Recovery of a Planted k-Densest Sub-Hypergraph

Joachim M. Buhmann[1], Luca Corinzia[1], Paolo Penna[1], and Wojciech Szpankowski[2]

[1]Department of Computer Science, ETH Zurich, Switzerland
{jbuhmann, luca.corinzia, paolo.penna}@inf.ethz.ch
[2]Department of Computer Science, Purdue University, USA
szpan@purdue.edu

***Abstract*—Recovery of a planted k-densest sub-hypergraph is a fundamental problem that appears in different contexts, e.g. community detection, average case complexity, and neuroscience applications. The underlying hypergraph parameters determine the geometry of the solution space and the statistical dependency between solutions. This captures whether the structured signal is highly localized, naturally suggesting a criterion to determine the boundary conditions for which the recovery is possible. In this work, we provide new information-theoretic upper and lower bounds for the recovery problem. These bounds apply to the whole spectrum of the hypergraph parameters, ranging from complex combinatorial search problems with high statistical dependency between solutions, to an extremely localized solution space, equivalent to the random energy model. The new bounds improve significantly prior bounds on most of the interesting regimes, and also provide the first results on partial recovery.**

I. Introduction

High dimensional inference problems play a key role in recent machine learning and data analysis applications. Typical scenarios exhibit problem dimensions that are comparable to the sample size, hence precluding effective estimation with no further structure imposed on the underlying signal, such as low rank or sparsity. Examples of such problems include sparse mean estimation, compressive sensing, sparse phase retrieval, low-rank matrix estimation, community detection, planted clique and densest subgraph recovery problems.

In this work we study the problem of recovering a *planted sparse k-densest sub-hypergraph* [1]. In the standard (graph) setting the problem is closely related to detecting a core structure in community detection and resembles the well-known planted clique problem and other community detection models like the stochastic block model (SBM). Recent work suggest that this problem is related to long term memory mechanism in the brain [2]. The hypergraph version introduces high order interactions between nodes which are rather natural in a number of real world situations, e.g. in modeling brain regions [3], [4], [5] and in computer vision applications [6].

Our goal is to understand the information-theoretic limits for this class of recovery problems, thereby establishing the regimes in which algorithms have the ability to recover a hidden structure or signal from noisy measurements and partial information. The setting is informally defined as follows: For N being the total number of nodes, the planted solution is defined by k nodes chosen uniformly at random, and all hyperedges between nodes of this set have *biased* weights, i.e., they are shifted from zero by some value, that defines the signal strength of our estimation problem. All hyperedge weights are perturbed by Gaussian *noise* with zero mean, and the task consists in recovering the planted set of nodes. The hardness and the feasibility of the task depends on the *signal-to-noise ratio (SNR)*, i.e., how strong is the bias compared to the noise level, denoted by γ throughout this work. We are interested in the probability of *exactly* and *partially* recovering the planted solutions in the information theoretic limit, hence we study how much the best possible statistical estimator of such a solution (the maximum likelihood estimator MLE) overlaps with the planted solution and whether this estimator can be calculated efficiently relative to the input size of the estimation problem.

One should note that the hyperedge cardinality h can vary between the two extremes $h = 1$ and $h = k$ of an interaction spectrum. Analysing these extremes yields insight on the information structure of the solution space. For $h = k$ the problem has the highest *localization*, i.e., there is *only one biased* solution and all other solutions are *unbiased* and statistically independent (see the **R**andom **E**nergy **M**odel [7]). At the opposite extreme, for $h = 1$, in addition to the planted solution, *several* other partially overlapping solutions are *biased*. In other words, for $h = 1$, the problem exhibits some statistical *dependencies* between solutions, which become more prominent for large h (e.g., for $h = o(N)$). A second reason for investigating different values of h, are algorithmic issues. Dependencies between solutions and hence the localization property in the solution space render computationally efficient algorithms information theoretically possible (in principle). To illustrate this, consider the following scenarios:

- For h and k very close to each other ($h = k - c_0$ with c_0 being a constant), exhaustive search through all $\binom{N}{k}$ solutions is both efficient and unavoidable since solutions are localized and (almost) statistically *independent*: the input size is $\binom{N}{h}$, and finding the optimum is equivalent to search through all these $\binom{N}{h} = poly(\binom{N}{h})$ edge weights.
- For $h = 1$, exhaustive search through all $\binom{N}{k}$ solutions is clearly inefficient, i.e., $\binom{N}{k}$ is exponentially larger then the input size $\binom{N}{h} = N$. Nonetheless the structure of the solution space is simple and selecting the k nodes with highest weights is efficient and optimally solves the problem.

- The intermediate sparse regime for $1 \ll h \ll k \ll N$ reveals insights into the information structure of sub-hypergraph recovery. In this regime the solution space can be again exponentially larger then the input size. In this regime, algorithms have to address the information theoretical localization effect in the solution space. i.e., two solutions that do not share any parameter are statistically independent.

II. Main Results

A. Setting

We consider *h-uniform hypergraphs* over N nodes, i.e. every subset of h nodes is an *(hyper)edge*. To every hyperedge we associate a weight that is a Gaussian random variable, defined according to the following process. Choose uniformly at random a set K_{planted} of k nodes. All hyperedges inside this set of k nodes have a *positive bias*, and all other are *unbiased*[1]. Formally:

Definition 1 (PkSH, also 2-hWSBM in [8])**.** *Let N, k, and h be positive integers, μ and σ two positive numbers. The couple $(K_{planted}, E)$, where $K_{planted}$ indicate the planted set of nodes and E the h tensor of weights, is drawn under the planted k sub-hypergraph model PkSH(μ, σ, N, k, h) if $|K_{planted}| = k$, is drawn uniformly at random from all the N nodes and the weights $E_{i_1 i_2 \cdots i_h}$ are random variables* conditional independent given $K_{planted}$ *and given by:*

$$E_{i_1 i_2 \cdots i_h} \sim \mu \, \mathbb{I}_{\{i_1 i_2 \cdots i_h \in K_{planted}\}} + \sigma \mathcal{N}(0,1)$$

where $\mathbb{I}_{\{P\}} = 1, 0$ iff predicate P is true or false.

Definition 2. *Exact recovery for the PkSH in Definition 1 is achieved if there exists an algorithm that takes the weight tensor E as input and outputs $\hat{K} = \hat{K}(E)$ such that*

$$\mathbb{P}[K_{planted} = \hat{K}] = 1 - o_N(1).$$

A k'-partial recovery is achieved if

$$\mathbb{P}[|K_{planted} \cap \hat{K}| = k'] = 1 - o_N(1).$$

Our bounds depend on the scale-normalized SNR γ as

$$\gamma := \frac{\hat{\mu}}{\hat{\sigma}} \cdot \sqrt{\frac{\binom{k}{h}}{k}}, \qquad \mu = \hat{\mu} \log N, \qquad \sigma = \hat{\sigma} \sqrt{\frac{\log N}{2}} \tag{1}$$

The scaling $\sqrt{\frac{\binom{k}{h}}{k}}$ is used to have non-trivial recovery thresholds located at finite values. Every solution consists of a subset of k nodes; its total weight is the sum of the weights of all $\binom{k}{h}$ hyperedges that it contains. If a solution shares k' nodes with the planted solution, it has exactly $\binom{k'}{h}$ biased hyperedges. It is convenient to partition all solutions according to k' and let $\mathsf{S}_{k'}$ be the set of all solutions which share k' nodes with the planted solution. Note that any solution S_i is itself a Gaussian random variable, whose bias depends on the number k' of nodes in the planted part as:

$$S_i \sim \mathcal{N}(\mu_{k'}, \sigma_k^2) \qquad \text{for all } S_i \in \mathsf{S}_{k'},$$

where

$$\mu_{k'} = \binom{k'}{h} \mu \qquad \sigma_k^2 = \binom{k}{h} \sigma^2 \ . \tag{2}$$

All parameters k, k', h depend on N, but the dependency is hidden in the notation for convenience.

[1]This model can be seen both as a planted densest sub-hypergraph or a two communities weighted SBM on hypergraph.

B. Results

Let us denote by $\beta_0, \beta_0', \beta_0'' \in [0,1]$ the constants satisfying respectively

$$\beta_0 := \lim_N \frac{\log k}{\log N} \quad \beta_0' := \lim_N \frac{\log(k-k')}{\log N} \quad \beta_0'' := \lim_N \frac{\log k'}{\log N} \tag{3}$$

where k' is the number of nodes detected in partial recovery. Note that these constants always exist since $k' < k \leq N$ and that $\beta_0'' \leq \beta_0$. We prove the following result:

Theorem 3 (Main Result)**.** *There exist an upper bound γ_{UB} and a lower bound γ_{LB} which determine whether recovery is possible or not possible, respectively:*
Upper bound: *For $\gamma > \gamma_{UB}^{(k')}$ it is* possible *to perform a $(k'+1)$-partial recovery; for $\gamma > \gamma_{UB} := \gamma_{UB}^{(k-1)}$ it is possible to perform exact recovery. These thresholds depend on the problem parameters (N, k, k', h) as follows:*

$$\gamma_{UB}^{(k')} := \sqrt{1 + \beta_0 - 2\beta_0' + \beta_0''} + \sqrt{\beta_0 - \beta_0' + \beta_0''} \tag{4}$$

$$\gamma_{UB} := \sqrt{1 + 2\beta_0} + \sqrt{2\beta_0} \tag{5}$$

Lower bound: *For $\gamma < \gamma_{LB}$, exact recovery is* impossible. *Specifically, the lower bound depends on the parameters (N, k, h) as follows:*

$$\gamma_{LB} := \max\left\{ \sqrt{\frac{1}{h}}, \sqrt{\frac{1 - \beta_0 - \epsilon}{2} \frac{k_\star}{k}} \right\} \tag{6}$$

for any $\epsilon > 0$ constant and where $k_\star$ is any sequence such that $\binom{k_\star}{h} / \binom{k}{h} \to 0$.

As we discuss below, our results improve prior bounds in [8] along two directions. First, they provide *tighter bounds* for a wide range of parameters, especially when k is not limited to be logarithmic in N. Second, we consider *partial* recovery, for which we provide the first asymptotic bounds for this problem.

Corollary 4 (partial recovery)**.** *For any constant $\rho_0 \in (0,1)$, it is possible to perform a $k' = \rho_0 k$-partial recovery if the SNR satisfies $\gamma > 1 + \sqrt{\beta_0}$. For $\alpha_0 := \lim_N \frac{\log k'}{\log k} \in (0,1)$ constant, then it is sufficient to have a SNR that satisfies $\gamma > \sqrt{1 - \beta_0 + \alpha_0 \beta_0} + \sqrt{\alpha_0 \beta_0}$.*

Proof. It follows from (4) with $\beta_0' = \beta_0$ and $\beta_0'' = \alpha_0 \beta_0$. □

Intuitively the first condition refers to a partial recovery of a linear fraction of the k planted nodes, while the second condition is a weaker partial recovery that aims at detecting some root of the planted part.

C. Comparison with prior bounds

For convenience here we report prior bounds on exact recoverability, rewritten according to our notation.

Theorem 5 (Theorem 5, [8]). *For any $2 \leq h \leq k$, exact recovery is impossible if $\gamma < \gamma_- = \sqrt{\frac{1}{h}}$ and possible if $\gamma > \gamma_+$, where the upper threshold is defined according to the different k and h regimes as:*

$$\gamma_+ = \begin{cases} \sqrt{2} & \frac{\binom{k}{h}}{k} = o(\log N) \\ 2\sqrt{\frac{(1+\log 2+\frac{1}{c})\binom{k}{h}}{k \log N}} & \frac{\binom{k}{h}}{k}/\log N \to c, c \in \mathbb{R}^+ \cup \{+\infty\} \\ 2\sqrt{\frac{(1+\log 2)\binom{k}{h}}{(1-\beta_0)k \log N}} & k \lessapprox N^{\beta_0},\ 0 < \beta_0 < 1 \end{cases} \tag{7}$$

Tighter bounds then eq. (7) are achieved if $\gamma_{LB} > \gamma_-$ and $\gamma_{UB} < \gamma_+$. We note from eq. (6) that the inequality $\gamma_{LB} \geq \gamma_-$ is obviously satisfied since $\gamma_{LB} \coloneqq \max\left\{\gamma_-, \sqrt{\frac{1-\beta_0-\epsilon}{2}\frac{k_\star}{k}}\right\}$, while the strict inequality is obtained whenever $\frac{k}{h} = o(k_*)$, and the condition for eq. (6), namely $\binom{k_*}{h}/\binom{k}{h} \to 0$ holds. This condition can be easily achieved on a wide spectrum of regimes, e.g. for $k_* = h$ and $k = o(h^2)$, while it is not always satisfied, e.g. for h constant. Regarding the upper bound, we can observe that the condition $\frac{\binom{k}{h}}{k} = o(\log N)$ implies $\beta_0 = 0$ for any $2 \leq h < k-1$, for which $\gamma_{UB} < \gamma_+$. In the regimes $h = \{k-1, k\}$ however, the condition $\frac{\binom{k}{h}}{k} = o(\log N)$ is always satisfied and hence $\gamma_{UB} = \sqrt{1+2\beta_0} + \sqrt{2\beta_0} < \sqrt{2}$ whenever $\beta_0 < \frac{1}{16}$. In the second regime for γ_+ we can first observe that the cases $h \in \{k-1, k\}$ are always excluded. For $h < k-1$ two conditions are possible: (i) $\frac{\binom{k}{h}}{k}/\log N \to +\infty$, in which case $\gamma_{UB} \ll \gamma_+$, and (ii) $\frac{\binom{k}{h}}{k}/\log N \to c \in \mathbb{R}^+$, in which case it follows that $\beta_0 = 0$ and hence $\gamma_{UB} < \gamma_+$. In the third regime of eq. (7) $k \approx N^{\beta_0}$ holds, hence it follow easily that $\gamma_{UB} \ll \gamma_+$. Note also that with $\beta_0 = 0$, γ_{UB} matches the critical SNR for the problem conjectured in [8].

III. Related work

High dimensional inference problems have been extensively studied recently in the statistics and computer science communities for showing interesting statistical and computational thresholds in the recoverability of a structured planted signal, where typical examples for structure include low rank and sparseness. The most widely studied planted model in literature is the *planted clique problem*, where a clique of size k is hidden in a Erdös-Rényi random graph (ER) of size N. This problem exhibits a *statistical-computational gap* [9], such that recover of the planted clique is information-theoretic possible for size $k \geq 2\log_2 N$, while the best known polynomial algorithms require $k = \Omega(N^{1/2})$ [10]. It is an open problem to assess whether the regime $k = o(\log N)$ is indeed intractable. Many variations of this problem have been introduced, allowing for generic values of the ER parameter, for random deletion of edges in the planted clique, and for weighted edges (i.e. planted densest subgraph problem [11]).

Similar statistical and computational thresholds have been observed also in the problem of sparse principal component analysis [12], [13] and in the stochastic block model (SBM) and its variations. The standard SBM exhibits no gap, with matching statistical and computational thresholds that have been found recently for the symmetric [14], [15] and the non-symmetric [16] model. The SBM extension to multi-community detection (also known as planted clustering problem) shows instead statistical-computational gaps [17]. Statistical and computational thresholds are still unknown for various SBM generalizations, like weighted-SBM (WSBM) [18], [19], [20] and SBM on (homogeneous) hypergraph (hSBM) [21], [22], that can model additional information of the problem that is expressed respectively by edge weights and higher order node interactions. Information-theoretic results are given in [23] for the homogeneous WSBM with equally sized communities, in [22], [24] for the spectral algorithms respectively on uniform and non-uniform hSBM, in [25] for the homogeneous equally sized community WSBM on hypergraphs and in [8] for the non-equally sized communities WSBM on hypergraphs.

Recent work has been performed with statistical physics methods to understand the nature of the information-theoretic and computational thresholds. Planted inference models are mapped to related disordered physical systems (e.g. spin glass, Potts and Mattis models [26]) such that the statistical and computational thresholds are mapped to the phase transitions of these systems [27], [28]. The hard phase of planted inference problems is conjectured to correspond to the spin-glass phase of related disorder systems, that is typically characterized by exponentially many clusters of solutions with close energy values and large energy walls between them [29], [30], [31], while the impossible and the easy regimes are mapped respectively to the paramagnetic and the ferromagnetic phases. This line of research was inspired by the seminal work done on error correcting codes using p-spin glass models [32], [33], [34] and its random energy model (REM) limit [35], [36], [37], with recent developments focused on recovery conditions for the spike Wigner model [38], [39], [40], [41], [42], stochastic block model [43], [44], generalized linear models [45]. These approaches have been mainly applied to dense scenarios, with the structure of the measurement imposed by the low dimensionality of the signal. The extension to sparseness scenarios have been developed recently in [46]. A review on the field is given in [47].

IV. Recovery via concentration bounds

To study the recoverability thresholds we analyze the behaviour of the maximum likelihood estimator of the planted set $K_{planted}$, that it can be easily identified as the k-densest sub-hypergraph (see Theorem 4 [8] for a proof). We consider the generic k'-partial recover for the MLE estimator. Exact recovery corresponds to the case $k' = k-1$. In the following most proofs are omitted and reported in the long version of the paper.

Lemma 6. *Let $P^{(k')}_{\mathsf{recover}}$ be the probability that the MLE recovers at least $k'+1$ nodes from the planted solution, and $P^{(k')}_{\mathsf{failrecover}} = 1 - P^{(k')}_{\mathsf{recover}}$ the probability that it fails in doing so. For any m let S_m denote the set of all solutions that share exactly m nodes with the planted solution S_{planted}. Denote in the following for any set A, $\max(\mathsf{A}) \coloneqq \max_{\mathsf{x}\in\mathsf{A}} \mathsf{x}$. Then the following holds:*

$$P^{(k')}_{\mathsf{failrecover}} \leq \sum_{m=0}^{k'} \mathbb{P}\left(S_{\mathsf{planted}} \leq \max(\mathsf{S}_m)\right) \tag{8}$$

$$P_{\mathsf{failrecover}} \coloneqq P^{(k-1)}_{\mathsf{failrecover}} \geq \mathbb{P}(S_{\mathsf{planted}} < \max(\mathsf{S}_{k'})) \tag{9}$$

for any $k' \in \{0,\dots,k-1\}$.

Proof. The upper bound in (8) follows since $S_{\mathsf{planted}} > S$ for all $S \in \mathsf{S}_0 \cup \mathsf{S}_1 \cup \cdots \cup \mathsf{S}_m$. Then the MLE must return S_{planted} or some solution in $\mathsf{S}_{m+1} \cup \cdots \cup \mathsf{S}_k$, and the inequality follow from the union bound on the probability $\mathbb{P}\left(\bigcup_{m=0}^{k'} \{S_{\mathsf{planted}} \leq \max(\mathsf{S}_m)\}\right)$. For the lower bound in (9) we can observe that if $S_{\mathsf{planted}} < \max(\mathsf{S}_{k'})$ for some $k' < k$, then the MLE cannot return S_{planted}, and thus it fails to exactly recover the planted solution. □

Intuitively speaking, our goal is to distinguish between the case in which recovery is possible from the one in which it is impossible, i.e., whether γ is in the regime such that

$$\mathbb{P}(P^{(k')}_{\mathsf{failrecover}}) \to 0 \qquad \text{or} \qquad \mathbb{P}(P^{(k')}_{\mathsf{failrecover}}) \to 1 \ .$$

We shall reduce this question to the study of the probabilities $\mathbb{P}(S_{\mathsf{planted}} < \max(\mathsf{S}_m))$, so to determine the values of γ for which

$$\mathbb{P}(S_{\mathsf{planted}} \leq \max(\mathsf{S}_m)) \to 0 \quad \text{or} \quad \mathbb{P}(S_{\mathsf{planted}} \leq \max(\mathsf{S}_m)) \to 1$$

where in the left scenario (recovery possible) we need these probabilities to go to zero *sufficiently fast* to apply the union bound (eq. (8)) over all different m.

A. Probability tools

We shall use the well-known inequalities on *tail distribution of Gaussians*: For any $X \sim \mathcal{N}(\mu,\sigma^2)$ and any $c > 0$, it holds that (see [48, Section 7.1])

$$(\frac{1}{c} - \frac{1}{c^2}) \cdot \frac{e^{-c^2/2\sigma^2}}{\sqrt{2\pi}} \leq \mathbb{P}(X > \mu + c) \leq \frac{1}{c} \cdot \frac{e^{-c^2/2\sigma^2}}{\sqrt{2\pi}} \ . \tag{10}$$

We typically deal with solutions (events) which are *dependent*, for which the following *Chung-Erdös inequality* and the well-known *union bound* will be used: Given n events $A_1,\dots A_n$, it holds that (see [49])

$$\frac{\left(\sum_{i=1}^n \mathbb{P}(A_i)\right)^2}{\sum_{i=1}^n \mathbb{P}(A_i) + \sum_{i\neq j} P(A_i \cap A_j)} \leq \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \tag{11}$$

$$\leq \sum_{i=1}^n \mathbb{P}(A_i) \tag{12}$$

where the first inequality is the *Chung-Erdös* bound and the second is the *union bound*.

B. Upper bound (approximate or exact recovery is possible)

In this section, we prove the upper bound on the signal-to-noise ratio in order to guarantee recovery. Recall that we consider the problem of recovering $k' \leq k$ nodes inside the planted solution. For the purpose of the analysis, we define the following quantities depending on $k' \in \{0,\dots,k-1\}$:

$$d(k') \coloneqq \binom{k}{h} - \binom{k'}{h}, \quad Q_{k'} \coloneqq \binom{N-k}{k-k'},$$
$$M_{k'} \coloneqq \binom{k}{k'}\binom{N-k}{k-k'} = \binom{k}{k-k'}\binom{N-k}{k-k'}$$

Note that, for each *fixed* subset of k' nodes of the planted solution, there are $Q_{k'}$ solutions that share these k' nodes with the planted solution. Moreover, there are exactly $M_{k'}$ solutions that share k' nodes with the planted solution. Each such solution sharing k' nodes with the planted solution differs in $d(k')$ edges with the planted solution. Finally, we let

$$t_\Delta = \sigma\sqrt{\Delta 2\log N} \ . \tag{13}$$

Lemma 7. *Fix an arbitrary subset $F \subset K_{planted}$ of k' nodes of the planted solution, with $k' \in \{0,\dots,k-1\}$, and let $\mathsf{S}^{(F)}_{k'}$ be the set of all solutions that share exactly this set of k' nodes with the planted solution. For any $S \in \mathsf{S}^{(F)}_{k'}$ let $S^{(-F)}$ denote the sum of the weights in the non-common part, that is the sum of $d(k') = \binom{k}{h} - \binom{k'}{h}$ edge weights in S and with none of their nodes contained in F. Denote by $\mathsf{S}^{(-F)}_{k'}$ the set of all the $S^{(-F)}$. For any $\Delta > 0$ and any $S^{(-F)}$ as above, it holds that*

$$\mathbb{P}\left(S^{(-F)} > t_\Delta\right) \leq p_{k',\Delta} \coloneqq \left(\frac{1}{N}\right)^{\frac{\Delta}{d(k')}} \frac{1}{\sqrt{\pi\Delta}\cdot 2\hat{\sigma}\log N} \ . \tag{14}$$

Moreover, the following holds:

$$\mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) > t_\Delta\right) \leq Q_{k'} \cdot p_{k',\Delta} \tag{15}$$

$$\mathbb{P}\left(S^{(-F)}_{\mathsf{planted}} < d(k')\mu - t_\Delta\right) \leq p_{k',\Delta} \tag{16}$$

where $\left|\mathsf{S}^{(-F)}_{k'}\right| = \binom{N-k}{k-k'} = Q_{k'}$.

Theorem 8. *For every $\eta \geq 0$ and for every $k' \in \{0,\dots,k-1\}$, let $\gamma^{(k')}_{UB_\eta} \coloneqq \sqrt{\frac{\binom{k}{h}}{kd(k')}} \cdot UB_\eta(k')$ where*

$$UB_\eta(k') \coloneqq \sqrt{\frac{\log M_{k'}}{\log N} + \eta} + \sqrt{\frac{\log\binom{k}{k'}}{\log N} + \eta} \ . \tag{17}$$

Then, for any $\gamma > \gamma^{(k')}_{UB_\eta}$ it holds that

$$\mathbb{P}\left(\max(\mathsf{S}_{k'}) > S_{\mathsf{planted}}\right) \in O\left(\frac{1}{N^\eta}\cdot\frac{1}{\hat{\sigma}\log k}\right) \ .$$

Proof Idea. By the union bound over the $\binom{k}{k'}$ possible fixed subsets F of k' nodes, we get $\mathbb{P}\left(\max(\mathsf{S}_{k'}) > S_{\mathsf{planted}}\right) \leq \binom{k}{k'} \cdot \mathbb{P}(\max(\mathsf{S}^{(-F)}_{k'}) > S^{(-F)}_{\mathsf{planted}})$. Moreover, for any t, we have $\mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) > S^{(-F)}_{\mathsf{planted}}\right) \leq \mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) > t\right) +$

$\mathbb{P}\left(t \geq S^{(-F)}_{\mathsf{planted}}\right)$. Calculations show that, since $\gamma > \gamma^{(k')}_{UB_\eta}$, there exists a particular t such that Lemma 7 implies that both these two probabilities go to 0 sufficiently fast. See Appendix A for the reminder of the proof. □

The constants β_0 and β'_0 in eq. (3) provide bounds on the fractions in eq. (17). Using the union bound over all $m = \{0, \dots, k'\}$ we can set $\eta = \beta''_0$ so that the right hand side of eq. (8) goes to 0. This leads to the upper bound in theorem 3.

Corollary 9. *For any $k' \in \{1, \dots, k\}$ and for every γ such that*

$$\gamma > \gamma^{(k')}_{UB} := \sqrt{1 + \beta_0 - 2\beta'_0 + \beta''_0} + \sqrt{\beta_0 - \beta'_0 + \beta''_0} \ .$$

it holds that

$$P^{(k')}_{\mathsf{recover}} \to 1 \ .$$

C. Lower bound (exact recovery is impossible)

To prove the lower bound on the recovery threshold, we make use of eq. (11) and of bounds on the number intersecting solutions at given k', given in the following:

Lemma 10. *For any $\rho_0 \in (0,1)$, for any k and any $k' \in \{0, \dots, k-1\}$ it holds that*

$$\frac{\log \binom{N-k}{k-k'} - \log\left((k-k')\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\right)}{\log N} \gtrsim \ell(k'')$$

where

$$\ell(k'') := \begin{cases} k''(1-\beta'_0) & \text{if } k'' \leq (k-k')\rho_0 \\ (k-k')(\rho_0 - \beta'_0) & \text{if } k'' > (k-k')\rho_0 \end{cases}$$

and $k'' \in \{1, k-k'-1\}$ is the cardinality of the intersection between non-planted solutions.

Lemma 11. *Given the set $\mathsf{S}^{(-F)}_{k'}$ as defined in Lemma 7, it satisfies*

$$\mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) > t_\Delta\right) \to 1 \tag{18}$$

for every $\Delta < d(k') \min\{LB(k'), (k-k')(1-\beta'_0)\}$ where

$$LB(k') := \min_{k'' \in I(k')} \frac{\ell(k'')}{2} \frac{D(k',k'')}{D(k',k'') - d(k')} \tag{19}$$

where $D(k',k'') = a(k'') + d(k') + \sqrt{8a(k'')(d(k') - a(k''))}$, $a(k'') = \binom{k'+k''}{h} - \binom{k'}{h}$, and $I(k') = \{\max(1, h-k'), \dots, k-k'-1\}$.

Proof Idea. The proof in based on the use of the Chung-Erdös bound (eq. (11)). For any two solutions S_i and S_j in $\mathsf{S}^{(F)}_{k'}$ we consider the corresponding $S^{(-F)}_i$ and $S^{(-F)}_j$ in $\mathsf{S}^{(-F)}_{k'}$, that is, the contribution of the hyperedges that are not in the fixed part F. We then consider the events $\mathbb{P}(A_i) := \mathbb{P}(S^{(-F)}_i > t)$ for a suitable t for the application of the Chung-Erdös bound. These $S^{(-F)}_i$ and $S^{(-F)}_j$ share k'' nodes outside the planted part and thus are *dependent* as long as $k'' \geq h - k'$. In this case, we consider the common part C_{ij} and the remaining part $\hat{S}^{(-F)}_i$ and $\hat{S}^{(-F)}_j$ as the hyperedges in both $S^{(-F)}_i$ and $S^{(-F)}_j$, and those that are only in one of the two, respectively. Then, we provide an upper bound on $\mathbb{P}(A_i \cap A_j) := \mathbb{P}(S^{(-F)}_i > t \cap S^{(-F)}_j > t)$. In particular, for $t = t_c + t_{k''}$ we have $\mathbb{P}(S^{(-F)}_i > t \cap S^{(-F)}_i > t) \leq \mathbb{P}(C_{ij} > t_c) + \mathbb{P}(\hat{S}^{(-F)}_i > t_{k''} \cap \hat{S}^{(-F)}_j > t_{k''})$. All these three random variables are Gaussians and $\hat{S}^{(-F)}_i$ and $\hat{S}^{(-F)}_j$ are independent (as they have no common nodes). Hence, $\mathbb{P}(\hat{S}^{(-F)}_i > t_{k''} \cap \hat{S}^{(-F)}_j > t_{k''}) = \mathbb{P}(\hat{S}^{(-F)}_i > t_{k''})\mathbb{P}(\hat{S}^{(-F)}_j > t_{k''})$, and all these probabilities can be bounded via the Gaussian tails (eq. (10)). The rest of the proof is devoted to optimize t and $t_{k''}$, for all possible k'', so that the upper bound on $\mathbb{P}(S^{(-F)}_i > t \cap S^{(-F)}_j > t)$ allows the fraction in the Chung-Erdös inequality to converge to 1 as desired. Intuitively, the term $\sum_{i \neq j} \mathbb{P}(A_i \cap A_j)$ is lower bounded using lemma 10 and is not too big compared to $\sum_i \mathbb{P}(A_i)$. The full proof is given in Appendix B. □

Theorem 12. *For any constant $\rho_0 \in (0,1)$ and for any $\gamma < \gamma^{(k')}_{LB} := \sqrt{\frac{\binom{k}{h} \min\{LB(k'), (k-k')(1-\beta'_0)\}}{kd(k')}}$ it holds that*

$$\mathbb{P}\left(S_{\mathsf{planted}} > \max(\mathsf{S}_{k'})\right) \to 0 \tag{20}$$

where $LB(k')$ is defined as in Lemma 11.

This theorem affirms under which regime the planted solution is "defeated" by some solution in $\mathsf{S}_{k'}$. By applying this result with $k' = 0$ we obtain the main result of theorem 3. The actual proof consists in showing that, for any $k_\star$ such that $\binom{k'+k_*}{h} / \binom{k}{h} \to 0$, Lemma 10 implies $LB(k') \geq (1/2)\min\{k_\star(1-\beta_0), (k-k')(\rho_0 - \beta_0)\}$ for any $\rho_0 \in (0,1)$.

Corollary 13. *For any constant $\rho_0 \in (0,1)$ and for any $k_\star$ such that $\binom{k_*}{h} / \binom{k}{h} \to 0$ the following holds: If $\gamma < \max\left\{\sqrt{\frac{1}{h}}, \sqrt{\frac{1-\beta_0-\epsilon}{2}\frac{k_\star}{k}}\right\}$, then $P_{\mathsf{recover}} \to 0$.*

V. Conclusions

Information-theoretic thresholds constitute a fruitful benchmarks for algorithm performance, establishing optimality conditions and impossibility results. Typical example of models that exhibits non-trivial information-theoretic and computational properties are high dimensional inference problems with structured signal, often sparse or low dimensional. Motivated by this, we study the information-theoretic limit for recovery of a planted k-densest sub-hypergraph. We provide new upper and lower bounds for exact and partial recovery that strictly improve on prior known bounds on a wide range of parameters. As future research directions, we plan to provide matching recoverability thresholds for the planted k-densest sub-hypergraph for any regime of hyperedge cardinality h, alongside an analysis of algorithmic and computational issues on the same model.

REFERENCES

[1] E. Chlamtac, M. Dinitz, C. Konrad, G. Kortsarz, and G. Rabanca, "The Densest k-Subhypergraph Problem," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016)*, vol. 60, 2016, pp. 6:1–6:19.

[2] R. A. Legenstein, "Long term memory and the densest k-subgraph problem," in *9th Innovations in Theoretical Computer Science Conference*, 2018.

[3] S. Gu, M. Yang, J. D. Medaglia, R. C. Gur, R. E. Gur, T. D. Satterthwaite, and D. S. Bassett, "Functional hypergraph uncovers novel covariant structures over neurodevelopment," *Human brain mapping*, vol. 38, no. 8, pp. 3823–3835, 2017.

[4] Z. Wang, J. Liu, N. Zhong, Y. Qin, H. Zhou, J. Yang, and K. Li, "A naive hypergraph model of brain networks," in *International Conference on Brain Informatics*. Springer, 2012, pp. 119–129.

[5] C. Zu, Y. Gao, B. Munsell, M. Kim, Z. Peng, Y. Zhu, W. Gao, D. Zhang, D. Shen, and G. Wu, "Identifying high order brain connectome biomarkers via learning on hypergraph," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2016, pp. 1–9.

[6] J.-M. Jolion and W. Kropatsch, *Graph based representations in pattern recognition*. Springer Science & Business Media, 2012, vol. 12.

[7] B. Derrida, "Random-energy model: An exactly solvable model of disordered systems," *Physical Review B*, vol. 24, no. 5, p. 2613, 1981.

[8] L. Corinzia, P. Penna, L. Mondada, and J. M. Buhmann, "Exact recovery for a family of community-detection generative models," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 415–419.

[9] J. Steinhardt, "Does robustness imply tractability? A lower bound for planted clique in the semi-random model," *arXiv preprint arXiv:1704.05120*, 2017.

[10] B. Barak, S. B. Hopkins, J. Kelner, P. Kothari, A. Moitra, and A. Potechin, "A nearly tight sum-of-squares lower bound for the planted clique problem," in *57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2016, pp. 428–437.

[11] E. Arias-Castro and N. Verzelen, "Community detection in random networks," *arXiv preprint arXiv:1302.7099*, 2013.

[12] Y. Deshpande and A. Montanari, "Information-theoretically optimal sparse pca," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 2197–2201.

[13] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Conference on Learning Theory*, 2013, pp. 1046–1066.

[14] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2016.

[15] E. Mossel, J. Neeman, and A. Sly, "Consistency thresholds for the planted bisection model," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 69–75.

[16] E. Abbe and C. Sandon, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, 2015, pp. 670–688.

[17] Y. Chen and J. Xu, "Statistical-computational phase transitions in planted models: The high-dimensional setting," in *International Conference on Machine Learning*, 2014, pp. 244–252.

[18] C. Aicher, A. Z. Jacobs, and A. Clauset, "Adapting the stochastic block model to edge-weighted networks," *arXiv preprint arXiv:1305.5782*, 2013.

[19] ——, "Learning latent block structure in weighted networks," *Journal of Complex Networks*, vol. 3, no. 2, pp. 221–248, 2014.

[20] T. P. Peixoto, "Nonparametric weighted stochastic block models," *Physical Review E*, vol. 97, no. 1, p. 012306, 2018.

[21] I. Chien, C.-Y. Lin, and I.-H. Wang, "Community detection in hypergraphs: Optimal statistical limit and efficient algorithms," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 871–879.

[22] D. Ghoshdastidar and A. Dukkipati, "Consistency of spectral partitioning of uniform hypergraphs under planted partition model," in *Advances in Neural Information Processing Systems*, 2014, pp. 397–405.

[23] V. Jog and P.-L. Loh, "Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence," *arXiv preprint arXiv:1509.06418*, 2015.

[24] D. Ghoshdastidar, A. Dukkipati *et al.*, "Consistency of spectral hypergraph partitioning under planted partition model," *The Annals of Statistics*, vol. 45, no. 1, pp. 289–315, 2017.

[25] C. Kim, A. S. Bandeira, and M. X. Goemans, "Community detection in hypergraphs, spiked tensor models, and sum-of-squares," in *2017 International Conference on Sampling Theory and Applications (SampTA)*. IEEE, 2017, pp. 124–128.

[26] D. L. Stein and C. M. Newman, *Spin glasses and complexity*. Princeton University Press, 2013, vol. 4.

[27] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*, ser. International Series of Monographs on Physics. Oxford University Press, 2001, no. 111.

[28] Y. Iba, "The nishimori line and bayesian statistics," *Journal of Physics A: Mathematical and General*, vol. 32, no. 21, p. 3875, 1999.

[29] R. Monasson and R. Zecchina, "Statistical mechanics of the random k-satisfiability model," *Physical Review E*, vol. 56, no. 2, p. 1357, 1997.

[30] M. Mézard, T. Mora, and R. Zecchina, "Clustering of solutions in the random satisfiability problem," *Physical Review Letters*, vol. 94, no. 19, p. 197205, 2005.

[31] D. Achlioptas and F. Ricci-Tersenghi, "On the solution-space geometry of random constraint satisfaction problems," in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*. ACM, 2006, pp. 130–139.

[32] N. Sourlas, "Spin-glass models as error-correcting codes," *Nature*, vol. 339, no. 6227, p. 693, 1989.

[33] P. Ruján, "Finite temperature error-correcting codes," *Physical review letters*, vol. 70, no. 19, p. 2968, 1993.

[34] O. Kinouchi, "Optimal decoding temperature for error-correcting codes," *J. Phys. A, Math. and Gen.*, vol. 25, pp. 6243–6250, 1992.

[35] B. Derrida, "The random energy model," *Physics Reports*, vol. 67, no. 1, pp. 29–35, dec 1980.

[36] M. Talagrand, *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models*. Springer Verlag, 2003.

[37] A. Bovier, *Statistical mechanics of disordered system. A mathematical perspective*. Cambridge University Press, 2006.

[38] F. Krzakala, J. Xu, and L. Zdeborová, "Mutual information in rank-one matrix estimation," in *2016 IEEE Information Theory Workshop (ITW)*. IEEE, 2016, pp. 71–75.

[39] M. Dia, N. Macris, F. Krzakala, T. Lesieur, L. Zdeborová *et al.*, "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula," in *Advances in Neural Information Processing Systems*, 2016, pp. 424–432.

[40] J. Barbier, M. Dia, N. Macris, F. Krzakala, and L. Zdeborová, "Rank-one matrix estimation: analysis of algorithmic and information theoretic limits by the spatial coupling method," *arXiv preprint arXiv:1812.02537*, 2018.

[41] J. Barbier and N. Macris, "The adaptive interpolation method for proving replica formulas. applications to the curie–weiss and wigner spike models," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 29, p. 294002, 2019.

[42] B. Aubin, B. Loureiro, A. Maillard, F. Krzakala, and L. Zdeborová, "The spiked matrix model with generative priors," *arXiv preprint arXiv:1905.12385*, 2019.

[43] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, 2016.

[44] J. Barbier, C. L. Chan, and N. Macris, "Mutual information for the stochastic block model by the adaptive interpolation method," *arXiv preprint arXiv:1902.07273*, 2019.

[45] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Optimal errors and phase transitions in high-dimensional generalized linear models," *Proceedings of the National Academy of Sciences*, vol. 116, no. 12, pp. 5451–5460, 2019.

[46] J. Barbier and N. Macris, "0-1 phase transitions in sparse spiked matrix estimation," *arXiv preprint arXiv:1911.05030*, 2019.

[47] A. S. Bandeira, A. Perry, and A. S. Wein, "Notes on computational-to-statistical gaps: predictions using statistical physics," *Portugaliae Mathematica*, vol. 75, no. 2, pp. 159–186, 2018.

[48] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 2008, vol. 2.

[49] W. Szpankowski, *Average case analysis of algorithms on sequences*. John Wiley & Sons, 2011, vol. 50.

In both appendices we indicate by $\gamma = \frac{\hat{\mu}}{\hat{\sigma}}$ and the relative thresholds values the *unnormalized* SNR, in order to avoid clutter in the proofs. Remember however that the SNR, indicated as γ in the main text, is normalized according to the scale $\frac{\binom{k}{h}}{k}$.

Appendix A
Proof for section IV-B

Proof of Lemma 7. Observe that each $S^{(-F)}$ consists of $d(k') = \binom{k}{h} - \binom{k'}{h}$ non-biased edges, and therefore $S^{(-F)} \sim \mathcal{N}(0, d(k')\sigma^2)$. Hence, by applying eq. (10) with $t = t_\Delta = \sigma\sqrt{\Delta 2 \log N}$ we have

$$\mathbb{P}\left(S^{(-F)} > t_\Delta\right) \overset{(10)}{\leq} \frac{1}{t_\Delta} \cdot \frac{e^{-t_\Delta^2/2d(k')\sigma^2}}{\sqrt{2\pi}} = \frac{e^{-(\Delta \log N)/d(k')}}{t_\Delta\sqrt{2\pi}} = \frac{N^{-\Delta/d(k')}}{\sqrt{\pi\Delta} \cdot 2\hat{\sigma}\log N}\ , \tag{21}$$

which proves eq. (14). By the union bound and eq. (14) we obtain eq. (15):

$$\mathbb{P}\left(\max(\mathsf{S}_{k'}^{(-F)}) > t\right) = \mathbb{P}\left(\bigcup_{S^{(-F)} \in \mathsf{S}_{k'}^{(-F)}} S > t\right) \leq Q_{k'} \cdot p_{k',\Delta}\ .$$

Finally, eq. (16) holds by observing that, since all $d(k')$ edges of $S_{\mathsf{planted}}^{(-F)}$ are biased, we have $S_{\mathsf{planted}}^{(-F)} \sim \mathcal{N}(d(k')\mu, d(k')\sigma^2)$. Therefore, by applying eq. (10) with $t = d(k')\mu + t_\Delta$ we get

$$\mathbb{P}\left(S_{\mathsf{planted}}^{(-F)} < d(k')\mu - t_\Delta\right) = \mathbb{P}\left(S_{\mathsf{planted}}^{(-F)} > d(k')\mu + t_\Delta\right) \overset{(10)}{\leq} \frac{1}{t_\Delta} \cdot \frac{e^{-t_\Delta^2/2d(k')\sigma^2}}{\sqrt{2\pi}}$$

and the remaining of the proof is as above in eq. (21). □

Proof of Theorem 8. For any $k' \in \{0, \dots k-1\}$ we have

$$\mathbb{P}\left(\max(\mathsf{S}_{k'}) > S_{\mathsf{planted}}\right) \leq \binom{k}{k'}\mathbb{P}\left(\max(\mathsf{S}_{k'}^{(F)}) > S_{\mathsf{planted}}\right) = \binom{k}{k'}\mathbb{P}\left(\max(\mathsf{S}_{k'}^{(-F)}) > S_{\mathsf{planted}}^{(-F)}\right) \tag{22}$$

We show below that $\gamma > \gamma_{UB_\eta}^{(k')}$ implies that there exists t such that

$$t_{\Delta'} < t < d(k')\mu - t_{\Delta''} \tag{23}$$

with

$$\frac{\Delta'}{d(k')} \geq \frac{\log M_{k'}}{\log N} + \eta \quad \text{and} \quad \frac{\Delta''}{d(k')} \geq \frac{\log\binom{k}{k'}}{\log N} + \eta\ . \tag{24}$$

Since $\mathbb{P}(X > Y) \leq \mathbb{P}(X > t) + \mathbb{P}(t \geq Y)$ for every t, we have

$$\begin{aligned}\mathbb{P}\left(\max(\mathsf{S}_{k'}^{(-F)}) > S_{\mathsf{planted}}^{(-F)}\right) &\leq \mathbb{P}\left(\max(\mathsf{S}_{k'}^{(-F)}) > t\right) + \mathbb{P}\left(t \geq S_{\mathsf{planted}}^{(-F)}\right) \\ &\leq \mathbb{P}\left(\max(\mathsf{S}_{k'}^{(-F)}) > t_{\Delta'}\right) + \mathbb{P}\left(d(k')\mu - t_{\Delta''} > S_{\mathsf{planted}}^{(-F)}\right) \\ &\leq Q_m \cdot p_{k',\Delta'} + p_{k',\Delta''} \end{aligned} \tag{25}$$

where the latter inequality follows from eq. (16) and eq. (15). Combining eq. (22) and eq. (25), we get

$$\begin{aligned}\mathbb{P}(\max(\mathsf{S}_{k'}) > S_{\mathsf{planted}}) &\leq \\ &\leq \binom{k}{k'}(Q_m \cdot p_{k',\Delta'} + p_{k',\Delta''}) \\ &= \binom{k}{k'}\left(\left(\frac{1}{N}\right)^{\frac{\Delta'}{d(k')}} \frac{Q_m}{\sqrt{\pi\Delta'} \cdot 2\hat{\sigma}\log N}\right) \\ &\quad + \binom{k}{k'}\left(\left(\frac{1}{N}\right)^{\frac{\Delta''}{d(k')}} \frac{1}{\sqrt{\pi\Delta''} \cdot 2\hat{\sigma}\log N}\right) \\ &= \left(\left(\frac{1}{N}\right)^{\frac{\Delta'}{d(k')} - \frac{\log M_{k'}}{\log N}} \frac{1}{\sqrt{\pi\Delta'} \cdot 2\hat{\sigma}\log N} + \right) \\ &\quad + \left(\left(\frac{1}{N}\right)^{\frac{\Delta''}{d(k')} - \frac{\log\binom{k}{k'}}{\log N}} \frac{1}{\sqrt{\pi\Delta''} \cdot 2\hat{\sigma}\log N}\right)\end{aligned}$$

where we used in the first equality eq. (14) and in the last equality the identity $x = N^{\frac{\log x}{\log N}}$ and the fact that $\binom{k}{h}Q_{k'} = \binom{k}{h}\binom{N-k}{k-k'} = M_{k'}$. Hence, by eq. (24) we can upper bound this probability as

$$\begin{aligned}\mathbb{P}(\max(\mathsf{S}_{k'}) > S_{\mathsf{planted}}) &\leq \\ &\leq \frac{1}{N^\eta}\frac{1}{\sqrt{\pi} \cdot 2\hat{\sigma}\log N}\left(\frac{1}{\Delta'} + \frac{1}{\Delta''}\right) \\ &\leq \frac{1}{N^\eta}\frac{1}{\sqrt{\pi} \cdot 2\hat{\sigma} \cdot d(k')}\left(\frac{1}{\log M_{k'}} + \frac{1}{\log\binom{k}{k'}}\right) \\ &\leq \frac{1}{N^\eta}\frac{2}{\sqrt{\pi} \cdot 2\hat{\sigma} \cdot \log k}\end{aligned}$$

where in the last inequality we used $d(k') \geq 1$, for $k' \in \{0, \dots, k-1\}$, and $M_{k'} \geq \binom{k}{h} \geq k$. To conclude the proof we show that $\gamma > \gamma_{UB_\eta}^{(k')}$ implies that there exists t such that eq. (23) and eq. (24) hold. We set $\Delta' = d(k')\left(\frac{\log M_{k'}}{\log N} + \eta\right)$ and $\Delta'' = d(k')\left(\frac{\log\binom{k}{k'}}{\log N} + \eta\right)$ so that eq. (24) holds. By plugging this into eq. (23), we can rewrite the inequality $t_{\Delta'} < d(k')\mu - t_{\Delta''}$ as follows:

$$\begin{aligned} d(k')\mu &> t_{\Delta'} + t_{\Delta''} \\ &\overset{(13)}{>} \sigma\sqrt{2\log N}(\sqrt{\Delta'} + \sqrt{\Delta''}) \end{aligned}$$

hence by the rescaling in eq. (1) we get the condition for the SNR:

$$\gamma > \frac{1}{d(k')}(\sqrt{\Delta'} + \sqrt{\Delta''}) = \sqrt{\frac{1}{d(k')}}\left(\sqrt{\frac{\log M_{k'}}{\log N} + \eta} + \sqrt{\frac{\log \binom{k}{k'}}{\log N} + \eta}\right) .$$

□

Proof of Corollary 9. We first show the following result:

Lemma 14. *For every k and $k' \in \{0, \dots, k-1\}$, the corresponding constants β_0 and β_0' in (3) satisfy the following:*

$$\gamma_{UB_\eta}^{(k')} \leq \left(\sqrt{1 + \beta_0 - 2\beta_0' + \eta} + \sqrt{\beta_0 - \beta_0' + \eta}\right)\sqrt{\frac{k-k'}{\binom{k}{h} - \binom{k'}{h}}} \tag{26}$$

$$\leq \left(\sqrt{1 + \beta_0 - 2\beta_0' + \eta} + \sqrt{\beta_0 - \beta_0' + \eta}\right)\sqrt{\frac{k}{\binom{k}{h}}} . \tag{27}$$

Proof of Lemma 14. Note that

$$\binom{k}{k'} = \binom{k}{k-k'} \leq e^{k-k'}\left(\frac{k}{k-k'}\right)^{k-k'} .$$

thus implying

$$\frac{\log \binom{k}{k'}}{\log N} \leq (k-k')\frac{1 + \log k - \log(k-k')}{\log N} \overset{(3)}{\approx} (k-k')(\beta_0 - \beta_0') .$$

Similarly

$$\binom{N-k}{k-k'} \leq e^{k-k'}\left(\frac{N-k}{k-k'}\right)^{k-k'}$$

thus implying

$$\frac{\log \binom{N-k}{k-k'}}{\log N} \leq (k-k')\frac{1 + \log(N-k) - \log(k-k')}{\log N} \overset{(3)}{\approx} (k-k')(1-\beta_0') .$$

Hence

$$\frac{\log M_{k'}}{\log N} = \frac{\log\left(\binom{k}{k'}\binom{N-k}{k-k'}\right)}{\log N} \lesssim (k-k')(1 + \beta_0 - 2\beta_0') .$$

By plugging this into the definition of $UB_\eta(k')$ (eq. (17)) we get

$$UB_\eta(k') = \sqrt{\frac{\log M_{k'}}{\log N} + \eta} + \sqrt{\frac{\log \binom{k}{k'}}{\log N} + \eta} \leq \sqrt{1 + \beta_0 - 2\beta_0' + \eta} + \sqrt{\beta_0 - \beta_0' + \eta} .$$

and hence, using the definition of $\gamma_{UB_\eta}^{(k')}$ given in theorem 8, we can conclude eq. (26). To conclude the proof we show that

$$\frac{k-k'}{\binom{k}{h} - \binom{k'}{h}} \leq \frac{k}{\binom{k}{h}} .$$

Simply observe that this inequality is equivalent to

$$\frac{k-k'}{k} \leq \frac{\binom{k}{h} - \binom{k'}{h}}{\binom{k}{h}} \Leftrightarrow \frac{\binom{k'}{h}}{\binom{k}{h}} \leq \frac{k'}{k} . \tag{28}$$

For $k' < h$ this inequality is trivially satisfied since $\binom{k'}{h} = 0$. Otherwise we can write the previous inequality as

$$\frac{\binom{k'}{h}}{\binom{k}{h}} = \frac{k'!}{h!\,(k'-h)!}\frac{h!\,(k-h)!}{k!} = \frac{k'(k'-1)\cdots(k'-h+1)}{k(k-1)\cdots(k-h+1)} \leq \frac{k'}{k}$$

which is satisfied for any $k' \leq k$ since all these terms satisfy $\frac{k'-i}{k-i} \leq 1$, for $1 \leq i \leq h-1$. □

Proof of Corollary 9: The idea is that by taking $\eta_0 = \beta_0'' + \epsilon_0$, where $\frac{\log k'}{\log N} \lesssim \beta_0''$, we can apply the union bound over all $m \leq k'$ to get that the right hand side of eq. (8) to go to 0, hence $P_{\mathsf{failrecover}}^{(k')} \to 0$ and $P_{\mathsf{recover}}^{(k')} \to 1$. Specifically, for γ satisfying $\gamma > \gamma_{UB_\eta}^{(m)}$ for all $m = \{0, \dots, k'\}$, we have $\mathbb{P}\left(\max(\mathsf{S}_m) > S_{\mathsf{planted}}\right) \in O\left(\frac{1}{N^\eta} \cdot \frac{1}{\hat{\sigma}\log k}\right)$ thus implying

$$P_{\mathsf{failrecover}}^{(k')} \in O\left(\frac{k'}{N^\eta} \cdot \frac{1}{\hat{\sigma}\log k}\right) .$$

Since $\frac{\log k'}{\log N} \lesssim \beta_0'$ we have $\frac{k'}{N^\eta} = \frac{2^{\log k'}}{2^{\eta \log N}} = 2^{\log k' - \eta \log N}$ with $\log k' - \eta \log N = \log N(\frac{\log k'}{\log N} - \eta) \approx -\epsilon_0 \log N$. Hence, $\frac{k'}{N^\eta} = o(1)$ and the probability $P_{\mathsf{failrecover}}^{(k')}$ tend to 0. □

Appendix B
Proof for section IV-C

Proof of Lemma 10. Using standard inequalities on the binomial coefficients,

$$a(\log b - \log a) \leq \log\binom{b}{a} \leq a(1 + \log b - \log a) , \tag{29}$$

and $\binom{b}{a} \le 2^b$ we have:

$$\log\binom{N-k}{k-k'} - \log\left(\binom{N-2k+k'}{k-k'-k''}(k-k')\binom{k-k'}{k''}\right) \ge$$
$$\log\binom{N-k}{k-k'} - \log\left(\binom{N-2k+k'}{k-k'-k''}2^{(k-k')}2^{(k-k')}\right) =$$
$$\log\binom{N-k}{k-k'} - \log\binom{N-2k+k'}{k-k'-k''} - 2(k-k')$$
$$\overset{(29)}{\ge} (k-k')\Big[\log(N-k) - \log(k-k')\Big]$$
$$- (k-k'-k'')\Big[1 + \log(N-2k+k') - \log(k-k'-k'')\Big]$$
$$- 2(k-k')$$
$$= (k-k')\Big[\log(N-k) - \log(k-k') - 3 - \log(N-2k+k')\Big]$$
$$+ (k-k'-k'')\log(k-k'-k'')$$
$$+ k''\Big[1 + \log(N-2k+k')\Big] \ . \tag{30}$$
$$= (k-k')\Big[\log(N-k) - \log(k-k') - 3 - \log(N-2k+k')\Big]$$
$$+ k''\Big[1 + \log(N-2k+k') - \log(k-k'-k'')\Big] +$$
$$+ (k-k')\log(k-k'-k'') \ . \tag{31}$$

If $k'' \le (k-k')\rho_0$ then $k-k'-k'' \ge (k-k')(1-\rho_0)$ and therefore

$$\frac{\log(k-k'-k'')}{\log N} \ge \frac{\log(k-k') + \log(1-\rho_0)}{\log N} \overset{(3)}{\to} \beta_0'$$

thus implying, together with eq. (31) and with $-\log(k-k'-k'') \ge -\log(k-k')$, that

$$\frac{\log\binom{N-k}{k-k'} - \log\left((k-k')\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\right)}{\log N} \gtrsim$$
$$(k-k')(-\beta_0') + k''(1-\beta_0') + (k-k')(\beta_0') =$$
$$k''(1-\beta_0') \ .$$

If $k'' > (k-k')\rho_0$ then, since $k-k'-k'' \ge 1$, by eq. (30) and eq. (3) we have

$$\frac{\log\binom{N-k}{k-k'} - \log\left((k-k')\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\right)}{\log N} \gtrsim$$
$$(k-k')(-\beta_0') + k'' > (k-k')(\rho_0 - \beta_0') \ .$$

□

Proof of Lemma 11. Observe that each $S_i^{(-F)} \in \mathsf{S}_{k'}^{(-F)}$ is the contribution of $d(k') = \binom{k}{h} - \binom{k'}{h}$ hyperedges, all of them unbiased. Therefore,

$$S_i^{(-F)} \sim \mathcal{N}(0, d(k')\sigma^2) \quad \text{where } d(k') = \binom{k}{h} - \binom{k'}{h} \ .$$

For $\delta = \sqrt{\Delta}$ and $t = \delta\sigma\sqrt{2\log N}$, by the left hand side of eq. (10), we have

$$P_{k'} := \mathbb{P}(S_i^{(-F)} > t) \ge \left(\frac{1}{N}\right)^{\frac{\delta^2}{d(k')}}\left(\frac{1}{\delta\hat{\sigma}\log N} - \frac{1}{(\delta\hat{\sigma}\log N)^2}\right)$$

We next show that

$$\mathbb{P}(\max(\mathsf{S}_{k'}^{(-F)}) > t) = \mathbb{P}\left(\bigcup_{S_i^{(-F)} \in \mathsf{S}_{k'}^{(-F)}} S_i^{(-F)} > t\right) \to 1 \ . \tag{32}$$

We use the Chung-Erdös inequality (eq. (11)) on the $M_{k'}$ random variables in $\mathsf{S}_{k'}^{(-F)}$, and our goal is to show the asymptotics below:

$$\mathbb{P}(\max(\mathsf{S}_{k'}^{(-F)}) > t) \ge \frac{\left(\sum_i P(A_i)\right)^2}{\sum_i P(A_i) + \sum_{i\ne j} P(A_i \cap A_j)}$$
$$= \frac{(M_{k'}P_{k'})^2}{M_{k'}P_{k'} + \sum_{i\ne j} P(A_i \cap A_j)}$$
$$\left(\frac{1}{M_{k'}P_{k'}} + \frac{\sum_{i\ne j} P(A_i \cap A_j)}{(M_{k'}P_{k'})^2}\right)^{-1} \to 1 \tag{33}$$

We shall prove below that the hypothesis $\Delta < d(k')(k-k')(1-\beta_0')$ implies $M_{k'}P_{k'} \to \infty$. Thus, in order to prove eq. (32), it is enough to show

$$\frac{\sum_{i\ne j} P(A_i \cap A_j)}{M_{k'}^2 P_{k'}^2} \to 1 \ . \tag{34}$$

Note that

$$\sum_{i\ne j} P(A_i \cap A_j) \le M_{k'}\Bigg((M_{k'}-1)P_k^2 +$$
$$+ (\bar{P}_{ij}^{(k'')} - P_k^2)\sum_{k''=\max\{1,h-k'\}}^{k-k'-1}\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\Bigg)$$

where $\bar{P}_{ij}^{(k'')}$ is an upper bound on $P(A_i \cap A_j)$ when the corresponding $S_i^{(-F)}$ and $S_j^{(-F)}$ share k'' nodes, for $\max\{1, h-k'\} \le k'' \le k-k'-1$ as:[2]

$$\mathbb{P}(A_i \cap A_j) := \mathbb{P}(S_i^{(-F)} > t \cap S_j^{(-F)} > t) \le \bar{P}_{ij}^{(k'')} \tag{35}$$

for all $S_i^{(-F)}, S_i^{(-F)} \in \mathsf{S}_{k'}^{(-F)}$ s.t. $|N(R_i) \cap N(R_j)| = k''$, where $N(S_i^{(-F)})$ and $N(S_j^{(-F)})$ denotes the set of $k-k'$ nodes of $S_i^{(-F)}$ and $S_j^{(-F)}$, respectively. Our goal is to show that

$$\sum_{k''=\max\{1,h-k'\}}^{k-k'-1}\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\bar{P}_{ij}^{(k'')} = o(M_{k'}P_{k'}^2) \ . \tag{36}$$

Note that $S_i^{(-F)}$ and $S_j^{(-F)}$ as above can be seen as the k'' common nodes plus the remaining $r := k-k'-k''$ nodes in each of them. For $t = t_{k''} + t_r$ where $t_{k''} := \beta t = \delta\beta\sigma\sqrt{2\log N}$ and $t_r := (1-\beta)t = \delta(1-\beta)\sigma\sqrt{2\log N}$, we derive the following type of upper bound:

$$\bar{P}_{ij}^{(k'')} \le \bar{P}_{k''} + \bar{P}_r^2$$

[2] Note that for $s < h$ the two solutions do not share any (hyper)edge and thus are independent.

where $\bar{P}_{k''}$ and $\bar{P}_r$ are given by considering the $a(k'')$ hyperedges in the common part (C_{ij}) and the $d(k') - a(k'')$ remaining ones (both $\hat{S}_j^{(-F)}$ and $\hat{S}_j^{(-F)}$), that is

$$\begin{aligned} C_{ij} \sim & \mathcal{N}(0, \sigma_C^2), \quad \text{for } \sigma_C^2 = a(k'')\sigma^2 \ , \\ \hat{S}_i^{(-F)}, \hat{S}_j^{(-F)} \overset{iid}{\sim} & \mathcal{N}(0, \sigma_R^2), \quad \text{for } \sigma_R^2 = [d(k') - a(k'')]\sigma^2 \ . \end{aligned}$$

Hence, the probabilities $\bar{P}_r$ and $\bar{P}_{k''}$ are given by

$$\begin{aligned} \mathbb{P}(C_{ij} > t_{k''}) &\leq \frac{e^{-t_{k''}^2/2\sigma_C^2}}{t_s} = \frac{e^{-(\delta\beta\sigma\sqrt{2\log N})^2/2a(k'')\sigma^2}}{\delta\beta\hat{\sigma}\log N} \\ &= \left(\frac{1}{N}\right)^{\frac{(\delta\beta)^2}{a(k'')}} \cdot \frac{1}{\delta\beta\hat{\sigma}\log N} =: \bar{P}_{k''} \ , \\ \mathbb{P}(\hat{S}_i^{(-F)} > t_r) &\leq \frac{e^{-t_r^2/2\sigma_R^2}}{t_r} \\ &= \frac{e^{-(\delta(1-\beta)\sigma\sqrt{2\log N})^2/2[d(k')-a(k'')]\sigma^2}}{\delta(1-\beta)\hat{\sigma}\log N} \\ &= \left(\frac{1}{N}\right)^{\frac{(\delta(1-\beta))^2}{d(k')-a(k'')}} \cdot \frac{1}{\delta(1-\beta)\hat{\sigma}\log N} =: \bar{P}_r \ . \end{aligned}$$

In order to have $\bar{P}_{k''} \approx \bar{P}_r^2$ we impose $\left(\frac{1}{N}\right)^{\frac{(\delta\beta)^2}{a(k'')}} \approx \left(\frac{1}{N}\right)^{\frac{(\delta(1-\beta))^2}{d(k')-a(k'')}}$ by equating the exponents:

$$\begin{aligned} \frac{(\delta\beta)^2}{a(k'')} &= \frac{2(\delta(1-\beta))^2}{d(k')-a(k'')} \quad \Leftrightarrow \\ \frac{1}{2}\left(\frac{d(k')}{a(k'')} - 1\right) &= \left(\frac{1-\beta}{\beta}\right)^2 \quad \Leftrightarrow \\ \sqrt{\frac{1}{2}\left(\frac{d(k')}{a(k'')} - 1\right)} &= \left(\frac{1}{\beta} - 1\right) \quad \Leftrightarrow \end{aligned}$$

that can be obtained with

$$\begin{aligned} \beta^2 &= \left(\frac{1}{1 + \sqrt{\frac{1}{2}\left(\frac{d(k')}{a(k'')} - 1\right)}}\right)^2 \\ &= \frac{1}{\frac{1}{2}(1 + \frac{d(k')}{a(k'')}) + 2\sqrt{\frac{1}{2}\left(\frac{d(k')}{a(k'')} - 1\right)}} \\ &= \frac{2a(k'')}{a(k'') + d(k') + \sqrt{8a(k'')[d(k') - a(k'')]}} \end{aligned}$$

and in particular, by also using $\delta^2 = \Delta$, we have

$$\frac{(\delta\beta)^2}{a(k'')} = \frac{2\delta^2}{a(k'') + d(k') + \sqrt{8a(k'')[d(k') - a(k'')]}}$$

thus implying

$$\bar{P}_r^2 \approx \bar{P}_{k''} \approx \left(\frac{1}{N}\right)^{\frac{2\delta^2}{D(k',k'')}}$$

for

$$D(k', k'') := a(k'') + d(k') + \sqrt{8a(k'')[d(k') - a(k'')]} \ .$$

Observe that

$$M_{k'}P_{k'}^2 = \binom{N-k}{k-k'} P_{k'}^2 \approx \left(\frac{1}{N}\right)^{\frac{2\delta^2}{d(k')} - \frac{\log\binom{N-k}{k-k'}}{\log N}} \ .$$

In order to have eq. (36) it is sufficient to have

$$\begin{aligned} \frac{2\delta^2}{D(k',k'')} - \frac{\log\left((k-k')\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\right)}{\log N} > \\ > \frac{2\delta^2}{d(k')} - \frac{\log\binom{N-k}{k+k'}}{\log N} \end{aligned}$$

that is

$$\begin{aligned} 2\delta^2\left(\frac{1}{d(k')} - \frac{1}{D(k',k'')}\right) < \\ < \frac{\log\binom{N-k}{k-k'} - \log\left((k-k')\binom{k-k'}{k''}\binom{N-2k+k'}{k-k'-k''}\right)}{\log N} \ . \end{aligned}$$

Lemma 10 provides a lower bound $\ell(k'')$ on the right hand side, and thus the above inequality holds if the following inequality holds:

$$\Delta < \frac{\ell(k'')}{2} \frac{d(k') \cdot D(k', k'')}{D(k', k'') - d(k')} \ ,$$

for all $k'' \in I(k') = \{\max(1, h - k'), \ldots, k - k' - 1\}$. To conclude the proof, we note that $M_{k'}P_{k'} \to \infty$ for

$$\frac{\delta^2}{d(k')} < \frac{\log\binom{N-k}{k-k'}}{\log N} \overset{(3)}{\approx} (k-k')(1-\beta_0')$$

which is the assumption $\Delta < d(k')(k-k')(1-\beta_0')$. This completes the proof. □

Proof of Theorem 12. We show that

$$\mathbb{P}\left(S_{\mathsf{planted}}^{(-F)} > \max(\mathsf{S}_{k'}^{(-F)})\right) \to 0 \tag{37}$$

where F is an arbitrarily fixed subset of k' nodes of the planted solution. Let us first observe that for any $\Delta' = \Theta\left(\frac{d(k')}{\log N}\right)$, it holds that $p_{k',\Delta'} \to 0$ and therefore

$$\mathbb{P}\left(S_{\mathsf{planted}}^{(-F)} > d(k')\mu + \sigma\sqrt{\Delta' 2\log N}\right) \leq p_{k',\Delta'} \to 0 \ .$$

Also notice that $\sigma\sqrt{\Delta' 2\log N} = \Theta\left(\sigma\sqrt{2d(k')}\right) = o(d(k')\mu)$ and therefore

$$d(k')\mu + \sigma\sqrt{\Delta' 2\log N} = (1 + o(1)) \cdot d(k')\mu \ .$$

For any Δ satifying the condition in Lemma 11, we also have

$$\mathbb{P}\left(\max(\mathsf{S}_{k'}^{(-F)}) \leq \sigma\sqrt{\Delta 2\log N}\right) \to 0 \ .$$

In particular, we can take $\Delta = d(k')\min\{LB(k'), (k-k')(1-\beta_0') - \rho_0$, where $LB(k')$ is defined as in Lemma 11 and $\rho_0 \in (0,1)$ is a constant. Suppose that, for Δ' and Δ as above, there exists t such that

$$d(k')\mu + \sigma\sqrt{\Delta' 2\log N} < t < \sigma\sqrt{\Delta 2\log N} \ . \tag{38}$$

Then we get both

$$\mathbb{P}\left(S^{(-F)}_{\mathsf{planted}} > t\right) \leq \mathbb{P}\left(S^{(-F)}_{\mathsf{planted}} > d(k')\mu + \sigma\sqrt{\Delta' 2\log N}\right) \leq p_{k',\Delta'} \to 0 \ .$$

$$\mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) < t\right) \leq \mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) \leq \sigma\sqrt{\Delta 2\log N}\right) \to 0 \ .$$

Using both inequalities above, we have

$$\mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) < S^{(-F)}_{\mathsf{planted}}\right) \leq \mathbb{P}\left(S^{(-F)}_{\mathsf{planted}} > t\right) + \mathbb{P}\left(\max(\mathsf{S}^{(-F)}_{k'}) < t\right)$$

and both quantities tend to 0 using the condition eq. (38) on t. Finally, we show that $\gamma < \gamma^{(k')}_{LB}$ implies that $d(k')\mu + \sigma\sqrt{\Delta' 2\log N} < \sigma\sqrt{\Delta 2\log N}$ so that we can find some t as above:

$$d(k')\mu + \sigma\sqrt{\Delta' 2\log N} < \sigma\sqrt{\Delta 2\log N} \overset{(1)}{\Leftrightarrow} \gamma < \frac{\sqrt{\Delta} - \sqrt{\Delta'}}{d(k')} \ .$$

Since $\Delta' = \Theta\left(\frac{d(k')}{\log N}\right)$ we have $\frac{\sqrt{\Delta'}}{d(k')} \approx 0$, while by the definition of Δ we have

$$\frac{\sqrt{\Delta}}{d(k')} = \sqrt{\frac{d(k')\min\{LB(k'), (k-k')(1-\beta'_0) - \rho_0}{d(k)^2}} \approx \sqrt{\frac{\min\{LB(k'), (k-k')(1-\beta'_0)}{d(k)}} = \gamma^{(k)}_{LB} \ .$$

Hence $\gamma < \gamma^{(k')}_{LB} \approx \frac{\sqrt{\Delta}-\sqrt{\Delta'}}{d(k')}$ implies the existence of t as above. This completes the proof. □

Proof of Corollary 13. We first prove the following:

Lemma 15. *For every k' and every $k_\star$ such that $\binom{k'+k_*}{h}/\binom{k}{h} \to 0$ the following holds. For every $\rho_0 \in (0,1)$ and for every $\gamma < \sqrt{\frac{\min(k_\star(1-\beta_0),(k-k')(\rho_0-\beta_0)}{2d(k')}}$*

$$\mathbb{P}\left(S_{\mathsf{planted}} > \max(\mathsf{S}_{k'})\right) \to 0 \ . \tag{39}$$

Proof. We prove that

$$LB(k') \geq (1/2)\min\left(k_\star(1-\beta_0), (k-k')(\rho_0-\beta_0)\right) \tag{40}$$

and thus $\gamma^{(k')}_{LB} \geq \sqrt{\frac{\min(k_\star(1-\beta_0),(k-k')(\rho_0-\beta_0)}{2d(k')}}$. For a generic k' and $k'' \in I(k')$, let us consider $a := a(k'') = \binom{k'+k''}{h} - \binom{k'}{h}$, $d := d(k') = \binom{k}{h} - \binom{k'}{h}$, and $D := D(k', k'')$ where we dropped the dependency on k', k'' for convenience. We show first that, by the hypothesis on $k_\star$, for all $k'' \leq k_\star$ we have $a = o(d)$. Observe that, for all $k'' \leq k_\star$ we have $a(k'') \leq a(k_\star)$, and thus

$$\frac{a(k'')}{d(k')} \leq \frac{a(k_\star)}{d(k')} = \frac{\binom{k+k_\star}{h} - \binom{k'}{h}}{\binom{k}{h} - \binom{k'}{h}} \leq \frac{\binom{k+k_\star}{h}}{\binom{k}{h}} \to 0 \ .$$

This implies

$$\frac{D}{D-d} = \frac{a+d+\sqrt{8a(d-a)}}{(a+d+\sqrt{8a(d-a)}) - d} \geq \frac{d}{a+\sqrt{8a(d-a)}} = \frac{d}{o(d)} \to +\infty \ .$$

Hence, $\frac{\ell(k'')D}{D-d} \to +\infty$ for such k'' since $\ell(k'') \geq (\rho_0 - \beta_0)$ by Lemma 10. For $k'' > k_\star$ we instead observe that, again using Lemma 10,

$$\ell(k'')\frac{D}{D-d} \geq \ell(k_\star) \geq \min\left(k_\star(1-\beta_0), (k-k')(\rho_0-\beta_0)\right) \ ,$$

which proves (40). □

The proof for Corollary 13 follows easily by taking $k' = 0$ in Lemma 15. □