CHALMERS

Spectral and probabilistic methods in planted random graph models

Master's Thesis in Engineering Mathematics and Computational Science

ANDERS MARTINSSON

Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden 2012

Thesis for the degree of
Master of Science in Engineering Mathematics and Computational Science

Spectral and probabilistic methods in planted random graph models

Author:
Anders Martinsson

Supervisors:
Assoc. Prof. Chiranjib Bhattacharyya
Asst. Prof. Devdatt Dubhashi

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2012

Spectral and probabilistic methods in planted random graph models
Anders Martinsson

Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden

Chalmers Reproservice
Gothenburg, Sweden 2012

Spectral and probabilistic methods in planted random graph models

Anders Martinsson
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden

Abstract

The problem of finding a maximum independent set in a graph is known to be NP-hard[1]. In fact, even various weaker forms of the maximum independent set problem, such as approximating its size within a large factor, are known to have no polynomial time solutions unless P=NP[2].

In this report, we consider a special case of the maximum independent set problem in which a large independent set is hidden in an otherwise random graph. This model is the graph complement of the random planted clique graph proposed by Jerrum[5] and Kučera[6]. We show that the Lovász number, as introduced by Lovász[17], of graphs generated according to this model is exactly the size of the planted independent set with high probability, provided the planted set is sufficiently large. This characterization of the Lovász number extends the technical Lemma by Feige and Krauthgamer[8]. We also derive a similar property for an upper bound on the Lovász number introduced by Luz[19].

After this, we discuss the problem of certifying a maximum independent set in this model. We present a method for reducing the problem of certifying a maximum independent set to certifying that a certain matrix is positive definite. This method is significantly faster than the algorithm presented by Feige and Krauthgamer, while still working in essentially the same parameter regimes.

Keywords: Independent set, planted clique, Erdös-Rényi graphs, Lovász number, certificate, eigenvalues.

Acknowledgements

This master's thesis project has been carried out under the supervision of Chiranjib Bhattacharyya and Devdatt Dubhashi. The project has been a completely new experience for me. I am very grateful to you for letting me take part in your work and all the time you have been willing to take out of your day to listen to my ramblings.

I would also like to thank Thomas Bååth Sjöblom, who was the opponent at my master's thesis seminar, for many insightful comments on this thesis and on the project in general.

Anders Martinsson, Gothenburg June 29, 2012

Contents

Chapter 1

Introduction

An *independent set* S of a simple undirected graph $G = (V, E)$ is a subset of the vertex set such that no pair of distinct vertices in S has an edge between them. Similarly, a *clique* S of G is a subset of the vertex set such that all pairs of distinct vertices in S have an edge between them. We define the independence number of G, $\alpha(G)$, as the size of the largest independent set of G.

The problem of computing $\alpha(G)$ for a general graph G (or more accurately the problem of deciding whether $\alpha(G) \geq k$ for some given k), is one of Karp's 21 NP-complete problems[1] and is thus one of the first problems to be shown to be NP-complete. In fact, it has been shown that even approximating $\alpha(G)$ within a factor of n^r for some $r > 0$ can not be done unless P=NP[2]. It is therefore unlikely that some polynomial algorithm for finding $\alpha(G)$ for a general graph ever will be found.

One method that has been used to tackle the problem of finding the independence number of a graph G is to express $\alpha(G)$ as the optimal value of some optimization problem and instead solve some appropriate relaxation. As $\alpha(G)$ is most naturally expressed as a maximization problem, these relaxations provide upper bounds on the independence number. Arguably the most famous such bound is known as Lovász ϑ-function or the Lovász number of a graph G. This function can be formulated as the optimal value of a semi-definite programming, SDP, problem and can thus be computed in polynomial time[3]. A definition of $\vartheta(G)$ will be presented in Chapter 3. While $\vartheta(G)$ has no guarantee of being close to $\alpha(G)$ for a general graphs, it has been shown that there is a large class of graphs where $\alpha(G) = \vartheta(G)$ including bipartite, chordal and perfect graphs[4].

In order to bridge the gap between worst case performance and these special classes of graphs, it is natural to ask what can be done for an "average" graph, one which is neither constructed to be particularly hard nor easy to solve. The most straight-forward way to formalize this question would be to consider some kind of random graph model.

The following random model was proposed by Jerrum[5] and Kučera[6] regarding to the maximum clique problem: A graph on the vertex set $\{1, \ldots n\}$ is constructed by tossing a fair coin for each of the $\binom{n}{2}$ possible edges and including the edge if the outcome is *heads*. A subset of k vertices is then picked at random with uniform probability and

a clique is planted on them by adding all edges between pairs of vertices in the subset.

As taking the graph complement maps cliques to independent sets and vice versa, solving the maximum clique problem for this graph and the maximum independent set problem for the complement graph are equivalent. To apply this model to the independent set problem we will thus consider the graph complement of the model. However, by convention we will keep the notion of planting cliques rather than independent sets.

Jerrum and Kučera posed independently the problem of finding for which n and k the planted set could be retrieved with high probability in polynomial time, given the graph. It can be shown that for k much larger than $\log n$, this set is the maximum clique with high probability, so the problem is equivalent to finding the maximum clique in the graph.

Kučera noted that for $k \geq C\sqrt{n \log n}$, the planted clique was the k vertices with highest degrees with high probability, allowing an easy means for retrieval. The problem of retrieving the clique for $k = o\left(\sqrt{n \log n}\right)$ was first solved by Alon, Krivelevich, and Sudakov[7]. They showed that, for $k \geq C\sqrt{n}$, the eigenvector corresponding to the second largest eigenvalue of the adjacency matrix is very concentrated on the planted clique with high probability, and used this to retrieve the clique. Their result has later been matched by various other techniques: Feige and Krauthgamer[8] showed that the clique could be recovered by inspecting the optimal solution of one formulation of $\vartheta(\bar{G})$. Feige and Ron[9] gave a very nice proof that iteratively removing the vertex with lowest degree would leave a large fraction of the clique. This algorithm has the advantage of running in $O(n^2)$ time, making it significantly faster than previously known methods, but their proof only guarantees a probability of success of $\frac{2}{3}$. Later, Dekel, Gurel-Gurevich, and Peres[10] presented an algorithm for recovering the planted set with high probability in $O(n^2)$ time based on sampling degrees.

Interestingly, the methods mentioned above all break down at $k = o(\sqrt{n})$ for seemingly unrelated reasons, and there is currently no known algorithm that has been proved to recover cliques of that size. It has however been noted by Alon et al. that the problem of recovering a hidden clique of size $k \geq C\sqrt{n}$ for any positive constant C can be solved in polynomial time. One approach to recover cliques of size $k \geq Cn^{1/3}(\log n)^4$ was proposed by Frieze and Kannan[11], but their method requires maximizing a cubic form, which currently has no known polynomial time implementations.

One method of special interest to this thesis is the algorithm presented by Feige and Krauthgamer. Using properties of Lovász ϑ-function their algorithm manages to provide a tight upper bound on the independence number, certifying optimality of found solutions. Feige and Krauthgamer also showed that their method retrieves the clique in an extension of the planted clique model in which an adversary is allowed to remove any edges not in the clique. While both of these properties could potentially be very useful in applications, computing Lovász ϑ-function requires solving an SDP problem, which is notoriously computer-intensive. To the author's knowledge, the fastest known algorithm approximates the Lovász number within an additive error of $\delta > 0$ in $O\left((\vartheta/\delta)^2 n^5 \ln n\right)$ time where ϑ denotes the Lovász number of the given graph[12]. Because of this, their method is simply infeasible in many cases, especially compared to the very elegant $O(n^2)$

solutions. It would therefore be of interest to reproduce these properties by some more practical algorithm.

In this report we will consider an extension of the planted clique model, where we allow the coin to be (possibly extremely) biased. We will investigate the behavior of two upper bounds of the independence number of the complement of these graphs, and how this relates to the problem of finding the maximum clique. After this, we discuss the problem of certifying maximum independent sets and propose a novel method for certifying optimality of the planted set in this model.

Chapter 2

Preliminaries

In this report, the word graph will always refer to simple undirected graphs. For a graph G, we will denote the complementary graph $\bar{G}$. For simplicity we will always assume $V = \{1, \dots n\}$. For a vertex i, $N(i)$ denotes the neighbors of i, i.e. the set of all $j \in V$ such that i and j are connected by an edge in G. An independent set S is said to be a *maximal independent set* if it is not a proper subset of some other independent set. If S is an independent set such that $|S| = \alpha(G)$, then S is a *maximum independent set*.

For vectors, we will use $\|\cdot\|$ to denote the Euclidean norm. Unless stated otherwise, all vectors are assumed to be column vectors.

The adjacency matrix A of a graph $G = (V, E)$ is the $n \times n$ matrix defined by $A_{ij} = 1$ if $ij \in E$ and $A_{ij} = 0$ if $ij \notin E$. Note that the adjacency matrix is always symmetric.

For any square diagonalizable $n \times n$ matrix M we will denote the eigenvalues by

$$\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_n(M). \tag{2.1}$$

By convention, we will in some cases write λ_{min} instead of λ_n. For a matrix M we define the matrix norm

$$\|M\| = \sup_{x,\, y} \frac{x^T M y}{\|x\|\|y\|}. \tag{2.2}$$

Recall that for symmetric matrices, this equals $\max_i |\lambda_i(M)|$.

The results derived will consider asymptotic properties of random graphs. This means that the random graph model considered will be likely to have said property provided the number of vertices n is sufficiently large. We will use the following notation:

- $f(n) = O(g(n))$ iff $\limsup_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| < \infty$
- $f(n) = \Omega(g(n))$ iff $g(n) = O(f(n))$
- $f(n) = o(g(n))$ iff $\limsup_{n\to\infty} \left|\frac{f(n)}{g(n)}\right| = 0$
- $f(n) \gg g(n)$ iff $g(n) = o(f(n))$
- $f(n) \ll g(n)$ iff $g(n) \gg f(n)$

In particular, if an event occurs with probability $1 - o(1)$ we say that the event occurs *with high probability*.

2.1 Random graphs models

We will use $G(n, p)$ to denote a random graph on n vertices where each of the $\binom{n}{2}$ potential edges are included independently with probability p. This model for generating random graphs is commonly referred to as the Erdős-Rényi model.

$G(n, p, k)$ denotes the random graph constructed by planting a clique on a random vertex set of size k in $G(n, p)$, i.e. picking k vertices at random with uniform distribution and adding all edges between these vertices. Equivalently, we can see $G(n, p, k)$ as the graph generated by first planting a clique on a random vertex set of size k on the empty graph of size n and then independently adding each remaining potential edge with probability p.

We use $\bar{G}(n, p, k)$ to denote the complement graph of $G(n, p, k)$. We will refer to the k chosen vertices as the *planted clique*, the *planted independent set* when considering the complement graph, or simply the *planted set*. To simplify notation, we will always assume that the planted set is the vertices $\{1, \dots k\}$.

2.2 Spectrum of random matrices

A central building block of the proofs in this thesis is bounding the eigenvalues of certain symmetric random matrices. In a famous article by Füredi and Komlós[13] it is shown that under certain conditions, all eigenvalues of a random symmetric matrix with independent entries are bounded in absolute value by $(2 + \epsilon)\sigma\sqrt{n}$ with high probability for any $\epsilon > 0$, where σ is the standard deviation of the individual elements in the matrix. The following Theorem is an improved result due to Vu[14]:

Theorem 2.1. *(Vu, 2007) There are constants C and C' such that the following holds. Let r_{ij}, $1 \leq i \leq j \leq n$ be independent random variables, each of which has mean 0 and variance at most σ^2 and is bounded in absolute value by K, where $\sigma \geq C' n^{-1/2} K (\ln n)^2$. Then with high probability*

$$\|R\| \leq 2\sigma\sqrt{n} + C(K\sigma)^{1/2} n^{1/4} \ln n. \tag{2.3}$$

■

To simplify the usage of this theorem we derive the following corollary:

Corollary 2.2. *Let C' be as above and assume $p = p(n)$ satisfies $p(1-p) \geq C'^2 n^{-1} (\ln n)^4$. Let $R = \{r_{ij}\}_{i,j=1}^n$ be any $n \times n$ symmetric random zero-mean matrix where r_{ij} for $i \geq j$ are independent random variables which are either constantly zero or assumes the value $p - 1$ with probability p and the value p with probability $1 - p$. Then*

$$\|R\| \leq 2\sqrt{np(1-p)} \left(1 + O\left(\left(\frac{(\ln n)^4}{np(1-p)}\right)^{1/4}\right)\right) \tag{2.4}$$

with high probability. ■

2.3 Other high concentration results

Besides eigenvalues, our results require bounding some additional random quantities. For example, it turns out that the number or neighbors the vertices in $V \setminus S$ has in S is very important for whether various methods manages to recognize the planted independent set or not.

As the model we consider includes edges not in the planted set independently, this type of quantity can be seen as a sum of independent random variables. For this case there are various known strong bounds on the tail distribution. In particular, in this thesis we use the following results:

Lemma 2.3. *[15] Let $X \sim Bin(n, p)$. For every $0 < a \leq np$ we have*

$$Pr\left[|X - np| \geq a\right] \leq 2e^{-\frac{a^2}{3kp}}. \tag{2.5}$$

■

Theorem 2.4. *(Hoeffding's inequality[16]) Let the random variables $X_1, \ldots, X_n$ be independent, with $a_k \leq X_k \leq b_k$ for each k, for suitable constants a_k, b_k. Let $S_n = \sum X_k$ and let $\mu = \mathbb{E}[S_n]$. Then for any $t \geq 0$,*

$$Pr\left[S_n - \mu \geq t\right] \leq e^{-2t^2/\sum(b_k - a_k)^2}. \tag{2.6}$$

■

Chapter 3

Lovász theta function

In 1979, László Lovász published an article in which he computed the Shannon zero-error capacity of some specific graphs[17]. Let us say we are given a graph where the vertices represents the letters in an alphabet and each edge represents a pair of letters that can be confused, and we want to choose as many strings in this alphabet as possible of a fix length such that no two strings can be confused. We denote the maximum number of strings of length k we can choose by $\alpha(G^k)$. For length 1, this is clearly the independence number of the graph, so $\alpha(G^1) = \alpha(G)$. For a general length k one could therefore suspect that the the maximum number of strings is $\alpha(G)^k$, but it turns out that one can do better in many cases. Instead, as k grows we can show that the "effective number of letters", $\sqrt[k]{\alpha(G^k)}$ will converge to some quantity $\Theta(G)$ which we call the Shannon zero-error capacity of G.

For given graphs, one can bound the Shannon capacity from below by constructing some good coding scheme. For instance, by choosing all strings that only consist of letters from a fix maximum independent set we see that $\alpha(G) \leq \Theta(G)$. The novelty of Lovász approach was that he presented an optimization problem which he showed was an upper bound on the Shannon capacity. This function has later been referred to as Lovász ϑ-function, or simply the Lovász number of the graph and has been well-studied. See [4] for a survey.

Lovász originally defined this function in terms of orthonormal representations of graphs. We say that two vertices i, j in a graph $G = (V, E)$ are adjacent iff $i = j$ or $ij \in E$. An *orthonormal representation* of a G is a system of unit vectors in a Euclidean space $(v_1, \ldots v_n)$ satisfying that if i and j are non-adjacent vertices of G then v_i and v_j are pairwise orthogonal. It is clear from this definition that any graph G has an orthonormal representation, one can for instance take a system of n pairwise orthonormal vectors. Define the *value* of an orthogonal representation $(u_1, \ldots u_n)$ to be

$$\min_{c} \max_{1 \leq i \leq n} \frac{1}{(c^T u_i)^2}, \tag{3.1}$$

where c goes over all unit vectors. Then $\vartheta(G)$ is defined as the minimum value of any orthonormal representation of G.

In his article, Lovász went on to show a number of equivalent formulations for $\vartheta(G)$. It can for instance be formulated as an SDP problem, which means that $\vartheta(G)$ can be computed in polynomial time with an arbitrarily small multiplicative or additive error[3].

To see that this function is an upper bound on the independence number, note that the vertices of an independent set is always represented by pairwise orthogonal vectors. Denoting an independent set by S, Bessel's inequality implies that for any normal vector c we have

$$\sum_{i \in S} (c^T u_i)^2 \leq 1 \tag{3.2}$$

hence for any orthonormal representation and any normal vector c there exists a vertex i such that $(c^T u_i)^2 \leq |S|^{-1}$ and thus the value of any orthonormal representation is at least $|S|$.

It was also shown by Lovász that $\vartheta(G)$ is a lower bound on minimum clique cover, which is known to be NP-hard. This means that Lovász ϑ-function has the property of being sandwiched between two NP-hard problems, while itself being computable in polynomial time[4].

3.1 The Lovász number of planted independent set graphs

Feige and Krauthgamer[8] presented a method for recovering the planted set of size $c\sqrt{n}$ based on Lovász ϑ-function. A central building block of their result is the following property of the Lovász number of random planted independent set graphs:

Lemma 3.1. *(Feige and Krauthgamer 2000) Let $G = G(n, \frac{1}{2}, k)$ where $k > c'\sqrt{n}$ for a large enough constant c'. Then $\vartheta(\bar{G}) = k$, with extremely high probability.* ■

The term *with extremely high probability* used here denotes a probability of at least $1 - e^{-n^r}$ for some $r > 0$.

The idea behind their algorithm is that if one removes one vertex from $G = G(n, \frac{1}{2}, k)$, the remaining graph will be distributed as $G(n-1, \frac{1}{2}, k-1)$ if the vertex was in the planted set and $G(n-1, \frac{1}{2}, k)$ otherwise. By choosing k large enough Lemma 3.1 implies that $\vartheta(\bar{G} \setminus \{v\})$ equals $k-1$ if v is in the planted set and k otherwise. By computing the Lovász number of these graphs within a error of, say, $\pm\frac{1}{3}$ we can therefore identify which vertices are in S. It should however be noted that the final algorithm presented by Feige et al. only computes Lovász ϑ-function and instead identifies the vertices in S by inspecting the optimal solution.

In their concluding remarks, Feige and Krauthgamer stated that they believed their result could be extended to work for $G(n, 1-p, k)$. In particular, for $p = o(1)$ they believed the corresponding bound on k would be $c'\sqrt{n/p}$.

In the next section we prove the following extension of their Lemma:

Theorem 3.2. *Let $G = G(n, 1-p, k)$ where $k > (2 + o(1))\sqrt{n(1-p)/p}$ and $p = p(n)$ satisfies $p(1-p) \gg n^{-1}(\ln n)^4$. Then $\vartheta(\bar{G}) = k$ with high probability.*

We will for simplicity not attempt to prove that this holds with extremely high probability. A more rigorous study of the probabilities involved would however be needed to show that this can be used to retrieve the planted set.

It is not hard to see from the from the definition of Lovász ϑ-function that it is decreasing, i.e. removing edges will always increase the Lovász number. This means that we can choose distributions such that $\vartheta\left(G(n,p)\right) \leq \vartheta\left(\bar{G}(n,1-p,k)\right)$ always holds. From this we see that $\vartheta\left(\bar{G}(n,1-p,k)\right)$ can only equal k if $k \geq \vartheta\left(G(n,p)\right)$. The Lovász number of Erdös-Rényi graphs have been studied by Coja-Oghlan[18]. In his article, he shows that for p as in Theorem 3.2 the Lovász number of $G(n,p)$ is with high probability bounded from below by $c\sqrt{n(1-p)/p}$ for some constant $c > 0$. This implies that the bound on k presented in Theorem 3.2 can at most be improved by a constant factor.

3.2 Proof of Theorem 3.2.

We will begin by presenting an outline the proof. This approach similar to the proof of Lemma 3.1.

We will use one of the alternative formulations of $\vartheta(G)$ given in Lovász original article[17]:

Theorem 3.3. *(Lovász 1979) Let G be a graph on vertices $\{1, \ldots n\}$. Then $\vartheta(G)$ is the minimum of the largest eigenvalue of any symmetric matrix $\{a_{ij}\}_{i,j=1}^{n}$, such that*

$$a_{ij} = 1, \qquad \text{if } i = j \text{ or if } i \text{ and } j \text{ are nonadjacent.} \tag{3.3}$$

■

It turns out that for $\bar{G}(n,1-p,k)$ it is possible to solve this optimization problem explicitly with high probability for sufficiently large k. Specifically, let A be the adjacency matrix of $\bar{G}$ and consider the matrix M defined by

$$M_{ij} = \begin{cases} 1 & \text{if } 1 \leq i,j \leq k \\ 1 - r_j A_{ij} & \text{if } 1 \leq i \leq k < j \leq n \\ 1 - r_i A_{ij} & \text{if } 1 \leq j \leq k < i \leq n \\ \frac{1}{p}\left(p - A_{ij}\right) & \text{if } k < i,j \leq n \end{cases} \tag{3.4}$$

where r_i is chosen such that $\sum_{j=1}^{k} M_{ij} = 0$ for all $i > k$. Equivalently, r_i must satisfy

$$k - S_i r_i = 0 \tag{3.5}$$

where S_i is the number of neighbors of i in the planted independent set. The case $S_i = 0$ for some $i > k$ may be resolved arbitrarily. Clearly, this matrix is feasible.

Let e_k denote the n-dimensional vector where the topmost k entries are 1 and the rest 0. By construction of M, it has e_k as an eigenvector with corresponding eigenvalue k. Since $\bar{G}$ contains an independent set of size k we know that $k \leq \alpha(\bar{G}) \leq \vartheta(\bar{G})$. This

means that if we can show that k is the maximum eigenvalue of M, then $\vartheta(\bar{G}) = k$. To prove this, it suffices to show that all but at most one eigenvalue of M is strictly less than k. The theorem now follows from Lemma 3.6. ■

Before we prove Lemma 3.6 we need some elementary machinery. As before, let A denote the adjacency matrix of $\bar{G}$ and consider the two matrices

$$U_{ij} = \begin{cases} 0 & \text{if } 1 \le i, j \le k \\ \frac{1}{p}(p - A_{ij}) & \text{if } 1 \le i \le k < j \le n \\ \frac{1}{p}(p - A_{ij}) & \text{if } 1 \le j \le k < i \le n \\ \frac{1}{p}(p - A_{ij}) & \text{if } k < i, j \le n \end{cases} \tag{3.6}$$

$$V_{ij} = \begin{cases} 0 & \text{if } 1 \le i, j \le k \\ \left(r_j - \frac{1}{p}\right)(p - A_{ij}) & \text{if } 1 \le i \le k < j \le n \\ \left(r_i - \frac{1}{p}\right)(p - A_{ij}) & \text{if } 1 \le j \le k < i \le n \\ 0 & \text{if } k < i, j \le n \end{cases} \tag{3.7}$$

where r_i are the same as in (3.4).

Lemma 3.4. *Let U be defined as above, then if $p(1-p) \gg n^{-1}(\ln n)^4$*

$$\lambda_1(U) \le (2 + o(1))\sqrt{\frac{n(1-p)}{p}} \tag{3.8}$$

with high probability.

Proof. Let D be the diagonal matrix where $D_{ii} = 0$ for $1 \le i \le k$ and $D_{ii} = 1$ for $k < i \le n$. The matrix $p(U - D)$ satisfies the conditions of Corollary 2.2. Hence

$$\lambda_1(U) \le \|U - D\| + \|D\| \le (2 + o(1))\sqrt{\frac{n(1-p)}{p}} \tag{3.9}$$

with high probability. ■

Lemma 3.5. *Let V be defined as above, then if $p(1-p) \gg n^{-1}(\ln n)^4$ and $k > 2\sqrt{n(1-p)/p}$*

$$\lambda_1(V) \le o\left(\sqrt{\frac{n(1-p)}{p}}\right) \tag{3.10}$$

with high probability.

Proof. Consider the $n \times n$ matrix V' defined by

$$V'_{ij} = \begin{cases} 0 & \text{if } 1 \le i, j \le k \\ \frac{1}{p}(p - A_{ij}) & \text{if } 1 \le i \le k < j \le n \\ \frac{1}{p}(p - A_{ij}) & \text{if } 1 \le j \le k < i \le n \\ 0 & \text{if } k < i, j \le n. \end{cases} \tag{3.11}$$

As in the previous proof, we note that pV' satisfies the conditions of Corollary 2.2. Hence $\|V'\| = O\sqrt{n(1-p)/p}$.

Let x be a unit eigenvector corresponding to the maximum eigenvalue of V. This means that

$$\begin{aligned}\lambda_1(V) &= x^T V x \\ &= 2\sum_{i=k+1}^{n} x_i \left(r_i - \frac{1}{p}\right) \sum_{j=1}^{k} (p - A_{ij})\, x_j.\end{aligned}$$

By Cauchy-Schwartz inequality, this is less than

$$\begin{aligned}&2\left(\sum_{i=k+1}^{n} x_i^2 \cdot \sum_{i=k+1}^{n} \left(r_i - \frac{1}{p}\right)^2 \left(\sum_{j=1}^{k} (p - A_{ij})\, x_j\right)^2\right)^{1/2} \\ &\leq 2\left(\sum_{i=k+1}^{n} (pr_i - 1)^2 \left(\sum_{j=1}^{k} \frac{1}{p}(p - A_{ij})\, x_j\right)^2\right)^{1/2} \\ &= 2\left(\sum_{i=k+1}^{n} (pr_i - 1)^2 \left(\sum_{j=1}^{k} V'_{ij} x_j\right)^2\right)^{1/2} \\ &\leq 2\max_i |pr_i - 1| \cdot \left(\sum_{i=k+1}^{n} \left(\sum_{j=1}^{k} V'_{ij} x_j\right)^2\right)^{1/2}.\end{aligned}$$

Since $V'_{ij} = 0$ for $i, j > k$, this equals

$$\begin{aligned}&2\max_i |pr_i - 1| \cdot \left(\sum_{i=k+1}^{n} \left(\sum_{j=1}^{n} V'_{ij} x_j\right)^2\right)^{1/2} \\ &\leq 2\max_i |pr_i - 1| \cdot \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{n} V'_{ij} x_j\right)^2\right)^{1/2} \\ &= 2\max_i |pr_i - 1| \cdot \|V'x\| \\ &= \max_i |pr_i - 1|\, O\left(\sqrt{\frac{n(1-p)}{p}}\right),\end{aligned}$$

where the last step follows from that $\|V'x\| \leq \|V'\| \cdot \|x\| = \|V'\|$.

By the definition of M, we know that $r_i = k/S_i$, where $S_i \sim Bin(k, p)$. Under the assumptions of this Lemma it is easy to verify that $kp \gg \sqrt{kp \ln n}$. Hence Lemma 2.3

with $a = \sqrt{6kp \ln n}$ implies that $r_i = p^{-1} + o(p^{-1})$ for all $k < i \leq n$ with probability $1 - O(\frac{1}{n})$. This means that $\max_i |pr_i - 1| = o(1)$ and thus $\lambda_1(V) \leq o\left(\sqrt{n(1-p)/p}\right)$ with high probability. ∎

Lemma 3.6. *Let M be defined as in (3.4) and assume $p(1-p) \gg n^{-1}(\ln n)^4$. Then there exists a function $f(n,p) = (2+o(1))\sqrt{n(1-p)/p}$ such that $\lambda_2(M) \leq f$ with high probability whenever $k > f$.*

Proof. Let e_k denote the n-dimensional vector where the first k elements are 1 and the rest 0. By the variational inequality

$$\begin{aligned}\lambda_2(M) &= \min_v \max_{|x|=1\, x\perp v} x^T M x \\ &\leq \max_{|x|=1\, x\perp e_k} x^T M x \\ &\leq \lambda_1(U) + \lambda_1(V) + \max_{|x|=1\, x\perp e_k} x^T(M-U-V)x.\end{aligned}$$

Recalling the definitions of M, U and V, see (3.4) (3.6) (3.7), we note that for all $x \perp e_k$

$$\begin{aligned}&x^T(M-U-V)x \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} x_i(M_{ij}-U_{ij}-V_{ij})x_j + 2\sum_{i=k+1}^{n}\sum_{j=1}^{k} x_i(M_{ij}-U_{ij}-V_{ij})x_j \\ &\quad + \sum_{i=k+1}^{n}\sum_{j=k+1}^{n} x_i(M_{ij}-U_{ij}-V_{ij})x_j \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} x_i x_j + 2\sum_{i=k+1}^{n}\sum_{j=1}^{k} x_i\left(1 - r_iA_{ij} - \frac{1}{p}(p-A_{ij}) - \left(r_i - \frac{1}{p}\right)(p-A_{ij})\right)x_j \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} x_i x_j + 2\sum_{i=k+1}^{n}\sum_{j=1}^{k} x_i(1-pr_i)x_j \\ &= \sum_{i=1}^{k} x_i \underbrace{\left(\sum_{j=1}^{k} x_j\right)}_{=0} + 2\sum_{i=k+1}^{n} x_i(1-pr_i)\underbrace{\left(\sum_{j=1}^{k} x_j\right)}_{=0} = 0\end{aligned}$$

so $\max_{|x|=1\, x\perp e_k} x^T(M-U-V)x = 0$. Thus if we assume that $k > 2\sqrt{n(1-p)/p}$, Lemma 3.4 and 3.5 implies that

$$\lambda_2(M) \leq \lambda_1(U) + \lambda_1(V) \leq (2+o(1))\sqrt{\frac{n(1-p)}{p}}. \tag{3.12}$$

The Lemma follows by letting f the maximum of this expression and $2\sqrt{n(1-p)/p}$. ∎

Chapter 4

A Convex Quadratic Relaxation of the Lovász number

While technically being solvable in polynomial time, SDP problems, such as $\vartheta(G)$, are notorious for being computer-intensive to solve. Hence, computing $\vartheta(G)$ can be infeasible for even moderately sized graphs. It would therefore be of great interest to find other bounds of the independence number which are faster to compute.

An alternative function was presented by Luz[19]. Assume G has at least one edge. We define

$$\upsilon(G) = \max_{\alpha \geq 0} 2e^T\alpha - \alpha^T \left(I + \frac{A}{-\lambda_{min}(A)} \right) \alpha, \tag{4.1}$$

where e denotes the vector of all ones.

This special case of a convex quadratic optimization problem can be seen as a support vector machine, SVM, problem. This type of problem has been studied in machine learning. It has been shown that the optimization step is solvable with arbitrary precision in $O(n^2)$ time[20], with potentially even faster convergence in practice. However, unless the minimum eigenvalue of A can be guessed for the specific graph, an implementation of $\upsilon(G)$ would besides this also need some to compute $\lambda_{min}(A)$.

Luz showed that this is an upper bound on the independence number of G. In fact, Luz and Schrijver[21] generalized $\upsilon(G)$ to a class of convex optimization problems and showed that all form upper bounds on $\vartheta(G)$. It could therefore also be of interest to use $\upsilon(G)$ as a lightweight characterization of $\vartheta(G)$.

4.1 Graph for which this function equals the independence number

Similarly to $\vartheta(G)$ there is a class of graphs satisfying $\alpha(G) = \upsilon(G)$. We will call these graphs *Q-graphs*. Luz gave the following characterization of such graphs:

Theorem 4.1. *(Luz 1995) Let G be a graph with at least one edge. Then, the equality $\alpha(G) = \upsilon(G)$ is true, if and only if for any maximum independent set S the following*

inequality holds:

$$-\lambda_{min}(A) \leq \min\{|N(i) \cap S| : i \notin S\}. \tag{4.2}$$

Proof sketch. Let S be any maximum independent set of G and let x denote the indicator vector of S, i.e. the vector where $x_i = 1$ if $i \in S$ and $x_i = 0$ otherwise. Since S is an independent set this means that $A_{ij}x_ix_j = 0$ for all i, j.

The key to proving the Theorem is to note that

$$2e^Tx - x^T\left(I + \frac{A}{-\lambda_{min}(A)}\right)x = 2|S| - (|S| + 0) = \alpha(G), \tag{4.3}$$

so $\upsilon(G) = \alpha(G)$ iff x is a global maximum of (4.1). Since this is a convex optimization problem over a polyhedral cone, necessary and sufficient conditions for optimality is the Karush-Kuhn-Tucker conditions which for x becomes exactly (4.2). ■

To make the criterion in 4.1 more tractable, we use the sufficiency of the KKT condtions again to derive the following statement:

Theorem 4.2. *Let G be a graph with at least one edge and let S be an independent set in G. If S satisfies 4.2 then S is a maximum independent set and $\alpha(G) = \upsilon(G) = |S|$.*■

4.2 Performance for planted independence set graphs

The contribution from this thesis here is to investigate in what regimes $G = \bar{G}(n, 1-p, k)$ is a Q-graph. We do this by applying Theorem 4.2 on $G = \bar{G}(n, 1-p, k)$ for S being the planted set. This yields the following result:

Theorem 4.3. *There exist constants C and C' such that for $p(1-p) \geq C'n^{-1}(\ln n)^4$ the graph $G = \bar{G}(n, 1-p(n), k)$ is a Q-graph with high probabilty if*

$$k \geq C \max\left(\left(\frac{n^2(1-p)\ln n}{p}\right)^{1/3}, n^{1/2}p^{-1/2}(\ln n)^{1/4}\right). \tag{4.4}$$

Unfortunately this lower bound on k is much larger than for Theorem 3.2. In fact, it turns out that for any k and p as in Theorem 4.3, the planted set can be retrieved by trivial means, such as sorting the vertices by degrees. This makes the applicability of $\upsilon(G)$ in this setting questionable.

4.3 Proof of Theorem 4.3.

Before proving Theorem 4.3 we will need a lower bound on the minimum eigenvalue of G.

Let A denote the adjacency matrix of G. Consider the matrix $A' = A + D$ where D is the diagonal matrix satisfying $D_{ii} = 0$ for $1 \leq i \leq k$ and $D_{ii} = p$ for $k < i \leq n$. Let $\hat{A}$ denote $A - \mathbb{E}A$. Furthermore, we define

$$\bar{\lambda} = -kp + \frac{3k^2p}{4n}. \tag{4.5}$$

Lemma 4.4. *Let A' and $\bar{\lambda}$ be as above. Then $\lambda_{min}(\mathbb{E}A') \geq \bar{\lambda}$.*

Proof. We begin by noting that all vectors in the column space of $\mathbb{E}A'$ are constant on S and $V \setminus S$ respectively. Let x be the unit vector having value $(\sin t)/\sqrt{k}$ on S and $(\cos t)/\sqrt{n-k}$ the complement of S. Clearly all unit vectors in the column space can be written on this form. Since all eigenvectors corresponding to non-zero eigenvalues must be in this vector space, the minimum eigenvalue of $\mathbb{E}A'$ is either 0 or the minimum of

$$\begin{aligned} x^T\mathbb{E}[A']x &= 2\sqrt{k(n-k)}p\sin t\cos t + (n-k)p\cos^2 t \\ &= \frac{(n-k)p}{2} + \sqrt{k(n-k)}p\sin 2t + \frac{(n-k)p}{2}\cos 2t \\ &\geq \frac{(n-k)p}{2} - \frac{p}{2}\sqrt{4k(n-k)+(n-k)^2} \\ &\geq \frac{(n-k)p}{2} - \frac{p}{2}\left(n + \frac{2nk-3k^2}{2n}\right) = -kp + \frac{3k^2p}{4n} \end{aligned}$$

where the last inequality follows by Taylor expanding the square root at n^2. Since the last expression is non-positive we conclude that $-kp + 3k^2p/4n$ is a lower bound on the eigenvalues of $\mathbb{E}A'$. ■

Lemma 4.5. *For any fix non-zero vector x and any $\delta > 0$ we have*

$$Pr\left[x^T\hat{A}x \leq -\delta\left\|x\right\|^2\right] \leq e^{-\delta^2}. \tag{4.6}$$

Proof. Consider the random variable

$$X = -x^T\hat{A}x = \sum_{i>j} -2x_ix_j\hat{A}_{ij}, \tag{4.7}$$

where we note that the sum goes over independent random variables. Since $\hat{A}_{ij}$ varies by at most by 1, $2x_ix_j\hat{A}_{ij}$ varies at most $|2x_ix_j|$ where

$$\sum_{i>j}(2x_ix_j)^2 = 2\sum_{i\neq j}x_i^2x_j^2 \leq 2\|x\|^4.$$

Thus Theorem 2.4 applied on X states that for any $t > 0$

$$Pr[X \geq t] \leq \exp\left(-\frac{t^2}{\|x\|^4}\right). \tag{4.8}$$

The statement is obtained by letting $t = \delta\|x\|^2$. ■

Lemma 4.6. *Let A be the adjacency matrix of $G = \bar{G}(n, 1-p, k)$ and let $\hat{A}$ denote $A - \mathbb{E}A$. Then with probability at least $1 - \frac{1}{n}$*

$$-\lambda_{min}(A) \leq kp - \frac{9k^2p}{16n} + \frac{\|\hat{A}\|^2}{kp} + \sqrt{\ln n} + p. \tag{4.9}$$

Proof. Let v be a normalized eigenvector corresponding to the minimum eigenvalue of $\mathbb{E}A'$ and let $u = x + y$ be any unit vector, where $x \parallel v$ and $y \perp v$. Lemma 4.4 implies that $x^T \mathbb{E}A' x \geq \bar{\lambda}\|x\|^2$. Since $\mathbb{E}A'$ is a rank 2 matrix and clearly $\mathrm{Tr}\,(\mathbb{E}A') \geq 0$, $\mathbb{E}A'$ can at most have one negative eigenvalue and thus $y^T \mathbb{E}[A']y \geq 0$. Furthermore, Lemma 4.5 states that $v^T \hat{A} v \geq -\sqrt{\ln n}$ with probability at least $1 - \frac{1}{n}$. Hence,

$$\begin{aligned}
\lambda_{min}(A) &= \min_{\|u\|=1} u^T \left(\mathbb{E}A' + \hat{A} - D\right) u \\
&\geq \min_{\|u\|=1} \bar{\lambda}\|x\|^2 + (2x+y)^T \hat{A} y + x^T \hat{A} x - u^T D u \\
&\geq \min_{x_1^2+x_2^2=1} \left(\bar{\lambda} x_1^2 - \|\hat{A}\|\sqrt{4x_1^2 + x_2^2}\, x_2\right) - \sqrt{\ln n} - p,
\end{aligned}$$

where we have used that $x \perp y$. By substituting $x_2 = \sqrt{4/3}\cos t$ in the first term we get

$$\begin{aligned}
\bar{\lambda}(1 - x_2^2) - \|\hat{A}\|\sqrt{4 - 3x_2^2}\, x_2 &= \frac{\bar{\lambda}}{3} - \frac{2}{3}\left(\bar{\lambda}\cos 2t + \sqrt{3}\|\hat{A}\|\sin 2t\right) \\
&\geq \frac{\bar{\lambda}}{3} - \frac{2}{3}\sqrt{\bar{\lambda}^2 + 3\|\hat{A}\|^2}
\end{aligned}$$

so by Taylor expanding the square root at k^2p^2 we conclude that

$$\begin{aligned}
\min_{x_1^2+x_2^2=1} \left(\bar{\lambda} x_1^2 - \|\hat{A}\|\sqrt{4x_1^2 + x_2^2}\, x_2\right) &\geq \frac{\bar{\lambda}}{3} - \frac{2}{3}\left(kp + \frac{\bar{\lambda}^2 + 3\|\hat{A}\|^2 - k^2p^2}{2kp}\right) \\
&= -kp + \frac{k^2 p}{n}\left(\frac{3}{4} - \frac{3k}{16n}\right) - \frac{\|\hat{A}\|}{kp}.
\end{aligned}$$

Hence with probability at least $1 - \frac{1}{n}$

$$\lambda_{min}(A) \geq -kp + \frac{k^2 p}{n}\left(\frac{3}{4} - \frac{3k}{16n}\right) - \frac{\|\hat{A}\|}{kp} - \sqrt{\ln n} - p. \tag{4.10}$$

■

Proof of Theorem 4.3. According to Theorem 4.2, to prove that G is a Q-graph it suffices to show that (4.2) is satisfied for S denoting the planted set.

For all $i \notin S$, $|N(i) \cap S|$ clearly has distribution $Bin(k, p)$. We use Lemma 2.3 to bound these uniformly. For $p \leq \frac{1}{2}$ we apply the lemma directly for $a = \sqrt{6kp \ln n}$. For $p > \frac{1}{2}$ we instead consider $Y = k - |N(i) \cap S|$ where clearly $Y \sim Bin(k, 1-p)$ and apply the lemma om Y for $a = \sqrt{6k(1-p)\ln n}$. We can summarise the result as

$$|N(i) \cap S| \geq kp - \sqrt{12kp(1-p)\ln n} \tag{4.11}$$

for all $i \notin S$ with probability at least $1 - O\left(\frac{1}{n}\right)$. Under the assumption on k it is easy to see that the conditions in Lemma 2.3 is satisfied for sufficiently large C'.

Putting (4.11) into (4.2) and applying Lemma 4.4 we get that G is a Q-graph with probability $1 - O(n^{-1})$ if

$$kp - \frac{9k^2}{16n} + \frac{\|A - \mathbb{E}A\|^2}{kp} + \sqrt{\ln n} + p \leq kp - \sqrt{12kp(1-p)\ln n}. \tag{4.12}$$

According to Corollary 2.2, $\|A - \mathbb{E}A\|^2 = O\left(p(1-p)n\right)$ with high probability for sufficiently large C'. Using this, we see that G is a Q-graph with high probability if

$$\frac{9k^2p}{16n} \geq \sqrt{12kp(1-p)\ln n} + \sqrt{\ln n} + O\left(\frac{n(1-p)}{k}\right). \tag{4.13}$$

Under the assumption on k,

$$\begin{aligned}
&\sqrt{12kp(1-p)\ln n} + \sqrt{\ln n} + O\left(\frac{n(1-p)}{k}\right) \\
&\quad = \left(\sqrt{12} + O\sqrt{\frac{n^2(1-p)}{k^3 p \ln n}}\right)\sqrt{kp(1-p)\ln n} + \sqrt{\ln n} \\
&\quad = \left(\sqrt{12} + O\left(\frac{1}{\ln n}\right)\right)\sqrt{kp(1-p)\ln n} + \sqrt{\ln n}.
\end{aligned}$$

Thus (4.13) is satisfied with high probability if for sufficiently large $C_1, C_2 > 0$, $k^2p/n \geq C_1\sqrt{kp(1-p)\ln n}$ and $k^2p/n \geq C_2\sqrt{\ln n}$. Solving these expressions for k we see that (4.13) is satisfies with high probability if for sufficiently large $C > 0$,

$$k \geq C\left(\frac{n^2(1-p)\ln n}{p}\right)^{1/3} \tag{4.14}$$

and

$$k \geq Cn^{1/2}p^{-1/2}(\ln n)^{1/4}. \tag{4.15}$$

■

Chapter 5

Certifying optimality of independent sets

Using Lemma 3.1, Feige and Krauthgamer[8] showed that it with extremely high probability is possible to certify the optimality of the maximum independent set in $\bar{G}(n, 1-p, k)$ in polynomial time. This can be done by computing $\vartheta(\bar{G})$. This is indeed an interesting property as it seems counter-intuitive to be possible to do without some kind of exhaustive search. In fact, it is not too hard to show that the problem of certifying maximum independent sets in a general graph is NP-hard.

Since the time the article by Feige and Krauthgamer was published, advancements have been made on the problem of retrieving the planted set and the algorithms proposed by Feige and Ron[9] and Dekel, Gurel-Gurevich, and Peres[10] are both considerably faster and simpler. However, to the author's knowledge no improvements have been made on the problem of certifying the maximum independent set.

Seeing that fast algorithms for retrieving the planted set is known, it makes sense to treat the problem of certifying optimality separately (the algorithm by Feige and Krauthgamer retrieves the planted set and certifies its optimality at the same time). We therefore consider the simplified problem where we are given a graph generated according to the planted independent set model and a candidate for the maximum independent set.

This modification can simplify the problem substantially. For example, from the proof of Theorem 3.2 can observe that there are instances where calculating $\vartheta(\bar{G})$ can be performed by solving the optimization problem in Theorem 3.3 explicitly, provided the planted set is known. For using Lovász ϑ-function to retrieve the planted set it is of course nonsensical to try to speed up the computation of $\vartheta(G)$ using this property, but it could be very useful for certifying optimality of the planted set.

Inspired by this, we devise a new algorithm for certifying optimality of independent sets in $\bar{G}(n, 1-p, k)$. Using the ideas of the proof of Theorem 3.2 we reduce the problem of certifying that an independent set is maximum to certifying that a certain matrix is positive definite, which is known to be possible to do in $O(n^3)$ time by means of Cholesky decomposition.

To conclude the chapter, we discuss another method for certifying independent sets, which reduces the problem of certifying the optimality of S to bounding $\alpha(G \setminus S)$. As $G \setminus S$ is an Erdös-Rényi graph, provided the input is correct, this could potentially be done in numerous ways and leaves room for various improvements. However, in its current form the author sees no benefit of using this approach as opposed to the first algorithm.

5.1 Reduction to certifying positive definiteness

The algorithm is given a graph $G = (V, E)$ and a candidate for the maximum independent set S. We will for simplicity assume that $V = \{1, \dots n\}$ and $S = \{1, \dots k\}$. It outputs *YES* if it manages to certify optimality, *NO* if it can disprove optimality, and *UNKNOWN* otherwise.

Let A denote the adjacency matrix of G. The algorithm is based on the properties of the matrix $N = N(r)$ defined by

$$N_{ij} = \begin{cases} 0 & \text{if } 1 \leq i, j \leq k \\ 1 - r_j A_{ij} & \text{if } 1 \leq i \leq k < j \leq n \\ 1 - r_i A_{ij} & \text{if } 1 \leq j \leq k < i \leq n \\ 1 - r A_{ij} & \text{if } k < i, j \leq n, \end{cases} \tag{5.1}$$

where r_i as above is chosen such that $\sum_{j=1}^{k} N_{ij} = 0$ for all $i > k$. We note that for this definition to be valid we require $|N(i) \cap S| > 0$ for all $k < i \leq n$. Assuming S is a maximal independent set this must hold.

Proposition 5.1. *For any real-valued r, $\max(\lambda_1(N), k)$ is an upper bound on the independence number of G.*

Proof. Let e_k as before denote the vector where the topmost k elements are 1 and the rest 0 and define $M = N + e_k e_k^T$. Since e_k is an eigenvector of N, this perturbation will keep all eigenvalues except one constant, which will change from 0 to k. This means that $\lambda_1(M) = \max(\lambda_1(N), k)$. Furthermore, M is a feasible matrix for the optimization problem in Theorem 3.3, so $\lambda_1(M) \geq \vartheta(G) \geq \alpha(G)$. ■

Corollary 5.2. *If S in an independent set and $\lambda_1(N) \leq k$, then S is a maximum independent set.* ■

Based on the properties of N an algorithm for certifying the optimality of S can be done in two steps. See Appendix for an implementation in Matlab.

1. Check that S is a maximal independent set.

2. If so, construct $N(r)$ and attempt to certify that $\lambda_1(N) \leq k$.

In the cases where S is not a maximal independent set, i.e. S is either not an independent set or not maximal we can clearly output *NO*. If the second step succeeds, Corollary 5.2 guarantees that S is optimal, so we can output *YES* is this case. If S is a maximal independent set but step 2 fails, we simply output *UNKNOWN*. The overall time complexity is dominated by verifying that $\lambda_1(N) \leq k$ which can be performed in $O(n^3)$ steps by Cholesky decomposition on $kI - N$.

So far we have made no assumptions on the input graph. Indeed, this algorithm is correct for any graph (and any r for that matter), but in general it has no guarantee of outputting something other than *UNKNOWN*. To conclude this section, we will show that if G is taken from $\bar{G}(n, 1-p, k)$ and S is the planted set, then the algorithm outputs *YES* with high probability in essentially the same regime for as state-of-the-art methods can retrieve the set.

For this to hold, we will need to choose r appropriately. Similarly to the proof of Theorem 3.2, one possible choice is $r = p^{-1}$. However, p may be unknown. In this case, we will show that we can also take $r = \left(k(n-k) + \binom{n-k}{2}\right) / |E|$.

Proposition 5.3. *Assume G is taken from $\bar{G}(n, 1-p, k)$ where $p(1-p) \gg n^{-1}(\ln n)^4$ and assume S in the planted set. Let $N = N(r)$ be as defined in (5.1) where r is either p^{-1} or $\left(k(n-k) + \binom{n-k}{2}\right) / |E|$. Then there exists a function $f(n,p) = (2+o(1))\sqrt{n(1-p)/p}$ such that $\lambda_1(N) \leq f$ with high probability whenever $|S| = k > f$.*

Proof. Let us define $M(r) = N(r) + e_k e_k^T$. We note that if $\lambda_2(M) < k$ then $\lambda_1(M) = k$ and thus $\lambda_1(N) = \max(\lambda_2(M), 0)$. This means that suffices to show that $\lambda_2(M) < 2\sqrt{n(1-p)/p}\,(1+o(1))$ with high probability. For $r = p^{-1}$ this is exactly Lemma 3.6, so it remains to show that this result can be extended to $r = \left(k(n-k) + \binom{n-k}{2}\right) / |E|$.

Using a simple perturbation argument we see that

$$\lambda_2\left(M(r)\right) \leq \lambda_2\left(M(p^{-1})\right) + \left|r - p^{-1}\right| \|A_{\bar{S}}\| \tag{5.2}$$

where $(A_{\bar{S}})_{ij}$ is given by A_{ij} if $k < i, j \leq n$ and 0 otherwise. By Lemma 3.6 we know that $\lambda_2\left(M(p^{-1})\right) \leq (2+o(1))\sqrt{n(1-p)/p}$ so it remains to show that the second term is $o\sqrt{n(1-p)/p}$ with high probability.

For $k = n$ the proposition trivially holds, so let us assume that $k < n$. Let $L = k(n-k) + \binom{n-k}{2}$ and note that in this case $L \geq n-1$. Since $|E| \sim Bin(L, p)$, Lemma 2.3 implies that $|E| = Lp + O\left(\sqrt{Lp \ln n}\right)$ with high probability. Furthermore, by Theorem 2.1 we know that $\|A_{\bar{S}} - \mathbb{E}A_{\bar{S}}\| + \|\mathbb{E}A_{\bar{S}}\| = O\left(\sqrt{np(1-p)} + (n-k)p\right)$ with high probability. Hence

$$\begin{aligned} |r - p^{-1}| \cdot \|A_{\bar{S}}\| &= O\left(\sqrt{\frac{\ln n}{Lp}}\right) \cdot O\left(\sqrt{np(1-p)} + (n-k)p\right) \\ &= O\left(\sqrt{(1-p)\ln n}\right) + O\left(\sqrt{p \ln n}\right) \ll \sqrt{\frac{n(1-p)}{p}}, \end{aligned}$$

with high probability. ∎

Corollary 5.4. *Assume G is taken from $\bar{G}(n, 1-p, k)$ where $k \geq (2+o(1))\sqrt{n(1-p)/p}$ and $p(1-p) \gg n^{-1}(\ln n)^4$ and assume S in the planted set. Then the certification algorithm outputs* YES *with high probability assuming $r = p^{-1}$ or $r = \left(k(n-k) + \binom{n-k}{2}\right) / |E|$.* ■

5.2 Reduction to bounding the independence number of the Erdös-Rényi graphs

An alternative approach to the problem of certifying optimality of a given independent is to use the following statements:

Proposition 5.5. *Let G be a graph and S a set of vertices. Then any independent set S' in G which is not a subset of S must satisfy*

$$\left|S'\right| \leq \max_{i \notin S} |S \setminus N(i)| + \alpha(G \setminus S). \tag{5.3}$$

Proof. For all S' which are not subsets of S there must exist a vertex i such that $i \in S'$ and $i \notin S$. As i shares an edge with all vertices in $S \cap N(i)$ we know that S' can at most contain $|S \setminus N(i)|$ vertices in S and as $S' \setminus S$ is an independent set in $G \setminus S$ we know that S' can at most contain $\alpha(G \setminus S)$ vertices not in S. The Proposition follows by taking the maximum over all $i \notin S$. ■

Corollary 5.6. *Let G be a graph. If S is an independent set in G satisfying*

$$\alpha(G \setminus S) \leq \min_{i \notin S} |S \cap N(i)| \tag{5.4}$$

then S is a maximum independent set. ■

This type of argument has been given by various sources to argue that the planted set is optimal with high probability. As seen from an algorithmic point of view, Corollary 5.6 states that if we are given $G = \bar{G}(n, \frac{1}{2}, k)$ and the planted set S, then we can prove the optimality of S by certifying that $\alpha(G \setminus S)$ is less than some easily computed quantity which with some reasonable assumptions on k is approximately $k/2$ with high probability. For $k > C\sqrt{n}$ for sufficiently large C, this can be done by computing either of $\vartheta(G \setminus S)$ and $\upsilon(G \setminus S)$ as both these quantities are $O(\sqrt{n})$ with high probability in this case, or use a similar technique to the one presented in the previous section.

This technique gives a nice relation between upper bounds on the independence number and certificates. While we could not use $\upsilon(G)$ directly to compute the size of a planted independent set in the same interval for k as $\vartheta(G)$ we can still use it to certify that the retrieved set is optimal. A natural question is what other possible methods we could use to bound the independence number. In particular: Is there an upper bound on the independence number which is $O(\sqrt{n})$ with high probability for $G(n, \frac{1}{2})$ which can be computed in $o(n^3)$ time? Is there an upper bound on the independence number which is $o(\sqrt{n})$ with high probability for $G(n, \frac{1}{2})$ which can be computed polynomial time? A positive answer to any of these questions would lead to interesting improvements over the algorithm presented in the previous section.

Chapter 6

Concluding Remarks

When discussing certifying optimality of independent sets we stated that we can certify positive definiteness of matrices using Cholesky decomposition. Strictly speaking this requires a more careful consideration of numerical stability. However, this should not be a problem in practice. Firstly, since independence numbers are always integer it follows that we can relax the constraint in Corollary 5.2 to $\lambda_1(N) < k + 1$ which means that we only need to rule out the cases where $kI - N$ are clearly not positive definite. Furthermore, as Proposition 5.3 is formulated we can choose k such that all eigenvalues of $kI - N$ are arbitrarily far from zero with high probability.

The proofs presented in this report have essentially only used the randomness of the graphs involved to bound certain spectral norms and compute concentration of quantities on the form $|S \cap N(i)|$. This suggests that these results could be generalized to a larger class of graphs. One possible generalization would be to instead of planting a random independent set in an Erdös-Rényi graph we plant a random independent set in a general graph. The concentration of $|S \cap N(i)|$ in this setting could probably be derived from the randomness of the planted set, which would imply that the only requirements on the base graph would be bounded spectral norms and bounds on the degrees. The ability to make these types of generalizations could be one big advantage of using spectral methods to approach the independent set problem.

In the section about the Lovász number of $\bar{G}(n, 1-p, k)$ we noted that $\vartheta(\bar{G})$ can only equal k if $k \geq \vartheta\left(G(n,p)\right)$, which is at most a constant factor below the bound on k in Theorem 3.2. The proof relied on an explicit guess on the optimal matrix from Theorem 3.3. A natural improvement would be to replace this guess of M by some version of the optimal matrix. It would therefore seem reasonable that the actual constraint on k is that it should be larger than $(1 + o(1))\,\vartheta\left(G(n,p)\right)$. A problem of showing this is that, since we in that case have no simple bound on the individual terms in the different S_i, we lose the concentration results for $\max |pr_i - 1|$. This problem could possibly be solved by answering one of the following questions. What happens if we only replace the lower right part of M for an optimal matrix for $G \setminus S$? Can anything be said about the optimization problem if we add the constraint that all elements in the matrix must be bounded by for example a large constant times p^{-1}? However, I currently see no way to

complete this modification of the proof.

One related question to this is if the same improvement can be made for a general base graph, i.e. can the needed size of k such that a random planted independent set in a general graph can be recovered be expressed in terms of the Lovász number of the graph. Similarly as for $G(n,p)$, we could replace the spectral norms of U and V' by eigenvalues of matrices related to the optimal solution in Theorem 3.2, but we are still faced with the problem that the terms in S_i have no simple bound.

Bibliography

[1] R. M. Karp. Reducibility Among Combinatorial Problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, and Mario Szegedy. Proof verification and hardness of approximation problems. *Foundations of Computer Science, IEEE Annual Symposium on*, 0:14–23, 1992.

[3] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1:169–197, 1981. 10.1007/BF02579273.

[4] Donald E. Knuth. The Sandwich Theorem. *Electr. J. Comb.*, 1, 1994.

[5] Mark Jerrum. Large Cliques Elude the Metropolis Process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[6] Luděk Kučera. Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57(2-3):193–212, February 1995.

[7] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Struct. Algorithms*, 13(3-4):457–466, October 1998.

[8] Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Struct. Algorithms*, 16(2):195–208, March 2000.

[9] Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. *DMTCS Proceedings*, 0(01), 2010.

[10] Yael Dekel, Ori Gurel-Gurevich, and Yural Peres. Finding Hidden Cliques in Linear Time with High Probability. *ArXiv e-prints*, October 2010.

[11] Alan M. Frieze and Ravi Kannan. A new approach to the planted clique problem. In Ramesh Hariharan, Madhavan Mukund, and V. Vinay, editors, *FSTTCS*, volume 2 of *LIPIcs*, pages 187–198. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2008.

[12] T.-H. Hubert Chan, Kevin L. Chang, and Rajiv Raman. An SDP Primal-Dual Algorithm for Approximating the Lovász-Theta Function. In *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory - Volume 4*, ISIT'09, pages 2808–2812, Piscataway, NJ, USA, 2009. IEEE Press.

[13] Z. Füredi and J. Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981.

[14] Van Vu. Spectral norm of random matrices. *Combinatorica*, 27:721–736, 2007. 10.1007/s00493-007-2190-z.

[15] Michael Molloy. The probabilistic method. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 15–20. Springer-Verlag, Berlin, 1998.

[16] Colin McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic methods for algorithmic discrete mathematics*, volume 16 of *Algorithms and Combinatorics*, pages 202–203. Springer-Verlag, Berlin, 1998.

[17] László Lovász. On the Shannon Capacity of a Graph. *IEEE Transactions on Information Theory*, 25(1):1–7, 1979.

[18] Amin Coja-Oghlan. The Lovász Number of Random Graphs. In Sanjeev Arora, Klaus Jansen, José Rolim, and Amit Sahai, editors, *Approximation, Randomization, and Combinatorial Optimization.. Algorithms and Techniques*, volume 2764 of *Lecture Notes in Computer Science*, pages 5–19. Springer Berlin / Heidelberg, 2003. 10.1007/978-3-540-45198-3_20.

[19] Carlos J. Luz. An upper bound on the independence number of a graph computable in polynomial-time. *Operations Research Letters*, 18(3):139 – 145, 1995.

[20] Don Hush, Patrick Kelly, Clint Scovel, and Ingo Steinwart. QP Algorithms with Guaranteed Accuracy and Run Time for Support Vector Machines. *JOURNAL OF MACHINE LEARNING RESEARCH*, 7:733–769, 2006.

[21] Carlos J. Luz and Alexander Schrijver. A convex quadratic characterization of the Lovász theta number. *Siam Journal on Discrete Mathematics*, 19:382–387, 2005.

Appendix A

Implementation of algorithm presented in Section 5.1

```
% Attempts to certify S as the maximum independent
% A — adjacency matrix
% S — indicator vector for maximum independent set
%
% Output:
%  1 — certified
%  0 — could not certify
% −1 — this is not the maximum independent set

function certified = certIndepset(A, S)
    n = size(A,1);
    k = sum(S);

    % Place independent set as first k elements
    [~, reorder] = sort(−S);
    A = A(reorder, reorder);

    if any(any(A(1:k,1:k)))
        certified = −1; % This is not an independent set
        return
    end

    rvec = k./sum(A(1:k,k+1:n));
    r    = ( k*(n−k) + (n−k)*(n−k−1)/2 )*2/sum(sum(A));

    if any(rvec==Inf)
        certified = −1; % This is not a maximal independent set
        return
    end

    % Construct N
    B = ones(k,n−k) − repmat(rvec,k,1).*A(1:k,k+1:n);
```

```
N = [zeros(k), B; B', ones(n-k) - r*A(k+1:n, k+1:n)];
    % Check if pd. Hopefully numerical instability won't be a problem..
[~, q] = chol(k * eye(n) - N);
if q==0
    certified = 1;
else
    certified = 0;
end
```