathcal{G}| - 2|\mathcal{F}|/\epsilon^2 \geq |\mathcal{G}|/2$. Therefore, by the weak ties property, for at most a β-fraction of the vertices $v_{S,(q_\ell)} \in C$, $S \cap R = \emptyset$.

Consider the remaining $(1 - \beta)$-fraction of vertices in C. Suppose that v assigns a value to some (g_1, g') for $g' \in R$: this value can only agree with one of the $2/\epsilon$ different assignments

to (g_1, g'). Therefore, in expectation, each of the $2/\epsilon$ vertices that assign different values for (g_1, g') is connected to at most a $(\beta + \epsilon/2)$-fraction of the vertices in C. This is a contradiction to C being a $(\beta + \epsilon, \beta)$-community. ◀

4.1.0.3 Completing the proof

Proof of Lemma 10. Suppose that at most a ϵ^3-fraction of the Label Cover constraints can be satisfied by any single assignment, and assume by contradiction that C is a $(\beta + \epsilon, \beta)$-community. By Claim 3, C induces at most $4/\epsilon$ assignments on each variable, so at most $O(\epsilon)$-fraction of the constraints are satisfied by any pair of assignments.

By Markov's inequality, for at least half of the subsets $X_i \subset A$, only an $O(\epsilon)$-fraction of the constraints that depend on X_i are satisfied. By Claim 2 at least $(1 - \beta)$-fraction of the vertices in C assign values to at least one such X_i. Consider any such vertex $v_{S,(q_\ell)}$ where $S \ni i$. By construction of the partitions (Lemma 7), each X_i shares approximately the same number of constraints with each Y_j. Therefore, for all but an $O(\epsilon)$-fraction of Y_j's, X_i and Y_j observe a violation - for all the assignments given by vertices in C to the variables in Y_j. In other words, $v_{S,(q_\ell)}$ cannot have edges to any vertex $v_{T,(r_\ell)}$ such that $T \ni j$, for a $(1 - O(\epsilon))$-fraction of $j \in [n_B/k_B]$. Finally, applying Claim 2 again, at most a β fraction of vertices in C do not contain any of those j's. This is a contradiction to $v_{S,(q_\ell)}$ having edges to $(\beta + \epsilon)$-fraction of the vertices in C. ◀

References

1 Scott Aaronson, Russell Impagliazzo, and Dana Moshkovitz. AM with multiple merlins. In *Computational Complexity (CCC), 2014 IEEE 29th Conference on*, pages 44–55. IEEE, 2014.

2 Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms*, 7(4):567–583, 1986. doi:10.1016/0196-6774(86)90019-2.

3 Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *ACM Conference on Electronic Commerce, EC '12, Valencia, Spain, June 4-8, 2012*, pages 37–54, 2012. doi:10.1145/2229012.2229020.

4 Yakov Babichenko, Christos H. Papadimitriou, and Aviad Rubinstein. Can almost everybody be almost happy? In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pages 1–9, 2016. doi:10.1145/2840728.2840731.

5 Maria-Florina Balcan, Christian Borgs, Mark Braverman, Jennifer T. Chayes, and Shang-Hua Teng. Finding endogenously formed communities. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 767–783, 2013. doi:10.1137/1.9781611973105.55.

6 Afonso S. Bandeira, Nicolas Boumal, and Vladislav Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 361–382, 2016. URL: http://jmlr.org/proceedings/papers/v49/bandeira16.html.

7 Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the 29th Confer-*

ence on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016, pages 383–416, 2016. URL: http://jmlr.org/proceedings/papers/v49/banks16.html.

8 Siddharth Barman. Approximating nash equilibria and dense bipartite subgraphs via an approximate version of caratheodory's theorem. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 361–369, 2015. doi:10.1145/2746539.2746566.

9 Shai Ben-David and Margareta Ackerman. Measures of clustering quality: A working set of axioms for clustering. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 121–128, 2008. URL: http://papers.nips.cc/paper/3491-measures-of-clustering-quality-a-working-set-of-axioms-for-clustering.

10 Umang Bhaskar, Yu Cheng, Young Kun Ko, and Chaitanya Swamy. Hardness results for signaling in bayesian zero-sum and network routing games. In *Proceedings of the 2016 ACM Conference on Economics and Computation, EC '16, Maastricht, The Netherlands, July 24-28, 2016*, pages 479–496, 2016. doi:10.1145/2940716.2940753.

11 Christian Borgs, Jennifer T. Chayes, Adrian Marple, and Shang-Hua Teng. An axiomatic approach to community detection. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pages 135–146, 2016. doi:10.1145/2840728.2840748.

12 Mark Braverman, Young Kun-Ko, Aviad Rubinstein, and Omri Weinstein. ETH hardness for densest-k-subgraph with perfect completeness. In *SODA*, 2017. To appear.

13 Mark Braverman, Young Kun-Ko, and Omri Weinstein. Approximating the best nash equilibrium in $n^{o(\log n)}$-time breaks the exponential time hypothesis. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 970–982, 2015. doi:10.1137/1.9781611973730.66.

14 Chris Calabro, Russell Impagliazzo, and Ramamohan Paturi. The complexity of satisfiability of small depth circuits. In *Parameterized and Exact Computation, 4th International Workshop, IWPEC 2009, Copenhagen, Denmark, September 10-11, 2009, Revised Selected Papers*, pages 75–85, 2009. doi:10.1007/978-3-642-11269-0_6.

15 Marco L. Carmosino, Jiawei Gao, Russell Impagliazzo, Ivan Mihajlin, Ramamohan Paturi, and Stefan Schneider. Nondeterministic extensions of the strong exponential time hypothesis and consequences for non-reducibility. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pages 261–270, 2016. doi:10.1145/2840728.2840746.

16 Yuxin Chen, Govinda M. Kamath, Changho Suh, and David Tse. Community recovery in graphs with locality. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 689–698, 2016. URL: http://jmlr.org/proceedings/papers/v48/chena16.html.

17 Yu Cheng, Ho Yee Cheung, Shaddin Dughmi, Ehsan Emamjomeh-Zadeh, Li Han, and Shang-Hua Teng. Mixture selection, mechanism design, and signaling. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1426–1445, 2015. doi:10.1109/FOCS.2015.91.

18 Argyrios Deligkas, John Fearnley, and Rahul Savani. Inapproximability results for approximate nash equilibria. *CoRR*, abs/1608.03574, 2016. URL: http://arxiv.org/abs/1608.03574.

19 Uriel Feige, Shafi Goldwasser, Laszlo Lovász, Shmuel Safra, and Mario Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM (JACM)*, 43(2):268–292, 1996.

20 Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 943–959, 2016. URL: http://jmlr.org/proceedings/papers/v49/florescu16.html.

21 Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

22 Bruce E. Hajek, Yihong Wu, and Jiaming Xu. Semidefinite programs for exact recovery of a hidden community. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1051–1095, 2016. URL: http://jmlr.org/proceedings/papers/v49/hajek16.html.

23 Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

24 Russell Impagliazzo and Ramamohan Paturi. On the complexity of k-sat. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001.

25 Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, 2001. doi:10.1006/jcss.2001.1774.

26 Jon M. Kleinberg. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 446–453, 2002. URL: http://papers.nips.cc/paper/2340-an-impossibility-theorem-for-clustering.

27 Richard J. Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *EC*, pages 36–41, 2003.

28 Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Learning communities in the presence of errors. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1258–1291, 2016. URL: http://jmlr.org/proceedings/papers/v49/makarychev16.html.

29 Pasin Manurangsi. Almost-Polynomial Ratio ETH-Hardness of Approximating Densest k-Subgraph with Perfect Completeness. *CoRR*, abs/1611.05991, 2016.

30 Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert Endre Tarjan. Clustering social networks. In *Algorithms and Models for the Web-Graph, 5th International Workshop, WAW 2007, San Diego, CA, USA, December 11-12, 2007, Proceedings*, pages 56–67, 2007. doi:10.1007/978-3-540-77004-6_5.

31 Ankur Moitra, William Perry, and Alexander S. Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 828–841, 2016. doi:10.1145/2897518.2897573.

32 Dana Moshkovitz and Ran Raz. Two-query PCP with subconstant error. *J. ACM*, 57(5), 2010. doi:10.1145/1754399.1754402.

33 Elchanan Mossel and Jiaming Xu. Density evolution in the degree-correlated stochastic block model. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1319–1356, 2016. URL: http://jmlr.org/proceedings/papers/v49/mossel16.html.

34 Aviad Rubinstein. Eth-hardness for signaling in symmetric zero-sum games. *CoRR*, abs/1510.04991, 2015. URL: http://arxiv.org/abs/1510.04991.

35 Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. In *To appear in FOCS*, 2016.

36 Jeanette P. Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff-hoeffding bounds for applications with limited independence. *SIAM J. Discrete Math.*, 8(2):223–250, 1995. doi:10.1137/S089548019223872X.

37 Nicolas Tremblay, Gilles Puy, Rémi Gribonval, and Pierre Vandergheynst. Compressive spectral clustering. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1002–1011, 2016. URL: `http://jmlr.org/proceedings/papers/v48/tremblay16.html`.

38 Twan van Laarhoven and Elena Marchiori. Axioms for graph clustering quality functions. *Journal of Machine Learning Research*, 15(1):193–215, 2014. URL: `http://dl.acm.org/citation.cfm?id=2627441`.