On the Hardness of Designing Public Signals*

Shaddin Dughmi
Department of Computer Science
University of Southern California
shaddin@usc.edu

August 2, 2018

Abstract

We use computational complexity as a lens to study the design of information structures in games of incomplete information. We focus on one of the simplest instantiations of the information structure design problem: Bayesian zero-sum games, and a principal who must design a public signal maximizing the equilibrium payoff of one of the players. In this setting, we show that optimal information structure design is computationally intractable, and in some cases hard to approximate, assuming that it is hard to recover a planted clique from an Erdős-Rényi random graph. Our result suggests that there is no "simple" characterization of optimal public-channel information structures in multi-player settings.

1 Introduction

Mechanism design is concerned with designing the rules of a game so as to effect desirable outcomes at equilibrium. This form of intervention, through the design of *incentives*, is fundamentally algorithmic in nature, and has led to a large body of work which examines the computational complexity of incentive-compatible mechanisms. This paper concerns a different, though arguably equally important, mode of intervention: through making available the right *information*.

A classic example illustrating the importance of information in games is Akerlof's "market for lemons" [1]. Each seller in this market is looking to sell a used car which is equally likely to be a "peach" (high quality) or a "lemon" (low quality). Prospective buyers value peaches at \$1000, and lemons at \$0, whereas sellers value keeping a peach at \$800 and a lemon at \$0. If both buyers and sellers are informed, then peaches are traded at some price between \$800 and \$1000, increasing the social welfare by \$200 per peach. When neither buyers or sellers are informed, every car (whether a peach or a lemon) is traded at some price between \$400 and \$500 (which are the buyer and seller's values for a car of unknown quality), yielding the same total increase in the social welfare. A strikingly different outcome occurs when only sellers are informed: each buyer would only be willing to pay \$500, driving the (informed) sellers of peaches out of the market, and leading to no gains from trade.

The previous example implicates informational asymmetry, and the resultant adverse selection, in the collapse of markets. These phenomena are not the focus of this paper. Instead, we consider

*A previous version of this paper, titled "On the Hardness of Signaling," appeared in the proceedings of the 55th Annual Symposium on Foundations of Computer Science (FOCS), 2014. We thank Aditya Bhaskara, Uriel Feige, Nicole Immorlica, Albert Jiang, and Shang-Hua Teng for helpful discussions. We also thank the anonymous FOCS and GEB reviewers for helpful feedback and suggestions. This work was supported by NSF CAREER Award CCF-1350900.

	Coop	Defect
Coop	$-1+\theta$ $-1+\theta$	0 $-5+\theta$
Defect	$-5+\theta$ 0	-4 -4

Figure 1: An incomplete information variant of the prisoners' dilemma.

the effects of publicly available information on the outcome of a game, motivated by the following counter-intuitive fact: even when all players have access to the same information, the optimal *information structure* may reveal some but not all the information available. For an illustrative example, consider the incomplete-information variant of the classical prisoners' dilemma shown in Figure 1, in which the game's payoffs are parametrized by a *state of nature* θ. When $\theta = 0$, this is the traditional prisoners' dilemma in which cooperation is socially optimal, yet the unique Nash equilibrium is the one where both players defect, making both worse off. When, however, the reward from cooperation is uncertain, the equilibrium depends on players' beliefs about θ. Assuming that θ is uniformly distributed in $[-3, 3]$, and that players are risk-neutral and know nothing about θ besides its distribution, they play as if θ equals its expectation and the defection equilibrium persists. On the other hand, if both players learn the realization of θ before making their decisions, they would cooperate at equilibrium when $\theta \geq 1$, and defect otherwise, improving both their expected utilities. What is surprising, however, is that neither the opaque nor the transparent information structure is optimal. Consider the following partially-informative scheme: when $\theta > -1$, both players receive the signal `High`, and otherwise receive the signal `Low`. On signal `High`, the posterior expected value of θ for both players is 1, inducing cooperation. On signal `Low`, the players defect. This information structure induces cooperation with greater probability than either of the opaque or transparent schemes, and thus improves the expected utility of both players.

Motivated by these intricacies, it is natural to seek characterizations of optimal public-channel information structures for a variety of natural games and objectives. Such characterizations are most useful when they are "simple", and hence translate to efficient algorithms. We therefore use computational complexity as a lens with which to shed light on the design of such information structures. When we view information structure design as an algorithmic task, we refer to it as *signaling* for convenience (slightly abusing well-established terminology), and refer to the algorithm implementing an information structure as a *signaling scheme*.

Since our results will be negative, we choose as our backdrop one of the simplest and most fundamental settings in game theory: 2-player zero-sum games. We consider incomplete-information zero-sum games, in which a state of nature θ determines the entries of the payoff matrix, and is drawn from a common-knowledge prior distribution λ represented explicitly. We adopt the perspective of a principal who has access to the realization of θ, and can *commit* to a *public signaling scheme* (a.k.a. public-channel information structure) for revealing (partial) information to all players about θ via a shared communication channel. Such a principal may be interested in maximizing some weighted combination of the players' utilities, or more generally some (possibly linear) function of their chosen mixed strategies. Most such natural classes of objective functions include, as a special case, the task of maximizing one player's utility at equilibrium. Since our results will be negative, we adopt that as the principal's objective, and analyze the complexity of the resulting

design problem.

Before describing our results and techniques, we elaborate on our chosen model and assumptions. First, we restrict attention to *public* signaling schemes: those which reveal the same information regarding θ to every player in the game. This restriction is justified in settings where private communication with individual players is either impractical or undesirable. Indeed, much of the work on signaling in auctions [23, 15, 29, 22] is restricted to a public channel. Moreover, since the initial publication of the conference version of this paper, public-channel information structure design has been studied in voting [4, 5] and in network routing [13].

Second, we prove our negative results for the special case of two-player zero-sum games. This is a simple and fundamental class of games, lending our impossibility results more "bite". Moreover, two-player zero-sum games exhibit additional properties which make a complexity-theoretic analysis of information structure design all the more instructive. To appreciate these, first observe that every public-channel information structure induces a subgame for each realization of the signal, and this subgame is payoff-equivalent to the zero-sum game of complete information gotten by taking the posterior expectation of the payoff matrix. Therefore, like in full-information zero-sum games, equilibrium payoffs in each subgame are unique, ridding our analysis of equilibrium selection considerations. Second, since equilibrium computation is computationally tractable in zero-sum games of complete information, the same holds for each subgame induced by a public-channel information structure. This "disentangles" the complexity of optimal information structure design from the complexity of equilibrium computation in the constituent subgames. Indeed, it is not hard to see that the optimal information structure design task must, in general, "inherit" the computational complexity of equilibrium computation in the underlying class of full-information games. By choosing a class of full-information games where equilibria are essentially unique and easy to compute, our results show that information structure design makes the game model fundamentally "more complex".

Results and Techniques

Our results are based on the conjectured hardness of the planted clique problem. Specifically, we assume that it is hard to recover a planted k-clique from an Erdős-Rényi random graph random graph $\mathcal{G}(n,p)$ for $p = \frac{1}{2}$ and some $k = k(n)$ satisfying $k = \omega(\log^2 n)$ and $k = o(\sqrt{n})$. This assumption is a common one in theoretical computer science, and is justified on two grounds: (a) No polynomial-time algorithms are known despite a large body of work targeting the problem for over two decades; (b) Several works have ruled out the most common algorithmic approaches for this problem. At the very least, there is widespread consensus that the problem appears to be beyond the reach of current techniques, justifying its use as a starting point for hardness results. See Section 3 for more details and a literature review of the planted clique problem.

We prove two main results assuming this conjecture in Section 4. For 2-player Bayesian zero-sum games which are explicitly represented via their normal form, we prove that no algorithm computes a player-optimal signaling scheme in polynomial time. We strengthen this to a hardness of approximation result, for an additive absolute constant, for an implicitly-represented zero-sum game which nevertheless permits efficient equilibrium computation.

Both our hardness results hold for a Bayesian zero-sum game in which the states of nature, as well as each player's pure strategies, are the nodes V of a graph G. Specifically, we construct such a game in which signaling schemes correspond to "fractional partitions" of the states of nature V. Roughly speaking, the first player's utility function favors signaling schemes which (fractionally) partition the graph into large yet highly connected clusters of nodes. If G is a random graph in which many k-cliques have been planted randomly throughout, so much so that the planted cliques

roughly *cover* the nodes of G, then the best signaling scheme roughly corresponds to a partition of G into the planted cliques. Moreover, we exhibit an algorithm which, given a near-optimal signaling scheme, recovers a constant fraction of the planted clique cover.

In order to base our results on the planted clique conjecture, in Section 3 we prove an "amplification lemma" extending the hardness of recovering a planted clique to recovering a constant fraction of a planted clique cover. This lemma, which may be of independent interest, appears surprising because its analogue for the "distinguishing version" of the planted cover and planted clique problems is false.

Prior Related Work

The study of the effects of information on strategic interactions has its roots in the early works of Akerlof [1] and Spence [44]. Hirshleifer [31] was the first to observe that more information sometimes leads to worse market outcomes, in contrast to earlier work by Blackwell [14] which implied that more information is always better for a single agent in a non-competitive environment. Since then, many works have examined the effects of additional information on players' equilibrium utilities. Lehrer et al. [38] showed that additional information improves players' utilities in *common interest* games — games where players have identical payoffs in each outcome, and Bassan et al. [10] exhibited a polyhedral characterization of games in which more information improves the individual utility of *every* player. The work of Peski [42] is related to ours, and considers *private-channel* information structures in zero-sum games. They show that maximizing one player's utility requires revealing as much information as possible to that player, and withholding as much as possible from his opponent.

A number of works have established striking characterizations of the space of equilibria attainable by varying the information structure, in natural game domains. Bergemann et al. [12] considers signaling in a fundamental buyer/seller price-discrimination game, and provides a polyhedral characterization of the space of potential equilibria and their associated payoffs. Bergemann and Morris [11] characterize the space of equilibria attainable by (private channel) information structures in general games, and relate it to the space of correlated equilibria. One implication of [11] is that designing private-channel information structures in explicitly-described normal-form games reduces to a simple linear program, similar to the program for computing a correlated equilibrium.

Despite appreciation of the importance of information in strategic interactions, it is only recently that researchers have started viewing the information structure of a game as a mathematical object to be designed, rather than merely an exogenous variable. Kamenica and Gentzkow [35] examine settings in which a sender must design a signaling scheme to convince a less informed receiver to take a desired action. Work in the CS community, including by Emek et al. [23], Bro Miltersen and Sheffet [15] and Guo and Deligkas [29], examines revenue-optimal public signaling in an auction setting, and presents polynomial-time algorithms and hardness results for computing it. Dughmi et al. [22] examine welfare-optimal public signaling in an auction setting under exogenous constraints, and presents polynomial-time algorithms and hardness results.

The original conference-version of this paper was the first to use computational complexity theory to shed light on the design and characterization of information structures in abstract game settings other than auctions. Since then, followup work has strengthened and extended our results, and the literature employing tools from computer science in the design and analysis of information structures has grown substantially. A high-level overview of followup work is relegated to the conclusion of this paper.

2 Preliminaries

2.1 Games

We consider *finite 2-player Bayesian zero-sum games*, each of which is partially specified by the following parameters.

- Nonnegative integers r and c, denoting the number of pure strategies of the row player and column player respectively.
- A finite family $\Theta = \{1, \ldots, M\}$ of *states of nature*, which we index by θ.
- A family of *payoff matrices* $\mathcal{A}^\theta \in \mathbb{R}^{r \times c}$, indexed by states of nature $\theta \in \Theta$.
- A *prior distribution* distribution $\lambda \in \Delta_M$ on the states of nature.

Naturally, $\mathcal{A}^\theta(i, j)$ denotes the payoff of the row player when the row player plays i, the column player plays j, and the state of nature is θ. The column player's payoff in the same situation is $-\mathcal{A}^\theta(i, j)$. The final ingredient of a Bayesian game, the information structure, is left unspecified and viewed as an object to be designed.

We use two different representations of these games. In the *explicit representation*, the matrices $\{A^\theta\}_\theta$, as well as the prior distribution λ, are given explicitly. We also relax this somewhat for our second result; specifically, we consider a game in which the individual matrices $\mathcal{A}^\theta$ are given *implicitly*, since one of the players has a number of strategies which is exponential in the natural description of the game. Nevertheless, a low-rank bilinear structure permits efficient computation of equilibria, and the value of the game, in the implicitly represented game we consider.[1]

2.2 Signaling Schemes

We consider a principal who can *commit* to a policy, or equivalently an algorithm, for revealing partial information regarding the state of nature θ to the players. Crucially, we require that the principal reveal the same information to both players in the game. A *public signaling scheme* is given by a set Σ of *signals*, and a (possibly randomized) map φ from states of nature Θ to signals Σ. Abusing notation, we use $\varphi(\theta, \sigma)$ to denote the probability of announcing signal $\sigma \in \Sigma$ when the state of nature is $\theta \in \Theta$. We restrict attention to signaling schemes with a finite set of signals Σ, and this is without loss of generality when Θ is finite. We elaborate on this after describing the convex decomposition interpretation of a signaling scheme.

We note that public signaling schemes are in one-to-one correspondence with *convex decompositions* of the prior distribution $\lambda \in \Delta_M$ — namely, distributions supported on the simplex Δ_M, and having expectation λ. Formally, a signaling scheme $\varphi : \Theta \to \Sigma$ corresponds to the convex decomposition

$$\lambda = \sum_{\sigma \in \Sigma} \alpha_\sigma \cdot x_\sigma,$$

where $\alpha_\sigma = \mathbf{Pr}[\varphi(\theta) = \sigma] = \sum_{\theta \in \Theta} \lambda(\theta)\varphi(\theta, \sigma)$, and $x_\sigma(\theta) = \mathbf{Pr}[\theta | \varphi(\theta) = \sigma] = \frac{\lambda(\theta)\varphi(\theta,\sigma)}{\alpha_\sigma}$. Note that $x_\sigma \in \Delta_M$ is the *posterior distribution* of θ conditioned on signal σ, and α_σ is the probability of signal σ. The proof of the converse direction, namely that every convex decomposition of λ corresponds to a signaling scheme, is elementary yet thought provoking, and hence left to the reader. We note

[1] Implicitly-represented 2-player zero-sum games often admit efficient algorithms for equilibrium computation. Fairly general conditions under which this is possible are well exposited by Immorlica et al. [32].

that this correspondence between public signaling schemes and convex decompositions is closely related to the "splitting lemma" of Aumann and Maschler [8].

We judge the quality of a signaling scheme by the outcome it induces signal by signal. Specifically, the principal is equipped with an objective function of the form $\sum_\sigma \alpha_\sigma \cdot f(x_\sigma)$, where $f : \Delta_M \to \mathbb{R}$ is some function mapping a posterior distribution to the quality of the equilibrium chosen by the players. For example, f may be the social welfare at the induced equilibrium, a weighted combination of players' utilities at equilibrium, or something else entirely. In this setup, one can show that there always exists a signaling scheme with a finite set of signals which maximizes our objective, so long as the states of nature are finitely many. The optimal choice of signaling scheme is related to the *concave envelope* f^+ of the function f.[2] Specifically, such a signaling scheme achieves $\sum_\sigma \alpha_\sigma \cdot f(x_\sigma) = f^+(\lambda)$. Application of Caratheodory's theorem to the hypograph of f, therefore, shows that $M + 1$ signals suffice.

Our impossibility results rule out algorithms for computing near optimal signaling schemes, almost agnostic to how such schemes are represented as output. For concreteness, the reader can think of a signaling scheme φ as represented by the matrix of pairwise probabilities $\varphi(\theta, \sigma)$. Since we only consider games where the states of nature, and therefore also the number of signals w.l.o.g., are polynomially many in the description size of the game, this is a compact representation. The representation of φ as a convex decomposition would do equally well, as both representations can be efficiently computed from each other.

We note that the fact that we can restrict attention to "small" signaling schemes, in particular those with no more than $M+1$ signals, implies that optimal signaling problem is an NP optimization problem under mild conditions. In particular, so long as the M states of nature are given explicitly as input, and the quality of equilibrium $f(x_\sigma)$ can be computed in polynomial time for each posterior belief x_σ, the objective $\sum_\sigma \alpha_\sigma \cdot f(x_\sigma)$ can also be computed in polynomial time from the convex-decomposition representation of a "small" signaling scheme. As will become clear, our chosen class of games, equilibrium concept, and objective function will permit evaluating $f(x_\sigma)$ in each subgame efficiently. This holds in both the explicit and implicitly described games we consider, rendering each an NP optimization problem.

2.3 Strategies, Equilibria, and Objectives

Given a Bayesian game, a public signaling scheme (α, x) with signals Σ induces $|\Sigma|$ sub-games, one for each signal. The subgame corresponding to signal $\sigma \in \Sigma$ is played with probability α_σ, and players' (common) beliefs regarding the state of nature in this subgame are given by the posterior distribution $x_\sigma \in \Delta_M$. The quality of a public signaling scheme in such a game is contingent on a choice of an *equilibrium concept* and an *objective function*.

Equilibrium Concept

An equilibrium concept distinguishes a mixed strategy profile for every posterior belief $x \in \Delta_M$. This permits defining an objective function on signaling schemes, as described in Section 2.2. In general Bayesian games, evaluating the quality of a signaling scheme may be complicated by issues of equilibrium selection — for example, general sum games often admit many Bayes-Nash equilibria. However, our restriction to two-player zero-sum games side-steps such complications: all standard equilibrium concepts are payoff-equivalent in zero-sum games.

[2] f^+ is the point-wise lowest concave function h for which $h(x) \geq f(x)$ for all x in the domain. Equivalently, the hypograph of f^+ is the convex hull of the hypograph of f.

Our restriction to zero-sum games avoids an additional complication, which arises for more general games. Equilibrium computation in full-information zero-sum games is *tractable*, permitting efficient computation of equilibrium in the Bayesian game, for every posterior distribution over the states of nature. This avoids "inheriting" the computational complexity of equilibrium computation into our optimal signaling problems, and therefore suggests that signaling introduces complexity beyond that of computing equilibria in the resulting subgames.

Formally, given a Bayesian zero-sum game $(\{\mathcal{A}^\theta\}_{\theta=1}^M, \lambda)$ as described in Section 2.1, and a signaling scheme corresponding to the convex decomposition (α, x) of the prior $\lambda \in \Delta_M$, this naturally induces a distribution over sub-games of the same form. Specifically, for each signal $\sigma \in \Sigma$, the Bayesian zero-sum game $(\{\mathcal{A}^\theta\}_{\theta=1}^M, x_\sigma)$ is played with probability α_σ. We use $\mathcal{A}^\sigma = \mathbf{E}_{\theta \sim x_\sigma}[A^\theta]$ to denote the matrix of posterior expected payoffs conditioned on signal σ. A *Bayesian Nash equilibrium* corresponds to an equilibrium of each sub-game $(\{\mathcal{A}^\theta\}_{\theta=1}^M, x_\sigma)$ in which players play as they would in the complete information zero-sum game $\mathcal{A}^\sigma$. In the sub-game corresponding to signal σ, the row player's payoff is simply his payoff in the complete information zero-sum game A^σ, namely

$$u_r(\mathcal{A}^\sigma) = \max_{y \in \Delta_r} \min_{j=1}^{c} (y^\intercal \mathcal{A}^\sigma)_j.$$

The expected payoff of the row player over the entire game is then given by

$$u_r(\alpha, x) = \sum_{\sigma \in \Sigma} \alpha_\sigma u_r(\mathcal{A}^\sigma)$$

Objective Function

We adopt u_r, as given above, as our objective function. By symmetry, this is technically equivalent to adopting objective function u_c, the utility of the column player. To justify this choice for our hardness results, observe that most natural classes of objective functions — say weighted combinations of players' utilities, or linear functions of players' mixed strategies — include u_r and u_c as special cases. In fact, in our two-player zero-sum setting the problem of maximizing a weighted combination of players' utilities is equivalent either to maximizing u_r or maximizing u_c, depending on which player is given the greater weight.

Additionally, we are motivated by the fact that the space of all payoff profiles achievable by signaling is spanned by the two schemes maximizing the utility of one of the players, in the following sense. Let φ_r be a signaling scheme maximizing the row player's utility, and φ_c be a signaling scheme maximizing the column player's utility. Moreover, let $u_r^{\max} = -u_c^{\min}$ denote the row player's maximum utility (equivalently, the negation of the column player's minimum utility), and similarly let $u_c^{\max} = -u_r^{\min}$ denote the column player's maximum utility (equivalently, the negation of the row player's minimum utility). If there exists a signaling scheme φ inducing a utility profile (u_r, u_c), then it can be formed as a convex combination of the two extreme signaling schemes φ_r and φ_c — specifically, by running φ_r with probability $\frac{u_r - u_r^{\min}}{u_r^{\max} - u_r^{\min}}$ and φ_c otherwise. Consequently, "mapping out" the space of possible payoff profiles, as well constructing a signaling scheme attaining a desired payoff profile, both reduce to computing the extreme schemes φ_r and φ_c.

2.4 Graphs, Clusters, and Density

An *undirected graph* G is a pair (V, E), where V is a finite set of *nodes* or *vertices*, and $E \subseteq \binom{V}{2}$ is a set of undirected *edges*. We usually use n to denote the number of vertices of a graph, and m to denote the number of edges. Given a graph $G = (V, E)$, a *cluster* is some $S \subseteq V$. Given a cluster

S, we define the intra-cluster edges $E(S)$ as those edges with both endpoints in S. The *induced subgraph* of S is the graph $H = (S, E(S))$. Moreover, given two clusters $S, T \subseteq V$, we define the inter-cluster edges $E(S,T)$ as the edges with at least one endpoint in each of S and T.

The *density* of G is the fraction of all potential edges in E — namely $\text{density}(G) = \frac{2|E|}{|V|(|V|-1)}$. More generally, the density of a cluster S in G is the density of the subgraph of G induced by S, specifically $\text{density}_G(S) = \frac{2|E(S)|}{|S|(|S|-1)}$. To precisely state and prove our results, we require a slightly different notion of density, defined between pairs of clusters. We define the *bi-density* between clusters S and T as the fraction of all pairs $(u,v) \in S \times T$ connected by an edge; formally, $\text{bi-density}(S,T) = \frac{1}{|S||T|} \cdot |\{(u,v) \in S \times T | \{u,v\} \in E\}|$. Equivalently, the bi-density between S and T can be thought of in terms of the adjacency matrix of A of G; specifically, $\text{bi-density}(S,T) = \frac{1}{|S||T|} \sum_{i \in S} \sum_{j \in T} A_{ij}$. Observe that density and bidensity are closely related, in that $\text{bi-density}(S,S) = \text{density}(S)(1 - \frac{1}{|S|})$.

2.5 Random Graphs

We make use of *Erdős-Rényi random graphs*. Given $n \in \mathbb{N}$ and $p \in [0,1]$, the random graph $\mathcal{G}(n,p)$ has vertices $V = \{1, \ldots, n\}$, and its edges include every $e \in \binom{V}{2}$ independently with probability p. Naturally, for $G \sim \mathcal{G}(n,p)$, every cluster S of G has expected density p. Moreover, every pair of clusters S and T has expected bi-density $p\left(1 - \frac{|S \cap T|}{|S||T|}\right)$.[3] We make use of the following standard probabilistic bound, proved in Appendix B.

Proposition 2.1. *Let $p \in (0,1)$ and $\alpha > 1$ be absolute constants (independent of n), and let $G \sim \mathcal{G}(n,p)$. There is an absolute constant $\beta = \beta(p,\alpha)$ such that the following holds with high probability for all clusters X and Y with $|X|, |Y| > \beta \log n$.*

$$\text{bi-density}_G(X,Y) \leq \alpha p$$

As in the above proposition, we make references throughout this paper to guarantees which hold *with high probability* over a family of graphs parametrized by the number of nodes n. By this we mean that the claimed property holds with probability at least $1 - 1/\operatorname{poly}(n)$, for some polynomial in the number of nodes.

3 Planting Cliques and Clique Covers

3.1 The Planted Clique Problem

For our hardness results, we assume that it is hard to recover a planted clique from an Erdős-Rényi random graph. Specifically, we consider the following problem for parameters $n, k \in \mathbb{N}$ and $p \in [0,1]$.

Definition 3.1 (The Planted Clique Problem **PCLIQUE**(n,p,k))**.** *Let $G \sim \mathcal{G}(n,p,k)$ be a random graph with vertices $[n] = \{1, \ldots, n\}$, constructed as follows:*

1. *Every edge is included in G independently with probability p.*
2. *A set $S \subseteq [n]$ with $|S| = k$ is chosen uniformly at random.*
3. *All edges with both endpoints in S are added to G.*

[3] The discrepancy from p is due to the absence of self-loops in the drawn graphs.

Given a sample from $G \sim \mathcal{G}(n, p, k)$, recover the planted clique S.

We refer to the parameter p as the *background density*, S as the *planted clique*, and k as the *size* of the clique. Typically, we will think of k as a function of n, and p as an absolute constant independent of n. Our results will hinge on the conjectured hardness of recovering the clique with constant probability. We use the following well-believed version of the planted-clique conjecture as our hardness assumption.

Assumption 3.2. *For some function $k = k(n)$ satisfying $k = \omega(\log^2 n)$ and $k = o(\sqrt{n})$, there is no probabilistic polynomial-time algorithm for* **PCLIQUE**$(n, \frac{1}{2}, k)$ *with constant success probability.*

By *constant success probability*, we mean that the probability of the algorithm recovering the clique is bounded below by a constant independent of n, over the random draw of the graph $G \sim \mathcal{G}(n, p, k)$ as well as the internal random coins of the algorithm.

The planted clique conjecture, as described in Assumption 3.2, is well believed by the theoretical computer science community. At the very least, there is widespread consensus that the problem is beyond the reach of current algorithmic techniques, justifying its use as a starting point for hardness results. Next, we briefly outline the history of this problem, including evidence of hardness.

The problem of recovering a planted clique, as well as the (no harder) problem of distinguishing a draw from $\mathcal{G}(n, p)$ from a draw from $\mathcal{G}(n, p, k)$, has been the subject of much work since it was introduced by Jerrum [33] and Kučera [37]. Almost all of this work has focused on $p = \frac{1}{2}$. On the positive side, there is a quasipolynomial time algorithm for recovering the clique when $k \geq 2 \log n$. The best known polynomial-time algorithms, on the other hand, can recover planted cliques of size $\Omega(\sqrt{n})$, through a variety of different algorithmic techniques (e.g. [2, 40, 17, 24, 6, 26, 18]).

Despite this extensive body of work, there are no known polynomial-time algorithms for recovering, or even detecting, planted cliques of size $k = o(\sqrt{n})$. There is evidence the problem is hard: Jerrum [33] ruled out Markov chain approach for $k = o(\sqrt{n})$; Feige and Krauthgamer [25] ruled out algorithms based on the Lovász-Schrijver family of semi-definite programming relaxations for $k = o(\sqrt{n})$; and recently Feldman et al. [27] defined a family of "statistical algorithms," and ruled out such algorithms for recovering planted cliques of size $k = o(\sqrt{n})$. In light of this evidence, the planted clique conjecture is frequently used by computer scientists as a hardness assumption for a variety of different applications; e.g. for proving security of cryptographic primitives [34], for establishing the hardness of testing some properties of distributions [3], and for showing hardness of (approximately) computing welfare-maximizing Nash equilibria [30, 41].

3.2 From Planting Cliques to Planting Covers

For our results, we construct games in which the random state of nature is a node in some graph G. Recalling that a signaling scheme is a kind of fractional partition of the states of nature, with each "fractional part" corresponding to a signal, we design our games so that the quality of a signal is proportional to the density of that part of the graph. Since a signaling scheme's quality is measured in aggregate over the entire distribution over states of nature, any hardness result must rule out recovering a family of dense clusters (one per signal) scattered throughout the graph, rather than merely a unique dense cluster as in the planted clique problem. In this section, we define the *planted clique-cover problem* in which many cliques are planted throughout the graph, and prove a kind "amplification lemma" extending the hardness of recovering a planted clique to recovering a constant fraction of the planted cover.

The planted clique-cover problem is parametrized by $n, k, r \in \mathbb{N}$ and $p \in [0, 1]$, and defined as follows.

Definition 3.3 (The Planted Clique Cover Problem **PCCOVER**(n,p,k,r))**.** *Let $G \sim \mathcal{G}(n,p,k,r)$ be a random graph on vertices $[n] = \{1,\dots,n\}$, constructed as follows:*

1. *Include every edge in G independently with probability p.*
2. *For $i = 1$ to r:*
   - *Choose $S_i \subseteq [n]$ with $|S_i| = k$ uniformly at random.*
   - *Add all edges with both endpoints in S_i to G.*

Given a sample from $G \sim \mathcal{G}(n,p,k,r)$, recover a constant fraction of the planted cliques $S_1,\dots,S_r$.[4]

As in the planted clique problem, we refer to the parameter p as the *background density*, $S_1,\dots,S_r$ as the *planted cliques*, k as the *size* of each clique, and r as the *number* of cliques. Moreover, we will think of k and r as functions of n, and p as an absolute constant independent of n. Note that the planted clique problem is the special case of the planted clique-cover problem when $r = 1$. For our results, we will use the planted clique-cover problem for $r = \Theta(\frac{n}{k})$, guaranteeing that a constant fraction of the nodes are in at least one of the planted cliques with high probability.

At first glance, it might appear that planting many cliques as we do here makes the problem easier than planted clique. Somewhat surprisingly — and we elaborate on why later — this is not the case. We exhibit a reduction from the planted clique problem to the planted clique-cover problem with an arbitrary parameter r, showing that indeed the planted clique-cover problem is no easier. We prove the following lemma.

Lemma 3.4. *Let k and r be arbitrary functions of n, and p be an arbitrary constant. If there is a probabilistic polynomial-time algorithm for* **PCCOVER**(n,p,k,r) *with constant success probability, then there is such an algorithm for* **PCLIQUE**(n,p,k)*.*

Proof. The reduction proceeds as follows: Given a graph $G \sim \mathcal{G}(n,p,k) = \mathcal{G}(n,p,k,1)$, we construct a graph $G' \sim \mathcal{G}(n,p,k,r)$ by planting $r-1$ additional k-cliques at random, as in Definition 3.3. In a sense, we "continue where planted clique left off" by adding $r-1$ more cliques placed randomly in the graph. Let S_1 denote the original planted clique, and $S_2,\dots,S_r$ be the "additional" cliques planted through the reduction.

The key observation is that the cliques $S_1,\dots,S_r$ are indistinguishable to any algorithm operating on a sample from $\mathcal{G}(n,p,k,r)$. This is by a symmetry argument: permuting the order in which cliques are planted does not change the distribution $\mathcal{G}(n,p,k,r)$. As a result, any algorithm which recovers a constant fraction of the planted cliques from G' with constant probability must recover each of $S_1,\dots,S_r$ with constant probability. In particular, applying an algorithm for the planted cover problem to the outcome of our reduction yields a list of cliques which includes S_1 — the original planted clique — with constant probability. This leads to an algorithm for the planted clique problem with constant success probability. □

Lemma 3.4 is surprising since it does not appear to hold for the *distinguishing* variants of **PCCOVER** and **PCLIQUE**. Specifically, when $k = n^\gamma$ for a constant $\gamma < \frac{1}{2}$, whereas it is conjectured that no algorithm can distinguish a sample from $\mathcal{G}(n,\frac{1}{2})$ from a sample from $\mathcal{G}(n,\frac{1}{2},k)$ with constant probability, a simple statistical test succeeds in distinguishing a sample from $\mathcal{G}(n,\frac{1}{2})$ from a sample from $\mathcal{G}(n,\frac{1}{2},k,n/k)$ with constant probability. The test in question simply counts the edges in the graph. The random graph $\mathcal{G}(n,\frac{1}{2})$ has $\frac{n^2}{4} \pm O(n)$ edges with constant probability approaching 1, whereas $\mathcal{G}(n,\frac{1}{2},k,n/k)$ has $\frac{n^2}{4} + \Omega(\frac{n}{k}\cdot k^2) = \frac{n^2}{4} + \Omega(n^{1+\gamma})$ edges in expectation — well above the confidence interval for the number of edges in $\mathcal{G}(n,\frac{1}{2})$.

[4] By this we mean that the algorithm should output a list of k-cliques in G, at least αr of which are in $\{S_1,\dots,S_r\}$, for some constant α independent of n.

3.3 Approximate Recovery

To simplify our hardness results, we show that producing a set of vertices with sufficient overlap with a planted clique is polynomial-time equivalent to recovering that entire clique, with high probability.

Lemma 3.5. *Let $\epsilon > 0$ and $p \leq \frac{1}{2}$ be constants, $k = k(n)$ satisfy $k = \omega(\log^2 n)$ and $k = o(\sqrt{n})$, and $r = O(\frac{n}{k})$. There is a probabilistic polynomial-time algorithm which takes as input $G \sim \mathcal{G}(n, p, k, r)$ and a cluster $T \subseteq [n]$, and outputs every planted k-clique S in G for which $|T \cap S| > \epsilon |T \cup S|$, with high probability.*

In other words, to recover a planted k-clique S it suffices to produce a set T of size $O(k)$ for which $|T \cap S| = \Omega(k)$. The analogous statement for the (single) planted clique problem is folklore knowledge. Our proof is similar, though requires some additional accounting because our planted cliques may overlap. We relegate the proof of Lemma 3.5 to Appendix A.

4 Hardness of Signaling in Zero-Sum Games

In this section, we show the intractability of optimal signaling in zero sum games, when the objective is maximizing one player's payoff at equilibrium. We prove two results which follow from Assumption 3.2: Hardness of optimal signaling for *explicit* zero sum games — those games with a polynomial number of strategies and matrices given explicitly, and hardness of approximation for *implicit* zero-sum games. Our former result, for explicit games, requires the real numbers in the payoff matrices to scale with the input size, and therefore does not qualify as a hardness of approximation result per se. Our latter result for implicitly-described games does rule out a constant additive approximation relative to range of possible payoffs, and crucially holds for a class of games in which equilibrium computation is *tractable*, in the sense described in Section 2.3.

Theorem 4.1. *Assumption 3.2 implies that there is no polynomial-time algorithm which computes a signaling scheme maximizing a player's payoff in an explicitly-represented 2-player Bayesian zero-sum game.*

Theorem 4.2. *Assumption 3.2 implies that there is an implicitly-described, yet tractable, class of Bayesian zero-sum games with payoffs in $[-1, 1]$, and an absolute constant ϵ, such that there is no polynomial-time algorithm for computing a signaling scheme which ϵ-approximately maximizes (in the additive sense) one player's payoff.*

For both our results, we use variants of the following *security game* involving two players, known as an *attacker* and a *defender*. In a security game, an attacker chooses a target to attack, and a defender chooses targets to defend. The attacker's payoff depends on both his choice of target, and the extent to which said target is protected by the defender. We consider a security game on a network, in which an attacker is looking to take down edges of the network, and a defender is looking to defend those edges. The state of nature determines a vulnerable node θ, and an attacker can only take down edges adjacent to θ by attacking its other endpoint. The defender, on the other hand, can select a small set of nodes and protect the edges incident on them from attack.

Definition 4.3 (Network Security Game)**.** *Instances of this Bayesian zero-sum security game are described by an undirected graph $G = (V, E)$ with n nodes and m edges, representing a communication network, an integer $d \geq 1$ equal to the number of nodes the defender can simultaneously protect, and a real number $\rho \geq 0$ equal to the utility gain to the defender for protecting a vulnerable*

or attacked node. States of nature correspond to vertices V. When the realized state of nature is $\theta = v$, this indicates that the edges incident on node v are vulnerable to attack. We assume that $\theta \sim \lambda$, where the prior λ is the uniform distribution over V. The attacker's strategies also correspond to the nodes V. The defender's strategies correspond to subsets of V of size at most d, representing the choice of nodes to defend. When the state of nature is $\theta \in V$, the attacker attacks $a \in V$, and the defender defends $D \subseteq V$ with $|D| \leq d$, the payoff of the attacker is defined as follows:

$$\mathcal{A}^\theta(a, D) = |\{(\theta, a)\} \cap E| - \rho|D \cap \{\theta, a\}|$$

When d is a constant, the size of the matrices $\mathcal{A}^\theta$ is polynomial in the representation size of the game, and we can think of the game as being represented *explicitly* as a set of n matrices $\mathcal{A}^\theta \in [-2\rho, 1]^{n \times \binom{n}{d}}$. However, when d is superconstant, the game matrices have a superpolynomial $\binom{n}{d}$ number of columns. Nevertheless, even then equilibrium computation is still *tractable* in the sense described in Section 2.3 — i.e., our representation permits computation of the equilibrium, and corresponding utilities, for every posterior distribution $x \in \Delta_n$ over states of nature. This is a consequence of the "low dimensional" nature of the defender's mixed strategy space, and linearity of players' utilities in this low-dimensional representation. Specifically, a mixed strategy of the defender can be summarized as a vector in the matroid polytope $\mathcal{P}_d = \{z \in [0,1]^n : \sum_{i=1}^n z_i \leq d\}$. A vector $z \in \mathcal{P}_d$ encodes the probability by which the defender defends each target. It is not hard to see that this representation is loss-less, in that (a) the vector z summarizing a defender's mixed strategy, together with the attacker's mixed strategy $y \in \Delta_n$, suffices for computing the payoff of each player in a subgame with beliefs x — specifically, using A to denote the adjacency matrix of graph G, the attacker's payoff is $x^\intercal A y - \rho(z^\intercal x + z^\intercal y)$; (b) Given a vector $z \in \mathcal{P}_d$, a corresponding mixed strategy of the defender with small support can be efficiently recovered as the convex decomposition of z into the corner points of the matroid polytope $\mathcal{P}_d$;[5] and (c) Given a vector $z \in \mathcal{P}_d$ summarizing the defender's mixed strategy in a subgame with beliefs x, a best response for the attacker can be computed efficiently.[6] Together, observations (a), (b), and (c) imply that an equilibrium of this game, as well as its optimal value, can be computed in time polynomial in n by linear programming.[7]

Theorems 4.1 and 4.2 follow from the two lemmas below, as well as Lemma 3.4. Specifically, Theorem 4.1 instantiates both lemmas below with $d = 1$ and $\rho = \frac{k}{150}$, and Theorem 4.2 sets $d = \frac{k}{150}$ and $\rho = 1$.

Lemma 4.4. *Let k and r be such that $r = \frac{3n}{k}$. For a network security game on $G \sim \mathcal{G}(n, \frac{1}{2}, k, r)$ with $d, \rho \geq 1$ satisfying $\rho d = \frac{k}{150}$, with high probability there is a signaling scheme attaining expected attacker utility at least* 0.8.

Lemma 4.5. *Let $\epsilon > 0$ be an absolute constant, k satisfy $k = \omega(\log^2 n)$ and $k = o(\sqrt{n})$, and $r = \Theta(\frac{n}{k})$. For a network security game on $G \sim \mathcal{G}(n, \frac{1}{2}, k, r)$ with with $d, \rho \geq 1$ satisfying $\rho d = \Theta(k)$, there is a probabilistic polynomial-time algorithm which, given any signaling scheme obtaining attacker utility at least $0.5 + \epsilon$, outputs an $\Omega(\epsilon)$ fraction of the planted cliques in G, with high probability.*

We emphasize that even though the graph G used to define our game is a random variable, players do know G. We resort to using a random graph $G \sim \mathcal{G}(n, \frac{1}{2}, k, r)$ not just because Clique

[5] This is the constructive version of Caratheodory's theorem, as shown by Grötschel et al. [28].

[6] Recall from Section 2.3 that (a) and (c) are equivalent to the analogous tasks (payoff and equilibrium computation, respectively) in the complete-information zero-sum game $\mathcal{A}^x = \mathbf{E}_{\theta \sim x} \mathcal{A}^\theta$.

[7] For more on equilibrium computation in implicitly-described zero-sum games, see [32].

recovery is conjectured to be hard on average (and hence also in the worst case) for such a graph, but also because such a graph exhibits structural properties (with high probability) which enable the algorithms used in our reduction and simplify our proofs. Naturally, our average-case arguments then imply the worst-case statements in Theorems 4.1 and 4.2.

4.1 Proof of Lemma 4.4

Denote $G = (V, E)$. We construct a signaling scheme which groups together states of nature in the same planted clique, and show that it obtains the claimed attacker utility at equilibrium. Let $S_1, \ldots, S_r$ be the planted k-cliques in G listed in some arbitrary order, and let $\widehat{S}_i = S_i \setminus \cup_{1 \le j < i} S_j$. Moreover, let $\widehat{S}_0 = V \setminus \cup_{i=1}^{r} S_i$ be all remaining vertices not in any of the planted cliques. Evidently, $\widehat{S}_0, \widehat{S}_1, \ldots, \widehat{S}_r$ form a partition of the vertices V. Our signaling scheme $\varphi : V \to \{0, \ldots, r\}$ is deterministic, and simply announces $\varphi(v) = i$ when $v \in \widehat{S}_i$. Represented as a convex decomposition, our signaling scheme is (α, x), where $\alpha_i = \frac{|\widehat{S}_i|}{n}$ and x_i is the uniform distribution over $\widehat{S}_i$.

To prove that φ obtains the desired bound with high probability, we need the fact that most nodes are in one of our planted cliques. Recall that, for each $i \neq 0$, S_i is a k-subset of V chosen independently and uniformly at random. Since $r = 3\frac{n}{k}$, in expectation a $1 - \frac{1}{e^3}$ fraction of the n vertices are in one of the planted cliques $S_1, \ldots, S_r$. This translates to a high-probability guarantee by a standard application of the Chernoff bound and the union bound. Specifically, with high probability we have that $\sum_{i=1}^{r} \alpha_i = \frac{1}{n} \sum_{i=1}^{r} |\widehat{S}_i| \ge 0.9$.

We can now analyze the utility of an attacker at equilibrium, given the signaling scheme φ described above. In the subgame corresponding to a signal i, the state of nature θ is uniformly distributed in $\widehat{S}_i$. Consider the mixed strategy of the attacker which simply chooses $a \in \widehat{S}_i$ uniformly at random. Since (θ, a) is a uniformly distributed pair in $\widehat{S}_i \times \widehat{S}_i$, the probability that (θ, a) share an edge is precisely bi-density$(\widehat{S}_i, \widehat{S}_i)$. Moreover, for any defender strategy $D \subseteq V$ with $|D| \le d$, the sum of $\mathbf{Pr}[a \in D]$ and $\mathbf{Pr}[\theta \in D]$ is at most $2d/|\widehat{S}_i|$. Therefore, any pure defender strategy D begets attacker utility $u^i(\varphi) \ge$ bi-density$(\widehat{S}_i, \widehat{S}_i) - \frac{2d\rho}{|\widehat{S}_i|}$. For $i \neq 0$, since $\widehat{S}_i$ forms a clique this is at least $1 - \frac{2d\rho+1}{|\widehat{S}_i|}$.

We can now complete the proof by bounding the expected utility of the attacker from signaling

scheme φ by his utility when he simply plays the uniform strategy over $\widehat{S}_i$ for each subgame i.

$$\begin{aligned}
u(\varphi) &= \sum_{i=0}^{r} \alpha_i u^i(\varphi) \\
&\geq \sum_{i=0}^{r} \alpha_i \left(\text{bi-density}(\widehat{S}_i, \widehat{S}_i) - \frac{2d\rho}{|\widehat{S}_i|} \right) \\
&= \left(\sum_{i=0}^{r} \alpha_i \, \text{bi-density}(\widehat{S}_i, \widehat{S}_i) \right) - \frac{2d(r+1)\rho}{n} \\
&\geq \left(\sum_{i=1}^{r} \alpha_i \, \text{bi-density}(\widehat{S}_i, \widehat{S}_i) \right) - \frac{2d(r+1)\rho}{n} \\
&= \left(\sum_{i=1}^{r} \alpha_i \left(1 - \frac{1}{|\widehat{S}_i|} \right) \right) - \frac{2d(r+1)\rho}{n} \\
&= \left(\sum_{i=1}^{r} \alpha_i \right) - \frac{2d(r+1)\rho + r}{n} \\
&\geq 0.9 - \frac{5dr\rho}{n} \\
&= 0.8
\end{aligned}$$

4.2 Proof of Lemma 4.5

We prove the Lemma by exhibiting an algorithm which, given as input a graph $G \sim \mathcal{G}(n, \frac{1}{2}, k, r)$ and a signaling scheme φ attaining the claimed attacker utility, outputs a family of clusters $\mathcal{T}$ from which a constant fraction of the planted cliques can be recovered in polynomial time. Specifically, we show that for a constant fraction of the planted cliques $S_1, \ldots, S_r$, there is a cluster $T \in \mathcal{T}$ which satisfies the requirements of Lemma 3.5. To simplify our proof, we assume that $\rho d/2$ is an integer, though unsurprisingly this assumption can easily be removed.

Consider Algorithm 1. In step (2), the algorithm "zeroes-out" those vertices which are "over represented" in either the attacker's strategy or the posterior distribution over states of nature. This is justified because, as will become clear in the proof, we designed the game's payoffs so that over represented nodes can easily be protected by the defender, and hence account for only a small fraction of the attacker's total utility. After zeroing-out large entries, the posteriors $\widehat{x}_\sigma$ and attacker strategies $\widehat{y}_\sigma$ have entries no larger than $1/\Theta(k)$, and therefore behave roughly as uniform distributions over $\Theta(k)$ vertices. Steps (3) and (4) are "cleanup steps", which convert the attacker's mixed strategy to a uniform distribution over $\frac{\rho d}{2} = \Theta(k)$ vertices T_σ. We will show that the resulting family $\mathcal{T}$ of clusters must overlap substantially with the planted cliques, in the sense of Lemma 3.5, for the utility of the attacker to exceed the background density of $\frac{1}{2}$ by a constant. We formalize all this through a sequence of propositions.

Proposition 4.6. *The attacker's utility $u(\varphi)$ from signaling scheme $\varphi = (\alpha, x)$ satisfies*

$$u(\varphi) \leq \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A \widehat{y}_\sigma$$

where A is the adjacency matrix of G, and $\widehat{x}_\sigma$ and $\widehat{y}_\sigma$ are as defined in step (2) of Algorithm 1.

Algorithm 1 Algorithm for Computing Approximation of Planted Cliques from a Signaling Scheme

Input: Graph $G = (V, E)$ with adjacency matrix $A \in \mathbb{R}^{n \times n}$

Input: A signaling scheme $\varphi : V \to \Sigma$, represented in convex decomposition form (α, x) where $\alpha \in \Delta_\Sigma$, and $x_\sigma \in \Delta_V$ for each $\sigma \in \Sigma$.

Output: A list $\mathcal{T}$ of subsets of V, each of size $\rho d/2$.

1: For each $\sigma \in \Sigma$, compute the attacker's minimax strategy $y_\sigma \in \Delta_V$ in the subgame corresponding to signal σ.

2: For each $\sigma \in \Sigma$, discard from each of x_σ and y_σ all entries greater than $\frac{2}{\rho d}$. Define

$$\widehat{x}_\sigma(v) = \begin{cases} x_\sigma(v) \text{ if } x_\sigma(v) \leq \frac{2}{\rho d}, \\ 0 \text{ otherwise.} \end{cases}$$

$$\widehat{y}_\sigma(v) = \begin{cases} y_\sigma(v) \text{ if } y_\sigma(v) \leq \frac{2}{\rho d}, \\ 0 \text{ otherwise.} \end{cases}$$

3: For each $\sigma \in \Sigma$, let $z_\sigma \in \Delta_V$ be an extreme-point solution of the following linear program. (Note that $\widehat{x}_\sigma$ is fixed)

$$\begin{array}{lll} \text{maximize} & (\widehat{x}_\sigma)^\intercal A z_\sigma & \\ \text{subject to} & ||z_\sigma||_\infty \leq \frac{2}{\rho d} & \\ & \sum_{v \in V} z_\sigma(v) \leq 1 & \\ & z_\sigma(v) \geq 0, & \text{for } v \in V. \end{array} \tag{1}$$

4: Let T_σ be the support of $z_\sigma \in \Delta_V$, for each $\sigma \in \Sigma$.

5: Output $\mathcal{T} = \{T_\sigma : \sigma \in \Sigma\}$.

Proof. We prove this bound signal-by-signal. Fix a signal $\sigma \in \Sigma$ with posterior x_σ and attacker strategy y_σ, and as shorthand use $x = x_\sigma$, $y = y_\sigma$, $\widehat{x} = \widehat{x}_\sigma$, and $\widehat{y} = \widehat{y}_\sigma$. We distinguish nodes which are *over represented* in x or y — specifically, the nodes $O = \left\{v : \max(x(v), y(v)) \geq \frac{2}{\rho d}\right\}$. There are at most ρd such nodes, since each of x and y, by virtue of being a probability distribution, can have at most $\rho d/2$ entries exceeding $2/\rho d$.

We consider a defender who chooses $D \subseteq O$ with $|D| = \min(d, |O|)$ uniformly at random. Observe that such a defender protects each node in O with probability at least $\frac{\min(d,|O|)}{|O|} \geq \frac{d}{\rho d} = \frac{1}{\rho}$. The utility of the attacker is equal to the probability that $\theta \sim x$ and $a \sim y$ share an edge, minus ρ times the quantity $\mathbf{Pr}[a \in D] + \mathbf{Pr}[\theta \in D]$, which can be bounded as follows.

$$\begin{aligned} u^\sigma(\varphi) &= x^\intercal A y - \rho(\mathbf{Pr}[\theta \in D] + \mathbf{Pr}[a \in D]) \\ &= x^\intercal A y - \rho(\mathbf{Pr}[\theta \in O]\,\mathbf{Pr}[\theta \in D | \theta \in O] + \mathbf{Pr}[a \in O]\,\mathbf{Pr}[a \in D | a \in O]) \\ &\leq x^\intercal A y - \rho.\frac{1}{\rho}(\mathbf{Pr}[\theta \in O] + \mathbf{Pr}[a \in O]) \\ &= x^\intercal A y - (x(O) + y(O)) \end{aligned} \tag{2}$$

Next, we let $\overline{x} = x - \widehat{x}$ and $\overline{y} = y - \widehat{y}$. Note that $x(O) = ||\overline{x}||_1$ and $y(O) = ||\overline{y}||_1$. We can bound

$x^\intercal Ay$ as follows.

$$\begin{aligned}
x^\intercal Ay &= (\widehat{x} + \overline{x})^\intercal A(\widehat{y} + \overline{y}) \\
&= \widehat{x}^\intercal A\widehat{y} + \widehat{x}^\intercal A\overline{y} + \overline{x}^\intercal A\widehat{y} + \overline{x}^\intercal A\overline{y} \\
&\leq \widehat{x}^\intercal A\widehat{y} + \widehat{x}^\intercal\overline{y} + \overline{x}^\intercal\widehat{y} + \overline{x}^\intercal\overline{y} \\
&\leq \widehat{x}^\intercal A\widehat{y} + \widehat{x}^\intercal\overline{y} + \overline{x}^\intercal\widehat{y} + 2\overline{x}^\intercal\overline{y} \\
&= \widehat{x}^\intercal A\widehat{y} + \overline{x}^\intercal y + \overline{y}^\intercal x \\
&\leq \widehat{x}^\intercal A\widehat{y} + ||\overline{x}||_1||y||_\infty + ||\overline{y}||_1||x||_\infty \qquad \text{(Holder's inequality)} \\
&\leq \widehat{x}^\intercal A\widehat{y} + ||\overline{x}||_1 + ||\overline{y}||_1 \\
&= \widehat{x}^\intercal A\widehat{y} + x(O) + y(O) \qquad (3)
\end{aligned}$$

Combining (2) with (3) completes the proof. □

Proposition 4.7.

$$u(\varphi) \leq \sum_{\sigma\in\Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A z_\sigma$$

Proof. This follows from Proposition 4.6 and the fact that that $\widehat{y}_\sigma$ is a feasible solution to linear program (1). □

Proposition 4.8. *z_σ is a uniform distribution over $\rho d/2$ vertices, namely T_σ.*

Proof. Recall that we assumed that $\rho d/2$ is an integer. Therefore, a simple argument shows that an extreme-point solution to LP (1) must set each variable either to 0 or to $\frac{2}{\rho d}$. In particular, the optimal solution must set precisely $\rho d/2$ entries of z_σ to $\frac{2}{\rho d}$, as needed. □

We now upper bound the contribution of edges outside the planted cliques to the utility of the attacker. Write the adjacency matrix A as $A = A^- + A^+$, where A^- are the background edges added in Step (1) of Definition 3.3, and A^+ are the clique edges added in Step (2) of Definition 3.3.

Proposition 4.9. *The following holds with high probability for all signals $\sigma \in \Sigma$.*

$$(\widehat{x}_\sigma)^\intercal A^- z_\sigma \leq 0.5 + \frac{\epsilon}{2}$$

Proof. Pick an arbitrary signal σ. Since $||\widehat{x}_\sigma||_\infty \leq \frac{2}{\rho d}$, we have that $(\widehat{x}_\sigma)^\intercal A^- z_\sigma$ is bounded from above by the value of the following linear program, in which z_σ is held fixed and $x \in \mathbb{R}^n$ is allowed to vary.

$$\begin{array}{ll}
\text{maximize} & x^\intercal A z_\sigma \\
\text{subject to} & ||x||_\infty \leq \frac{2}{\rho d} \\
& \sum_{v\in V} x(v) \leq 1 \\
& x(v) \geq 0, \qquad \text{for } v \in V.
\end{array} \qquad (4)$$

By an argument similar to in Proposition 4.8, at optimality x is a uniform distribution over some $\rho d/2$ vertices R. Therefore, since z_σ is also a uniform distribution over the $\rho d/2$ vertices T_σ (Proposition 4.8), the value of the above linear program is equal to bi-density(R, T_σ). Since $\rho d/2 = \Theta(k) = \omega(\log n)$, Proposition 2.1 implies the claimed bound. □

We can now wrap up the proof of Lemma 4.5 using the above propositions. We will show that, on average over our planted cliques $S_1, \ldots, S_r$, there is some $T \in \mathcal{T}$ output by Algorithm 1 overlapping with a constant fraction of the clique vertices. As notation, we let A^i denote the adjacency matrix of the clique S_i, for $i = 1, \ldots, r$. Note that $A^+ \leq \sum_{i=1}^r A^i$. First, we show that foreground edges contribute $\Omega(\epsilon)$ to the outer product $(\widehat{x}_\sigma)^\intercal A z_\sigma$, on average over signals $\sigma \in \Sigma$.

$$\begin{aligned}
0.5 + \epsilon &\leq u(\varphi) \\
&\leq \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A z_\sigma && \text{(Proposition 4.7)} \\
&= \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal (A^- + A^+) z_\sigma \\
&\leq (0.5 + \frac{\epsilon}{2}) \sum_{\sigma \in \Sigma} \alpha_\sigma + \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A^+ z_\sigma && \text{(Proposition 4.9)} \\
&= 0.5 + \frac{\epsilon}{2} + \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A^+ z_\sigma
\end{aligned}$$

Therefore $\sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A^+ z_\sigma \geq \frac{\epsilon}{2}$. Next, we "break up" this sum of outer products into the constituent contributions of each planted clique.

$$\begin{aligned}
\frac{\epsilon}{2} &\leq \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A^+ z_\sigma \\
&\leq \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal (\sum_{i=1}^r A^i) z_\sigma \\
&= \sum_{i=1}^r \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma^\intercal A^i z_\sigma \\
&\leq \sum_{i=1}^r \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma(S_i) z_\sigma(S_i) \\
&= \sum_{i=1}^r \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma(S_i) \frac{|T_\sigma \cap S_i|}{|T_\sigma|}
\end{aligned}$$

Finally, we show that the average planted clique is well represented by some $T \in \mathcal{T}$.

$$\begin{aligned}
\frac{\epsilon}{2} &\leq \sum_{i=1}^{r} \sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma(S_i) \frac{|T_\sigma \cap S_i|}{|T_\sigma|} \\
&\leq \sum_{i=1}^{r} \left(\sum_{\sigma \in \Sigma} \alpha_\sigma \widehat{x}_\sigma(S_i) \right) \left(\max_{T \in \mathcal{T}} \frac{|T \cap S_i|}{|T_\sigma|} \right) \\
&\leq \sum_{i=1}^{r} \frac{|S_i|}{n} \left(\max_{T \in \mathcal{T}} \frac{|T \cap S_i|}{|T_\sigma|} \right) && \text{(Since } \textstyle\sum_\sigma \alpha_\sigma \widehat{x}_\sigma(v) \leq \sum_\sigma \alpha_\sigma x_\sigma(v) = \lambda_v = \frac{1}{n}\text{)} \\
&= \frac{k}{n} \sum_{i=1}^{r} \left(\max_{T \in \mathcal{T}} \frac{|T \cap S_i|}{|T_\sigma|} \right) \\
&= O(1) \cdot \frac{1}{r} \sum_{i=1}^{r} \left(\max_{T \in \mathcal{T}} \frac{|T \cap S_i|}{|T_\sigma|} \right) && \text{(Since } r = \Theta(\tfrac{n}{k})\text{)} \\
&= O(1) \cdot \frac{1}{r} \sum_{i=1}^{r} \left(\max_{T \in \mathcal{T}} \frac{|T \cap S_i|}{k} \right) && \text{(Since } |T_\sigma| = \rho d/2 = \Theta(k)\text{)}
\end{aligned}$$

Therefore $\frac{1}{r} \sum_{i=1}^{r} \left(\max_{T \in \mathcal{T}} \frac{|T \cap S_i|}{k} \right) \geq \Omega(\epsilon)$; i.e., the average planted clique intersects at least one of the sets in $\mathcal{T}$ in a constant fraction $\Omega(\epsilon)$ of its vertices. This implies that an $\Omega(\epsilon)$ fraction of the planted cliques intersect at least one of the sets in $\mathcal{T}$ in an $\Omega(\epsilon)$ fraction of its vertices, by a simple counting argument. Since each set in $\mathcal{T}$ is of size $\Theta(k)$, this satisfies the requirements of Lemma 3.5 for a constant fraction of the planted cliques $S_1, \ldots, S_r$, completing the proof.

5 Conclusions and Followup Work

This paper was the first to use complexity theory to shed light on the design and characterization of information structures in abstract game settings other than auctions. Since the initial publication of the conference version of this paper, followup work has strengthened and extended our results, and the literature on computational aspects of information in games has diversified and flourished. For a somewhat recent survey, see [19].

Assuming the planted clique conjecture, our impossibility result for explicit zero sum games rules out polynomial-time signaling algorithms with additive error on the order of $\frac{1}{\text{polylog } n}$ relative to the range of player payoffs, where n is the number of a player strategies. Essentially, this rules out a *fully polynomial time approximation scheme* (FPTAS) for optimal signaling in zero-sum games, but not a PTAS: an ϵ additive approximation algorithm for every constant ϵ independent of n. Two papers have since strengthened our results: Bhaskar et al. [13] rule out a PTAS assuming the planted clique conjecture, and also rule out an FPTAS based on the more mainstream assumption that $P \neq NP$. Rubinstein [43] rules out a PTAS based on the exponential time hypothesis, which is also considered a more mainstream hardness assumption than the planted clique conjecture.

These hardness results strongly suggest that, for multi-agent signaling problems over a shared communication channel, domain-specific structure must be leveraged in order to yield simple and near-optimal algorithm. Cheng et al. [16] identify two "smoothness" parameters which govern the complexity of these public-channel information structure design problems, and design approximation schemes with guarantees parametrized accordingly. This framework leads to polynomial-

time approximation schemes for a number of public-channel information structure design problems. Moreover, it also yields a *quasipolynomial-time approximation scheme (QPTAS)*[8] for public signaling in explicitly-represented normal-form constant-player games (a class which includes the explicitly-represented Bayesian zero-sum games discussed in this paper) and most natural objective functions. A related result is by Bhaskar et al. [13], who design a PTAS for signaling in zero-sum games which satisfy a different smoothness condition.

Beyond public-channel multi-agent information structure design problems, the CS/Econ community has examined algorithms for signaling in a number of fundamental domains. Dughmi and Xu [20] design efficient algorithms for optimal and near-optimal signaling in the case of a single decision-making agent — this is the popular *Bayesian Persuasion* model of Kamenica and Gentzkow [35]. Very recently, Babichenko and Barman [9] and Dughmi and Xu [21] examine the natural generalization of the Bayesian persuasion model to multiple agents introduced by Arieli and Babichenko [7], and derive efficient approximation algorithms for *private* signaling in this setting. Other information structure design models to which the tools of of algorithms and complexity have been deployed include multi-armed bandits [36, 39], Stackelberg games [45, 46], and network routing [13].

References

[1] George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, pages 488–500, 1970.

[2] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 594–598. Society for Industrial and Applied Mathematics, 1998.

[3] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505. ACM, 2007.

[4] Ricardo Alonso and Odilon Câmara. Persuading voters. *American Economic Review*, 106(11): 3590–3605, November 2016.

[5] Ricardo Alonso and Odilon Câmara. Political disagreement and information in elections. *Games and Economic Behavior*, 100:390 – 412, 2016. ISSN 0899-8256.

[6] Brendan PW Ames and Stephen A Vavasis. Convex optimization for the planted k-disjoint-clique problem. *Mathematical Programming*, 143(1-2):299–337, 2014.

[7] Itai Arieli and Yakov Babichenko. Private bayesian persuasion. *Available at SSRN 2721307*, 2016.

[8] R.J. Aumann and M. Maschler. *Repeated Games with Incomplete Information*. MIT Press, 1995. ISBN 9780262011471.

[9] Yakov Babichenko and Siddharth Barman. Computational aspects of private bayesian persuasion. In *Proceedings of the 8th ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, 2017.

[8] A QPTAS is an algorithm, parametrized by $\epsilon > 0$, which computes a solution within ϵ of the optimal in time $\exp(\text{polylog}(n, \frac{1}{\epsilon}))$.

[10] Bruno Bassan, Olivier Gossner, Marco Scarsini, and Shmuel Zamir. Positive value of information in games. *International Journal of Game Theory*, 32(1):17–31, 2003.

[11] Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016.

[12] Dirk Bergemann, Benjamin Brooks, and Stephen Morris. The limits of price discrimination. *The American Economic Review*, 105(3):921–957, 2015.

[13] 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 17th ACM Conference on Economics and Computation (EC)*, 2016.

[14] David Blackwell. Comparison of experiments. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 93–102, 1951.

[15] Peter Bro Miltersen and Or Sheffet. Send mixed signals: earn more, work less. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, pages 234–247, 2012.

[16] Yu Cheng, Ho Yee Cheung, Shaddin Dughmi, Ehsan Emamjomeh-Zadeh, Li Han, and Shang-Hua Teng. Mixture selection, mechanism design, and signaling. In *Proceedings of the 56th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

[17] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability & Computing*, 19(2):227, 2010.

[18] Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. In *ANALCO*, pages 67–75. SIAM, 2011.

[19] Shaddin Dughmi. Algorithmic information structure design: A survey. *SIGecom Exchanges*, 15(2), 2017.

[20] Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. In *Proceedings of the 47th ACM Symposium on Theory of Computing (STOC)*, 2016.

[21] Shaddin Dughmi and Haifeng Xu. Algorithmic persuasion with no externalities. In *Proceedings of the 18th ACM Conference on Economics and Computation (EC)*, 2017.

[22] Shaddin Dughmi, Nicole Immorlica, and Aaron Roth. Constrained signaling in auction design. In *Proceedings of the 25th ACM Symposium on Discrete Algorithms (SODA)*, 2014.

[23] Yuval Emek, Michal Feldman, Iftah Gamzu, Renato Paes Leme, and Moshe Tennenholtz. Signaling schemes for revenue maximization. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, 2012.

[24] Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semi-random graph. *Random Structures and Algorithms*, 16(2):195–208, 2000.

[25] Uriel Feige and Robert Krauthgamer. The probable value of the lovász–schrijver relaxations for maximum independent set. *SIAM Journal on Computing*, 32(2):345–370, 2003.

[26] Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. *DMTCS Proceedings*, (01): 189–204, 2010.

[27] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pages 655–664. ACM, 2013.

[28] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[29] Mingyu Guo and Argyrios Deligkas. Revenue maximization via hiding item attributes. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 157–163. AAAI Press, 2013.

[30] Elad Hazan and Robert Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM Journal on Computing*, 40(1):79–91, 2011.

[31] Jack Hirshleifer. The private and social value of information and the reward to inventive activity. *The American Economic Review*, 61(4):561–574, 1971.

[32] Nicole Immorlica, Adam Tauman Kalai, Brendan Lucier, Ankur Moitra, Andrew Postlewaite, and Moshe Tennenholtz. Dueling algorithms. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 215–224, 2011.

[33] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[34] Ari Juels and Marcus Peinado. Hiding cliques for cryptographic security. *Designs, Codes and Cryptography*, 20(3):269–280, 2000.

[35] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[36] Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

[37] Luděk Kučera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2):193–212, 1995.

[38] Ehud Lehrer, Dinah Rosenberg, and Eran Shmaya. Signaling and mediation in games with common interests. *Games and Economic Behavior*, 68(2):670–682, 2010.

[39] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582. ACM, 2015.

[40] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 529–537, 2001.

[41] Lorenz Minder and Dan Vilenchik. Small clique detection and approximate nash equilibria. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 673–685. Springer, 2009.

[42] Marcin Peski. Comparison of information structures in zero-sum games. *Games and Economic Behavior*, 62(2):732–735, 2008.

[43] Aviad Rubinstein. Eth-hardness for signaling in symmetric zero-sum games. *arXiv preprint arXiv:1510.04991*, 2015.

[44] Michael Spence. Job market signaling. *The quarterly journal of Economics*, 87(3):355–374, 1973.

[45] H. Xu, Z. Rabinovich, S. Dughmi, and M. Tambe. Exploring information asymmetry in two-stage security games. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[46] Haifeng Xu, Rupert Freeman, Vincent Conitzer, Shaddin Dughmi, and Milind Tambe. Signaling in bayesian stackelberg games. In *Proceedings of the 15th International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, 2016.

A Proof of Lemma 3.5

Let $G = (V, E)$. We fix one of the planted k-cliques $S \subseteq V$, and prove Lemma 3.5 for every T satisfying $|T \cap S| > \epsilon |T \cup S|$. Our proof has two steps. First, we show that if we can sample uniformly from $T \cap S$, then we can recover planted clique S in polynomial time with high probability. Then we show that we can simulate sampling from $T \cap S$ by combining sampling from T with some brute force enumeration in polynomial time.

A.1 Step 1

Assume we can sample uniformly from $T \cap S$. Our algorithm for recovering S is as follows.

- Sample: Let R be a sample of $200 \log n$ vertices from $T \cap S$.
- Filter 1: Let $\widetilde{S}$ be all the common neighbors of R.
- Filter 2: Let $\widehat{S}$ be those nodes in $\widetilde{S}$ with at least $k - 1$ neighbors in $\widetilde{S}$.

It is clear that nodes in the k-clique S survive both filtration steps, and therefore $\widehat{S} \supseteq S$. We will show that in fact $\widehat{S} = S$, with high probability. For this, we need the following propositions bounding the "connectivity" of nodes in $V \setminus S$ into S and $T \cap S$. We partition the edges of G into *background edges* E^-, added in step (1) of Definition 3.3, and *foreground edges* E^+ added in step (2).

Proposition A.1. *For every T, the number of vertices v outside S with $|E^-(v, S \cap T)| > 0.6|T \cap S|$ is $O(\log n)$, with high probability.*

Proof. Since $p \leq 0.5$ and $|T \cap S| = \Omega(k) = \omega(\log n)$, this follows directly from Proposition 2.1. □

Proposition A.2. *There are no vertices v outside S with $|E^-(v, S)| > 0.6k$, with high probability.*

Proof. This is by a straightforward application of the Chernoff bound and the union bound. □

Proposition A.3. *Every vertex v outside S has $|E^+(v, S)| = O(\log^2 n) = o(k)$, with high probability.*

Proof. This follows from two facts which hold with high probability: (a) no two planted cliques overlap in more than $O(\log n)$ vertices, and (b) no vertex is in more than $O(\log n)$ planted cliques. We show both next.

For (a), let $S_1, \ldots, S_r$ be the planted cliques (one of which is our S). Fix S_i and S_j with $i \neq j$. The probability a vertex v is in both S_i and S_j is $(\frac{k}{n})^2 < \frac{1}{n}$. A simple application of the Chernoff bound implies that $|S_i \cap S_j| < 4 \log n$ with probability at least $1 - 1/n^4$.[9] By the union bound and the fact that $r < n$, this holds for all pairs $i \neq j$ with probability at least $1 - 1/n^2$.

For (b), observe that for a fixed vertex v the events $\{v \in S_i\}_{i=1}^r$ are independent Bernoulli trials with probability $\frac{k}{n}$ each. Since the number of these events is $r = O(\frac{n}{k})$, the Chernoff bound implies that at most $O(\log n)$ of these events hold with high probability $1 - 1/\operatorname{poly}(n)$. Taking a union bound over all vertices v completes the proof.

□

[9] This also requires showing negative association of the relevant indicator variables — one per node of the graph, indicating membership in both S_i and S_j. We omit the (straightforward) details.

We can now show that no vertex v outside S survives both filtration steps, with high probability. First, combining Propositions A.1 and A.3, we know that for all but at most $O(\log n)$ nodes $v \notin S$ we have that $|E(v, T \cap S)| = E^-(v, T \cap S) + E^+(v, T \cap S) \leq 0.6|T \cap S| + o(k) \leq 0.7|T \cap S|$ (for sufficiently large k). Since R is a random sample of $200 \log n$ vertices from $|T \cap S|$, a simple application of the Chernoff bound shows that $E(v, R) \leq 1.1 \frac{|R|}{|T \cap S|} E(v, T \cap S)$ for every vertex v, with high probability. Therefore, for those vertices with $|E(v, T \cap S)| \leq 0.7|T \cap S|$, none of them connect to more than $0.8|R|$ vertices in our sample R, with high probability. Therefore, at most $O(\log n)$ vertices outside S survive the first filtration step — i.e., $|\widetilde{S} \setminus S| = O(\log n)$ — with high probability.

Second, since $|\widetilde{S} \setminus S| = O(\log n)$, every node $v \in \widetilde{S} \setminus S$ has $|E(v, \widetilde{S})| \leq |E(v, S)| + O(\log n) = |E(v, S)| + o(k)$. Proposition A.2 implies that $|E(v, S)| \leq 0.6k$ with high probability, and therefore no vertex $v \in \widetilde{S} \setminus S$ survives the second filtration step (for sufficiently large k).

A.2 Step 2

In step 1, we assumed the ability to sample $200 \log n$ vertices uniformly at random from $T \cap S$. However, since S is a-priori unknown, this is impossible in the most naive sense. Nevertheless, such sampling can be simulated efficiently as follows: Sample roughly $\frac{200}{\epsilon} \log n$ vertices uniformly from T, and attempt the algorithm of step 1 on every subset of roughly $200 \log n$ of the sampled vertices. This runs in polynomial time. Moreover, since $|T \cap S| \geq \epsilon |T|$, with high probability it is the case that roughly $200 \log n$ of the sampled vertices lie in $T \cap S$, and are distributed uniformly therein.

B Omitted Preliminaries

Probabilistic Bounds on Random Graphs

Proposition B.1. *Let $G \sim \mathcal{G}(n, p)$, and let A be the adjacency matrix of G. For any family $P \subseteq [n] \times [n]$ of pairs of vertices, and $\alpha \geq 1$,*

$$\mathbf{Pr}\left[\frac{1}{|P|} \sum_{(i,j) \in P} A_{ij} \geq \alpha p\right] \leq 2 \exp\left(-\frac{(\alpha - 1)^2 p |P|}{12\alpha}\right)$$

Proof. For an unordered pair $e \in \binom{[n]}{2}$, let x_e denote the indicator variable for the event $e \in E(G)$. By definition of $\mathcal{G}(n, p)$, the indicator variables x_e are independent Bernoulli random variable with parameter p. Observe that

$$\sum_{(i,j) \in P} A_{ij} = \sum_{(i,j) \in X \times Y} x_{\{i,j\}} \tag{5}$$

Since our edges are undirected, whereas the above sum is over ordered pairs, the number of times any of our indicator variables x_e appears in the above sum varies. Specifically, for $e = \{i, j\}$, the variable x_e appears twice if both (i, j) and (j, i) are in P, once if only one of (i, j) and (j, i) is in P, and zero times otherwise. Let F_1 be the set of edges appearing at least once, and F_2 be the set of edges appearing twice. Observe that $|F_2| \leq |F_1| \leq |P|$ and $|F_1| + |F_2| \leq |P|$. We can now simplify expression (5).

$$\sum_{(i,j) \in P} A_{ij} = \sum_{e \in F_1} x_e + \sum_{e \in F_2} x_e \tag{6}$$

Observe that each of $m_1 = \sum_{e \in F_1} x_e$ and $m_2 = \sum_{e \in F_2} x_e$ is the sum of independent Bernoulli random variables with parameter p. Using the Chernoff bound (Proposition B.3, Equation (8)), we can bound the probability that m_1 is too large.

$$\begin{aligned}
\mathbf{Pr}[m_1 \geq p|F_1| + 0.5(\alpha-1)p|P|] &= \mathbf{Pr}[m_1 \geq \mathbf{E}[m_1] + 0.5(\alpha-1)p|P|] \\
&\leq \exp\left(-\frac{0.25(\alpha-1)^2p^2|P|^2}{(2p|F_1| + 0.5(\alpha-1)p|P|)}\right) \\
&= \exp\left(-\frac{0.25(\alpha-1)^2p|P|^2}{2|F_1| + 0.5(\alpha-1)|P|}\right) \\
&\leq \exp\left(-\frac{0.25(\alpha-1)^2p|P|^2}{2|F_1| + \alpha|P|}\right) \\
&\leq \exp\left(-\frac{0.25(\alpha-1)^2p|P|^2}{2\alpha|F_1| + \alpha|P|}\right) \\
&\leq \exp\left(-\frac{0.25(\alpha-1)^2p|P|^2}{3\alpha|P|}\right) \\
&= \exp\left(-\frac{(\alpha-1)^2p|P|}{12\alpha}\right)
\end{aligned}$$

An essentially identical calculation yields a similar bound for m_2.

$$\mathbf{Pr}[m_2 \geq p|F_2| + \frac{1}{2}(\alpha-1)p|P|] \leq \exp\left(-\frac{(\alpha-1)^2p|P|}{12\alpha}\right)$$

Using the union bound, we complete the proof.

$$\begin{aligned}
\mathbf{Pr}\left[\frac{1}{|P|}\sum_{(i,j)\in P} A_{ij} > \alpha p\right] &= \mathbf{Pr}[m_1 + m_2 > \alpha p|P|] \\
&\leq \mathbf{Pr}[m_1 + m_2 > p(|F_1| + |F_2|) + (\alpha-1)p|P|] \\
&\leq \mathbf{Pr}[m_1 > p|F_1| + \frac{1}{2}(\alpha-1)p|P|] + \mathbf{Pr}[m_2 > p|F_2| + \frac{1}{2}(\alpha-1)p|P|] \\
&\leq 2\exp\left(-\frac{(\alpha-1)^2p|P|}{12\alpha}\right)
\end{aligned}$$

□

Proposition B.2 (Proposition 2.1 in main body)**.** *Let $p \in (0,1)$ and $\alpha > 1$ be absolute constants (independent of n), and let $G \sim \mathcal{G}(n,p)$. There is an absolute constant $\beta = \beta(p,\alpha)$ such that the following holds with high probability for all clusters X and Y with $|X|, |Y| > \beta \log n$.*

$$\text{bi-density}_G(X,Y) \leq \alpha p$$

Proof. Consider two cluster sizes $k, \ell \in [n]$. Given a fixed pair of clusters $|X|$ and $|Y|$ with $|X| = k$ an $|Y| = \ell$, Proposition B.1 implies that

$$\begin{aligned}
\mathbf{Pr}\left[\text{bi-density}(X,Y) \geq \alpha p\right] &= \mathbf{Pr}\left[\frac{1}{|X||Y|}\sum_{(i,j)\in X\times Y} A_{ij} \geq \alpha p\right] \\
&\leq 2\exp\left(-\frac{(\alpha-1)^2pk\ell}{12\alpha}\right)
\end{aligned}$$

In particular, since p and α are constants, there is a constant $c > 0$ such that

$$\mathbf{Pr}\left[\text{bi-density}(X, Y) \geq \alpha p\right] \leq \exp(-ck\ell).$$

There are at most n^k clusters of size k, and n^ℓ clusters of size ℓ. Therefore, the probability that any pair of clusters X and Y with $|X| = k$ and $|Y| = \ell$ satisfy bi-density$(X, Y) \geq \alpha p$ is, by the union bound, at most

$$n^{k+\ell} \exp(-ck\ell) = \exp((k + \ell) \log n - ck\ell).$$

When $k, \ell \geq \beta \log n$ for some constant β, this probability is at most

$$\exp((k + \ell) \log n - ck\ell) \leq \exp(2\beta \log^2 n - \beta^2 c \log^2 n).$$

We can choose the constant β large enough so that this probability is $\exp(-\Omega(\log^2 n))$ for every pair of integers $k, \ell \geq \beta \log n$. Taking a union bound over all pairs of integers $k, \ell \in [n]$ with $k, \ell \geq \beta \log n$, of which there are at most n^2, completes the proof. □

Tail Bounds

We use the following convenient forms of the Chernoff bound

Proposition B.3 (Chernoff Bounds)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mu = \sum_{i=1}^n \mathbf{E}[X_i]$. Let $X = \sum_{i=1}^n X_i$ be their sample sum. For every $\alpha \geq 1$ the following holds.*

$$\mathbf{Pr}[X > \alpha\mu] \leq \exp\left(-\frac{(\alpha - 1)^2 \mu}{\alpha + 1}\right) \tag{7}$$

Equivalently, for every $t \geq 0$, the following holds.

$$\mathbf{Pr}[X > \mu + t] \leq \exp\left(-\frac{t^2}{2\mu + t}\right) \tag{8}$$