Signal Recovery in Random Graphical Models with Local Dependencies

by

Chenghao Guo

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2026

Authored by: Chenghao Guo
Department of Electrical Engineering and Computer Science
May 15, 2026

Certified by: Guy Bresler
Associate Professor, Thesis Supervisor

Certified by: Yury Polyanskiy
Professor, Thesis Supervisor

Accepted by: Leslie Kolodziejski
Joseph F. and Nancy P. Keithley Professor in EE
Graduate Officer, Department of Electrical Engineering and Computer Science

Signal Recovery in Random Graphical Models with Local Dependencies

by

Chenghao Guo

Submitted to the Department of Electrical Engineering and Computer Science on May 15, 2026 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

Community recovery and signal recovery in random graphical models aim to infer latent structure from noisy relational data. Much of the classical theory is built on independent random graph models, especially Erdős–Rényi graphs and their planted variants. These models make sharp analysis possible, but they miss a central feature of many networks: edges are often locally correlated through clustering, intersecting communities, hyperedges, or latent geometry. This thesis studies recovery in random graphical models with local dependencies, using local structure as a bridge between independent noise and globally constrained latent graph models.

The thesis develops a hierarchy of dependence mechanisms. Independent models have product-measure edge noise. Latent-feature and geometric models couple many edges through hidden labels, features, or positions. Between these extremes are locally dependent models, generated from an independent base by bounded-footprint updates such as triangle insertions, clique insertions, and signed triangle relations. This viewpoint leads to two complementary notions of comparison: total-variation equivalence of graph distributions, and computational equivalence of inference problems through polynomial-time reductions preserving both null and planted laws.

A main theme is that local dependence is not merely an ad hoc perturbation of independence. It emerges naturally as the effective form of latent dependence in high-dimensional regimes. Random intersection graphs can be interpreted as geometric graphs over the discrete cube, and in suitable constant-density regimes they are close in total variation to random graphs with triangles. Balanced high-dimensional random geometric graphs exhibit an analogous phenomenon: below the Erdős–Rényi indistinguishability threshold, their first surviving correction is an explicit local signed triangle-sampling chain. Thus both discrete latent-feature geometry and continuous spherical geometry approach independence through local triangle corrections.

The thesis then asks whether such local dependence changes the difficulty of global recovery. For planted dense subgraph and planted clique problems in a random graph with triangles, it develops algorithmic decorrelation maps that remove local triangle dependence while approximately preserving the planted signal. In the parameter regimes studied, these reductions show that local triangle dependence does not affect the computational difficulty of global recovery: the planted problem in the locally dependent ambient graph is computationally equivalent to its Erdős–Rényi counterpart.

The final part of the thesis studies local recovery, where the latent signals themselves are small generative structures. A random hypergraph is observed only through its graph projection, in which each hyperedge appears as a clique. The thesis proves sharp all-or-nothing thresholds for partial and exact recovery of random d-uniform hypergraphs from unweighted and weighted projections, and extends the recovery framework to heterogeneous hypergraphs with multiple hyperedge sizes. Together, these results show that local dependencies are rich enough to capture realistic higher-order network structure while remaining structured enough for sharp statistical and computational analysis.

Thesis supervisor: Guy Bresler
Title: Associate Professor

Thesis supervisor: Yury Polyanskiy
Title: Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Community recovery and signal recovery in random graphical models seek to infer latent structure from noisy observations. In its most classical form, the observation is a graph on n vertices, and the signal is a planted community, clique, dense subgraph, block assignment, or collection of higher-order relations. In network science, this question appears through the study of community structure in empirical networks [1]; in statistics and machine learning, it is formalized through models such as the stochastic block model and its variants [2, 3]; and in average-case complexity, it appears through planted problems such as planted clique, planted dense subgraph, sparse PCA, and submatrix detection [4–7]. These problems have become central testbeds for understanding both statistical limits and statistical-computational gaps [8–11].

Much of the sharpest theory is built on independent random graph models. The canonical example is the Erdős–Rényi graph $\mathrm{G}(n, p)$, in which the edge indicators are mutually independent. Independent noise is analytically powerful: likelihoods factor over edges, concentration inequalities can be applied directly, and average-case reductions can often be implemented by entrywise randomization. This product structure underlies much of the existing theory of planted clique, planted dense subgraph, and related average-case inference problems [4–6, 12]. It also provides the baseline against which more realistic dependent models are compared.

The limitation is that independence is rarely an accurate description of observed graph data. Real networks exhibit clustering, transitivity, overlapping groups, spatial constraints, and higher-order interactions. In social networks, the presence of two ties in a triangle often makes the third more likely [13, 14]; in collaboration networks, pairwise edges are frequently projections of group events such as coauthorship [15]; and in biological networks, graph structure is shaped by functional organization and shared latent mechanisms rather than independent pairwise noise [16]. These effects are not merely cosmetic. They can create dense local motifs, alter spectral structure, introduce dependencies among nearby edges, and produce pairwise observations that are lossy projections of higher-order data.

A broad response to this mismatch is to introduce latent-variable graph models. The stochastic block model introduces vertex labels and edge probabilities determined by community membership [2, 3]. Random intersection graphs introduce latent features, with an edge appearing when two vertices share at least one feature; a single feature can induce an entire clique [17]. Random geometric graphs introduce latent positions, with edges determined by geometric proximity or correlation in an ambient space [17, 18]. These models are more realistic than product-measure graphs, but they typically introduce dependencies through hidden global, vertex-level, or feature-level structure. Conditioned on the latent variables the model may be simple, but after marginalizing over the latent variables, the observed edge array is no longer a product measure. This creates a gap between the tractability of independent-noise theory and the structure of dependent observations.

The central premise of this thesis is that local dependencies provide a useful bridge across this gap. A local dependency is a dependence generated by a bounded-size motif, latent feature, hyperedge, or short-range constraint. Examples include triangles added to an Erdős–Rényi graph, cliques induced by bounded latent intersection features, pairwise projections of hyperedges, and signed triangle relations arising from high-dimensional geometry. Local dependencies are expressive enough to model clustering and higher-order interactions directly, but structured enough to admit couplings, total-variation comparisons, and algorithmic reductions. They therefore provide an intermediate class between independent random graphs and globally constrained latent graph models.

This bridge is not only conceptual; it is also mathematical. Random graphs with triangles provide an explicit locally dependent perturbation of $\mathrm{G}(n, p)$, and in suitable regimes they approximate random intersection graphs in total variation [19]. High-dimensional random geometric graphs exhibit a different instance of the same phenomenon: although the latent geometry creates global dependence, the first surviving correction below the Erdős–Rényi indistinguishability threshold is captured by a local triangle-sampling mechanism [17, 20]. Thus local dependencies can arise both as direct models of motifs and as effective descriptions of more global latent structures.

This thesis studies two complementary recovery problems within this framework. The first is a global recovery problem. A single macroscopic signal, such as a planted clique or planted dense subgraph, is embedded in a dependent ambient graph. The question is whether local dependence in the noise changes the computational complexity of detecting or recovering the planted signal. The random graph with triangles is the main test case: it contains explicit local dependence, yet admits algorithmic decorrelation maps that transform the dependent ambient graph back toward an independent Erdős–Rényi baseline while approximately preserving the planted structure [19].

The second is a local recovery problem. Here the latent signal is not one large planted set but many small generative objects, such as the hyperedges of a random hypergraph. The observation is a graph projection: each hyperedge is replaced by a clique, and only pairwise co-occurrence is observed. This setting captures the fact that higher-order interactions are often recorded through pairwise data, as in coauthorship, communication, and other group-interaction networks [15, 21–23]. The recovery question is whether the original hyperedges can be reconstructed from their pairwise shadows, and how the answer depends on the hyperedge density, the uniformity d, the availability of edge multiplicities, and the presence of heterogeneous hyperedge sizes [24–26].

A recurring theme is that dependence must be compared at two levels. At the level of distributions, the relevant notion is statistical equivalence, most often expressed as vanishing total variation distance. If two graph distributions are close in total variation, then no statistical procedure can distinguish them asymptotically. At the level of inference tasks, the relevant notion is computational equivalence, expressed through polynomial-time randomized reductions that preserve both the null distribution and the planted distribution. These notions are related but distinct: a total-variation approximation identifies when two observation models are statistically interchangeable, while an average-case reduction transfers algorithmic upper and lower bounds between inference problems [6–9, 19].

The thesis develops this viewpoint through three families of results. First, it shows that local dependencies can mediate between independent and globally constrained graph models: random graphs with triangles approximate random intersection graphs in suitable regimes, and a local triangle-sampling chain captures the middle-scale correction to high-dimensional random geometric graphs [17, 19, 20]. Second, it develops algorithmic decorrelation for global recovery, giving average-case reductions between planted dense subgraph problems in independent and locally dependent ambient graphs [19]. Third, it characterizes local recovery from graph projections, proving sharp thresholds for reconstructing random hypergraphs from unweighted and weighted projections and extending the recovery viewpoint to heterogeneous hypergraphs [24–26]. Together, these results support the thesis that local dependencies form a mathematically tractable bridge between idealized independent noise models and realistic dependent graphical observations.

1.2 The Dependency Hierarchy

The motivation above suggests organizing random graphical models by how dependence is generated. This is different from organizing them by the visual structures they produce. A graph may contain many local motifs, such as triangles or cliques, because these motifs were inserted by local operations; but the same motifs may also arise after marginalizing over latent labels, latent features, or latent geometric positions. These mechanisms lead to different comparison tools.

We use the hierarchy $\mathcal{C}_{\mathrm{ind}} \longrightarrow \mathcal{C}_{\mathrm{loc}} \longrightarrow \mathcal{C}_{\mathrm{glob}}$ as a methodological guide rather than as a literal inclusion of probability spaces. The class $\mathcal{C}_{\mathrm{ind}}$ consists of independent-noise baselines. The class $\mathcal{C}_{\mathrm{glob}}$ consists of models whose dependencies are induced by macroscopic latent variables or global constraints. The class $\mathcal{C}_{\mathrm{loc}}$ is the intermediate class that this thesis isolates: graph models generated from an independent base by bounded-footprint local updates. Thus $\mathcal{C}_{\mathrm{loc}}$ is not merely the class of graphs that empirically exhibit clustering. It is the class of models that admit a controlled local-kernel representation, formalized later in Definition 3.8.

At the independent end are product-measure graph models, collected in $\mathcal{C}_{\mathrm{ind}}$. The canonical example is the Erdős–Rényi graph $\mathrm{G}(n, p)$, in which the edge indicators are mutually independent. Planted clique and planted dense subgraph models add a structured signal to this independent background. The product structure of $\mathrm{G}(n, p)$ is responsible for much of its analytic tractability: likelihoods factor over edges, concentration inequalities can be applied directly, and average-case reductions can often be implemented by entrywise randomization. These models serve as the baseline for both the statistical and computational comparisons in the thesis.

At the other end are latent or globally constrained models, collected in $\mathcal{C}_{\mathrm{glob}}$. In these models, edges are coupled through latent variables that affect many potential observations at once. In the stochastic block model, a latent vertex labeling controls the edge probabilities across a macroscopic collection of vertex pairs [2, 3]. In a random intersection graph, latent features induce cliques among all vertices carrying the same feature [17]. In a random geometric graph, latent positions impose metric or spherical constraints on the entire edge array [17, 18]. Conditional on the latent variables these models may have simple edge-generation rules, but after marginalizing over the latent variables the observed graph law is generally not a product measure. Moreover, a latent feature or a geometric constraint can involve an unbounded number of vertices, so such a model is not automatically local in the bounded-footprint sense used in this thesis.

The middle class $\mathcal{C}_{\mathrm{loc}}$ consists of graph models generated by local dependencies. The formal definition is a sequential local-kernel representation. Starting from a product base graph or signed graph, the model performs active updates on vertex sets V_r of bounded size, $|V_r| \leq d$, and each update changes only the induced motif on V_r. The update is allowed to be random and may depend on the current local motif. Thus locality is not the same as monotonicity: an update may add a triangle, insert a clique, resample a bounded induced subgraph, or enforce a signed relation among existing edge signs. The definition measures locality through the read and write loads of the updates, and distinguishes bounded-load local models from sparse local models.

This definition makes $\mathcal{C}_{\mathrm{loc}}$ an intermediate model class between latent-feature models and independent models. A random intersection graph can be viewed as a union of feature-induced cliques: for each feature a, the vertices carrying a form a set C_a, and the feature contributes the clique $\binom{C_a}{2}$ to the observed graph. If the feature sizes are truncated at r, or if features larger than r are negligible in total variation, then the resulting model is a degree-r monotone local-kernel model. The independent Erdős–Rényi baseline corresponds to retaining only the effective size-two contribution; degree 3 local dependence retains the first nontrivial triangle correction; higher local degrees retain larger feature-induced cliques. Thus, as the allowed local degree increases, local-kernel models represent latent-feature models in progressively broader parameter regimes. Fixed-degree $\mathcal{C}_{\mathrm{loc}}$ does not contain every random intersection graph, because a latent feature can have unbounded support, but it provides a controlled hierarchy of bounded-degree approximations between G and the full latent-feature law.

Several models in the thesis fall naturally into $\mathcal{C}_{\mathrm{loc}}$. The random graph with triangles $\mathrm{RGT}(n, p, p')$ is a degree-3 monotone local-kernel model: it starts from $\mathrm{G}(n, p)$ and, for each active triple, inserts the corresponding triangle. Its read load is $(n-2)p'$, so the regime $p' = O(1/n)$ is bounded-load local and $p' = o(1/n)$ is sparse local [19]. Graph projections of independent d-uniform hypergraphs are degree-d monotone local-kernel models: each active hyperedge inserts the clique on its d vertices. Under the parameterization $p_d = n^{-d+1+\delta+o(1)}$, the read load per graph edge is $n^{-1+\delta+o(1)}$, which is sparse local throughout the reconstruction regime $0 < \delta < 1$ [24, 25]. The ordered signed triangle-relation chain arising in the middle-scale approximation to random geometric graphs is also degree-3 local, but it is not monotone: it reads two previously exposed signs and replaces the current sign by their product [20].

This convention clarifies the role of random intersection graphs. A random intersection graph can create local cliques, but if a latent feature can contain an unbounded number of vertices, then the model is not itself a bounded-degree local-kernel model. It becomes part of the local framework either under additional restrictions, such as bounded feature sizes, or through an asymptotic approximation by an explicit local model. One example is the total-variation approximation of certain random intersection graph regimes by random graphs with triangles [17, 19]. More generally, replacing the full feature-induced clique process by its degree-r truncation gives a sequence of local models that interpolates between the independent baseline and the full latent-feature model. This is the sense in which local dependence is a bridge rather than a synonym for latent dependence.

The hierarchy is therefore best read in two complementary directions. As a modeling hierarchy, $\mathcal{C}_{\text{ind}} \longrightarrow \mathcal{C}_{\text{loc}} \longrightarrow \mathcal{C}_{\text{glob}}$ moves from independent noise to local motif-generated dependence and then to latent or geometric dependence. As an analysis strategy, the thesis often moves in the reverse direction: a globally dependent model may be approximated by a local model, and a locally dependent model may then be compared to an independent baseline. This is the bridge viewpoint. In the balanced high-dimensional random geometric graph, the global spherical dependence has a middle-scale regime in which the first surviving correction to $\mathrm{G}(n, 1/2)$ is an explicit local signed triangle update [20]. In random intersection graphs, latent-feature dependence can be approximated by local triangle dependence in suitable regimes [17, 19].

Two notions of comparison accompany this hierarchy. The first is statistical equivalence of graph distributions. For probability measures P_n and Q_n on the same graph space,

$$d_{\mathrm{TV}}(P_n, Q_n) = \sup_A |P_n(A) - Q_n(A)|.$$

We say that P_n and Q_n are statistically equivalent if

$$d_{\mathrm{TV}}(P_n, Q_n) = o(1).$$

This is the appropriate notion when the goal is to show that a dependent observation model and a simpler model are asymptotically indistinguishable by any statistical procedure.

The second notion is computational equivalence of inference problems. Suppose

$$P = (P_{0,n}, P_{1,n}), \qquad Q = (Q_{0,n}, Q_{1,n})$$

are two sequences of testing or recovery problems, where the subscript 0 denotes the null distribution and the subscript 1 denotes the planted distribution. A randomized polynomial-time map A_n is an average-case reduction from P to Q if

$$d_{\mathrm{TV}}(A_n(P_{0,n}), Q_{0,n}) = o(1), \qquad d_{\mathrm{TV}}(A_n(P_{1,n}), Q_{1,n}) = o(1).$$

Reductions in both directions give computational equivalence in the relevant parameter regime. This distinction is important because a distributional approximation compares observation laws directly, while a computational reduction must also preserve the planted signal.

The same local dependencies play different roles in the two main recovery problems. In global recovery, the latent signal is a macroscopic object, such as a planted clique or dense subgraph, and local dependencies belong to the ambient noise. The central algorithmic task is to decorrelate this dependent noise without destroying the planted signal [19]. In local recovery, the latent signals themselves are local objects, such as hyperedges, and the observation is a pairwise projection. The central statistical task is to determine when these local generative motifs can be reconstructed from their shadows [24–26]. This is why local dependencies are not only a modeling feature, but also the organizing tool connecting the global and local recovery problems of the thesis.

Table 1.1: Dependency hierarchy used in the thesis.

Class	Dependency type	Representative models
$\mathcal{C}_{\mathrm{ind}}$	Product-measure edge noise.	$\mathrm{G}(n,p)$; planted clique; planted dense subgraph.
$\mathcal{C}_{\mathrm{loc}}$	Bounded-footprint local kernels with controlled read/write load; degree controls the maximum motif size.	RGT; graph projections of d-uniform and heterogeneous hypergraphs; bounded-size feature truncations; signed triangle-relation chains.
$\mathcal{C}_{\mathrm{glob}}$	Latent labels, latent features, latent geometry, or other macroscopic constraints coupling many edges.	Stochastic block models; random intersection graphs; random geometric graphs; thresholded Wishart signs.

1.3 Main Contributions

1.3.1 Local Dependencies as a Bridge between Latent Geometry and Independence

The first contribution is a distributional bridge between latent geometric graph models, locally dependent graph models, and independent Erdős–Rényi baselines. In this thesis, "latent geometry" includes both continuous geometric models and discrete feature-space models. A random intersection graph can be written as a random geometric graph over the discrete cube: sample latent feature vectors

$$X_1, \dots, X_n \in \{0,1\}^m, \qquad (X_i)_a \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(\delta),$$

and connect i and j if and only if

$$\langle X_i, X_j \rangle \geq 1.$$

Thus the random intersection graph is obtained by thresholding the inner product of random 0-1 feature vectors at 1. Its dependence is latent-geometric in the sense that all edges incident to a vertex are coupled through the same latent vector, while each feature may create a clique among all vertices carrying that feature.

This representation also explains why local models naturally interpolate between random intersection graphs and Erdős–Rényi graphs. A feature whose support has size 2 contributes a single edge and is absorbed into an effective independent edge density. A feature whose support has size 3 contributes a triangle. More generally, a feature whose support has size r contributes an r-clique, which is a degree-r local update. Truncating the latent-feature model to features of size at most r therefore gives a degree-r local-kernel model. As r increases, this truncated local model captures larger feature-induced cliques and approximates the full random intersection graph in more regimes. The degree-3 truncation is the first nontrivial correction beyond independence, and it is precisely the random graph with triangles in the constant-edge-density regime.

Theorem 1.1 (Random intersection graphs and local triangle dependence, informal)**.** *Let $G^{\mathrm{int}} \sim \mathrm{RIG}(n, m, \delta)$, and assume the constant-edge-density scaling*

$$m\delta^2 = \Theta(1).$$

If

$$m \gg n^{5/2},$$

then

$$d_{\mathrm{TV}}\big(\mathcal{L}(G^{\mathrm{int}}), \mathrm{RGT}(n, p_m, p'_m)\big) = o(1),$$

where

$$p'_m = 1 - \exp\big\{-m\delta^3(1-\delta)^{n-3}\big\},$$

and

$$p_m = 1 - \exp\big\{-m\delta^2 + (n-2)m\delta^3(1-\delta)^{n-3}\big\}.$$

Moreover, since the indistinguishability threshold between $\mathrm{RIG}(n, m, \delta)$ *and the corresponding Erdős–Rényi graph occurs at the* $m \asymp n^3$ *scale, there is an intermediate regime*

$$n^{5/2} \ll m \ll n^3$$

in which the random intersection graph is close in total variation to a locally dependent RGT *model but not to the independent Erdős–Rényi baseline.*

This result explains why triangles are the correct local correction for random intersection graphs in the constant-edge-density regime. The latent feature model may generate cliques of many sizes, but in the range above, size-two features contribute to the effective independent edge density, size-three features contribute the leading local triangle correction, and larger feature-induced cliques become negligible in total variation [17, 19]. Thus an apparently latent-global model admits an explicit local description before it becomes fully independent.

The second local approximation result concerns continuous latent geometry. Consider the balanced spherical random geometric graph $\mathrm{RGG}(n, D, 1/2)$, equivalently the sign pattern

$$G^{\mathrm{geo}}_{ij} = \mathrm{sign}\langle Z_i, Z_j\rangle, \qquad 1 \le i < j \le n,$$

where $Z_1, \ldots, Z_n$ are independent Gaussian vectors in $\mathbb{R}^D$. It is known that the threshold for asymptotic indistinguishability from $\mathrm{G}(n, 1/2)$ occurs at the $D \asymp n^3$ scale, and that signed triangle statistics govern the leading distinction [17]. The thesis develops the next level of this picture: below the n^3 threshold but above n^2 up to logarithmic factors, the first surviving correction is local.

Theorem 1.2 (Middle-scale local approximation, informal)**.** *Let G^{geo} denote the balanced signed random geometric graph, and let G_6 denote the explicit local triangle-sampling chain obtained from independent signs by processing edges in lexicographic order and occasionally replacing the current edge by the product of two earlier edges sharing an anchor. If*

$$D \gg n^2 \log^{3/2} n,$$

then

$$d_{\mathrm{TV}}(\mathcal{L}(G^{\mathrm{geo}}), \mathcal{L}(G_6)) = o(1).$$

The proof begins from an exact Gram–Schmidt coupling of the geometric sign array, isolates fragile edges whose signs can be changed by lower-order triangle terms, and then passes through a sequence of local graph models. The final comparison uses an anchor-based polymer expansion to replace weighted local updates by an unweighted local triangle-sampling rule [20]. This result shows that a globally geometric dependence structure can, in a middle-scale regime, be represented by an explicit local combinatorial mechanism.

Together, the random intersection graph and random geometric graph results give two complementary instances of the same principle. Discrete latent geometry on $\{0,1\}^m$ and continuous spherical geometry in $\mathbb{R}^D$ both approach independence through local triangle corrections. Local triangle dependencies are therefore not merely an artificial perturbation of Erdős–Rényi graphs; they arise as effective descriptions of latent-feature and latent-geometric graph models.

1.3.2 Algorithmic Decorrelation for Global Recovery

The second contribution is a set of algorithmic reductions showing that planted dense subgraph and planted clique problems in an RGT ambient graph are computationally equivalent to their Erdős–Rényi counterparts in a nontrivial parameter regime [19].

The difficulty is that the dependent ambient graph contains many local triangle structures. These structures can obscure the planted signal and invalidate reductions designed for independent noise. The thesis develops an algorithmic decorrelation map that removes the triangle dependence while preserving the planted signal distribution up to vanishing total variation error.

At the null level, the forward map from $\mathrm{G}(n,p)$ to $\mathrm{RGT}(n,p,p')$ adds random triangles. The reverse map is more delicate. It views the triangle-addition procedure as a Markov kernel and implements an approximate time reversal: given an observed graph G, it samples a latent triangle set from the posterior triangle distribution and then resamples the edges covered by those triangles. The main challenge is to prove that this same reverse map also works when a dense subgraph has been planted.

Theorem 1.3 (Computational equivalence of PDS and planted RGT, informal)**.** *Let $0 < p, q < 1$ be constants, let $p' \leq 1/(n \log n)$, and let*

$$k = o\Big(n^{1/4} \log^{-17/4} n\Big).$$

There exist randomized polynomial-time maps $A_\triangle$ and A such that $A_\triangle$ maps both the null and planted Erdős–Rényi planted dense subgraph distributions to their planted RGT *analogues up to $o(1)$ total variation error, and A maps both the null and planted* RGT *distributions back to the corresponding Erdős–Rényi planted dense subgraph distributions up to $o(1)$ total variation error. The induced planted edge-density maps compose to the identity up to $o(1)$.*

The proof introduces a perturbation-insensitivity framework for average-case reductions. The framework shows that a map between unplanted distributions also maps planted distributions correctly if its output is sufficiently insensitive to single-edge perturbations. Establishing this condition for the reverse triangle-removal map requires new total-variation estimates for dependent Gibbs-type distributions over triangle sets, low-influence bounds, likelihood-ratio concentration, and mixing estimates for Glauber dynamics.

This contribution gives a rigorous sense in which local triangle dependence does not change the computational complexity of the planted dense subgraph problem in the stated regime. It also provides a general template for transforming dependent noise into independent noise while preserving planted structure.

1.3.3 Sharp Local Recovery Thresholds for Hypergraph Projections

The third contribution concerns local recovery. Let H be a random d-uniform hypergraph on $[n]$, with each d-set included independently with probability

$$p = (c + o(1))n^{-d+1+\delta}.$$

The graph projection $\mathrm{Proj}(H)$ connects two vertices if they lie in a common hyperedge. The recovery problem is to reconstruct H from $\mathrm{Proj}(H)$, or from the weighted projection $\mathrm{Proj}_{\mathrm{W}}(H)$ that records edge multiplicities.

The thesis establishes sharp thresholds for both partial recovery and exact recovery. In partial recovery, the goal is to recover all but a vanishing fraction of the hyperedges, measured by normalized symmetric-difference loss. In exact recovery, the goal is to recover the entire hypergraph with probability $1 - o(1)$.

Theorem 1.4 (Partial recovery from graph projections, informal)**.** *For $d \geq 3$, partial recovery of a random d-uniform hypergraph from its unweighted projection has an all-or-nothing threshold at*

$$\delta_{\mathrm{part}}(d) = \frac{d-1}{d+1}.$$

If $\delta < \delta_{\mathrm{part}}(d)$, the optimal partial recovery loss is $o(1)$, and this is achieved by the maximum clique cover estimator. If $\delta > \delta_{\mathrm{part}}(d)$, the optimal partial recovery loss is $1 - o(1)$. The same partial recovery threshold holds for weighted projections.

For exact recovery, the answer depends on whether multiplicity information is observed. In the unweighted projection model, the bottleneck is the appearance of ambiguous graphs: projected graphs with two distinct minimum hypergraph preimages. In the weighted projection model, these ambiguities are partly removed by edge multiplicities.

Theorem 1.5 (Exact recovery from graph projections, informal)**.** *For unweighted projections, exact recovery has threshold*

$$\delta_{\mathrm{exact}}(d) = \begin{cases} \dfrac{2d-4}{2d-1}, & 3 \leq d \leq 4, \\ \dfrac{d-1}{d+1}, & d \geq 5. \end{cases}$$

Equivalently, the formula $(2d-4)/(2d-1)$ agrees with $(d-1)/(d+1)$ at $d = 5$. Below the threshold, the probability of exact recovery is $1 - o(1)$; above it, the probability of exact recovery is $o(1)$. For weighted projections, the exact recovery threshold is $(d-1)/(d+1)$ for every $d \geq 3$.

These results resolve the partial and exact recovery thresholds for d-uniform random hypergraphs from graph projections, including all-or-nothing behavior. The achievability side relies on clique-cover and MAP estimators, together with structural results showing that relevant two-connected components remain bounded in the recoverable regime. The converse side uses overlap arguments for partial recovery and ambiguous-graph obstructions for exact recovery [24, 25].

1.3.4 Heterogeneous Hypergraph Recovery

The final contribution extends the local recovery viewpoint beyond the uniform setting. In a heterogeneous random hypergraph, hyperedges of degrees

$$2 \le d_1 < d_2 < \cdots < d_k$$

appear independently, with degree-d_j hyperedges included with probability

$$p_j = n^{1-d_j+\delta_j+o(1)}.$$

The observed graph is the projection of the union of all degrees, and the goal is to recover the hyperedges of a target degree d_j.

A simple estimator selects maximal cliques of size d_j in the projected graph. The analysis shows that this estimator succeeds when the dominant false-clique mechanisms are asymptotically dominated by the true degree-d_j hyperedges.

Theorem 1.6 (Heterogeneous achievability, informal)**.** *Let*

$$\delta_* = \max_{r\in[k]} \delta_r.$$

Fix j with $d_j \ge 3$. If

$$\delta_* < \frac{d_j - 2}{d_j} + \frac{2\delta_j}{d_j(d_j-1)},$$

then the maximal-clique estimator recovers the degree-d_j hyperedges with vanishing normalized expected symmetric-difference error:

$$\mathbb{E}\Big[|E(H_j) \triangle E(\widehat{H}_j)|\Big] = o(1)\,\mathbb{E}[|E(H_j)|].$$

This theorem is an achievability result rather than a full threshold theorem. It shows that the d-uniform threshold $(d-1)/(d+1)$ extends naturally to a heterogeneous setting through a domination condition involving the largest projected noise exponent. It also identifies the all-edges noise configuration as the key obstruction controlled by the maximal-clique estimator [26].

Taken together, the contributions of the thesis show how locally dependent random graphical models can be compared, decorrelated, and inverted. The global results demonstrate that local dependencies need not change the computational complexity of planted-signal recovery when suitable reductions are available. The local results identify exactly when higher-order communities can be reconstructed from their pairwise shadows. Both directions support the same message: local dependencies are rich enough to model realistic network effects, but structured enough to permit sharp statistical and computational analysis.

Chapter 2

Related Work

This chapter situates the thesis within the literature on community recovery, average-case complexity, dependent random graph models, and higher-order network inference. The emphasis is on prior work and on the limitations that motivate the thesis. Detailed statements of the thesis results are deferred to the later technical chapters.

The organization follows the modeling progression introduced in Chapter 1. Section 2.1 reviews independent-noise baselines, including stochastic block models, planted clique, planted dense subgraph, and algorithmic lower-bound frameworks. Section 2.2 discusses average-case reductions and the extent to which they explain statistical-computational gaps. Section 2.3 reviews dependent graph models, including motif-based, intersection, and geometric models. Section 2.4 turns to higher-order network models, hypergraph stochastic block models, and graph projections. Section 2.5 places hypergraph projection recovery in the broader context of inverse problems and all-or-nothing reconstruction phenomena. Finally, Section 2.6 summarizes the gaps in the literature and explains how the thesis addresses them.

2.1 Independent-Noise Models and Planted Inference

2.1.1 Community detection and stochastic block models

Community detection is a central problem in network science and statistics. In empirical network analysis, the goal is often to identify groups of vertices with unusually dense internal connectivity or shared structural roles. Modularity-based methods and related approaches formalize this intuition at the algorithmic level [1]. In probabilistic modeling, the stochastic block model gives one of the simplest and most influential statistical formulations: vertices receive latent labels, and edges are generated with probabilities depending on the labels [2, 3].

The stochastic block model already displays several themes that recur throughout this thesis: latent structure, noisy pairwise observations, and thresholds for weak or exact recovery. However, most of the classical stochastic block model literature retains conditional edge independence given the labels. This conditional product structure is a major reason the model admits sharp analyses of likelihoods, spectral methods, semidefinite relaxations, and belief-propagation-type algorithms.

For the purposes of this thesis, the limitation is that conditional independence is still an independence assumption. It captures community heterogeneity but not the local dependence created by motifs, overlapping features, hyperedges, or geometric constraints. Thus stochastic block models provide a foundational baseline for recovery, but they do not by themselves explain how inference changes when the noise distribution has dependent edges.

2.1.2 Planted clique and planted dense subgraph

The planted clique problem is one of the canonical average-case problems in theoretical computer science. In its standard form, a clique of size k is planted in an otherwise independent $\mathrm{G}(n, 1/2)$ graph. Although detection and recovery are information-theoretically possible well below the $k \asymp \sqrt{n}$ scale, no polynomial-time algorithm is known below that scale, and the corresponding planted clique conjecture has become a standard hardness assumption [4, 5, 10, 12].

The planted dense subgraph problem generalizes planted clique by replacing the planted complete subgraph with a planted subgraph of elevated edge density. This generality is important for reductions because many statistical problems contain a weak planted signal rather than a hard combinatorial constraint. Planted dense subgraph and related dense subgraph problems have been used to understand computational barriers in community detection, submatrix detection, sparse PCA, and tensor problems [6, 27–31].

The strength of this literature is that it gives a precise independent-noise benchmark for statistical-computational gaps. The gap for the present thesis is that the benchmark is almost always formulated with product-measure ambient noise. In a dependent ambient graph, local structures such as triangles or feature-induced cliques may create dense subgraphs unrelated to the planted signal. It is therefore not automatic that the planted clique or planted dense subgraph heuristics transfer to dependent graph models.

2.1.3 Restricted algorithmic models and lower-bound evidence

A large body of work gives evidence for computational barriers by proving lower bounds against restricted classes of algorithms. Statistical query lower bounds rule out broad families of algorithms that access the data only through noisy expectations [4, 32, 33]. Sum-of-squares lower bounds show that powerful semidefinite programming hierarchies fail for planted clique and related hidden-structure problems below conjectured thresholds [5, 10, 12]. Low-degree polynomial methods provide a predictive framework for the power of polynomial-time algorithms in many high-dimensional testing problems [10, 11]. Approximate message passing gives sharp algorithmic predictions in many matrix and generalized linear models, and its relationship with low-degree methods has been clarified in rank-one estimation settings [34, 35]. Other evidence comes from first-order method lower bounds, circuit lower bounds, and overlap-gap arguments [36–40].

These frameworks have also been compared to one another. Statistical query lower bounds and low-degree lower bounds are closely connected in several settings [41]; low-degree methods have been related to reconstruction on trees [42]; and hierarchy-based methods such as Kikuchi and sum-of-squares provide further evidence for barriers in tensor and clustering problems [43, 44].

The common limitation is that these frameworks are most mature for product or conditionally product distributions. In a locally dependent graph, the natural orthogonal basis, likelihood factorization, or message-passing state evolution may no longer be available in a standard form. This motivates the reduction-based and distribution-comparison viewpoint used in the thesis: rather than proving a new lower bound separately for every algorithmic class in every dependent model, one can try to relate the dependent inference problem to an independent benchmark.

2.2 Average-Case Reductions and Statistical-Computational Gaps

2.2.1 Average-case complexity and distributional reductions

Average-case reductions compare distributions of inputs rather than individual worst-case instances. This is substantially more delicate than worst-case reduction because a valid reduction must map not only a support set but an entire probability law. The foundations of average-case complexity emphasize exactly this point: a reduction must preserve the distributional structure of the problem, including the way typical instances are sampled [8, 9, 45].

In statistical inference, an average-case reduction usually compares a pair of distributions (P_0, P_1) with another pair (Q_0, Q_1), where P_0, Q_0 are null distributions and P_1, Q_1 are planted distributions. To transfer hardness, a randomized polynomial-time map must send P_0 close to Q_0 and P_1 close to Q_1, typically in total variation distance. This requirement is more stringent than preserving a statistic or a marginal distribution: the entire observation law must be controlled.

The gap that remains is methodological. Distributional reductions are powerful, but the known techniques are highly sensitive to independence. Entrywise randomization, rejection kernels, Gaussianization, and secret-leakage constructions are easiest to analyze when the input noise is a product measure. Dependent graph models require reductions that control correlations among many nearby entries simultaneously.

2.2.2 Reductions from planted clique and planted dense subgraph

Planted clique and planted dense subgraph have served as source problems for a large web of conditional hardness results. Reductions from planted clique give evidence for computational barriers in sparse PCA [46, 47], submatrix detection [28], community detection and planted dense subgraph variants [6, 29], sparse CCA [48], and tensor problems [30, 31]. More general reduction frameworks establish broad universality results for planted sparse structure and related high-dimensional models [6, 7, 49, 50].

This literature has two major achievements. First, it shows that many apparently different statistical problems share a common computational barrier. Second, it gives a language for transferring phase diagrams across problems, thereby explaining why the same exponents recur in sparse estimation, detection, and recovery.

The limitation is that most of these reductions start with independent noise and produce another model whose noise is either independent or generated by an entrywise transformation of independent noise. They usually do not remove dependence from an observed graph. Thus, before this thesis, there was no general reason that a planted problem in a locally dependent ambient graph should be computationally equivalent to its Erdős–Rényi analogue.

2.2.3 Cryptographic hardness and alternative assumptions

A complementary line of work bases statistical hardness on cryptographic assumptions rather than on planted clique. Examples include reductions from variants of learning with errors or related pseudorandomness assumptions to learning and estimation problems [51–58]. These works are important because they supply hardness evidence that does not rely on one particular average-case graph conjecture.

However, cryptographic reductions also usually aim to create hidden structure inside a distribution whose random part is controlled by a pseudorandom generator or an entrywise construction. They do not directly address the graph-theoretic issue central to this thesis: the observed noise itself may contain local dependencies caused by motifs, hyperedges, or latent features. The open methodological question is how to compare inference problems when the difference between them is not just the entry distribution, but the dependence structure among entries.

2.2.4 Null distributions versus planted distributions

A recurring subtlety in average-case reductions is the distinction between matching null distributions and matching planted distributions. It is sometimes possible to construct a map that transforms one unplanted random graph distribution into another. For planted inference, this is not enough. The same map must also preserve the planted signal, or at least transform it into the correct planted signal in the target model.

This issue becomes more difficult in dependent graphs. A transformation that changes local motifs may interact with the planted set, especially if the planted set changes the local motif counts near its boundary or inside the signal. Consequently, dependent-noise reductions require a form of stability: the transformation must be insensitive enough to the planted perturbation while still strong enough to alter the ambient dependence. Existing reduction frameworks identify the importance of distributional accuracy, but they provide few general tools for this reverse direction from dependent noise to independent noise.

2.3 Dependent Graph Models

2.3.1 Clustering, motifs, and transitivity

Empirical networks often contain far more triangles and other small motifs than an Erdős–Rényi graph with the same edge density. In social networks, the strength of weak ties and the prevalence of triadic closure point to local dependence among edges [13, 14]. In collaboration networks, a single group event can generate a clique in the pairwise coauthorship graph [15].

Biological and technological networks also exhibit structure driven by shared mechanisms rather than independent pairwise links [16].

Motif-based random graph models attempt to represent these phenomena directly. A particularly simple example is obtained by starting from an Erdős–Rényi graph and adding random triangles, producing a graph with explicit local edge dependence [19]. Such models are useful because the dependence mechanism is transparent: correlations arise from bounded-size objects.

The gap is that motif models occupy an intermediate position in the literature. They are more realistic than independent-edge models, but much less structured than full latent-space or exponential-family network models. Classical planted-inference theory is not designed for them, while general dependent-network theory often does not give the fine total-variation estimates needed for average-case reductions.

2.3.2 Random intersection graphs and overlapping communities

Random intersection graphs model overlapping features or memberships. Each vertex is assigned a random subset of latent features, and two vertices are connected if they share a feature. This mechanism is natural for affiliation networks, collaboration networks, and settings in which observed pairwise edges arise from hidden groups. A single feature may create a clique among all vertices carrying that feature, so the observed graph contains local clustering after the features are marginalized out [17].

The theory of random intersection graphs connects latent features, clustering, and Erdős–Rényi approximation. In sparse or high-dimensional regimes, the observed graph may approach an independent-edge model; in other regimes, feature-induced dependencies remain detectable [17]. Variants with community structure further enrich the model by combining overlapping features with latent labels.

The limitation is that random intersection graphs are not automatically local in a bounded-motif sense. If a latent feature is carried by many vertices, it creates a large clique and therefore couples many edges at once. Thus an intersection graph may exhibit local-looking motifs while being generated by latent objects of unbounded size. This distinction motivates the search for regimes where latent-feature dependence can be approximated by an explicit local model, rather than assuming that all feature models are local by definition.

2.3.3 Random geometric graphs and latent geometry

Random geometric graphs generate edges from latent positions in a metric or spherical space [18]. In high-dimensional spherical models, the balanced case can be viewed as the sign pattern of a Wishart matrix: vertices correspond to random vectors, and an edge sign is determined by the sign of their inner product. This representation connects random geometric graphs to random matrix theory and to tests for latent geometry.

A line of work studies when a random geometric graph is distinguishable from an Erdős–Rényi graph with the same edge density. In high dimensions, the geometric constraints weaken, but they do not disappear uniformly across all statistics. Triangle-based statistics play an important role in detecting the residual dependence created by latent geometry [17, 59]. These results give a sharp statistical comparison between geometric and independent random graphs in several regimes.

The remaining gap is interpretive and algorithmic. Detection results identify when geometry can be distinguished from independence, but they do not necessarily provide an explicit graph process that explains how the geometric law first deviates from Erdős–Rényi. For the thesis, this distinction matters because an explicit local description is needed to connect global latent dependence with locally dependent graph models.

2.3.4 Distributional comparison for dependent graphs

Total variation distance and contiguity are natural tools for comparing random graph distributions. If two graph laws are close in total variation, then every test and estimator has asymptotically the same distribution under the two models. This is a stronger statement than matching finitely many moments or showing that a particular statistic has the same limit.

For independent graphs, total-variation calculations often reduce to likelihood-ratio second moments or product-measure estimates. For dependent graphs, such calculations are substantially more difficult because the likelihood ratio may involve overlapping motifs or latent variables. Existing results for random intersection and geometric graphs show that total-variation comparisons are possible, but they also show that the answer can be sensitive to intermediate regimes [17, 59].

The gap is that total-variation equivalence of unplanted distributions does not automatically yield equivalence of planted inference problems. Moreover, for globally dependent models, even finding the right intermediate comparison distribution can be nontrivial. This motivates the thesis's use of local dependence as a comparison layer between independent baselines and latent dependent models.

2.4 Higher-Order Networks and Graph Projections

2.4.1 Hypergraphs as models of higher-order interaction

Hypergraphs provide a natural language for interactions involving more than two vertices. They arise in coauthorship networks, project membership data, group communication, biological complexes, and other settings where the fundamental unit of interaction is a group rather than a pair [15, 60]. In such settings, replacing each group interaction by all of its pairwise links can be useful computationally but may lose information about which pairs came from the same group event.

This tension is pervasive in network analysis. Most classical graph algorithms operate on pairwise edges, so higher-order data are often transformed into graphs through clique expansion or similarity matrices. Such transformations make the data compatible with existing graph methods, but they also obscure the original hyperedge structure.

The gap is therefore an information question: when does projecting a hypergraph to a graph preserve the higher-order information, and when does it irreversibly merge distinct hypergraph explanations? This question is separate from, and often stronger than, community recovery: even if one can recover vertex labels from a projected graph, it does not follow that one can recover the individual hyperedges.

2.4.2 Hypergraph stochastic block models

The hypergraph stochastic block model extends the stochastic block model to higher-order interactions. Hyperedges are generated with probabilities depending on the labels of all vertices in the hyperedge. This model has motivated spectral methods, exact recovery analyses, weak recovery thresholds, and non-uniform extensions [61–65].

A recurring approach in this literature is to reduce the hypergraph to a pairwise similarity matrix. The entry W_{ij} records how many hyperedges contain both i and j, and graph-based algorithms are then applied to W. This approach is appealing because it imports tools from graph community detection and spectral graph theory. It also raises a natural question: does the similarity matrix lose information compared to the original hypergraph?

Prior work on HSBM shows that pairwise projections or similarity matrices can be algorithmically powerful for label recovery [63–65]. The limitation is that label recovery does not fully characterize the projection channel. For a complete understanding of higher-order data reduction, one also needs to know when the hyperedges themselves are recoverable from the pairwise observation.

2.4.3 Empirical hypergraph reconstruction from projected graphs

Recent work has studied hypergraph reconstruction from network data directly. Posterior-sampling approaches place a prior on hypergraphs and attempt to sample plausible hypergraphs conditioned on the observed projected graph [21, 22]. Scoring methods use auxiliary hypergraph samples or learned structural information to identify likely hyperedges in a projected graph [23]. Related work studies hypergraph recovery from the perspective of relational learning and representation learning [66].

These works are important because they show that hypergraph reconstruction is not merely a technical abstraction: it is a practical problem arising from the way higher-order data are often recorded. They also highlight the diversity of possible observation models, including uncertain pairwise observations, weighted hyperedges, and data-dependent priors.

The limitation is that empirical reconstruction methods usually do not give sharp information thresholds for simple random hypergraph models. They may perform well on real or simulated data, but they do not by themselves answer when reconstruction is statistically possible, when it is impossible, or whether there is a sharp transition between the two.

2.5 Reconstruction Phenomena and Inverse Problems

2.5.1 Exact recovery, partial recovery, and fidelity transitions

Recovery problems are often studied under different fidelity requirements. Exact recovery asks for the entire latent object to be reconstructed with probability $1 - o(1)$. Partial recovery, or almost exact recovery, asks for a vanishing fraction of errors under an appropriate normalized loss. The distinction is important because exact recovery can be blocked by rare local ambiguities, while partial recovery may still be possible if those ambiguities affect only a negligible fraction of the latent structure.

All-or-nothing phenomena have been identified in several high-dimensional inference problems. In sparse linear regression, graph matching, planted constraint satisfaction, and generalized linear models, the best achievable overlap or reconstruction fidelity can jump sharply from nearly perfect to essentially trivial [67–71]. These results provide a conceptual backdrop for studying projection recovery: a projected graph may either retain almost all of the latent hypergraph information or almost none of it, depending on the density and observation model.

The gap is that hypergraph projection has a distinct combinatorial structure. Ambiguity arises from clique covers, overlapping hyperedges, and multiple preimages of the same projected graph. Thus existing all-or-nothing results suggest the type of transition one might expect, but they do not determine the thresholds or mechanisms for graph projections of random hypergraphs.

2.5.2 Projection as a planted inverse problem

Hypergraph projection recovery can be viewed as an inverse problem. In the weighted case, the observation records a linear image of the hyperedge indicator vector under the edge-hyperedge incidence matrix. In the unweighted case, the observation is a nonlinear, thresholded version of the same incidence relation. This places projection recovery near planted constraint satisfaction and generalized linear inverse problems, while retaining a strongly graph-theoretic flavor.

The literature on planted CSPs and inverse problems provides tools for studying posterior overlap, entropy, and threshold behavior [69, 71, 72]. However, the projection map has special features not present in generic random linear models: each hyperedge maps to a clique, different hyperedges can share projected edges, and the feasible preimages of an observed graph are clique covers.

The gap is that the combinatorial geometry of clique covers must be analyzed directly. Generic inverse-problem theory does not identify which local projected configurations create ambiguity, nor does it determine when such configurations proliferate in a random projected graph.

2.6 Synthesis and Thesis Positioning

The prior literature leaves three main gaps.

First, independent-noise models provide sharp benchmarks for planted clique, planted dense subgraph, stochastic block models, and many related inference problems. Average-case reductions and restricted-algorithm lower bounds give substantial evidence for statistical-computational gaps in these models. What is missing is a comparable theory for planted inference when the ambient noise has local dependencies. Existing reductions largely change entry distributions or embed independent noise into other independent-noise problems; they rarely remove dependence from the observed noise while preserving a planted signal.

Second, dependent graph models such as random intersection graphs and random geometric graphs capture important network effects, including overlapping memberships and latent geometry. Yet their dependence structures are often global or latent, making them difficult to compare directly with independent baselines. Existing testing results identify when such models are distinguishable from Erdős–Rényi, but they do not always provide an explicit local graph mechanism explaining the first nontrivial deviation from independence.

Third, higher-order network models and hypergraph stochastic block models motivate the use of pairwise projections and similarity matrices. Empirical reconstruction methods show that inferring hyperedges from pairwise data is useful, and HSBM results show that pairwise reductions can be powerful for label recovery. What remains is a sharp understanding of the projection channel itself: when the original hyperedges can be recovered, how exact and partial recovery differ, and how the answer changes with weighted observations or heterogeneous interaction sizes.

The thesis addresses these gaps by developing a common framework centered on local dependencies. For distributional comparison, it studies when dependent graph laws can be approximated in total variation by explicit local models, connecting latent-feature and geometric models to locally dependent graph processes [17, 20]. For global recovery, it develops reductions that compare planted inference in independent and locally dependent ambient graphs, thereby addressing the missing dependent-to-independent direction in average-case reductions [19]. For local recovery, it analyzes graph projections of random hypergraphs as inverse problems and characterizes when the latent hyperedges are recoverable from their pairwise shadows [24–26].

Thus the contribution of the thesis is not to replace the independent-noise theory, but to extend its reach. Independent models remain the benchmark; locally dependent models provide the intermediate comparison layer; and globally dependent or higher-order models supply the motivating network structures. The following chapters formalize this viewpoint and develop the corresponding distributional comparisons, reductions, and reconstruction algorithms.

Chapter 3

Models, Metrics, and Dependency Hierarchy

3.1 Notation and Asymptotic Conventions

Throughout the thesis, $[n] = \{1, \dots, n\}$. Graphs are simple and undirected unless stated otherwise. For a graph G, let $V(G)$ and $E(G)$ denote its vertex and edge sets. When the vertex set is clear, we identify G with its edge set or edge-indicator vector. For a hypergraph H, let $E(H)$ denote its hyperedge set. We write

$$\binom{[n]}{r} = \{S \subseteq [n] : |S| = r\}.$$

All asymptotics are as $n \to \infty$. Other parameters, such as the geometric dimension, feature dimension, planted size, or hyperedge density, may depend on n. An event holds with high probability if its probability is $1 - o(1)$.

For a random variable X, let $\mathcal{L}(X)$ denote its law. For probability measures P, Q on the same measurable space, the total variation distance is

$$d_{\mathrm{TV}}(P, Q) = \sup_{A} |P(A) - Q(A)|.$$

When the random variables are clear, we also write

$$d_{\mathrm{TV}}(X, Y) = d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Y)).$$

To avoid overloading notation, D denotes the dimension of a geometric latent space, while d denotes hyperedge uniformity. For signed graph models, an edge variable takes values in $\{-1, +1\}$. The balanced signed independent-edge law is the product law in which all edge signs are independent Rademacher random variables; it is equivalent to $\mathrm{G}(n, 1/2)$ after identifying $+1$ with edge presence and -1 with edge absence.

3.2 Ambient Graph Models

Definition 3.1 (Erdős–Rényi graph)**.** The Erdős–Rényi graph $\mathrm{G}(n, p)$ is the random graph on vertex set $[n]$ in which each edge $e \in \binom{[n]}{2}$ is included independently with probability p.

Definition 3.2 (Random graph with triangles)**.** The random graph with triangles $\mathrm{RGT}(n, p, p')$ is generated by first sampling $G_0 \sim \mathrm{G}(n, p)$. Then, independently for every triple $T \in \binom{[n]}{3}$, with probability p' we add all three edges of the triangle on T. Equivalently,

$$E(G) = E(G_0) \cup \bigcup_{T : X_T = 1} \binom{T}{2}, \qquad X_T \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p').$$

The marginal edge density is

$$1 - (1-p)(1-p')^{n-2}.$$

The regime $p' = O(1/n)$ keeps this density bounded away from one.

Definition 3.3 (Random intersection graph)**.** The random intersection graph $\mathrm{RIG}(n, m, \alpha)$ has vertex set $[n]$ and latent feature set $[m]$. Each vertex $i \in [n]$ is assigned each feature $a \in [m]$ independently with probability α. Vertices i and j are adjacent if and only if their feature sets intersect:

$$\{i, j\} \in E(G) \quad \Longleftrightarrow \quad S_i \cap S_j \neq \emptyset.$$

The marginal edge density is

$$1 - (1 - \alpha^2)^m.$$

A single latent feature creates a clique among all vertices carrying that feature, so the observed graph has edge dependencies induced by latent local communities.

Definition 3.4 (Spherical random geometric graph)**.** Let $\Theta_1, \ldots, \Theta_n$ be i.i.d. uniform points on the sphere S^{D-1}. The spherical random geometric graph $\mathrm{RGG}(n, D, p)$ connects i and j if

$$\langle \Theta_i, \Theta_j \rangle \geq \theta_{p,D},$$

where $\theta_{p,D}$ is chosen so that

$$\mathbb{P}\big(\langle \Theta_1, \Theta_2 \rangle \geq \theta_{p,D}\big) = p.$$

In the balanced case $p = 1/2$, $\theta_{p,D} = 0$, and it is often convenient to work with signs

$$Y_{ij} = \mathrm{sign}\langle Z_i, Z_j \rangle, \qquad Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} N(0, I_D),$$

where the graph edge $\{i, j\}$ is present exactly when $Y_{ij} = 1$. This model has global dependence induced by the latent geometry.

3.3 Hypergraphs and Graph Projections

Definition 3.5 (d-uniform random hypergraph)**.** For $d \geq 3$, the random d-uniform hypergraph $H_d(n, p)$ has vertex set $[n]$, and each d-set $h \in \binom{[n]}{d}$ is included independently with probability p. We frequently use the parameterization

$$p = (c + o(1)) n^{-d+1+\delta}, \qquad c \in (0, \infty),$$

so that the expected degree of a fixed vertex is

$$\binom{n-1}{d-1} p = \left(\frac{c}{(d-1)!} + o(1) \right) n^{\delta}.$$

The sparse reconstruction regime of primary interest has $0 < \delta < 1$.

Definition 3.6 (Graph projection and weighted projection)**.** For a hypergraph $H = ([n], E(H))$, its graph projection is the simple graph

$$\mathrm{Proj}(H) = \left([n], \left\{\{i,j\} \in \binom{[n]}{2} : \exists h \in E(H) \text{ with } \{i,j\} \subseteq h\right\}\right).$$

Equivalently, each hyperedge h is replaced by the clique $\binom{h}{2}$. The weighted projection $\mathrm{Proj}_{\mathrm{W}}(H)$ records edge multiplicities:

$$\mathrm{Proj}_{\mathrm{W}}(H)_{ij} = |\{h \in E(H) : \{i,j\} \subseteq h\}|.$$

Definition 3.7 (Heterogeneous random hypergraph)**.** Fix degrees

$$2 \leq d_1 < d_2 < \cdots < d_k.$$

For each $r \in [k]$, each d_r-set is included independently with probability

$$p_r = n^{1-d_r+\delta_r+o(1)}, \qquad 0 < \delta_r < 1.$$

All hyperedges, across all degrees, are independent. We write H_r for the sub-hypergraph consisting of degree-d_r hyperedges, and $H = \bigcup_{r=1}^{k} H_r$ for the full heterogeneous hypergraph.

3.4 Graph Models Generated by Local Dependencies

The models above show that local dependence can enter in several ways. In RGT, a local motif monotonically inserts a triangle. In a graph projection, a local motif monotonically inserts a clique. In the middle-scale approximation to high-dimensional geometry, a signed triangle relation may replace the current edge sign by a function of two previously exposed signs. Thus the thesis uses a local-kernel definition rather than a purely monotone motif-insertion definition.

For a finite vertex set V, write

$$\Omega_V^{\mathrm{ind}} = \{0,1\}^{\binom{V}{2}}, \qquad \Omega_V^{\mathrm{sgn}} = \{-1,+1\}^{\binom{V}{2}}.$$

An element of either space is a local motif on V. The indicator space is used for ordinary graphs, while the signed space is used for balanced signed graph models.

Definition 3.8 (Graph model generated by local dependencies of degree r)**.** Fix $r \geq 2$ independent of n, and let

$$\Omega_{[n]} = \{0,1\}^{\binom{[n]}{2}} \quad \text{or} \quad \Omega_{[n]} = \{-1,+1\}^{\binom{[n]}{2}}.$$

A random graph or signed graph $G \in \Omega_{[n]}$ is generated by local dependencies of degree r if it can be realized by the following sequential local-kernel process.

First sample a product base graph $G^{(0)}$. In the ordinary graph case this is typically $G^{(0)} \sim \mathrm{G}(n, p_0)$. In the signed case this is typically a product Rademacher sign array.

Then perform R_n local update steps. At step s, choose a footprint

$$V_s \subseteq [n], \qquad |V_s| \leq r,$$

which may be random. Let $K_s(x, \cdot)$ be a Markov kernel from local motifs on V_s to local motifs on V_s. Given the history up to time $s-1$, sample

$$H_s \sim K_s\big(G^{(s-1)}[V_s], \cdot\big),$$

and set

$$G^{(s)}[V_s] = H_s, \qquad G_e^{(s)} = G_e^{(s-1)} \quad \text{for all } e \notin \binom{V_s}{2}.$$

The output is $G = G^{(R_n)}$.

The read load and write load of such a representation are

$$\Lambda_{\text{read}}(n) := \max_{e \in \binom{[n]}{2}} \mathbb{E} \sum_{s=1}^{R_n} \mathbf{1}\{e \in \binom{V_s}{2}\},$$

and

$$\Lambda_{\text{write}}(n) := \max_{e \in \binom{[n]}{2}} \mathbb{E} \sum_{s=1}^{R_n} \mathbf{1}\{e \in \binom{V_s}{2}\} \sup_x K_s\big(x, \{y : y_e \neq x_e\}\big).$$

The representation is bounded-load local if $\Lambda_{\text{read}}(n) = O(1)$, and sparse local if $\Lambda_{\text{read}}(n) = o(1)$.

Remark 3.9 (Why kernels rather than insertions). The update kernel K_s may depend on the current local motif $G^{(s-1)}[V_s]$. Thus a local update may add a triangle, add a clique, resample a bounded induced subgraph, or impose a signed relation among existing signs. What makes the operation local is the bounded footprint $|V_s| \leq r$, together with a controlled read or write load, not monotonicity.

Definition 3.10 (Monotone motif insertion)**.** A monotone motif insertion model of degree r is the special case of Definition 3.8 in which each active update has a fixed edge shadow

$$E_s \subseteq \binom{V_s}{2}, \qquad |V_s| \leq r,$$

and

$$K_s(x, \cdot) = \delta_{x \cup E_s}(\cdot).$$

Thus each active motif only adds edges. Random graphs with triangles and graph projections of independent hypergraphs are monotone motif insertion models.

Example 3.11 (Random graph with triangles as a degree-3 local-kernel model). The model $\mathrm{RGT}(n, p, p')$ is generated by local dependencies of degree 3. Start with $G^{(0)} \sim \mathrm{G}(n, p)$. Independently for every triple $T \in \binom{[n]}{3}$, let $B_T \sim \mathrm{Ber}(p')$. For each active triple T with $B_T = 1$, take $V_s = T$ and use the monotone kernel

$$K_T(x, \cdot) = \delta_{x \cup \binom{T}{2}}(\cdot).$$

The output satisfies

$$E(G) = E(G^{(0)}) \cup \bigcup_{T:B_T=1} \binom{T}{2}.$$

For a fixed edge e, there are $n-2$ triangles containing e, so

$$\Lambda_{\text{read}}(n) = (n-2)p'.$$

Thus $p' = O(1/n)$ is the bounded-load regime.

Example 3.12 (Ordered signed triangle-relation chain). The ordered signed triangle-relation chain is generated by local dependencies of degree 3, but it is not a monotone insertion model. Work on

$$\Omega_{[n]} = \{-1, +1\}^{\binom{[n]}{2}},$$

and start from independent balanced signs

$$Y_{ij}^{(0)} = X_{ij}, \qquad X_{ij} \overset{\text{i.i.d.}}{\sim} \text{Rad}(0).$$

Process top edges (i, j), $i < j$, in lexicographic order. For each top edge, choose

$$L_{ij} \in \{0, 1, \ldots, i-1\}$$

with

$$\mathbb{P}(L_{ij} = a) = \theta_{aij}, \qquad 1 \le a < i,$$

and

$$\mathbb{P}(L_{ij} = 0) = 1 - \sum_{a<i} \theta_{aij}.$$

If $L_{ij} = 0$, there is no active update. If $L_{ij} = a \ge 1$, perform an update on

$$V_s = \{a, i, j\}$$

with deterministic kernel

$$\Phi_{aij}(x)_{ai} = x_{ai}, \qquad \Phi_{aij}(x)_{aj} = x_{aj}, \qquad \Phi_{aij}(x)_{ij} = x_{ai}x_{aj}.$$

Equivalently, the update enforces

$$Y_{ai}Y_{aj}Y_{ij} = 1$$

while leaving the lower signs Y_{ai} and Y_{aj} unchanged. In the uniform case $\theta_{aij} = \lambda$, the read load is $O(n\lambda)$. The signed triangle chain appearing later in the thesis has $\lambda = \Theta(D^{-1/2})$, so it is sparse local when $D \gg n^2$.

Example 3.13 (Graph projections of random hypergraphs). Let $H \sim H_d(n, p_d)$ be an independent d-uniform random hypergraph. Its graph projection is generated by local dependencies of degree d. Start from the empty graph. For each active hyperedge $h \in E(H)$, take $V_s = h$ and use the monotone kernel

$$K_h(x, \cdot) = \delta_{x \cup \binom{h}{2}}(\cdot).$$

The output is

$$\mathrm{Proj}(H) = \bigcup_{h \in E(H)} \binom{h}{2}.$$

For a fixed graph edge e, the expected number of active d-hyperedges containing it is

$$\binom{n-2}{d-2} p_d.$$

Under

$$p_d = n^{-d+1+\delta+o(1)},$$

this load is

$$\binom{n-2}{d-2} p_d = n^{-1+\delta+o(1)}.$$

Thus, in the sparse reconstruction regime $0 < \delta < 1$, random hypergraph projections are sparse local graph models.

For a heterogeneous random hypergraph with a degree-two component, the degree-two hyperedges form an Erdős–Rényi base graph, while the higher-degree hyperedges add independent local cliques.

Remark 3.14 (Local dependence versus arbitrary latent dependence). Definition 3.8 is intentionally narrower than the class of all latent-variable graph models. A latent feature that creates an unbounded clique is not a bounded-degree local dependency unless the feature size is itself bounded or the model is approximated by bounded-size active footprints. Similarly, a low-dimensional geometric graph is not itself generated by bounded-degree local kernels. The point of the later equivalence results is that, in certain asymptotic regimes, such latent or geometric models can be approximated by local-kernel models.

3.5 Planted and Reconstruction Problems

3.5.1 Global planted-signal problems

The global recovery problems in this thesis involve a macroscopic planted structure embedded in an ambient graph. In this setting, the dependence structure belongs to the ambient noise distribution.

Definition 3.15 (Planted dense subgraph)**.** The planted dense subgraph distribution $\mathrm{PDS}(n,p,k,q)$ is generated by first choosing $S \subset [n]$ uniformly with $|S| = k$. Starting from $G \sim \mathrm{G}(n,p)$, erase all edges inside S, and then include each edge inside S independently with probability q. The null model is $\mathrm{G}(n,p)$. The planted clique model is the special case in which the planted subgraph is forced to be complete.

Definition 3.16 (Planted dense subgraph in RGT)**.** The planted RGT distribution $\mathrm{RGT}(n,p,p',k,q)$ is generated by sampling an ambient graph from $\mathrm{RGT}(n,p,p')$, choosing $S \subset [n]$ uniformly with $|S| = k$, erasing the edges inside S, and resampling each edge inside S independently with probability q. The corresponding null model is $\mathrm{RGT}(n,p,p')$.

A detection algorithm distinguishes the null model from the planted model. A recovery algorithm estimates the planted vertex set S.

3.5.2 Local hypergraph recovery

The local recovery problems in this thesis involve many small latent structures. In this setting, the local motifs are not merely noise; they are the objects to be reconstructed.

Problem 3.17 (Hypergraph recovery from projection)**.** Given $\mathrm{Proj}(H)$, or $\mathrm{Proj}_{\mathrm{W}}(H)$, estimate the latent hyperedge set $E(H)$.

Definition 3.18 (Exact recovery)**.** An estimator A achieves exact recovery for a random hypergraph H from its projection if

$$\mathbb{P}\big(A(\mathrm{Proj}(H)) = H\big) = 1 - o(1).$$

The weighted version replaces $\mathrm{Proj}(H)$ by $\mathrm{Proj}_{\mathrm{W}}(H)$.

Definition 3.19 (Partial recovery loss)**.** For an estimator $\widehat{H}$ of a random d-uniform hypergraph H, define the normalized symmetric-difference loss

$$\ell(\widehat{H}, H) = \frac{\mathbb{E}\big[|E(\widehat{H}) \triangle E(H)|\big]}{\mathbb{E}\big[|E(H)|\big]}.$$

Partial recovery with vanishing loss means

$$\ell(\widehat{H}, H) = o(1).$$

For heterogeneous hypergraphs, when the target degree is d_r, the analogous criterion is

$$\mathbb{E}\big[|E(\widehat{H}_r) \triangle E(H_r)|\big] = o(1)\, \mathbb{E}\big[|E(H_r)|\big].$$

3.6 Equivalence Metrics

3.6.1 Statistical equivalence

Definition 3.20 (Statistical equivalence)**.** Two sequences of probability measures P_n, Q_n on the same observation space are statistically equivalent if

$$d_{\mathrm{TV}}(P_n, Q_n) = o(1).$$

Total variation equivalence is the strongest distributional comparison used in this thesis. It implies that for every event A_n,

$$|P_n(A_n) - Q_n(A_n)| = o(1).$$

Consequently, any test, estimator, or statistic has asymptotically the same performance under P_n and Q_n, up to an $o(1)$ error term.

3.6.2 Average-case reductions

Definition 3.21 (Average-case reduction in total variation)**.** Let

$$P = (P_{0,n}, P_{1,n}) \quad \text{and} \quad Q = (Q_{0,n}, Q_{1,n})$$

be two sequences of hypothesis testing problems. A randomized polynomial-time map A_n is an average-case reduction from P to Q if

$$d_{\mathrm{TV}}(A_n(P_{0,n}), Q_{0,n}) = o(1), \qquad d_{\mathrm{TV}}(A_n(P_{1,n}), Q_{1,n}) = o(1).$$

For recovery problems, the analogous requirement is imposed on the joint law of the observation and the latent signal, or on conditional laws after fixing the planted object.

Such a reduction transfers algorithms from the target problem to the source problem. If the target problem has a polynomial-time algorithm, then the source problem has a polynomial-time algorithm obtained by composing the reduction with that target algorithm.

Definition 3.22 (Computational equivalence)**.** Two parameterized inference problems are computationally equivalent in a given parameter regime if there are randomized polynomial-time average-case reductions in both directions, with parameter maps that preserve the relevant signal strength up to $o(1)$ perturbations.

The distinction between statistical and computational equivalence is central. Statistical equivalence compares observation laws directly. Computational equivalence compares inference tasks through efficient transformations that must preserve both the ambient noise distribution and the planted or latent signal.

Chapter 4

Local Dependencies as a Bridge Between Independent and Global Models

The preceding chapter fixed the ambient models, recovery tasks, equivalence metrics, and the local-kernel definition of bounded-degree graph dependence. We now use those definitions to explain the main structural role of local dependencies in the thesis. The point is not merely that local motifs create dependence. Rather, local kernels occupy an intermediate scale: they are rich enough to approximate latent and geometric graph laws, yet explicit enough to support sharp total variation comparisons.

The refinement in Definition 3.8 is essential in this chapter. A local dependency is not required to be a monotone insertion of a fixed motif. At a local update, the kernel may inspect the current induced motif and resample it according to a distribution depending on that motif. This distinction is what allows the framework to include both clique insertions and the signed triangle-relation chain that appears in the middle-scale approximation to high-dimensional random geometric graphs.

4.1 Local Kernels as Controlled Perturbations

We use the following terminology throughout the chapter.

Definition 4.1 (Motif-local perturbation)**.** Fix a representation of a graph or signed graph G by the sequential local-kernel process of Definition 3.8. A motif-local perturbation is one active update in this representation: at step r, the process chooses a footprint V_r, with $|V_r| \leq d$, reads the current local motif $G^{(r-1)}[V_r]$, samples a new local motif from

$$K_r\big(G^{(r-1)}[V_r], \cdot\big),$$

and leaves all other edges unchanged. The kernel K_r may depend on the current local motif. A monotone motif insertion is the special case in which K_r deterministically adds a fixed edge shadow.

This terminology emphasizes the object that changes from model to model: the local kernel. In RGT, the kernel adds a triangle. In a hypergraph projection or random intersection graph, the kernel adds a clique. In the signed triangle-relation chain, the kernel reads two lower signs and replaces the top sign by their product. Thus local dependence should be viewed as a bounded-footprint perturbation mechanism rather than as a list of fixed subgraphs.

The results in this chapter are all total variation comparisons. The chapter has two complementary goals. First, it identifies when explicit bounded-size clique perturbations are washed out by an Erdős–Rényi baseline, and when they remain statistically visible. Second, it shows that latent hidden-variable graph laws can collapse, in suitable asymptotic regimes, to explicit local-kernel laws. The two examples of the second type are random intersection graphs, where the hidden variables are latent features, and random geometric graphs, where the hidden variables are geometric positions.

4.2 From Independent Graphs to Explicit Local Dependence

The first bridge is from independent graphs to explicit local-kernel perturbations. The random graph with triangles is the simplest example, but the same total variation picture holds for arbitrary fixed-size clique insertions. This motivates the following model.

Definition 4.2 (Random graph with bounded-size cliques)**.** Fix an integer $L \geq 2$ and parameters

$$\boldsymbol{\pi}_n = (\pi_{3,n}, \ldots, \pi_{L,n}), \qquad 0 \leq \pi_{s,n} \leq 1.$$

The random graph with cliques

$$\mathrm{RGC}_{\leq L}(n, p, \boldsymbol{\pi}_n)$$

is generated as follows. First sample $G_0 \sim \mathrm{G}(n,p)$. Then, independently for every $s \in \{3, \ldots, L\}$ and every $S \in \binom{[n]}{s}$, with probability $\pi_{s,n}$ add the full clique on S. When $L = 2$, the vector $\boldsymbol{\pi}_n$ is empty and the model is simply $\mathrm{G}(n,p)$.

The marginal edge density of $\mathrm{RGC}_{\leq L}(n, p, \boldsymbol{\pi}_n)$ is

$$q_n = 1 - (1-p)\prod_{s=3}^{L}(1-\pi_{s,n})^{\binom{n-2}{s-2}}.$$

Let

$$M_n = \sum_{s=3}^{L}\binom{n}{s}\pi_{s,n}$$

denote the expected number of inserted cliques, up to constant-factor errors in the sparse regime. The next theorem generalizes the triangle perturbation threshold to all fixed clique sizes.

Theorem 4.3 (Constant-size clique perturbations: independent or locally visible)**.** *Fix $L \geq 3$ and $p \in (0,1)$. Assume*

$$\sum_{s=3}^{L} n^{s-2}\pi_{s,n} = O(1).$$

Let q_n be the marginal edge density defined above. Then:

$$d_{\mathrm{TV}}\Big(\mathrm{RGC}_{\le L}(n,p,\boldsymbol{\pi}_n),\mathrm{G}(n,q_n)\Big)=o(1)\qquad \textit{if } M_n\ll n^{3/2},$$

while

$$d_{\mathrm{TV}}\Big(\mathrm{RGC}_{\le L}(n,p,\boldsymbol{\pi}_n),\mathrm{G}(n,q_n)\Big)=1-o(1)\qquad \textit{if } M_n\gg n^{3/2}.$$

Equivalently, the transition occurs at

$$\sum_{s=3}^{L} n^s\pi_{s,n}\asymp n^{3/2}.$$

In the second regime, the separating statistic is the centered signed triangle statistic

$$T(G)=\sum_{\{i,j,k\}\in\binom{[n]}{3}}(A_{ij}-q_n)(A_{ik}-q_n)(A_{jk}-q_n),$$

where A is the adjacency matrix of G.

For a single clique size $r\ge 3$, Theorem 4.3 says that, under $n^{r-2}\pi_n=O(1)$,

$$d_{\mathrm{TV}}\left(\mathrm{RGC}_r(n,p,\pi_n),\mathrm{G}\left(n,1-(1-p)(1-\pi_n)^{\binom{n-2}{r-2}}\right)\right)=\begin{cases}o(1), & \pi_n\ll n^{-r+3/2},\\ 1-o(1), & \pi_n\gg n^{-r+3/2}.\end{cases}$$

The case $r=3$ recovers the random graph with triangles threshold $p'\asymp n^{-3/2}$ from Appendix A.1 of [19].

The reason the same centered triangle statistic detects all fixed-size clique insertions is that every inserted s-clique contains $\binom{s}{3}$ triangles. Thus the first non-product footprint of a bounded clique perturbation is already visible at Fourier degree three. Below the $M_n\asymp n^{3/2}$ scale, the total number of such low-order correlations is too small to distinguish the model from an Erdős–Rényi graph with the same edge density. Above that scale, these local correlations aggregate into a statistically visible signed triangle count.

The proof of Theorem 4.3 is deferred to Appendix A. The upper bound follows the Poissonization and one-planted-constant-clique total variation argument used for the triangle case in [19]. The lower bound computes the mean and variance of the centered signed triangle statistic; the contribution of a fixed clique layer enters through the probability that a given triple is contained in at least one inserted clique.

4.3 Approximating Random Intersection Graphs with Local Dependence

The next two sections both approximate hidden-variable graph laws by explicit local-dependence models. In the present section, the hidden variables are latent features in a random intersection graph. In the following section, the hidden variables are geometric positions in a random geometric graph. In both cases, the conclusion is a total variation approximation by a bounded-footprint local-kernel model.

A random intersection graph is generated by latent features: each feature induces a clique on the vertices carrying it. In the sparse feature regime, most features have small realized size. Size-2 features contribute independent-looking edges, size-3 features contribute triangles, size-4 features contribute 4-cliques, and so on. If sufficiently many clique sizes are retained explicitly, the remaining tail of larger features can be absorbed into an Erdős–Rényi base layer in total variation.

Theorem 4.4 (Random intersection graphs reduce to bounded clique kernels)**.** *Fix $L \geq 2$. Let $G \sim \mathrm{RIG}(n, m, \alpha)$, assume constant edge density*

$$m\alpha^2 = \Theta(1),$$

and suppose

$$m \gg n^{2+\frac{1}{L-1}}.$$

For $s \geq 2$, define

$$\lambda_s = m\alpha^s(1-\alpha)^{n-s}.$$

For $3 \leq s \leq L$, set

$$\pi_{s,n} = 1 - e^{-\lambda_s}.$$

Finally define

$$p_L = 1 - \exp\left(-m\alpha^2 + \sum_{s=3}^{L} \binom{n-2}{s-2} \lambda_s\right),$$

where the sum is empty when $L = 2$. Then

$$d_{\mathrm{TV}}\Big(\mathrm{RIG}(n, m, \alpha), \mathrm{RGC}_{\leq L}(n, p_L, (\pi_{3,n}, \ldots, \pi_{L,n}))\Big) = o(1).$$

The parameter p_L is chosen so that the total missing-edge probability matches the random intersection graph:

$$(1 - p_L) \prod_{s=3}^{L} (1 - \pi_{s,n})^{\binom{n-2}{s-2}} = \exp(-m\alpha^2).$$

Thus the explicit clique layers account for all feature-induced cliques of sizes $3, \ldots, L$, while the base Erdős–Rényi layer absorbs size-2 features together with the edge-density contribution of the omitted tail.

The case $L = 2$ gives an Erdős–Rényi approximation to the random intersection graph when $m \gg n^3$. The case $L = 3$ gives the random graph with triangles approximation

$$\mathrm{RIG}(n, m, \alpha) \approx_{d_{\mathrm{TV}}} \mathrm{RGT}\left(n, 1 - \exp\big(-m\alpha^2 + (n-2)m\alpha^3(1-\alpha)^{n-3}\big), 1 - \exp\big(-m\alpha^3(1-\alpha)^{n-3}\big)\right)$$

when $m \gg n^{5/2}$, matching the original RIG-to-RGT window from [19]. Keeping larger clique sizes improves the approximation range: for example $L = 4$ gives an Erdős–Rényi plus triangle plus 4-clique local model when $m \gg n^{7/3}$.

Conceptually, Theorem 4.4 says that the latent feature model has a local normal form. The latent features are not directly observed, but in the relevant sparse-feature regimes their observable effect is exhausted, up to $o(1)$ total variation error, by an explicit finite list of local clique kernels.

4.4 Approximating Random Geometric Graphs with Local Dependence

We now turn from latent features to latent geometry. The balanced spherical random geometric graph is globally constrained by an underlying point configuration. In very high dimension, this global dependence disappears and the model becomes close to $\mathrm{G}(n, 1/2)$. In the middle-scale window, however, the first correction below independence is not an arbitrary global constraint: it is a local signed triangle relation.

Let $G_{\text{geo}}(n, D) \in \{-1, +1\}^{\binom{[n]}{2}}$ denote the signed balanced spherical random geometric graph:

$$(G_{\text{geo}})_{ij} = \operatorname{sign}\langle Z_i, Z_j \rangle, \qquad Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} N(0, I_D).$$

Let $\mathsf{LTSC}(n, \lambda)$ denote the ordered signed local triangle-relation chain. The chain starts from independent balanced signs and processes edges in lexicographic order. When processing edge (i, j), it chooses

$$L \in \{0, 1, \dots, i-1\},$$

with

$$\mathbb{P}(L = \ell) = \lambda \quad (1 \le \ell < i), \qquad \mathbb{P}(L = 0) = 1 - \lambda(i-1).$$

If $L = 0$, the chain keeps a fresh independent sign; if $L = \ell \ge 1$, it sets

$$Y_{ij} = Y_{i\ell} Y_{j\ell}.$$

Thus a local update enforces the signed triangle relation

$$Y_{i\ell} Y_{j\ell} Y_{ij} = 1.$$

Theorem 4.5 (Middle-scale geometry is locally generated)**.** *There is an explicit update rate*

$$\lambda_D \asymp D^{-1/2}$$

such that, whenever

$$D \gg n^2 \log^{3/2} n,$$

one has

$$d_{\text{TV}}\Big(\mathcal{L}\big(G_{\text{geo}}(n, D)\big), \mathsf{LTSC}(n, \lambda_D)\Big) = o(1).$$

Theorem 4.5 is the main theorem of [20]. Its proof proceeds through the comparison chain

$$\mathcal{L}\big(G_{\text{geo}}(n, D)\big) \approx G_1 \approx G_2 \approx G_3 \approx G_4 \approx G_{5,\text{reg}} \approx G_6,$$

where G_6 is the local triangle-sampling chain. The argument combines an exact Gram–Schmidt coupling, a fragile-edge reduction, replacement of Gaussian weights by uniform and then unweighted sampling, and a polymer expansion for the final weighted-to-unweighted comparison.

The Fourier intuition is as follows. For a balanced sign array $x \in \{\pm 1\}^{\binom{[n]}{2}}$, write

$$\chi_F(x) = \prod_{e \in F} x_e$$

for a Boolean Fourier monomial. The independent sign model has zero Fourier coefficient on every nonempty F. In the signed random geometric graph, the first local coefficient that survives below the $D \asymp n^3$ Erdős–Rényi testing scale is the degree-3 triangle coefficient

$$\mathbb{E}\big[(G_{\text{geo}})_{ij}(G_{\text{geo}})_{ik}(G_{\text{geo}})_{jk}\big],$$

which is of order $D^{-1/2}$. The local triangle-relation update is designed precisely to match this coefficient: when Y_{ij} is replaced by $Y_{i\ell}Y_{j\ell}$, the triangle product $Y_{i\ell}Y_{j\ell}Y_{ij}$ becomes 1, while one-edge marginals remain balanced. Choosing $\lambda_D \asymp D^{-1/2}$ matches the degree-3 Fourier correction of the random geometric graph to first order. The content of Theorem 4.5 is that, once this signed triangle coefficient is matched, the remaining higher-order geometric constraints are negligible in total variation throughout the regime $D \gg n^2 \log^{3/2} n$.

This is the clearest example in the thesis of a latent geometric graph being approximated by explicit local dependence. The original model is defined through all latent positions simultaneously, but its middle-scale law is asymptotically reproduced by a bounded-footprint signed local kernel.

Corollary 4.6 (Three geometric regimes)**.** *For the balanced signed random geometric graph, the following qualitative picture holds.*

1. *In the independent regime, above the $D \asymp n^3$ testing scale, the law is asymptotically indistinguishable from independent balanced edge signs.*

2. *In the middle-scale regime*
$$n^2 \log^{3/2} n \ll D \ll n^3,$$
*the law is not independent, but its first surviving dependence is captured by a degree-*3 *local signed triangle-relation chain.*

3. *Below this middle-scale window, higher-order geometric constraints may survive, and the degree-*3 *local approximation no longer describes the full dependence structure.*

This corollary is a synthesis of the high-dimensional random geometric graph testing threshold and Theorem 4.5; see [17, 20, 59]. It should be read as a dependency hierarchy, not as a complete classification of all lower-dimensional regimes.

Conjecture 4.7 (Higher-order local Fourier approximation of random geometric graphs)**.** *For every fixed $r \geq 4$, there exists an explicit signed local-kernel model*

$$\mathsf{LK}^{(\leq r)}(n, D)$$

constructed from independent balanced signs by bounded-footprint updates on at most r vertices, such that its local Fourier coefficients

$$\mathbb{E}\prod_{e\in F} Y_e$$

match the corresponding coefficients of $G_{\text{geo}}(n, D)$, to the first non-vanishing order, for every edge set F supported on at most r vertices. Moreover, after matching these higher-order local Fourier coefficients, one should obtain

$$d_{\text{TV}}\Big(\mathcal{L}\big(G_{\text{geo}}(n, D)\big), \mathsf{LK}^{(\leq r)}(n, D)\Big) = o(1)$$

in dimension regimes strictly below the present $D \gg n^2 \operatorname{polylog} n$ window.

Equivalently, the local triangle-relation chain should be the first member of a hierarchy: matching the degree-3 Fourier coefficient by adding signed triangle relations gives the middle-scale approximation in Theorem 4.5; matching higher-order Fourier coefficients by adding higher-order local signed kernels should approximate the same latent geometric graph in smaller dimensions.

4.5 The Total Variation Bridge Principle

The results above support the following organizing principle.

Principle 4.8 (Total-variation local-kernel bridge). A locally dependent graph model can play two distinct total variation roles.

First, it can be an explicit perturbation of an independent model. In this role, one asks whether the perturbation is statistically negligible or statistically visible after matching the edge density.

Second, it can be an asymptotic normal form for a latent hidden-variable model. In this role, one proves that a graph generated by hidden features or latent geometry is close in total variation to a graph generated by explicit bounded-footprint local kernels.

The first role appears in Theorem 4.3: bounded-size clique perturbations are asymptotically invisible below the $M_n \asymp n^{3/2}$ scale and visible above it through a local signed triangle statistic. The second role appears in Theorems 4.4 and 4.5: random intersection graphs and balanced random geometric graphs, although generated from latent variables, admit total variation approximations by explicit local-dependence models in suitable regimes.

This bridge principle clarifies the thesis arc. The independent model is not discarded; it remains the baseline. Local kernels quantify the first ways in which independence can fail. Latent feature and latent geometric models, in appropriate asymptotic regimes, may collapse onto these local kernels. The next chapters exploit this structure in two directions: global planted recovery in locally dependent ambient graphs and local recovery of latent motifs from graph projections.

Chapter 5

Global Recovery via Algorithmic Decorrelation

5.1 Overview

The preceding chapters introduced a hierarchy of random graph models, ranging from the independent baseline $\mathrm{G}(n,p)$ to locally dependent models such as the random graph with triangles $\mathrm{RGT}(n,p,p')$. This chapter studies a global planted-signal problem in such a locally dependent ambient graph. The signal is a planted dense subgraph, with planted clique as the motivating endpoint in which all planted edges are forced to be present.

In the independent ambient model, the planted dense subgraph problem compares

$$H_0 :\ G \sim \mathrm{G}(n,p), \qquad H_1 :\ G \sim \mathrm{PDS}(n,p,k,q).$$

The dependent analogue replaces the independent ambient noise by $\mathrm{RGT}(n,p,p')$:

$$H_0 :\ G \sim \mathrm{RGT}(n,p,p'), \qquad H_1 :\ G \sim \mathrm{RGT}(n,p,p',k,q).$$

Here p' controls the strength of the local triangle dependence. Even when p' is small enough that the edge density remains bounded away from one, the graph contains many correlated local motifs. These motifs can produce spurious dense local structure, so an algorithm designed for product-measure noise need not apply directly.

The goal of algorithmic decorrelation is to remove the local dependence while preserving the global planted signal. The main result in this chapter is that, in the sparse-triangle regime

$$p' \le \frac{1}{n \log n},$$

planted dense subgraph in RGT is computationally equivalent to planted dense subgraph in G, up to $o(1)$ total variation error, for planted sizes

$$k = o\Big(n^{1/4} \log^{-17/4} n\Big).$$

The forward reduction adds triangle dependencies. The reverse reduction removes triangle dependencies by sampling from the posterior distribution of the triangles that may have generated the observed graph, and then resampling the graph edges explained by those triangles.

The proof is based on a perturbation-insensitivity principle. If a randomized map sends one null distribution to another and is insensitive to changing a single input edge, then the same map sends the corresponding planted distributions to one another, with an error proportional to the number of planted edges. For planted dense subgraph, the number of planted edges is $\binom{k}{2}$, which explains why the single-edge perturbation error must be $o(k^{-2})$.

The formal results in this chapter follow [19]. They complement the local recovery results of Chapter 6: here the local motifs are nuisance dependence to be decorrelated, whereas in hypergraph projection problems the local motifs are the latent objects to be recovered.

5.2 Planted Dense Subgraph and Planted Clique in RGT

We use the notation of Chapter 3. For a fixed set $S \subset [n]$ with $|S| = k$, let

$$\mathrm{PDS}(n, p, S, q)$$

denote the distribution obtained by sampling $G \sim \mathrm{G}(n, p)$, erasing all edges in $\binom{S}{2}$, and then resampling every edge in $\binom{S}{2}$ independently from $\mathrm{Ber}(q)$. The distribution $\mathrm{PDS}(n, p, k, q)$ is the mixture over a uniformly random $S \in \binom{[n]}{k}$.

Definition 5.1 (Planted RGT model)**.** For a fixed set $S \subset [n]$ with $|S| = k$, the planted RGT distribution

$$\mathrm{RGT}(n, p, p', S, q)$$

is obtained by sampling an ambient graph $G \sim \mathrm{RGT}(n, p, p')$, erasing all edges in $\binom{S}{2}$, and then resampling every edge in $\binom{S}{2}$ independently from $\mathrm{Ber}(q)$. The distribution $\mathrm{RGT}(n, p, p', k, q)$ is the mixture over a uniformly random planted set $S \in \binom{[n]}{k}$.

The planted clique model corresponds formally to the endpoint $q = 1$, where the planted subgraph is forced to be complete. The total-variation equivalence theorem below is stated for the planted dense subgraph regime $0 < p < q < 1$, with p and q fixed constants, which is the parameter range proved in the source result. We therefore use planted clique primarily as the motivating global-recovery problem and as the standard hardness benchmark, while the formal statements in this chapter are for planted dense subgraph unless explicitly noted otherwise.

5.3 A General Perturbation Framework

We first state the abstract theorem that underlies both reductions. For an edge $e \in \binom{[n]}{2}$, write

$$G + e = G \cup \{e\}, \qquad G - e = G \setminus \{e\}.$$

For a random graph H, write $H_{\sim e}$ for the edge vector with coordinate e removed. If $\mathcal{A}$ is a randomized map from graphs to graphs, then $\mathcal{A}(G)_{\sim e}$ denotes the output with edge e omitted.

Let Res_e^r be the operation that resamples edge e independently from $\mathrm{Ber}(r)$, leaving every other edge unchanged. For an edge sequence

$$\mathbf{e} = (e_1, \dots, e_K)$$

and probabilities

$$\mathbf{r} = (r_1, \dots, r_K),$$

define

$$\mathrm{GPS}_{\mathbf{e},\mathbf{r}} = \mathrm{Res}_{e_K}^{r_K} \circ \cdots \circ \mathrm{Res}_{e_1}^{r_1}.$$

For planted dense subgraph on a fixed S, $\mathbf{e}$ is an ordering of $\binom{S}{2}$ and $r_i = q$ for every i.

Theorem 5.2 (Perturbation-insensitivity criterion)**.** *Let P_0 and P_0' be two distributions on graphs with vertex set $[n]$. Let $\mathcal{A}$ be a randomized map such that*

$$\mathcal{A}(P_0) = P_0'.$$

For $0 \le i \le K$, define

$$P_{1,i} = \mathrm{Res}_{e_i}^{r_i} \circ \cdots \circ \mathrm{Res}_{e_1}^{r_1}(P_0), \qquad P_{1,0} = P_0.$$

Suppose there is a set of graphs $\mathcal{G}$ such that

$$\mathbb{P}_{G \sim P_{1,i}}(G \in \mathcal{G}) \ge 1 - \delta \qquad \text{for every } 0 \le i \le K.$$

Assume that for every $G \in \mathcal{G}$ and every edge $e = e_i$ in the planted sequence, there exist $p_e^-, p_e^+ \in [0,1]$ such that the following conditions hold.

First,

$$d_{\mathrm{TV}}(\mathcal{A}(G-e)_{\sim e}, \mathcal{A}(G+e)_{\sim e}) \le \varepsilon. \tag{C1}$$

Second,

$$d_{\mathrm{TV}}(\mathcal{A}(G-e), \mathcal{A}(G-e)_{\sim e} \times \mathcal{A}(G-e)_e) \le \varepsilon, \tag{C2-}$$

and

$$d_{\mathrm{TV}}(\mathcal{A}(G+e), \mathcal{A}(G+e)_{\sim e} \times \mathcal{A}(G+e)_e) \le \varepsilon. \tag{C2+}$$

Third,

$$\left|\mathbb{P}(e \in \mathcal{A}(G-e)) - p_e^-\right| \le \varepsilon, \qquad \left|\mathbb{P}(e \in \mathcal{A}(G+e)) - p_e^+\right| \le \varepsilon. \tag{C3}$$

Define

$$s_i = r_i p_{e_i}^+ + (1-r_i) p_{e_i}^-, \qquad \mathbf{s} = (s_1, \dots, s_K).$$

Then

$$d_{\mathrm{TV}}\big(\mathcal{A}\big(\,\mathrm{GPS}_{\mathbf{e},\mathbf{r}}(P_0)\big), \mathrm{GPS}_{\mathbf{e},\mathbf{s}}(P_0')\big) = O\big(K(\varepsilon + \delta)\big).$$

The theorem says that it is enough to understand how $\mathcal{A}$ reacts to a single edge perturbation. Conditions (C1)–(C3) imply that applying $\mathcal{A}$ and resampling the next planted edge approximately commute. Iterating over the K planted edges gives the stated total variation bound.

Algorithm 1 Forward triangle-addition map

Require: Graph G, triangle probability p'

for all $T = \{i, j, k\} \in \binom{[n]}{3}$ **do**

  With probability p', add the three edges in $\binom{T}{2}$ to G

end for

return G

5.4 Forward Map: Adding Local Dependencies

The forward map adds local triangle dependencies. Starting from an independent graph, we independently add every triangle with probability p'. This maps $\mathrm{G}(n, p)$ exactly to $\mathrm{RGT}(n, p, p')$.

Let $\mathcal{A}_{\triangle}^{p'}$ denote Algorithm 1. Then

$$\mathcal{A}_{\triangle}^{p'}(\mathrm{G}(n, p)) = \mathrm{RGT}(n, p, p')$$

as an exact equality in distribution.

For a planted edge inside S, triangle addition may force the edge to be present. Hence a planted edge originally distributed as $\mathrm{Ber}(q)$ has transformed marginal

$$f_{p'}(q) = 1 - (1 - q)(1 - p')^{n-2} = q + (1 - q)\bigl(1 - (1 - p')^{n-2}\bigr). \tag{5.1}$$

When $p' \le 1/(n \log n)$, this satisfies

$$f_{p'}(q) = q + O(np') = q + o(1).$$

The nontrivial point is that the forward map also preserves the planted model, even though the added triangles can correlate planted edges with one another and with edges outside S. Applying Theorem 5.2 gives, uniformly over fixed S in the parameter regime below,

$$d_{\mathrm{TV}}\Bigl(\mathcal{A}_{\triangle}^{p'}\bigl(\mathrm{PDS}(n, p, S, q)\bigr), \mathrm{RGT}(n, p, p', S, f_{p'}(q))\Bigr) = o(1).$$

Since the bound is uniform in S, the same statement holds after mixing over a uniformly random planted set.

5.5 Reverse Map: Removing Local Dependencies

The reverse map is the decorrelation step. It views triangle addition as a Markov transition kernel and implements its time reversal under the RGT null.

Let $X \in \{0, 1\}^{\binom{[n]}{3}}$ denote a set of sampled triangles. Write

$$|X| = \sum_{T \in \binom{[n]}{3}} X_T, \qquad E(X) = \bigcup_{T : X_T = 1} \binom{T}{2}, \qquad e(X) = |E(X)|.$$

Algorithm 2 Reverse decorrelation map

Require: Graph G, parameters p, p'

Sample a triangle set X from μ_G
Let $E(X)$ be the set of graph edges covered by sampled triangles
for all $e \in E(X)$ **do**
    Resample e independently from $\mathrm{Ber}(p)$
end for
return the resulting graph

Definition 5.3 (Posterior triangle distribution)**.** For a graph G, define the triangle distribution μ_G on $\{0,1\}^{\binom{[n]}{3}}$ by

$$\mu_G(x) = \frac{1}{Z_G} \left(\frac{p'}{1-p'} \right)^{|x|} p^{-e(x)} \mathbf{1}\{E(x) \subseteq E(G)\},$$

where

$$Z_G = \sum_{x : E(x) \subseteq E(G)} \left(\frac{p'}{1-p'} \right)^{|x|} p^{-e(x)}$$

is the normalizing constant.

This is the posterior law of the triangle set conditional on observing G after the forward triangle-addition map, when the preimage graph is distributed as $\mathrm{G}(n,p)$.

Let $\mathcal{A}_{\mathrm{rev}}^{p,p'}$ denote Algorithm 2 with exact sampling from μ_G. Under the null,

$$\mathcal{A}_{\mathrm{rev}}^{p,p'}(\mathrm{RGT}(n,p,p')) = \mathrm{G}(n,p)$$

as an exact equality in distribution. The algorithmic implementation samples approximately from μ_G using Glauber dynamics. For $p' \ll 1/n$ and fixed p, the Glauber chain on the triangle variables mixes in $O(n^3 \log n)$ steps, so the reverse map can be implemented in randomized polynomial time with an additional $o(1)$ total variation error.

The reverse perturbation analysis is carried out at

$$p'_\star = \frac{1}{n \log n}.$$

For $p' < p'_\star$, the decorrelation map first adds triangles with probability

$$p'_\Delta = \frac{p'_\star - p'}{1-p'} \tag{5.2}$$

so that the combined triangle probability is $p'_\star$, and then applies $\mathcal{A}_{\mathrm{rev}}^{p,p'_\star}$. This yields a polynomial-time reverse reduction for every $p' \le p'_\star$.

5.6 Equivalence of Ambient and Planted Models

We now state the main equivalence theorem in the notation of this thesis. The theorem is stated conditionally on a fixed planted set S, and therefore also holds after mixing over a uniformly random planted set. Because the maps do not use S, the same total-variation conclusions apply to the joint law of (S, G), which is the form needed for recovery.

Theorem 5.4 (RGT–ER equivalence for planted dense subgraph)**.** *Let $0 < p < q < 1$ be fixed constants. Let*

$$p' \le \frac{1}{n \log n}, \qquad k = o\Big(n^{1/4} \log^{-17/4} n\Big).$$

Then there are randomized polynomial-time maps

$$\mathcal{A}_{\triangle}^{p'} \qquad \textit{and} \qquad \mathcal{A}_{\mathrm{dec}}^{p,p'}$$

with the following properties.

First, the forward triangle-addition map satisfies

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\triangle}^{p'}(\mathrm{G}(n,p)), \mathrm{RGT}(n,p,p')\Big) = 0,$$

and, for every fixed $S \subset [n]$ with $|S| = k$,

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\triangle}^{p'}\big(\,\mathrm{PDS}(n,p,S,q)\big), \mathrm{RGT}(n,p,p',S,f_{p'}(q))\Big) = o(1),$$

where

$$f_{p'}(q) = q + (1-q)\big(1 - (1-p')^{n-2}\big) = q + o(1).$$

Second, the reverse decorrelation map satisfies

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\mathrm{dec}}^{p,p'}(\mathrm{RGT}(n,p,p')), \mathrm{G}(n,p)\Big) = o(1),$$

and, for every fixed $S \subset [n]$ with $|S| = k$,

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\mathrm{dec}}^{p,p'}\big(\,\mathrm{RGT}(n,p,p',S,q)\big), \mathrm{PDS}(n,p,S,g_{p'}(q))\Big) = o(1),$$

where $g_{p'}(q) = q + o(1)$.

More explicitly, when $p' = p'_\star = 1/(n \log n)$,

$$g_{p'_\star}(q) = q\, p_e,$$

where

$$p_e = \mathbb{E}_{G \sim \mathrm{RGT}(n,p,p'_\star)} \mathbb{P}\big(e \in \mathcal{A}_{\mathrm{rev}}^{p,p'_\star}(G+e)\big) = 1 - O(np'_\star) = 1 + o(1).$$

When $p' < p'_\star$, the reverse map first raises the triangle density from p' to $p'_\star$ using (5.2), and then applies $\mathcal{A}_{\mathrm{rev}}^{p,p'_\star}$. In that case

$$g_{p'}(q) = \left[q + (1-q)\left(1 - \left(1 - \frac{p'_\star - p'}{1-p'}\right)^{n-2}\right)\right] p_e = q + o(1).$$

Consequently,

$$f_{p'}(g_{p'}(q)) = q + o(1), \qquad g_{p'}(f_{p'}(q)) = q + o(1).$$

The same conclusions hold after mixing over a uniformly random planted set S, giving reductions between $\mathrm{PDS}(n,p,k,q)$ and $\mathrm{RGT}(n,p,p',k,q)$ with the corresponding parameter maps.

5.7 Consequences for Recovery and Hardness

Theorem 5.4 transfers algorithms in both directions.

First, any polynomial-time recovery algorithm for the independent model transfers to the dependent model. Given

$$G \sim \mathrm{RGT}(n, p, p', S, q),$$

apply $\mathcal{A}_{\mathrm{dec}}^{p,p'}$ and then run the algorithm designed for

$$\mathrm{PDS}(n, p, S, g_{p'}(q)).$$

Since the joint law of $(S, \mathcal{A}_{\mathrm{dec}}^{p,p'}(G))$ is $o(1)$-close to the independent planted model, the recovery success probability changes by at most $o(1)$.

Second, any polynomial-time recovery algorithm for the dependent model transfers back to the independent model. Given

$$G \sim \mathrm{PDS}(n, p, S, q),$$

apply $\mathcal{A}_{\triangle}^{p'}$ and then run the algorithm designed for

$$\mathrm{RGT}(n, p, p', S, f_{p'}(q)).$$

Again, the success probability changes by at most $o(1)$.

Thus, in the proved regime

$$p' \le \frac{1}{n \log n}, \qquad k = o\left(n^{1/4} \log^{-17/4} n\right),$$

local triangle dependence does not define a new polynomial-time recovery problem distinct from the independent baseline. It can be added or removed efficiently while preserving the global planted signal.

The theorem also transfers conditional hardness statements. For instance, a planted dense subgraph hardness conjecture in $\mathrm{G}(n, p)$ transfers to the corresponding RGT problem, because an efficient algorithm for the dependent problem would compose with the forward map to solve the independent problem. Conversely, any conjectured hardness for the dependent version transfers back through the reverse decorrelation map.

This chapter does not prove an unconditional computational lower bound. It proves a distribution-preserving equivalence. Hardness conclusions are therefore conditional on the baseline planted clique or planted dense subgraph conjecture, and the quantitative range is exactly the range in Theorem 5.4. The source work conjectures that the equivalence should extend to planted sizes up to the natural $\widetilde{O}(\sqrt{n})$ scale, but the theorem proved here establishes the $o(n^{1/4} \log^{-17/4} n)$ regime.

5.8 Technical Overview

5.8.1 Sensitivity to Single-Edge Perturbations

The main technical object is the posterior triangle distribution μ_G. If $X \sim \mu_G$, define

$$Y_e = \mathbf{1}\{e \in E(X)\}, \qquad e \in \binom{[n]}{2}.$$

The vector Y records which graph edges are explained by the sampled posterior triangles.

The key perturbation estimate compares posterior edge-coverage laws under $G - e$ and $G + e$. For graphs in a high-probability class $\mathcal{G}_1$, one proves

$$\mathcal{L}_{G-e}(Y_{\sim e}) = \mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 0), \tag{5.3}$$

and

$$d_{\mathrm{TV}}(\mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 0), \mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 1)) = O\left(\frac{\log^{17/2} n}{\sqrt{n}}\right). \tag{5.4}$$

The identity (5.3) is a conditioning identity: removing edge e is equivalent to conditioning that no posterior triangle covers e. The difficult part is (5.4): conditioning on $Y_e = 1$ has only a small effect on the posterior edge coverage away from e.

This estimate implies the perturbation conditions for the reverse map. Since the reverse map only resamples edges in $E(X)$, total variation cannot increase when passing from posterior edge coverage Y to the output graph.

5.8.2 Low Influences and Mixing

The good graph class is

$$\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2.$$

The first part, $\mathcal{G}_1$, is a uniform two-star density condition: every pair of vertices has at least a constant multiple of n common neighbors. This gives enough local redundancy to simulate the edge effect of triangles containing a perturbed edge by using auxiliary triangles not containing that edge.

The second part, $\mathcal{G}_2$, controls the posterior marginal $\mu_{G+e}(Y_e = 1)$ uniformly over all edges. This is needed for condition (C3) in the perturbation theorem.

The proof uses the fact that μ_G is marginally small. Under arbitrary conditioning, the probability of any fixed triangle variable being one is $O(p')$. The graphical model for μ_G has one variable per triangle, and two triangle variables interact only if the corresponding triangles share an edge. Its maximum degree is $O(n)$. Since $p' \ll 1/n$, marginal smallness gives low influences and rapid Glauber mixing.

The low-influence estimates show that conditioning on local edge-coverage events has limited effect away from those edges. The mixing estimate shows that sampling from μ_G can be implemented in randomized polynomial time.

5.8.3 Deferred Proofs

The full proof of Theorem 5.4 is given in Appendix B. The appendix proves the perturbation theorem, verifies the forward triangle-addition map, analyzes the posterior triangle distribution μ_G, proves the reverse single-edge perturbation estimate, proves the high-probability good graph conditions, and completes the polynomial-time implementation.

Chapter 6

Local Recovery from Hypergraph Projections

6.1 Overview

The preceding chapters studied global recovery problems in which a macroscopic planted structure is hidden in an ambient random graph. This chapter turns to a local recovery problem. The latent objects are many small higher-order interactions, represented by hyperedges, and the observation is only their pairwise shadow. A d-uniform hyperedge h appears in the graph projection as the clique $\binom{h}{2}$. The statistical question is whether the original hyperedges can be reconstructed from the union of these projected cliques.

This problem gives a complementary view of local dependencies. In global planted recovery, local dependencies in the ambient graph are nuisance structure that must be decorrelated. Here, the local dependencies are the signal itself: the observation is a graph whose edges are locally dependent because they arise from common hyperedges, and the task is to invert this projection.

The chapter contains three main results.

First, for d-uniform random hypergraphs, partial recovery has a sharp all-or-nothing threshold at

$$\delta_{\text{part}}(d) = \frac{d-1}{d+1}.$$

Below this threshold, the maximum clique cover estimator recovers all but an $o(1)$ fraction of hyperedges in expectation. Above it, even the Bayes-optimal estimator has normalized loss $1 - o(1)$. The same threshold holds for weighted projections.

Second, exact recovery from the unweighted projection has threshold

$$\delta_{\text{ex}}(d) = \begin{cases} \dfrac{2d-4}{2d-1}, & d = 3, 4, \\ \dfrac{d-1}{d+1}, & d \geq 5. \end{cases}$$

For $d = 3, 4$, exact recovery fails below the partial recovery threshold because of finite ambiguous projected graphs with two equally likely minimum preimages. For $d \geq 5$, the exact and partial thresholds coincide. For weighted projections, exact recovery has threshold $(d-1)/(d+1)$ for all $d \geq 3$.

Third, for heterogeneous random hypergraphs with several possible hyperedge sizes, maximal cliques of a target size d_j recover $1-o(1)$ of the degree-d_j hyperedges under an explicit density condition.

All technical proofs are deferred to Appendix C. The chapter gives the model definitions, the estimators, the theorem statements, and the proof mechanisms.

6.2 The d-Uniform Reconstruction Problem

6.2.1 Model and recovery criteria

Fix $d \geq 3$. Let

$$H \sim H_d(n,p), \qquad p = (c+o(1))n^{-d+1+\delta}, \qquad c \in (0,\infty),$$

and observe

$$\Psi = \mathrm{Proj}(H).$$

Thus each d-set $h \in \binom{[n]}{d}$ is included independently with probability p, and the projected graph contains the edge $\{u,v\}$ if and only if u and v are contained in at least one selected hyperedge. The weighted projection $\mathrm{Proj}_{\mathrm{W}}(H)$ records the number of hyperedges containing each graph edge.

Definition 6.1 (Exact recovery)**.** An estimator A achieves exact recovery from the unweighted projection if

$$\mathbb{P}\big(A(\mathrm{Proj}(H)) = H\big) = 1 - o(1).$$

The weighted version replaces $\mathrm{Proj}(H)$ by $\mathrm{Proj}_{\mathrm{W}}(H)$.

Definition 6.2 (Partial recovery loss)**.** For an estimator $\widehat{H}$, define the normalized symmetric-difference loss

$$\ell(\widehat{H}, H) = \frac{\mathbb{E}\big[|E(\widehat{H}) \triangle E(H)|\big]}{\mathbb{E}\big[|E(H)|\big]}.$$

Partial recovery with vanishing loss means

$$\ell(\widehat{H}, H) = o(1).$$

Since

$$\mathbb{E}[|E(H)|] = \binom{n}{d} p = \Big(\frac{c}{d!} + o(1)\Big)\, n^{1+\delta},$$

this normalization measures the expected fraction of true hyperedges that are missed or spuriously added.

Algorithm 3 Maximum clique cover estimator

Require: Projected graph G, uniformity d

$\widehat{E} \leftarrow \emptyset$
for all $S \in \binom{[n]}{d}$ **do**
  if $G[S]$ is a clique **then**
    Add S to $\widehat{E}$
  end if
end for
return $\widehat{H} = ([n], \widehat{E})$

6.2.2 Maximum clique cover

Every true hyperedge creates a d-clique in the projection. The most direct estimator therefore returns every d-clique of the projected graph.

Let
$$H_c = \left([n], \left\{S \in \binom{[n]}{d} : \Psi[S] \text{ is a clique}\right\}\right)$$
be the output of Algorithm 3. The estimator H_c has no false negatives, because
$$E(H) \subseteq E(H_c).$$

Its loss is controlled by fake d-cliques in the projection.

Theorem 6.3 (Partial recovery threshold for d-uniform projections)**.** *Let $H \sim H_d(n, p)$, where*
$$p = (c + o(1))n^{-d+1+\delta}, \qquad c \in (0, \infty).$$
For both $\mathrm{Proj}(H)$ *and* $\mathrm{Proj}_{\mathrm{W}}(H)$*, partial recovery has the all-or-nothing threshold*
$$\delta_{\mathrm{part}}(d) = \frac{d-1}{d+1}.$$
If
$$\delta < \frac{d-1}{d+1},$$
then Algorithm 3 achieves
$$\ell(H_c, H) = o(1).$$
If
$$\delta > \frac{d-1}{d+1},$$
then every estimator has loss $1 - o(1)$*; equivalently, the Bayes-optimal partial recovery loss is* $1 - o(1)$*.*

Proof idea. Let

$$q_n = \mathbb{P}\left(\binom{[d]}{2} \subseteq E(\mathrm{Proj}(H))\right)$$

be the density of projected d-cliques. A d-clique can arise either from a true hyperedge, with probability p, or by covering each of its $\binom{d}{2}$ graph edges using other hyperedges. The leading fake-clique contribution has order

$$\left(\frac{c\,n^{\delta-1}}{(d-2)!}\right)^{\binom{d}{2}}.$$

Thus

$$q_n = (1+o(1))\left[p + \left(\frac{c\,n^{\delta-1}}{(d-2)!}\right)^{\binom{d}{2}}\right],$$

and the fake-clique density is negligible compared with p exactly when

$$\delta < \frac{d-1}{d+1}.$$

This proves achievability for the maximum clique cover estimator.

For the converse, one compares two independent posterior samples

$$H, H' \mid \mathrm{Proj}(H)$$

and shows that, above the threshold,

$$\mathbb{E}[|E(H) \cap E(H')|] = o(1)\,\mathbb{E}[|E(H)|].$$

A posterior-overlap identity then implies that the Bayes-optimal normalized loss is $1 - o(1)$. The weighted-projection converse follows by the same overlap argument after truncating edge multiplicities. See Appendix C.2.

6.3 Exact Recovery

6.3.1 MAP estimator and minimum preimages

For exact recovery, the optimal estimator is the maximum a posteriori rule. Since $p = o(1)$, the posterior probability of a preimage decreases with its number of hyperedges. Thus MAP is equivalent to finding a minimum-cardinality preimage.

Definition 6.4 (Minimum preimage)**.** For a projected graph G, a d-uniform hypergraph H' is a minimum preimage if

$$\mathrm{Proj}(H') = G \qquad \text{and} \qquad |E(H')| = \min\{|E(K)| : \mathrm{Proj}(K) = G\}.$$

For weighted projections, Proj is replaced by $\mathrm{Proj}_{\mathrm{W}}$.

Algorithm 4 MAP estimator via minimum preimages

Require: Projected graph G, uniformity d

1: Find a minimum-cardinality d-uniform hypergraph H' such that $\mathrm{Proj}(H') = G$

2: **return** H'

6.3.2 Two-connected components

The MAP problem decomposes over a notion of component adapted to clique covers. Let H_c be the d-clique hypergraph of the projection. Two hyperedges of H_c are two-neighbors if they intersect in at least two vertices. The two-connected components of H_c are the connected components of the graph whose vertices are hyperedges of H_c, with adjacency given by two-neighborhood.

Below the partial recovery threshold, all two-connected components of H_c have bounded size with high probability. Consequently, MAP can be implemented in polynomial time for fixed d by solving the minimum-preimage problem separately on each component.

6.3.3 Ambiguous graphs

Definition 6.5 (Ambiguous graph)**.** A graph G is ambiguous if it has at least two distinct minimum d-uniform hypergraph preimages.

Ambiguous graphs are local obstructions to exact recovery. If a two-connected component of H_c has an ambiguous projection, then two different hypergraphs have the same projection and the same posterior probability up to tie-breaking. No estimator can reliably distinguish between them.

The obstruction determining the unweighted exact threshold for $d = 3, 4$ is the following family. Let

$$u, w, v_1, \dots, v_{d-1}$$

be distinct vertices, and for each $i \in [d-1]$, let S_i^u and S_i^w be disjoint sets of size $d-2$, disjoint from all other listed vertices. Let $\mathcal{A}_d$ be the d-uniform hypergraph with hyperedges

$$\{u, v_1, \dots, v_{d-1}\},$$

$$\{u, v_i\} \cup S_i^u, \qquad 1 \le i \le d-1,$$

and

$$\{w, v_i\} \cup S_i^w, \qquad 1 \le i \le d-1.$$

Then $\mathrm{Proj}(\mathcal{A}_d)$ has two minimum preimages of the same size: one contains $\{u, v_1, \dots, v_{d-1}\}$, and the other replaces it by $\{w, v_1, \dots, v_{d-1}\}$. The hypergraph $\mathcal{A}_d$ has $2d-1$ hyperedges and $2d^2 - 5d + 5$ vertices, giving the ambiguity exponent

$$\delta_{\mathrm{amb}}(d) = d - 1 - \frac{2d^2 - 5d + 5}{2d - 1} = \frac{2d - 4}{2d - 1}.$$

Theorem 6.6 (Exact recovery threshold for d-uniform projections)**.** *Let $H \sim H_d(n, p)$, where*

$$p = (c + o(1))n^{-d+1+\delta}, \qquad c \in (0, \infty),$$

and let $\Psi = \mathrm{Proj}(H)$. *Define*

$$\delta_{\mathrm{ex}}(d) = \begin{cases} \dfrac{2d-4}{2d-1}, & d = 3, 4, \\ \dfrac{d-1}{d+1}, & d \geq 5. \end{cases}$$

If

$$\delta < \delta_{\mathrm{ex}}(d),$$

then MAP achieves exact recovery:

$$\mathbb{P}\big(A_{\mathrm{MAP}}(\Psi) = H\big) = 1 - o(1).$$

Moreover, for fixed d, MAP is computable in polynomial time with high probability.

If

$$\delta > \delta_{\mathrm{ex}}(d),$$

then exact recovery is impossible: for every estimator A,

$$\mathbb{P}\big(A(\Psi) = H\big) = o(1).$$

Proof idea. Below $(d-1)/(d+1)$, the d-clique hypergraph H_c decomposes into bounded-size two-connected components. MAP can therefore be solved componentwise. If no ambiguous component appears, each component has a unique minimum preimage, and the posterior penalty for any non-minimum preimage is a factor $O(p) = o(1)$. Thus MAP succeeds.

The extremal ambiguous graph $\mathrm{Proj}(\mathcal{A}_d)$ appears at exponent $(2d-4)/(2d-1)$, and a deterministic extremal lemma shows that no ambiguous graph can appear at a lower exponent. Hence below $\delta_{\mathrm{ex}}(d)$, no ambiguous component appears with high probability.

For impossibility, there are two mechanisms. If $\delta > (d-1)/(d+1)$, partial recovery already has loss $1 - o(1)$, so exact recovery is impossible. If $d = 3, 4$ and

$$\frac{2d-4}{2d-1} < \delta < \frac{d-1}{d+1},$$

many vertex-disjoint copies of $\mathrm{Proj}(\mathcal{A}_d)$ appear. Each copy contributes a binary posterior ambiguity, so the posterior mass of any single exact preimage tends to zero. See Appendix C.3.

6.4 Weighted Projections

The weighted projection records, for each graph edge, how many hyperedges contain it. This additional multiplicity information removes the low-density ambiguous obstruction responsible for the gap between partial and exact recovery when $d = 3, 4$. It does not move the partial recovery threshold.

Theorem 6.7 (Exact recovery from weighted projections)**.** *Let* $H \sim H_d(n, p)$, *where*

$$p = (c + o(1))n^{-d+1+\delta}.$$

For every $d \geq 3$, exact recovery from $\mathrm{Proj}_{\mathrm{W}}(H)$ *has threshold*

$$\frac{d-1}{d+1}.$$

That is, if

$$\delta < \frac{d-1}{d+1},$$

then weighted MAP recovers H with probability $1 - o(1)$*. If*

$$\delta > \frac{d-1}{d+1},$$

then every estimator from $\mathrm{Proj}_{\mathrm{W}}(H)$ *succeeds with probability* $o(1)$*.*

Proof idea. The achievability proof again uses bounded two-connected components below $(d-1)/(d+1)$. The weighted observation distinguishes the unweighted ambiguous configurations $\mathrm{Proj}(\mathcal{A}_d)$: the two candidate preimages have the same unweighted projection but different edge multiplicities. A weighted ambiguity therefore requires a denser local configuration, and a deterministic degree argument shows that no weighted ambiguous component appears below $(d-1)/(d+1)$. The converse follows because weighted partial recovery has loss $1 - o(1)$ above the same threshold. See Appendix C.4.

6.5 Application to Hypergraph Stochastic Block Models

The hypergraph stochastic block model is often observed through a similarity matrix or weighted projection rather than through the hypergraph itself. In the usual exact-recovery HSBM scaling,

$$p = \Theta(n^{-d+1} \log n),$$

the density lies far below the reconstruction thresholds of this chapter. Therefore the original hypergraph can be recovered from its similarity matrix or weighted projection with high probability before running a community-recovery algorithm. In this sparse regime, using the pairwise similarity matrix does not lose statistical information relative to observing the full hypergraph.

6.6 Heterogeneous Hypergraphs

6.6.1 Model

We now consider the heterogeneous random hypergraph model of Definition 3.7. Fix degrees

$$2 \leq d_1 < d_2 < \cdots < d_k,$$

and let each d_r-set appear independently with probability

$$p_r = n^{1-d_r+\delta_r+o(1)}, \qquad 0 < \delta_r < 1.$$

Algorithm 5 Maximal clique estimator for degree d_j

Require: Projected graph ψ, target degree d_j

$\widehat{E}_j \leftarrow \emptyset$
for all maximal cliques C of ψ **do**
  if $|C| = d_j$ **then**
    Add C to $\widehat{E}_j$
  end if
end for
return $\widehat{H}_j = ([n], \widehat{E}_j)$

Let H_r denote the sub-hypergraph of degree-d_r hyperedges, and let

$$H = \bigcup_{r=1}^{k} H_r.$$

We observe only the unweighted projection

$$\psi = \mathrm{Proj}(H).$$

The goal is to recover the hyperedges of one target degree d_j. An estimator $\widehat{H}_j$ achieves partial recovery for degree d_j if

$$\mathbb{E}\Big[|E(H_j) \triangle E(\widehat{H}_j)|\Big] = o(1)\, \mathbb{E}[|E(H_j)|].$$

6.6.2 Maximal clique estimator

A true degree-d_j hyperedge creates a clique of size d_j. In a heterogeneous model, however, a clique of size d_j may be contained in the projection of a larger hyperedge. The maximal clique estimator avoids such obvious false positives by returning only maximal cliques of the target size.

Theorem 6.8 (Heterogeneous partial recovery achievability)**.** *Consider the heterogeneous random hypergraph model of Definition 3.7. Fix j with $d_j \geq 3$, and let*

$$\delta_* = \max_{r \in [k]} \delta_r.$$

If

$$\delta_* < \frac{d_j - 2}{d_j} + \frac{2\delta_j}{d_j(d_j - 1)},$$

then Algorithm 5 recovers $1 - o(1)$ of the d_j-hyperedges in expectation:

$$\mathbb{E}\Big[|E(H_j) \triangle E(\widehat{H}_j)|\Big] = o(1)\, \mathbb{E}[|E(H_j)|].$$

Table 6.1: Recovery thresholds and algorithms for hypergraph projection problems.

Model	Observation	Objective	Threshold / guarantee
d-uniform	$\mathrm{Proj}(H)$	partial recovery	$\delta_{\mathrm{part}}(d) = \dfrac{d-1}{d+1}$
d-uniform	$\mathrm{Proj}(H)$	exact recovery	$\delta_{\mathrm{ex}}(d) = \begin{cases} \dfrac{2d-4}{2d-1}, & d = 3, 4, \\ \dfrac{d-1}{d+1}, & d \geq 5 \end{cases}$
d-uniform	$\mathrm{Proj}_{\mathrm{W}}(H)$	exact recovery	$\dfrac{d-1}{d+1}$
heterogeneous	$\mathrm{Proj}(H)$	degree-d_j partial recovery	$\delta_* < \dfrac{d_j - 2}{d_j} + \dfrac{2\delta_j}{d_j(d_j-1)}$

Proof idea. By symmetry it suffices to analyze a fixed d_j-set S. A false positive requires S to be a maximal clique in the projection without being a true degree-d_j hyperedge. Such a clique must be assembled from smaller pieces, and the dominant obstruction is the all-edges cover, where each of the $\binom{d_j}{2}$ graph edges is supplied separately. The density condition is exactly the condition that this fake-clique exponent is smaller than the true degree-d_j hyperedge exponent $1 - d_j + \delta_j$.

A false negative requires S to be a true hyperedge but not maximal in the projection. This can only happen if some outside vertex is connected to every vertex of S. The probability of such an outside-star extension is negligible under the same density condition. The full cover-exponent calculation and the false-positive/false-negative estimates are given in Appendix C.5.

6.7 Comparison of Local Recovery Results

Chapter 7

Synthesis, Limitations, and Open Problems

7.1 Synthesis of the Thesis

This thesis began from a tension between two views of random graphical models. On the one hand, the classical theory of average-case inference is built around product noise models such as Erdős–Rényi graphs, where edges are independent and the planted signal is the only structured object. On the other hand, many networks of interest exhibit clustering, latent features, higher-order relations, or geometry; in such models, even the null distribution is not a product measure. The central claim developed throughout the thesis is that these two views are not disjoint. Local dependencies provide a bridge between independent random graph models and globally constrained graphical models.

This chapter returns to the dependency hierarchy introduced in Section 1.2. The hierarchy

$$\mathcal{C}_{\mathrm{ind}} \longrightarrow \mathcal{C}_{\mathrm{loc}} \longrightarrow \mathcal{C}_{\mathrm{glob}}$$

is not meant as a literal inclusion of probability spaces, but as a guide to how dependencies are generated and compared. In $\mathcal{C}_{\mathrm{ind}}$, edges are formed independently. In $\mathcal{C}_{\mathrm{loc}}$, correlations are induced by bounded-footprint or controlled local objects such as triangles, latent features, signed triangle updates, or hyperedges. In $\mathcal{C}_{\mathrm{glob}}$, correlations arise from macroscopic latent variables or constraints such as geometry. The results of the thesis show that, in appropriate high-dimensional or sparse regimes, a model that appears globally constrained can often be reproduced by a model whose dependence is local. Thus the correct comparison is not simply "independent versus dependent," but rather a more refined comparison between the scale of dependence in the noise and the scale of the signal.

The first part of the thesis studies this bridge at the level of distributions. Random graphs with triangles provide an explicit locally dependent perturbation of $\mathrm{G}(n, p)$, and in suitable regimes they approximate random intersection graphs in total variation [17, 19]. High-dimensional random geometric graphs give a second instance of the same phenomenon: in the balanced middle-scale regime, the first surviving correction below the Erdős–Rényi indistinguishability threshold is captured by an explicit local signed triangle-sampling mechanism [20]. These distributional results are summarized in Theorems 1.1 and 1.2. They show that local dependencies can arise both as direct models of motifs and as effective descriptions of latent-feature and latent-geometric graph models.

The second part of the thesis studies global recovery. Here a single macroscopic signal, such as a planted clique or planted dense subgraph, is embedded in a dependent ambient graph. The additional triangle process in RGT creates strong local correlations and destroys many tools that rely on edge independence. Nevertheless, the planted signal is a coherent object involving many edges, whereas the ambient dependencies are generated by bounded-size motifs. The algorithmic decorrelation results exploit precisely this separation of scales: they transform a locally dependent instance into an independent-noise instance while approximately preserving the planted structure in the parameter regimes where the reduction applies [19]. This is the content of the computational equivalence result summarized in Theorem 1.3.

The third part of the thesis studies local recovery. In hypergraph projection recovery, the local motifs are not nuisance correlations to be removed; they are the objects to be recovered. The observation is the graph projection $\mathrm{Proj}(H)$, in which a pair of vertices is adjacent if they appear together in at least one hyperedge. The projection loses the identity of the generating hyperedges, so the task is to invert a many-to-one local generative mechanism. The thesis identifies when this local structure remains visible in the graph projection and when distinct hypergraphs produce essentially indistinguishable observations [24, 25]. The uniform recovery thresholds are summarized in Theorems 1.4 and 1.5, and the heterogeneous achievability result is summarized in Theorem 1.6.

Taken together, these directions support a single conceptual message. Local dependencies are not merely a technical nuisance. They are an organizing principle. For global recovery, they can often be separated from the planted signal by decorrelation or reduction. For local recovery, they are the signal itself, and the main difficulty is to determine when the projection has erased too much information. The same mathematical objects—triangles, cliques, overlaps, neighborhoods, and small subgraph counts—therefore play two different roles depending on the inference task. They are noise features in the planted clique problem and generative features in hypergraph recovery.

7.2 Statistical Equivalence Versus Computational Equivalence

A recurring theme of the thesis is that several notions of equivalence must be kept separate. The strongest notion used here is statistical equivalence of model sequences: two distributions P_n and Q_n are equivalent if

$$d_{\mathrm{TV}}(P_n, Q_n) = o(1).$$

This is a distributional statement. Once it is proved, every test, estimator, threshold, and impossibility result transfers automatically between the two models, because no statistical procedure can distinguish the two observations with asymptotically nonvanishing advantage.

The thesis uses this notion in the structural part of the work. The random intersection graph approximation and the high-dimensional geometric approximation are statements about when apparently dependent graph models can be approximated, in total variation, by localized or independent baselines [17, 19, 20]. In these results, the conclusion is not merely that a few graph statistics match. Rather, the entire law of the observed graph is asymptotically indistinguishable from the comparison model in the stated regime. This justifies treating the local model as a faithful surrogate for the original model for all statistical purposes.

Computational equivalence is a different statement. Two inference problems are computationally equivalent when there are polynomial-time reductions between them that preserve success and failure probabilities. Such a result may hold even when the raw graph distributions are not unconditionally close in total variation. The algorithmic decorrelation theorem for planted dense subgraph and planted clique in locally dependent noise is of this type. It constructs randomized polynomial-time transformations between independent-noise and RGT-noise formulations. The key point is task preservation: the maps must both decorrelate the ambient graph and preserve the planted set. Thus the theorem is not only a statement about null distributions; it is a statement about planted and unplanted ensembles together [19].

The third level is conjectural transfer. When a reduction maps a classical hard problem such as planted clique to a dependent-noise recovery problem, it proves a conditional lower bound: an efficient algorithm for the dependent problem would imply an efficient algorithm for the classical problem. The reduction is unconditional, but the final hardness interpretation depends on the planted clique conjecture or on related evidence from low-degree, statistical-query, and sum-of-squares lower bounds [4–7, 12]. This distinction is important. The thesis proves reductions and equivalences; it does not prove unconditional lower bounds for all polynomial-time algorithms in regimes where the conclusion is explicitly based on standard average-case hardness assumptions.

The following table summarizes the role of each type of result.

This separation clarifies the logical structure of the thesis. Total variation equivalence is a model-level statement. Algorithmic decorrelation is a task-level statement. Conditional hardness is a consequence of a task-level reduction plus an external conjecture. The thesis uses all three, but for different purposes.

7.3 Global Recovery and Local Recovery

The thesis also separates two forms of recovery that are often grouped together under the phrase "community recovery." In global recovery, there is one planted object of macroscopic size. The planted clique problem is the canonical example: the goal is to recover a single set $S \subseteq [n]$ with $|S| = K$, and success means identifying this set. The ambient graph may contain many local irregularities, but these irregularities are not themselves the target. They are obstacles that may obscure the global signal.

In this regime, the central question is whether the local dependencies in the noise can imitate the global signal. Triangles in the null graph create clustering and can generate spurious eigenvectors, so purely spectral methods may fail for reasons unrelated to the planted clique. However, a triangle is a bounded-size dependency, while the clique is a coherent structure across many vertices and many edges. Algorithmic decorrelation formalizes the idea that these two types of structure can be separated. The decorrelation map removes or neutralizes the local excess-triangle structure while maintaining enough of the planted clique

Table 7.1: Statistical and computational comparisons used in the thesis.

Setting	**Type of statement**	**Consequence**
High-dimensional geometric and latent-feature graph models	Total variation equivalence between the original model and a localized or independent surrogate, in the relevant asymptotic regime.	All statistical procedures and thresholds transfer at the level of observation laws.
Planted dense subgraph and planted clique in random graphs with triangle dependencies	Polynomial-time algorithmic decorrelation and average-case reductions.	Algorithms and conditional lower-bound evidence transfer between independent and dependent-noise planted formulations.
Uniform hypergraph recovery from graph projections	Direct information-theoretic and algorithmic analysis of an inverse problem.	Partial and exact recovery thresholds are governed by whether local generative motifs remain identifiable after projection.
Heterogeneous hypergraph recovery	Achievability through maximal-clique extraction under density inequalities.	Recovery is guaranteed in a natural sparse regime, but matching converses remain open in general.
Hardness below algorithmic thresholds	Conditional consequence of reductions from planted clique or related average-case assumptions.	The conclusion is evidence of computational hardness, not an unconditional impossibility theorem for every polynomial-time algorithm.

for recovery or detection to transfer to the independent model [19].

Local recovery is different. In hypergraph projection recovery, there is no single massive planted set. Instead, the signal consists of many small or moderate-size hyperedges, each of which creates a clique in the observed graph projection. The inference problem is therefore not to find one global object but to determine which observed cliques were generated by true hyperedges and which arose accidentally from overlaps among smaller or different hyperedges. The obstacles are combinatorial rather than spectral: two different collections of hyperedges may cast the same pairwise shadow, and dense overlap can make maximal cliques cease to correspond to true generative motifs.

For d-uniform random hypergraphs with

$$p = (c + o(1))n^{-d+1+\delta},$$

partial recovery from the unweighted projection has an all-or-nothing threshold at

$$\delta_{\mathrm{part}}(d) = \frac{d-1}{d+1},$$

and the same threshold holds for weighted projections, as summarized in Theorem 1.4 [25]. Exact recovery from the unweighted projection has threshold

$$\delta_{\mathrm{exact}}(d) = \begin{cases} \dfrac{2d-4}{2d-1}, & 3 \le d \le 4, \\ \dfrac{d-1}{d+1}, & d \ge 5, \end{cases}$$

while exact recovery from the weighted projection has threshold $(d-1)/(d+1)$ for every $d \ge 3$, as summarized in Theorem 1.5 [24, 25]. These thresholds have a different character from planted clique lower bounds: above the threshold, the projection itself loses the relevant information.

The heterogeneous model extends this local recovery viewpoint beyond a single arity. If

$$2 \le d_1 < \cdots < d_k, \qquad p_r = n^{1-d_r+\delta_r+o(1)}, \qquad 0 < \delta_r < 1,$$

and $\delta_* = \max_r \delta_r$, then the maximal-clique estimator recovers degree-d_j hyperedges with vanishing normalized expected symmetric-difference error whenever

$$\delta_* < \frac{d_j - 2}{d_j} + \frac{2\delta_j}{d_j(d_j-1)}$$

for a target degree $d_j \ge 3$, as summarized in Theorem 1.6 [26]. This is an achievability theorem rather than a full converse, but it shows how the d-uniform obstruction generalizes to a multi-scale projection problem.

This contrast explains why the methods in the two parts of the thesis look different. For global recovery, local motifs are removed by a randomized transformation. For local recovery, local motifs are recovered by neighborhood intersections, maximal-clique structure, clique-cover arguments, MAP analysis, and overlap bounds. For global recovery, the main reduction compares dependent noise to independent noise. For local recovery, the main comparison is between the true hypergraph and alternative hypergraphs with the same or nearly the same projection. For global recovery, the key scale is the planted size K. For local recovery, the key scales are the hyperedge arities, the hyperedge probabilities, and the density of collisions in the projection.

7.4 Limitations

The results of the thesis are asymptotic and regime-dependent. They identify sharp phenomena in several natural models, but they do not give a complete theory of all dependent random graphical models. The main limitations are as follows.

(i) **Parameter regimes.** The equivalence and recovery theorems are proved in specific scaling windows for edge probabilities, motif probabilities, hyperedge probabilities, geometric dimension, and planted set sizes. Outside these windows, the same intuition may remain valid, but the proofs may fail. In particular, the current RGT decorrelation theorem applies to planted sizes below the conjectural $\sqrt{n}$ planted clique scale, and the high-dimensional geometric approximation is proved in a middle-scale regime above n^2 up to logarithmic factors [19, 20]. Critical windows where two error probabilities balance often require sharper constants than the methods in the thesis provide.

(ii) **Uniformity and bounded arity.** The hypergraph results are strongest for fixed d-uniform models or for heterogeneous models with a bounded set of arities. If the maximum hyperedge size grows with n, or if the arity distribution is heavy-tailed, then the local projection geometry can change substantially. In such regimes, maximal cliques may be produced by mixtures of smaller hyperedges, and the clean bijections used in bounded-arity settings need not hold.

(iii) **Exact versus partial recovery.** For d-uniform noiseless projections, the thesis now has sharp partial and exact recovery thresholds. This does not settle the corresponding questions in heterogeneous and noisy settings. Exact recovery is especially sensitive to rare local obstructions, while partial recovery is governed by typical false-clique density. The separation between these two criteria becomes more pronounced once hyperedges of different sizes or corrupted pairwise observations are introduced.

(iv) **Sparse and dense boundaries.** The sparse regime can be information-theoretically simple because hyperedge shadows are isolated, but it can also become statistically uninformative if too few pairwise observations are generated. The dense regime can become difficult because many local motifs overlap, causing the graph projection to contain large cliques that do not correspond to single hyperedges. The methods in the thesis work best between these extremes, where the projection is informative but not so dense that all local structure has merged.

(v) **Computational cost of MAP.** The maximum a posteriori estimator is the natural statistical benchmark for exact recovery. In the uniform noiseless projection model, it can be implemented efficiently in the relevant recoverable regimes using bounded two-connected components and local structure [24, 25]. This does not imply that MAP remains efficient in richer models. For heterogeneous hypergraphs, noisy projections, growing arities, or additional latent structure, the posterior optimization problem may no longer decompose into small components.

(vi) **Model specificity of decorrelation.** The algorithmic decorrelation results are developed for structured local dependencies, especially triangle-generated dependence. The construction uses detailed information about how the dependent graph is generated, including the posterior distribution of latent triangle sets. A general dependent graph with arbitrary clustering or arbitrary motif correlations may not admit the same posterior-sampling or edge-resampling map.

(vii) **Dependence on conjectural hardness.** The algorithmic lower bounds for dependent-noise planted clique should be interpreted conditionally. The reductions show that an efficient algorithm in the dependent model would imply an efficient algorithm for the corresponding independent planted clique problem. The conclusion that no such algorithm exists depends on the standard planted clique hardness conjecture and related bodies of evidence, rather than on an unconditional lower bound.

(viii) **Noiseless observation assumptions.** The hypergraph projection results primarily study the clean projection map. In applications, observed graphs may contain false positive edges, missing edges, or additional pairwise relations not generated by the hypergraph. Even small independent edge noise can destroy exact clique structure, so robust recovery requires new arguments beyond the noiseless projection analysis.

These limitations do not undermine the main message of the thesis, but they indicate where the current theory is still incomplete. The results show that local dependencies can be analyzed rigorously and can sometimes be removed, simulated, or inverted. They do not yet provide a universal calculus for all forms of dependence.

7.5 Open Problems

Open Problem 1. Sharper algorithmic decorrelation.

Extend RGT decorrelation to larger planted sets, denser triangle regimes, or more general local motif distributions. The existing decorrelation paradigm relies on a controlled separation between the planted signal and the local dependencies in the null model. A sharper theory would identify the largest class of motif-generated random graphs for which one can construct a polynomial-time map to an independent or simpler planted model.

One concrete direction is to replace triangle-specific corrections with a general posterior-sampling procedure for bounded-size motifs. Given a graph generated by first sampling local motifs and then revealing their pairwise shadows, can one efficiently sample a decorrelated graph whose law is close to an independent baseline while preserving a planted structure? A positive answer would turn algorithmic decorrelation into a reusable reduction tool rather than a model-specific construction.

A second direction is to understand the failure modes. Decorrelation may fail when the planted signal itself changes local motif statistics too strongly, or when the motif density is high enough that many explanations of the same graph become statistically plausible. Characterizing these obstructions would clarify whether the current restrictions are proof artifacts or genuine computational barriers.

Open Problem 2. General local dependency universality.

Characterize which local motif models approximate high-dimensional geometric or latent-feature graph models in total variation. The thesis supports the view that local dependencies can simulate some globally generated graph laws in appropriate regimes, but a general universality theorem is still missing.

Such a theorem should specify both necessary and sufficient conditions. On the geometric side, one would like to know which finite collections of signed subgraph statistics capture the relevant dependence of a high-dimensional random geometric graph. On the latent-feature side, one would like to know when a random intersection graph can be replaced by an explicit motif model without remembering the latent bipartite representation. The answer may involve a combination of moment matching, entropy bounds, small-subgraph conditioning, and graph limit ideas.

A satisfactory theory would also distinguish positive and negative local correlations. Triangle injection models positive clustering, but geometric constraints can create more subtle signed dependencies among nearby edges. A universal local model should be flexible enough to encode both types.

Open Problem 3. **Heterogeneous hypergraph converses.**

Determine optimal thresholds for recovery in heterogeneous hypergraphs, especially when recovering intermediate degrees rather than the largest degree. In heterogeneous models, hyperedges of different sizes interact through the same graph projection. A clique of size d in the projection may be a true d-hyperedge, the shadow of a larger hyperedge, or the accidental union of several smaller hyperedges. This creates a multi-scale identifiability problem that is absent from the clean d-uniform setting.

The main open challenge is to prove matching converses for the achievability condition in Theorem 1.6. Achievability can often be obtained by greedy extraction or maximal-clique tests under suitable sparsity assumptions [26]. Proving that recovery is impossible below the corresponding threshold is harder, because one must construct alternative hypergraphs that produce the same or statistically indistinguishable projections. For intermediate arities, the obstruction may come neither from the smallest nor from the largest hyperedges, but from collisions across scales.

A complete answer would give, for each arity d_j, the exact threshold at which d_j-hyperedges become recoverable in the presence of all other arities. This would turn heterogeneous projection recovery into a genuinely multi-parameter phase diagram.

Open Problem 4. **Noisy projections.**

Determine partial and exact recovery thresholds when the graph projection is corrupted by edge noise. In the clean model, a hyperedge produces a complete clique in the observed graph. With missing edges, a true hyperedge may no longer appear as a clique. With added edges, false cliques may appear. Both effects break the exact combinatorial structure used by maximal-clique and neighborhood-intersection algorithms.

A noisy theory should distinguish several noise models: independent edge deletions, independent edge additions, semirandom noise, and noise whose rate depends on vertex degrees or hyperedge sizes. The exact recovery threshold will depend on whether the algorithm must recover every hyperedge, all large hyperedges, or a vanishing-error fraction of hyperedges. In many regimes, exact recovery may be impossible while partial recovery remains feasible.

The central technical problem is to replace exact clique certificates with robust certificates. Candidate tools include common-neighborhood statistics, dense subgraph tests, error-correcting codes for clique shadows, and local likelihood ratios. The goal is a theory that reduces to the noiseless thresholds when the noise rate is zero and degrades smoothly as noise increases.

Open Problem 5. Efficient exact recovery beyond known regimes.

Clarify when MAP can be efficiently implemented and whether statistical-computational gaps appear in dense hypergraph projection recovery. The MAP estimator is the natural benchmark for exact recovery, but its naive implementation is generally exponential. The algorithms in the thesis show that local structure can make exact recovery efficient in certain regimes, but the boundary of efficient recovery is not yet known.

One possibility is that the statistical threshold and the computational threshold coincide for broad classes of projection recovery problems. Another possibility is that dense heterogeneous projections exhibit genuine statistical-computational gaps: the hypergraph is identifiable in principle, but every known polynomial-time method fails because the projection contains too many overlapping candidate explanations. Distinguishing these possibilities requires both better algorithms and sharper lower-bound evidence.

Promising algorithmic directions include local-to-global peeling procedures, semidefinite or sum-of-squares relaxations, belief propagation on factor graphs, and parameterized algorithms that exploit bounded arity or bounded overlap. On the lower-bound side, one would like reductions or low-degree evidence showing that certain dense projection regimes are computationally hard despite being information-theoretically identifiable. Resolving this question would complete the parallel between the global planted-signal problems and the local hypergraph-recovery problems studied in this thesis.

Appendix A

Deferred Proofs for Chapter 4

This appendix contains the proofs explicitly deferred in Chapter 4. The proof of Theorem 4.3 follows the argument in [**BGP23**]; the one-triangle comparison used in the upper bound is the standard planted-clique comparison appearing in [**BBN20**]. The non-reproved bridge theorems in the chapter are source theorems from [**BGP23**] and [**BBG26MiddleScale**], as recorded in Section A.3 below.

A.1 Observable Control by Local Writes

Proof of Proposition **??**. We work on the probability space on which the local-kernel construction is defined:

$$G^{(0)}, G^{(1)}, \dots, G^{(R_n)}.$$

At step r, only edges in $\binom{V_r}{2}$ may change. Therefore, by the assumed edge-Lipschitz property of f,

$$\left|f(G^{(r)}) - f(G^{(r-1)})\right| \le \sum_{e\in\binom{[n]}{2}} c_e\,\mathbf{1}\{G_e^{(r)} \neq G_e^{(r-1)}\}.$$

Since $G_e^{(r)} = G_e^{(r-1)}$ for every $e \notin \binom{V_r}{2}$, this is equivalently

$$\left|f(G^{(r)}) - f(G^{(r-1)})\right| \le \sum_{e\in\binom{V_r}{2}} c_e\,\mathbf{1}\{G_e^{(r)} \neq G_e^{(r-1)}\}.$$

Summing over $r = 1, \dots, R_n$ and applying the triangle inequality gives

$$\left|f(G^{(R_n)}) - f(G^{(0)})\right| \le \sum_{r=1}^{R_n} \sum_{e\in\binom{V_r}{2}} c_e\,\mathbf{1}\{G_e^{(r)} \neq G_e^{(r-1)}\}.$$

Changing the order of summation,

$$\left|f(G) - f(G^{(0)})\right| \le \sum_{e\in\binom{[n]}{2}} c_e \sum_{r=1}^{R_n} \mathbf{1}\{e \in \binom{V_r}{2}\}\mathbf{1}\{G_e^{(r)} \neq G_e^{(r-1)}\}.$$

The inner sum is exactly W_e. Thus

$$\left|f(G) - f(G^{(0)})\right| \le \sum_{e \in \binom{[n]}{2}} c_e W_e.$$

Taking expectations yields

$$\mathbb{E}\left|f(G) - f(G^{(0)})\right| \le \sum_{e \in \binom{[n]}{2}} c_e \, \mathbb{E} W_e.$$

Finally,

$$\left|\mathbb{E} f(G) - \mathbb{E} f(G^{(0)})\right| \le \mathbb{E}\left|f(G) - f(G^{(0)})\right|,$$

so

$$\left|\mathbb{E} f(G) - \mathbb{E} f(G^{(0)})\right| \le \sum_{e \in \binom{[n]}{2}} c_e \, \mathbb{E} W_e.$$

If $c_e \le c_n$ for every edge e, then

$$\sum_{e \in \binom{[n]}{2}} c_e \, \mathbb{E} W_e \le c_n \sum_{e \in \binom{[n]}{2}} \mathbb{E} W_e,$$

which proves the final displayed inequality. □

A.2 Random Graphs with Triangles and the $n^{-3/2}$ Window

We first prove the one-triangle comparison input used in the upper bound. This lemma is a self-contained version of the comparison used in [**BBN20**] and [**BGP23**].

Lemma A.1 (One-triangle insertion with matched edge density)**.** *Fix $r \in (0,1)$ bounded away from 0 and 1. Let P_n be the law obtained by sampling $G \sim \mathrm{G}(n,r)$, choosing an unordered triple $T \in \binom{[n]}{3}$ uniformly and independently, and forcing the three edges of T to be present. Let*

$$N = \binom{n}{2}, \qquad M = \binom{n}{3}, \qquad r_+ = r + (1-r)\frac{3}{N}.$$

Then

$$d_{\mathrm{TV}}\big(P_n, \mathrm{G}(n, r_+)\big) = O(n^{-3/2}).$$

Proof. Let $Q = \mathrm{G}(n, r_+)$. For a fixed triple T, let P_T be the law in which the three edges of T are forced to be present and all other edges are independent with density r. Then

$$P_n = \frac{1}{M} \sum_{T \in \binom{[n]}{3}} P_T.$$

We bound $\chi^2(P_n \| Q)$.

Write

$$\delta = r_+ - r = (1-r)\frac{3}{N}.$$

For a single edge, define the likelihood factors relative to $\mathrm{Bern}(r_+)$:

$$h(x) = \frac{\mathbf{1}\{x=1\}}{r_+}, \qquad g(x) = \left(\frac{r}{r_+}\right)^x \left(\frac{1-r}{1-r_+}\right)^{1-x}.$$

Here h is the likelihood factor for a forced edge, while g is the likelihood factor for an ordinary $\mathrm{Bern}(r)$ edge. A direct calculation gives

$$\mathbb{E}_{\mathrm{Bern}(r_+)} h^2 = \frac{1}{r_+},$$

$$\mathbb{E}_{\mathrm{Bern}(r_+)} hg = \frac{r}{r_+},$$

and

$$\mathbb{E}_{\mathrm{Bern}(r_+)} g^2 = \frac{r^2}{r_+} + \frac{(1-r)^2}{1-r_+} = 1 + \frac{(r_+-r)^2}{r_+(1-r_+)}.$$

Set

$$a = 1 + \frac{\delta^2}{r_+(1-r_+)}.$$

Let T, T' be two unordered triples. Let

$$s = s(T,T') := |\binom{T}{2} \cap \binom{T'}{2}|.$$

Since two distinct triples either share no edge or share exactly one edge, $s \in \{0,1,3\}$, where $s=3$ means $T=T'$. By independence of edges under Q,

$$\mathbb{E}_Q\left[\frac{dP_T}{dQ}(G)\frac{dP_{T'}}{dQ}(G)\right] = r_+^{-s}\left(\frac{r}{r_+}\right)^{6-2s} a^{N-6+s}.$$

Denote this quantity by F_s. If T, T' are chosen independently and uniformly from $\binom{[n]}{3}$, then

$$\chi^2(P_n \| Q) + 1 = \mathbb{E}_{T,T'} F_{s(T,T')}.$$

The probabilities of the three overlap types are

$$\mathbb{P}(s=3) = \frac{1}{M} = O(n^{-3}),$$

$$\mathbb{P}(s=1) = \frac{3(n-3)}{M} = \frac{9}{N} + O(n^{-3}),$$

and

$$\mathbb{P}(s=0) = 1 - \mathbb{P}(s=1) - \mathbb{P}(s=3).$$

We now expand the terms. Since r is bounded away from 0 and 1, and since $\delta = O(N^{-1})$, all constants below are uniform in n. First,

$$\frac{r}{r_+} = 1 - \frac{\delta}{r} + O(\delta^2),$$

and

$$a^{N-6} = 1 + \frac{N\delta^2}{r(1-r)} + O(N^{-2}).$$

Because $\delta = (1-r)3/N$, this gives

$$a^{N-6} = 1 + \frac{9(1-r)}{rN} + O(N^{-2}).$$

Therefore

$$F_0 = \left(\frac{r}{r_+}\right)^6 a^{N-6} = 1 - \frac{18(1-r)}{rN} + \frac{9(1-r)}{rN} + O(N^{-2}) = 1 - \frac{9(1-r)}{rN} + O(N^{-2}).$$

Equivalently, using $h = 3/N$,

$$F_0 = 1 - \frac{3(1-r)}{r}h + O(N^{-2}).$$

Next, for $s = 1$,

$$F_1 = r_+^{-1}\left(\frac{r}{r_+}\right)^4 a^{N-5} = \frac{1}{r} + O(N^{-1}),$$

so

$$F_1 - F_0 = \frac{1-r}{r} + O(N^{-1}).$$

Finally, $F_3 = O(1)$, because r_+ is bounded away from 0 and 1.

Combining the three cases,

$$\chi^2(P_n\|Q) + 1 = F_0 + \mathbb{P}(s=1)(F_1 - F_0) + \mathbb{P}(s=3)(F_3 - F_0).$$

Using

$$\mathbb{P}(s=1) = \frac{9}{N} + O(n^{-3}), \qquad \mathbb{P}(s=3) = O(n^{-3}),$$

we get

$$\chi^2(P_n\|Q) + 1 = \left(1 - \frac{9(1-r)}{rN}\right) + \frac{9}{N}\frac{1-r}{r} + O(n^{-3}) = 1 + O(n^{-3}).$$

Thus

$$\chi^2(P_n\|Q) = O(n^{-3}).$$

Pinsker's inequality gives

$$d_{\mathrm{TV}}(P_n, Q) \le \sqrt{\frac{1}{2}\chi^2(P_n\|Q)} = O(n^{-3/2}).$$

$\square$

Proof of Theorem 4.3. Let

$$G \sim \mathrm{RGT}(n, p, p'), \qquad q = 1 - (1-p)(1-p')^{n-2}.$$

Throughout the proof $p \in (0,1)$ is fixed and $p' = O(1/n)$, so q is bounded away from 0 and 1. Let $A = (A_{ij})$ be the adjacency matrix of the observed graph. For a triple $t = \{i, j, k\}$, define

$$\tau_t = (A_{ij} - q)(A_{ik} - q)(A_{jk} - q),$$

and define the centered triangle statistic

$$T(A) = \sum_{1 \le i < j < k \le n} \tau_{\{i,j,k\}}.$$

The Erdős–Rényi calculation. If $A \sim \mathrm{G}(n, q)$, the centered edge variables $A_{ij} - q$ are independent and have mean zero. Hence

$$\mathbb{E}_{\mathrm{G}(n,q)} T = 0.$$

For two triples $t \neq t'$, the product $\tau_t \tau_{t'}$ contains at least one centered edge variable appearing to the first power, unless $t = t'$. By independence and centering,

$$\mathbb{E}_{\mathrm{G}(n,q)}[\tau_t \tau_{t'}] = 0 \qquad (t \neq t').$$

Therefore

$$\mathrm{Var}_{\mathrm{G}(n,q)}(T) = \binom{n}{3} \mathbb{E}(A_{12} - q)^2 (A_{13} - q)^2 (A_{23} - q)^2.$$

The three edge variables are independent, and

$$\mathbb{E}(A_{12} - q)^2 = q(1-q).$$

Thus

$$\mathrm{Var}_{\mathrm{G}(n,q)}(T) = \binom{n}{3} q^3 (1-q)^3 = O(n^3).$$

The expectation under RGT. Let X_{ijk} be the indicator that the triangle on $\{i, j, k\}$ is added by the triangle layer. Conditional on $X_{ijk} = 1$, all three edges $(i, j), (i, k), (j, k)$ are present, so

$$\mathbb{E}[\tau_{\{i,j,k\}} \mid X_{ijk} = 1] = (1-q)^3.$$

Conditional on $X_{ijk} = 0$, no remaining triangle-addition variable can affect two of the three edges simultaneously. Hence the three edges of $\{i, j, k\}$ are conditionally independent. Their common conditional edge density is

$$q_0 = 1 - (1-p)(1-p')^{n-3},$$

and therefore

$$\mathbb{E}[\tau_{\{i,j,k\}} \mid X_{ijk} = 0] = (q_0 - q)^3.$$

Since

$$q_0 - q = -(1-p)(1-p')^{n-3}p' = O(p'),$$

we have

$$\mathbb{E}\tau_{\{i,j,k\}} = p'(1-q)^3 + (1-p')O((p')^3).$$

Because $p' = O(1/n)$,

$$\binom{n}{3} O((p')^3) = O(1).$$

Thus

$$\mathbb{E}_{\mathrm{RGT}(n,p,p')} T = \binom{n}{3} p'(1-q)^3 + O(1).$$

The variance under RGT**.** We prove

$$\mathrm{Var}_{\mathrm{RGT}(n,p,p')}(T) = O(n^3).$$

For two triples t, t', the covariance $\mathrm{Cov}(\tau_t, \tau_{t'})$ is zero when t and t' are vertex-disjoint, because then the edge sets and all triangle-addition variables that can affect them are disjoint.

It remains to consider pairs that share vertices. If $t = t'$, then

$$|\operatorname{Cov}(\tau_t, \tau_t)| \le \mathbb{E}\tau_t^2 = O(1),$$

and there are $O(n^3)$ such pairs.

Suppose next that two distinct triples share an edge. By relabeling, take $t = \{1,2,3\}$ and $t' = \{1,2,4\}$. Let

$$E_1 = \{12, 13, 23, 14, 24\}$$

be the set of edges appearing in $\tau_t \tau_{t'}$. Let $\mathcal{T}_1$ be the set of triangles whose edge set contains at least two edges of E_1. Explicitly,

$$\mathcal{T}_1 = \{123, 124, 134, 234\}.$$

Condition on all triangle-addition indicators in $\mathcal{T}_1$. Given this conditioning, every remaining triangle-addition variable affects at most one edge of E_1. Therefore the five edge indicators in E_1 are conditionally independent.

For an edge $e \in E_1$ not forced by the conditioned variables, its conditional centered mean is $O(p')$. Indeed, removing or fixing a constant number of possible triangle additions changes the edge density by at most $O(p')$, while the unconditional edge density is q. Also

$$\mathbb{E}\big[(A_e - q)^2 \mid \mathcal{T}_1\big] = O(1)$$

uniformly.

If at least two triangles in $\mathcal{T}_1$ are active, the probability of this conditioning event is $O((p')^2)$, and the contribution to $\mathbb{E}[\tau_t \tau_{t'}]$ is $O((p')^2)$. If at most one triangle in $\mathcal{T}_1$ is active, then among the four non-shared edges

$$13, 23, 14, 24,$$

at least two are not forced by the conditioning. The conditional expectation of the product therefore contains at least two factors of size $O(p')$, while all remaining factors are $O(1)$. This gives

$$\mathbb{E}[\tau_t \tau_{t'}] = O((p')^2).$$

Since $\mathbb{E}\tau_t = O(p')$, the covariance is also $O((p')^2)$. There are $O(n^4)$ ordered pairs of triples sharing an edge, so their total contribution is

$$O(n^4 (p')^2) = O(n^2),$$

because $p' = O(1/n)$.

Finally suppose two triples share exactly one vertex. By relabeling, take $t = \{1, 2, 3\}$ and $t' = \{1, 4, 5\}$. Let E_2 be the six edges appearing in the two triangle monomials:

$$E_2 = \{12, 13, 23, 14, 15, 45\}.$$

Let $\mathcal{T}_2$ be the finite set of triangles containing at least two edges from E_2. There are $O(1)$ such triangles. Conditioning on all triangle indicators in $\mathcal{T}_2$, the six edge indicators in E_2 become conditionally independent, because every remaining triangle-addition variable affects at most one of these six edges.

If at least two triangles in $\mathcal{T}_2$ are active, the conditioning event has probability $O((p')^2)$. If at most one is active, then at least three edges in E_2 are not forced, and the conditional product contains at least three factors of size $O(p')$. In either case

$$\mathbb{E}[\tau_t \tau_{t'}] = O((p')^2).$$

Again $\mathbb{E}\tau_t = O(p')$, so the covariance is $O((p')^2)$. There are $O(n^5)$ ordered pairs of triples sharing exactly one vertex, and hence their total contribution is

$$O(n^5 (p')^2) = O(n^3),$$

because $p' = O(1/n)$.

Combining the three cases gives

$$\mathrm{Var}_{\mathrm{RGT}(n,p,p')}(T) = O(n^3).$$

The distinguishing regime. If $p' \gg n^{-3/2}$, then

$$\mathbb{E}_{\mathrm{RGT}} T = \binom{n}{3} p'(1-q)^3 + O(1) \gg n^{3/2}.$$

Under both $\mathrm{G}(n, q)$ and $\mathrm{RGT}(n, p, p')$, the standard deviation of T is $O(n^{3/2})$. Therefore the test

$$\phi(A) = \mathbf{1}\left\{ T(A) \geq \frac{1}{2} \binom{n}{3} p'(1-q)^3 \right\}$$

has vanishing type-I and type-II error by Chebyshev's inequality. Hence

$$d_{\mathrm{TV}}\big(\mathrm{RGT}(n, p, p'), \mathrm{G}(n, q)\big) = 1 - o(1) \qquad \text{if } p' \gg n^{-3/2}.$$

The indistinguishable regime. Assume now that $p' \ll n^{-3/2}$. We use the one-triangle comparison from Lemma A.1.

Poissonize the triangle additions. Let

$$\theta = -\log(1-p'), \qquad M_\triangle \sim \mathrm{Pois}\left(\binom{n}{3}\theta\right).$$

Starting from $\mathrm{G}(n,p)$, add $M_\triangle$ independently and uniformly chosen triangles, with replacement. For each fixed triple, the number of times it is selected is $\mathrm{Pois}(\theta)$, independently over triples. Thus the probability that it is selected at least once is $1 - e^{-\theta} = p'$, and this Poissonized construction has exactly the law $\mathrm{RGT}(n,p,p')$.

Condition on $M_\triangle = m$. Let G_m be the graph obtained from $\mathrm{G}(n,p)$ after m independent uniformly chosen triangle insertions. Let

$$q_m = 1 - (1-p)\left(1 - \frac{3}{\binom{n}{2}}\right)^m.$$

This is the marginal edge density after m such insertions. Applying Lemma A.1 one insertion at a time and using the triangle inequality gives

$$d_{\mathrm{TV}}\big(\mathcal{L}(G_m), \mathrm{G}(n,q_m)\big) \le Cmn^{-3/2},$$

where C depends only on the fixed lower and upper bounds for the edge densities. Averaging over $M_\triangle$,

$$\mathbb{E}\big[CM_\triangle n^{-3/2}\big] = Cn^{-3/2}\binom{n}{3}\theta = O(n^{3/2}p') = o(1).$$

Thus

$$d_{\mathrm{TV}}\big(\,\mathrm{RGT}(n,p,p'),\, \mathbb{E}_{M_\triangle}\,\mathrm{G}(n, q_{M_\triangle})\big) = o(1),$$

where the expectation denotes the mixture over the random parameter $q_{M_\triangle}$.

It remains to replace the random edge density $q_{M_\triangle}$ by the deterministic density q. Since

$$\mathbb{E}M_\triangle = \binom{n}{3}\theta,$$

we have

$$1 - q = (1-p)e^{-\theta(n-2)} = (1-p)(1-p')^{n-2},$$

which agrees with the marginal edge density in $\mathrm{RGT}(n,p,p')$.

The map

$$m \mapsto q_m = 1 - (1-p)\left(1 - \frac{3}{\binom{n}{2}}\right)^m$$

is Lipschitz with constant $O(n^{-2})$. Hence

$$\mathbb{E}|q_{M_\triangle} - q| \le Cn^{-2}\mathbb{E}|M_\triangle - \mathbb{E}M_\triangle| + O(n^{-2}),$$

where the $O(n^{-2})$ term accounts for replacing $(1-3/\binom{n}{2})^{\mathbb{E}M_\triangle}$ by $e^{-\theta(n-2)}$. Since

$$\mathrm{Var}(M_\triangle) = \binom{n}{3}\theta = O(n^3p'),$$

we get

$$\mathbb{E}|q_{M_\triangle} - q| = O\left(\sqrt{\frac{p'}{n}}\right) + O(n^{-2}).$$

For two Erdős–Rényi graphs with edge probabilities a, b bounded away from 0 and 1, Pinsker's inequality applied to the product measures gives

$$d_{\mathrm{TV}}(\mathrm{G}(n,a), \mathrm{G}(n,b)) \le Cn|a-b|.$$

Therefore

$$\mathbb{E}d_{\mathrm{TV}}(\mathrm{G}(n,q_{M_\triangle}), \mathrm{G}(n,q)) \le Cn\,\mathbb{E}|q_{M_\triangle} - q| = O(\sqrt{np'}) + O(n^{-1}) = o(1),$$

because $p' \ll n^{-3/2}$.

Combining the Poissonized comparison with this deterministic-density replacement gives

$$d_{\mathrm{TV}}\big(\mathrm{RGT}(n,p,p'), \mathrm{G}(n,q)\big) = o(1),$$

as claimed. □

A.3 Source Notes for the Imported Bridge Theorems

The remaining bridge statements in Chapter 4 are used as source theorems rather than reproved in the thesis text.

Algorithmic decorrelation for planted RGT. Theorem **??** is the two-way average-case reduction between planted dense subgraph in $\mathrm{G}(n,p)$ and planted dense subgraph in $\mathrm{RGT}(n,p,p')$ proved in [**BGP23**]. The forward map adds independent triangle motifs. The reverse map implements the posterior time reversal of the triangle-addition kernel. The proof in [**BGP23**] verifies that these maps preserve the planted edge resampling operation by a perturbation-insensitivity criterion, together with influence, concentration, and mixing estimates for the triangle posterior.

Random intersection graphs and random graphs with triangles. Theorem 4.4 is the random-intersection-to-random- triangles approximation from [**BGP23**]. In the constant-edge-density scaling $m\alpha^2 = \Theta(1)$, the proof Poissonizes the latent features, separates the contributions by realized feature size, and identifies the effective size-2 contribution with an independent edge kernel and the effective size-3 contribution with a triangle kernel. The resulting parameters are

$$p_\triangle = 1 - \exp\big(-m\alpha^3(1-\alpha)^{n-3}\big),$$

and

$$p_{\text{edge}} = 1 - \exp\big(- m\alpha^2 + (n-2)m\alpha^3(1-\alpha)^{n-3}\big),$$

so that

$$1 - (1 - p_{\text{edge}})(1 - p_{\triangle})^{n-2} = 1 - \exp(-m\alpha^2),$$

matching the RIG marginal edge density at the relevant scale. The total variation error is $o(1)$ for $m \gg n^{5/2}$.

Middle-scale random geometric graphs. Theorem 4.5 is the main theorem of [**BBG26MiddleScale**]. It proves that, for the balanced signed spherical random geometric graph,

$$(G_{\text{geo}})_{ij} = \operatorname{sign}\langle Z_i, Z_j\rangle,$$

the law is $o(1)$-close in total variation to the ordered local triangle-sampling chain whenever

$$D \gg n^2 \log^{3/2} n.$$

The proof proceeds through the comparison chain

$$\mathcal{L}(G_{\text{geo}}) \approx G_1 \approx G_2 \approx G_3 \approx G_4 \approx G_{5,\text{reg}} \approx G_6,$$

using the exact Gram–Schmidt coupling, fragile-edge localization, uniformization of fragile innovations, the local noisy-majority to sampling identity, and a final weighted-to-unweighted comparison by an anchor-polymer expansion. The theorem is imported in Chapter 4 as the source result identifying the middle-scale geometric correction with a degree-3 local signed triangle-relation chain.

Appendix B

Technical Proofs for Algorithmic Decorrelation

This appendix gives the deferred proofs for Chapter 5. The proofs follow the algorithmic decorrelation framework of [19], with notation aligned to Chapter 3. Throughout the appendix, $p, q \in (0,1)$ are fixed constants, $p' \le 1/(n \log n)$, and all asymptotics are as $n \to \infty$. Constants implicit in $O(\cdot)$ may depend on p and q, but not on n, k, p', or the planted set S.

For a graph G and an edge $e \in \binom{[n]}{2}$, write

$$G + e = G \cup \{e\}, \qquad G - e = G \setminus \{e\}.$$

For a triangle configuration $X \in \{0,1\}^{\binom{[n]}{3}}$, write

$$E(X) = \bigcup_{T : X_T = 1} \binom{T}{2}, \qquad e(X) = |E(X)|.$$

When $X \sim \mu_G$, define the edge-coverage vector

$$Y_f = \mathbf{1}\{f \in E(X)\}, \qquad f \in \binom{[n]}{2}.$$

B.1 Total Variation Preliminaries

Lemma B.1 (Data processing). *Let P, Q be probability measures on a common space, and let K be a Markov kernel. Then*

$$d_{\mathrm{TV}}(KP, KQ) \le d_{\mathrm{TV}}(P, Q).$$

Proof. For every measurable output event A,

$$(KP)(A) - (KQ)(A) = \int K(x, A)\, d(P - Q)(x).$$

Taking the supremum over A gives the claim. $\square$

Lemma B.2 (Mixture bound)**.** *Let X, X' and Y be random variables such that the conditional laws $\mathcal{L}(X \mid Y)$ and $\mathcal{L}(X' \mid Y)$ are defined on a common space. Then*

$$d_{\mathrm{TV}}(X, X') \le \mathbb{E}_Y\big[d_{\mathrm{TV}}\big(\mathcal{L}(X \mid Y), \mathcal{L}(X' \mid Y)\big)\big].$$

Proof. This is convexity of total variation:

$$\mathcal{L}(X) = \mathbb{E}_Y\, \mathcal{L}(X \mid Y), \qquad \mathcal{L}(X') = \mathbb{E}_Y\, \mathcal{L}(X' \mid Y).$$

□

Lemma B.3 (Decomposition through an intermediate coordinate)**.** *For pairs of random variables (X, Y) and (X', Y'),*

$$d_{\mathrm{TV}}(Y, Y') \le d_{\mathrm{TV}}(X, X') + d_{\mathrm{TV}}\big(Y, \mathbb{E}_{Z\sim\mathcal{L}(X)}\mathcal{L}(Y' \mid X' = Z)\big).$$

Proof. Insert the intermediate law

$$\mathbb{E}_{Z\sim\mathcal{L}(X)}\mathcal{L}(Y' \mid X' = Z).$$

The distance between this law and $\mathcal{L}(Y')$ is at most $d_{\mathrm{TV}}(X, X')$ by Lemma B.1. The result follows by the triangle inequality. □

Lemma B.4 (Likelihood-ratio concentration implies total variation)**.** *Let $Q \ll P$. If*

$$\left|\frac{dQ}{dP}(X) - 1\right| \le \varepsilon$$

with P-probability at least $1 - \delta$, then

$$d_{\mathrm{TV}}(P, Q) \le 2\varepsilon + 2\delta.$$

Proof. Let $A = \{|dQ/dP - 1| > \varepsilon\}$. Then $P(A) \le \delta$. On A^c,

$$\frac{dQ}{dP} \ge 1 - \varepsilon,$$

so

$$Q(A^c) \ge (1-\varepsilon)P(A^c) \ge 1 - \varepsilon - \delta.$$

Thus $Q(A) \le \varepsilon + \delta$, and

$$d_{\mathrm{TV}}(P, Q) = \mathbb{E}_P\left|\frac{dQ}{dP} - 1\right| \le P(A) + Q(A) + \varepsilon \le 2\delta + 2\varepsilon.$$

□

Lemma B.5 (Repeated Markov-kernel perturbation)**.** *Let K be a Markov kernel and let X be a random variable on its state space. Then, for every integer $m \ge 1$,*

$$d_{\mathrm{TV}}(X, K^m X) \le m\, d_{\mathrm{TV}}(X, KX).$$

Proof. By the triangle inequality and Lemma B.1,

$$d_{\mathrm{TV}}(X, K^m X) \le \sum_{j=0}^{m-1} d_{\mathrm{TV}}(K^j X, K^{j+1} X) \le m\, d_{\mathrm{TV}}(X, KX).$$

□

B.2 The Perturbation Framework

Let Res_e^r denote resampling edge e from $\mathrm{Ber}(r)$, independently of all other randomness.

Lemma B.6 (Single-edge commutation)**.** *Under the assumptions of Theorem 5.2, for every $1 \le i \le K$,*

$$d_{\mathrm{TV}}\big(\mathcal{A} \circ \mathrm{Res}_{e_i}^{r_i}(P_{1,i-1}), \mathrm{Res}_{e_i}^{s_i} \circ \mathcal{A}(P_{1,i-1})\big) = O(\varepsilon + \delta),$$

where

$$s_i = r_i p_{e_i}^+ + (1 - r_i) p_{e_i}^-.$$

Proof. Fix i, and write $e = e_i$, $r = r_i$, $s = s_i$, $p^+ = p_e^+$, and $p^- = p_e^-$. Define an auxiliary map $\mathcal{A}_e$ by

$$\mathcal{A}_e(G) = \mathrm{Res}_e^{p^- \mathbf{1}\{e \notin G\} + p^+ \mathbf{1}\{e \in G\}} \big(\mathcal{A}(G - e)_{\sim e}\big).$$

Thus $\mathcal{A}_e$ samples the output away from e using input $G - e$, and then independently inserts edge e with probability p^- or p^+ depending on whether $e \notin G$ or $e \in G$.

If $G \in \mathcal{G}$ and $e \notin G$, then Conditions (C2−) and (C3) imply

$$d_{\mathrm{TV}}(\mathcal{A}(G), \mathcal{A}_e(G)) = O(\varepsilon).$$

If $G \in \mathcal{G}$ and $e \in G$, then Conditions (C2+) and (C3) first give

$$\mathcal{A}(G) \approx_{O(\varepsilon)} \mathcal{A}(G + e)_{\sim e} \times \mathrm{Ber}(p^+),$$

and Condition (C1) replaces $\mathcal{A}(G + e)_{\sim e}$ by $\mathcal{A}(G - e)_{\sim e}$, again with $O(\varepsilon)$ total variation loss. Hence, for every $G \in \mathcal{G}$,

$$d_{\mathrm{TV}}(\mathcal{A}(G), \mathcal{A}_e(G)) = O(\varepsilon).$$

Averaging over $G \sim P_{1,i-1}$ and over $G \sim P_{1,i}$, and using the assumption that both laws put mass at least $1 - \delta$ on $\mathcal{G}$, gives

$$d_{\mathrm{TV}}\big(\mathcal{A}(P_{1,i-1}), \mathcal{A}_e(P_{1,i-1})\big) = O(\varepsilon + \delta),$$

and

$$d_{\mathrm{TV}}\big(\mathcal{A}(P_{1,i}), \mathcal{A}_e(P_{1,i})\big) = O(\varepsilon + \delta).$$

Now sample $G \sim P_{1,i-1}$, and let $G' = \mathrm{Res}_e^r(G)$. Conditional on $G - e$, the final edge inserted by $\mathcal{A}_e(G')$ has probability

$$rp^+ + (1 - r)p^- = s.$$

Therefore

$$\mathcal{A}_e(P_{1,i}) = \mathrm{Res}_e^s \big(\mathcal{A}(P_{1,i-1} - e)_{\sim e}\big).$$

Similarly,

$$\mathrm{Res}_e^s \circ \mathcal{A}_e(P_{1,i-1}) = \mathrm{Res}_e^s \big(\mathcal{A}(P_{1,i-1} - e)_{\sim e}\big).$$

Combining the two $O(\varepsilon + \delta)$ approximation bounds proves the lemma. ∎

Proof of Theorem 5.2. For $0 \le i \le K$,

$$P_{1,i} = \mathrm{Res}_{e_i}^{r_i}(P_{1,i-1}).$$

Lemma B.6 gives

$$d_{\mathrm{TV}}\big(\mathcal{A}(P_{1,i}), \mathrm{Res}_{e_i}^{s_i} \circ \mathcal{A}(P_{1,i-1})\big) = O(\varepsilon + \delta).$$

Iterating over $i = 1, \dots, K$ and using the triangle inequality yields

$$d_{\mathrm{TV}}\big(\mathcal{A}(P_{1,K}), \mathrm{Res}_{e_K}^{s_K} \circ \cdots \circ \mathrm{Res}_{e_1}^{s_1} \circ \mathcal{A}(P_0)\big) = O(K(\varepsilon + \delta)).$$

Since $\mathcal{A}(P_0) = P_0'$, the second law is $\mathrm{GPS}_{\mathbf{e},\mathbf{s}}(P_0')$. This proves the theorem. ☐

B.3 Forward Triangle Addition

Definition B.7 (Uniform two-star density)**.** A graph G is c-uniformly two-star dense if for every pair $u, v \in [n]$,

$$|\{w \in [n] \setminus \{u,v\} : \{u,w\}, \{v,w\} \in E(G)\}| \ge c(n-2).$$

Lemma B.8 (Two-star density under RGT and intermediate planted laws)**.** *For fixed $p \in (0,1)$, $G \sim \mathrm{G}(n,p)$ is $p^2/2$-uniformly two-star dense with probability $1 - e^{-\Omega(n)}$. The same conclusion holds for $G \sim \mathrm{RGT}(n,p,p')$. Moreover, if $o(n)$ edges are modified, the graph remains $p^2/3$-uniformly two-star dense for all sufficiently large n.*

Proof. For a fixed pair u, v, the number of common neighbors in $\mathrm{G}(n,p)$ is $\mathrm{Bin}(n-2, p^2)$. Chernoff's inequality gives

$$\mathbb{P}\left(\mathrm{Bin}(n-2,p^2) \le \frac{p^2}{2}(n-2)\right) \le e^{-\Omega(n)}.$$

A union bound over $O(n^2)$ pairs proves the statement for $\mathrm{G}(n,p)$. Triangle addition only adds edges, so it preserves the property for $\mathrm{RGT}(n,p,p')$. Changing $o(n)$ edges changes the common-neighbor count of any fixed pair by at most $o(n)$, so the lower bound $p^2(n-2)/2$ weakens to $p^2(n-2)/3$ for all large n. ☐

Lemma B.9 (Forward single-edge perturbation)**.** *Let G be c-uniformly two-star dense, let $e = \{u,v\} \notin G$, and let $Z = \mathcal{A}_{\triangle}^{p'}(G)$ with $p' \le 1/n$. Then*

$$d_{\mathrm{TV}}\big(\mathcal{L}(Z_{\sim e}), \mathcal{L}(Z_{\sim e} \mid Z_e = 0)\big) = O(n^{-1/2} \log^{5/2} n).$$

Proof. Let T_e be the set of triangles containing e, and let $X \in \{0,1\}^{\binom{[n]}{3}}$ be the independent triangle indicators used by $\mathcal{A}_{\triangle}^{p'}$. Let X' agree with X off T_e and satisfy $X'_T = 0$ for every $T \in T_e$. Then

$$Z_{\sim e} \overset{d}{=} G \cup E(X) \setminus \{e\}, \qquad (Z_{\sim e} \mid Z_e = 0) \overset{d}{=} G \cup E(X') \setminus \{e\}.$$

We construct an auxiliary triangle set X^{aux} such that

$$G \cup E(X' \vee X^{\mathrm{aux}}) \setminus \{e\} \stackrel{d}{=} G \cup E(X) \setminus \{e\}.$$

For $a, b \in [n]$, write

$$W_G(a,b) = \{w : \{a,w\}, \{b,w\} \in E(G)\}.$$

By uniform two-star density, for each $w \notin \{u,v\}$ one can choose subsets

$$\overline{W}_G(u,w) \subseteq W_G(u,w), \qquad \overline{W}_G(v,w) \subseteq W_G(v,w),$$

each of size at least $cn/3$, so that the resulting candidate triangles have bounded overlap. This follows by greedily discarding at most a constant fraction of candidates involved in mutual conflicts.

For each $w \notin \{u,v\}$, independently with probability p', sample

$$w_1 \sim \mathrm{Unif}(\overline{W}_G(u,w)), \qquad w_2 \sim \mathrm{Unif}(\overline{W}_G(v,w)),$$

and set

$$X^{\mathrm{aux}}_{\{u,w,w_1\}} = X^{\mathrm{aux}}_{\{v,w,w_2\}} = 1.$$

The two added auxiliary triangles create, away from e, exactly the edges that the triangle $\{u,v,w\}$ would have created.

It remains to compare X' and $X' \vee X^{\mathrm{aux}}$. Let K be the kernel that chooses one w, w_1, and w_2 as above and adds the two triangles $\{u,w,w_1\}$ and $\{v,w,w_2\}$. If $M \sim \mathrm{Bin}(n-2, p')$, then $X' \vee X^{\mathrm{aux}}$ differs from $K^M(X')$ only through collisions among the selected w's, which have total probability $O(np'^2)$.

Let P be the law of X' and Q the law of $K(X')$. For a configuration x, let $A(x)$ be the set of triples (w, w_1, w_2) that could have generated x under K. Since X' is a product Bernoulli field,

$$\frac{dQ}{dP}(x) = \frac{|A(x)|}{\mathbb{E}|A(X')|},$$

and

$$\mathbb{E}|A(X')| = \Theta(n^3 p'^2).$$

The variable $|A(X')|$ is a sum over $w \notin \{u,v\}$ of products of two binomial counts. With probability $1 - O(n^{-2})$, all such counts are $O(\log n)$; on this event, changing one w-summand changes $|A(X')|$ by at most $O(\log^2 n)$. McDiarmid's inequality gives

$$||A(X') - \mathbb{E}|A(X')|| = O(n^{3/2} p' \log^{5/2} n)$$

with probability $1 - O(n^{-1})$. Hence

$$\left| \frac{dQ}{dP}(X') - 1 \right| = O\left(\frac{\log^{5/2} n}{n^{3/2} p'} \right)$$

with probability $1 - O(n^{-1})$, and Lemma B.4 gives

$$d_{\mathrm{TV}}(K(X'), X') = O\left(\frac{1}{n} + \frac{\log^{5/2} n}{n^{3/2} p'} \right).$$

By Lemma B.5 and $\mathbb{E}M = O(np')$,

$$d_{\mathrm{TV}}(K^M(X'), X') = O(n^{-1/2}\log^{5/2} n).$$

The conclusion follows by data processing from triangle configurations to graph edge sets. $\square$

Theorem B.10 (Forward map preserves the planted signal)**.** *Let $p'_1 < p'_2 \le 1/n$, and set*

$$p'_\Delta = \frac{p'_2 - p'_1}{1 - p'_1}.$$

Let $\mathcal{A}_\triangle^{p'_\Delta}$ add every triangle independently with probability p'_Δ. If

$$k = o(n^{1/4}\log^{-5/4} n),$$

then, for every fixed $S \subset [n]$ with $|S| = k$,

$$d_{\mathrm{TV}}\Big(\mathcal{A}_\triangle^{p'_\Delta}(\mathrm{RGT}(n, p, p'_1, S, q)), \mathrm{RGT}(n, p, p'_2, S, q')\Big) = o(1),$$

where

$$q' = q + (1-q)\big(1 - (1 - p'_\Delta)^{n-2}\big).$$

Proof. The null map is exact because two independent rounds of triangle addition with probabilities p'_1 and p'_Δ result in triangle probability

$$p'_1 + (1 - p'_1)p'_\Delta = p'_2.$$

We apply Theorem 5.2 to the planted edge sequence $\binom{S}{2}$. The good set $\mathcal{G}$ is the class of $p^2/3$-uniformly two-star dense graphs. By Lemma B.8, every intermediate planted distribution lies in $\mathcal{G}$ with probability $1 - e^{-\Omega(n)}$, since $\binom{k}{2} = o(n)$.

For the forward map, Condition (C1) holds with zero error: changing input edge e changes only output coordinate e. Condition (C3) holds exactly with

$$p_e^- = 1 - (1 - p'_\Delta)^{n-2}, \qquad p_e^+ = 1.$$

Condition (C2+) holds exactly because e is always present in $\mathcal{A}_\triangle(G + e)$. Condition (C2−) follows from Lemma B.9. Therefore Theorem 5.2 applies with

$$\varepsilon = O(n^{-1/2}\log^{5/2} n), \qquad \delta = e^{-\Omega(n)}.$$

Since $K = \binom{k}{2} = o(n^{1/2}\log^{-5/2} n)$, we have $K\varepsilon = o(1)$.

The transformed planted density is

$$qp_e^+ + (1-q)p_e^- = q + (1-q)\big(1 - (1 - p'_\Delta)^{n-2}\big),$$

which proves the theorem. $\square$

B.4 The Reverse Kernel and the Triangle Posterior

Recall that

$$\mu_G(x) = \frac{1}{Z_G}\left(\frac{p'}{1-p'}\right)^{|x|} p^{-e(x)}\mathbf{1}\{E(x) \subseteq E(G)\}.$$

Lemma B.11 (The reverse kernel is the time reversal of triangle addition)**.** *Under exact posterior sampling,*

$$\mathcal{A}_{\mathrm{rev}}^{p,p'}(\mathrm{RGT}(n,p,p')) = \mathrm{G}(n,p).$$

Proof. Let $G_0 \sim \mathrm{G}(n,p)$, let X be the independent triangle set used by the forward map, and let $G = G_0 \cup E(X)$. For a candidate preimage H, Bayes' rule gives

$$\mathbb{P}(G_0 = H \mid G) \propto \sum_x \mathbf{1}\{G = H \cup E(x)\}\left(\frac{p'}{1-p'}\right)^{|x|}\left(\frac{p}{1-p}\right)^{|E(H)|}.$$

The reverse algorithm first samples $X \sim \mu_G$, and then resamples each edge in $E(X)$ from $\mathrm{Ber}(p)$. Its probability of outputting H is proportional to

$$\sum_x \mathbf{1}\{G = H \cup E(x)\}\left(\frac{p'}{1-p'}\right)^{|x|} p^{-e(x)} p^{e(x)-|E(G)|+|E(H)|}(1-p)^{|E(G)|-|E(H)|}.$$

After dropping factors depending only on G, this matches the posterior law of $G_0 \mid G$. Hence the reverse transition samples from the conditional law of the original $\mathrm{G}(n,p)$ preimage. $\square$

Lemma B.12 (Marginal smallness)**.** *For every graph G, the distribution μ_G is $O(p')$-marginally small. More precisely, for any triangle T and any feasible conditioning of all other triangle variables,*

$$\mu_G(X_T = 1 \mid X_{\sim T}) \le C_p p'.$$

If $T \subseteq G$, the same conditional probability is at least $c_p p'$.

Proof. Let x^+ and x^- be configurations agreeing off T, with $x_T^+ = 1$ and $x_T^- = 0$. Whenever both are feasible,

$$\frac{\mu_G(x^+)}{\mu_G(x^-)} = \frac{p'}{1-p'} p^{-(e(x^+)-e(x^-))}.$$

The covered-edge increment $e(x^+) - e(x^-)$ lies in $\{0, 1, 2, 3\}$, so the likelihood ratio is between constant multiples of p', because p is fixed and $p' = o(1)$. This gives the stated upper and lower bounds. $\square$

Lemma B.13 (Posterior edge coverage in two-star dense graphs)**.** *If G is c-uniformly two-star dense and $e \in E(G)$, then*

$$\mu_G(Y_e = 1) = \Theta(np').$$

Proof. Let T_e be the set of triangles in G containing e. Uniform two-star density gives $|T_e| = \Theta(n)$. Enumerate $T_e = \{T_1, \dots, T_m\}$. By Lemma B.12,

$$\mu_G(Y_e = 1) = \sum_{\ell=1}^{m} \mu_G(X_{T_\ell} = 1 \mid X_{T_1} = \cdots = X_{T_{\ell-1}} = 0) = \Theta(mp') = \Theta(np').$$

$\square$

B.5 Influence, Concentration, and Mixing

For a distribution P on $\{0,1\}^{\mathcal{X}}$, define the marginal influence

$$\mathrm{Inf}^{\mathrm{m}}_{A\to B} = \sup_{x_A, x'_A} d_{\mathrm{TV}}\big(P(X_B \in \cdot \mid X_A = x_A), P(X_B \in \cdot \mid X_A = x'_A)\big).$$

The pinned influence $\mathrm{Inf}_{A\to B}$ is defined similarly but also allows arbitrary pinning of variables outside $A \cup B$.

Lemma B.14 (Low influence for marginally small graphical models)**.** *Let P be a distribution on $\{0,1\}^N$ whose graphical model has maximum degree Δ. Suppose P is q-marginally small with $q\Delta = o(1)$. Then influences decay along the graphical model. In particular, if two variables i, j have graph distance r, then*

$$\mathrm{Inf}_{i\to j} = O(q^r),$$

with constants depending only on the local path-counting constants of the graphical model.

Proof. The proof is a disagreement-percolation argument. Couple two conditional samples that differ only at coordinate i. A disagreement can affect coordinate j only if it propagates along a path in the graphical model from i to j. At every step, after arbitrary pinning, marginal smallness implies that the probability of creating a new one-disagreement is $O(q)$. The number of possible next steps is at most Δ, and the assumption $q\Delta = o(1)$ makes the resulting branching process subcritical. Summing over paths of length at least r gives $O(q^r)$, because the triangle-intersection graph used below has uniformly bounded local path multiplicities once the path endpoints are fixed. □

Lemma B.15 (Edge-level marginal influence)**.** *Let G be c-uniformly two-star dense and $p' = o(1/n)$. Under the edge-coverage law $\mathcal{L}_G(Y)$, for any $A \subseteq E(G)$ with $|A| \le cn/4$ and any $e \notin A$,*

$$\mathrm{Inf}^{\mathrm{m}}_{A\to e} = O\left(\frac{|A|}{n^3 p'^2}\right).$$

If $e, e' \in E(G)$ do not share a vertex, then

$$\mathrm{Inf}^{\mathrm{m}}_{e'\to e} = O\left(\frac{1}{n^3 p'}\right).$$

Proof. Let T_A be the set of triangles containing at least one edge of A, and let T_e be the set of triangles containing e. Conditioning on Y_A affects Y_e only through triangle variables in T_A and their influence on T_e.

First consider triangles in $T_A \cap T_e$, which can directly cover e. Since G is c-uniformly two-star dense, each conditioned edge event $Y_a = 1$ has probability $\Theta(np')$, and each particular triangle responsible for it has conditional probability $O(1/(n^2p'))$. Summing over the at most $O(|A|)$ triangles that can directly overlap T_e gives the contribution

$$O\left(\frac{|A|}{n^3 p'^2}\right).$$

For the remaining triangles, the effect must propagate through the triangle-intersection graphical model. Lemma B.14 gives an additional factor $O(p')$ for each graph-distance step. Summing over the $O(n|A|)$ triangles in T_A and the $O(n)$ triangles in T_e, and using $p' = o(1/n)$, yields the same bound

$$O\left(\frac{|A|}{n^3 p'^2}\right).$$

If e and e' do not share a vertex, then $T_e \cap T_{e'} = \emptyset$. The direct-overlap contribution is absent, and the first possible influence path has one additional step. The same summation gives

$$O\left(\frac{1}{n^3 p'}\right).$$

□

Corollary B.16 (Posterior marginal stability)**.** *If G is c-uniformly two-star dense, then for any two distinct edges e, e',*

$$|\mu_G(Y_e = 1) - \mu_{G+e'}(Y_e = 1)| = \widetilde{O}(1/n).$$

Proof. If $e' \in G$, the claim is trivial. Otherwise,

$$\mu_G = \mu_{G+e'}(\,\cdot \mid Y_{e'} = 0).$$

Hence

$$|\mu_G(Y_e = 1) - \mu_{G+e'}(Y_e = 1)| \le \mathrm{Inf}^{\mathrm{m}}_{e' \to e}.$$

The conclusion follows from Lemma B.15; the incident case uses the first bound with $|A| = 1$, and the non-incident case uses the second bound. □

Lemma B.17 (Glauber mixing and concentration)**.** *Let P be q-marginally small on $\{0,1\}^N$, and suppose its graphical model has maximum degree Δ with $q\Delta = o(1)$. Then Glauber dynamics mixes in $O(N \log N)$ steps. Moreover, if f is L-Lipschitz in Hamming distance and $X \sim P$, then*

$$|f(X) - \mathbb{E}f(X)| = O\Big(\sqrt{Nq}\, L \log N\Big)$$

with high probability.

Proof. Couple two Glauber chains started at configurations differing at one coordinate i. Use the same update coordinate. If the update coordinate is i, the chains coalesce. If it is not adjacent to i, the conditional laws are identical. If it is adjacent to i, both conditional probabilities of value one are at most q, so a maximal coupling creates a new disagreement with probability at most $2q$. Therefore the expected Hamming distance contracts by

$$1 - \frac{1 - O(q\Delta)}{N} = 1 - \Omega(1/N).$$

Path coupling gives $O(N \log N)$ mixing. The concentration bound follows from the standard martingale argument along the Glauber chain: run the chain for $T = O(N \log N)$ steps from a fixed initial condition, expose the updates one at a time, and apply bounded differences with conditional variance $O(qL^2)$ per effective update. This gives fluctuations $O(\sqrt{Tq}\, L) = O(\sqrt{Nq}\, L \log N)$, and the mixing error is absorbed into the high-probability term. □

Corollary B.18 (Sampling from μ_G)**.** *For $p' = o(1/n)$ and fixed p, Glauber dynamics for μ_G mixes in $O(n^3 \log n)$ steps.*

Proof. The number of triangle variables is $N = \binom{n}{3}$. The triangle graphical model has maximum degree at most $3(n-2)$. Lemma B.12 gives $q = O(p')$, so $q\Delta = o(1)$. Lemma B.17 applies. $\square$

B.6 Reverse Single-Edge Perturbation

Throughout this section fix an edge $e = \{u, v\} \notin G$, and assume G is c-uniformly two-star dense. Let $T(e)$ be the set of triangles in $G + e$ containing e. For a triangle configuration X, write $X_{T(e)}$ for its coordinates on $T(e)$ and $X_{\sim T(e)}$ for all other triangle coordinates.

Lemma B.19 (Conditioning identity)**.** *For $e \notin G$,*

$$\mathcal{L}_{G-e}(Y_{\sim e}) = \mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 0).$$

Proof. A triangle configuration sampled from μ_{G+e} and conditioned on $Y_e = 0$ uses no triangle covering e. Equivalently, its support is exactly the set of triangle configurations feasible in $G - e = G$, with the same Gibbs weight. Projecting to $Y_{\sim e}$ proves the identity. $\square$

Lemma B.20 (Non-containing triangles are insensitive)**.** *Let $p' = 1/(n \log n)$. Let*

$$X \sim \mu_{G+e}(\,\cdot \mid Y_e = 0) = \mu_G, \qquad X^+ \sim \mu_{G+e}(\,\cdot \mid Y_e = 1).$$

Then

$$d_{\mathrm{TV}}(X_{\sim T(e)}, X^+_{\sim T(e)}) = O\left(\frac{\log^2 n}{\sqrt{n}}\right).$$

Proof. First compare $X_{\sim T(e)}$ with the marginal of μ_{G+e} without conditioning on $Y_e = 1$. This marginal is a mixture over $X_{T(e)}$. For two independent copies $X_{T(e)}$ and $\widetilde{X}_{T(e)}$, the chi-square divergence of this mixture from μ_G is bounded by

$$\mathbb{E} \min \left\{ \left(1 + C \sup_{A' \subseteq A \cup B,\ f \in (A \cup B) \setminus A'} \mathrm{Inf}^{\mathrm{m}}_{A' \to f}\right)^{|B|} p^{-|A \cap B|}, p^{-|A| - |B|} \right\} - 1,$$

where

$$A = E(\widetilde{X}_{T(e)}) \setminus \{e\}, \qquad B = E(X_{T(e)}) \setminus \{e\}.$$

This identity is obtained by expanding the chi-square divergence of a mixture of Gibbs measures and using the edge-coverage influence bound to control the conditional normalizing constants.

By Lemma B.12, $|X_{T(e)}|$ and $|\widetilde{X}_{T(e)}|$ are stochastically dominated by $\mathrm{Bin}(n, Cp')$. Hence $|A|, |B| = O(\log n)$ with probability $1 - o(n^{-1})$, and $A \cap B = \emptyset$ except with probability $O(\log^2 n/n)$. On this event, Lemma B.15 gives

$$\sup_{A' \subseteq A \cup B,\ f \in (A \cup B) \setminus A'} \mathrm{Inf}^{\mathrm{m}}_{A' \to f} = \widetilde{O}(1/n).$$

Therefore the chi-square divergence is $O(\log^2 n/n)$, and Pinsker's inequality gives

$$d_{\mathrm{TV}}\big(X_{\sim T(e)}, \mu_{G+e}(X_{\sim T(e)} \in \cdot)\big) = O\left(\frac{\log n}{\sqrt{n}}\right).$$

Finally, by Lemma B.13,

$$\mu_{G+e}(Y_e = 1) = \Theta(np') = \Theta(1/\log n).$$

Since μ_{G+e} is a mixture of the laws conditioned on $Y_e = 0$ and $Y_e = 1$, passing from the unconditioned comparison to the $Y_e = 1$ comparison costs a factor $O((np')^{-1}) = O(\log n)$. This proves the displayed bound. □

Define the reweighted triangle distribution

$$\mu^*_{G+e}(x) = \frac{1}{Z^*_{G+e}}\left(\frac{p'}{1-p'}\right)^{|x|} p^{-|E(x)\setminus\{e\}|}\mathbf{1}\{E(x) \subseteq E(G+e)\}.$$

Let $X^* \sim \mu^*_{G+e}$, and let Y^* be its edge-coverage vector.

Lemma B.21 (Productization by reweighting)**.** *Conditional on $X^*_{\sim T(e)}$, the variables $\{X^*_T : T \in T(e)\}$ are independent. Moreover,*

$$\mu^*_{G+e} = \frac{p\,\mu_{G+e}(Y_e = 1)\mu_{G+e}(\,\cdot \mid Y_e = 1) + \mu_{G+e}(Y_e = 0)\mu_{G+e}(\,\cdot \mid Y_e = 0)}{p\,\mu_{G+e}(Y_e = 1) + \mu_{G+e}(Y_e = 0)}.$$

Consequently,

$$d_{\mathrm{TV}}(\mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 0), \mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 1)) \le O(\log n)\, d_{\mathrm{TV}}(Y_{\sim e}, Y^*_{\sim e}),$$

where $Y \sim \mathcal{L}_{G+e}(\,\cdot \mid Y_e = 0)$.

Proof. The only difference between μ^*_{G+e} and μ_{G+e} is the treatment of the edge e. If $Y_e = 1$, then μ^*_{G+e} has an extra factor p relative to μ_{G+e}; if $Y_e = 0$, the weights coincide. This gives the displayed mixture identity.

Conditional on $X^*_{\sim T(e)}$, the triangles in $T(e)$ share only the edge e, and e has been removed from the exponent $p^{-|E(x)\setminus\{e\}|}$. Hence the conditional weight factorizes over $T \in T(e)$, proving conditional independence.

Solving the mixture identity for $\mu_{G+e}(\cdot \mid Y_e = 1)$ and using $\mu_{G+e}(Y_e = 1) = \Theta(1/\log n)$ gives the factor $O(\log n)$. □

Lemma B.22 (Auxiliary replacement)**.** *Let $X \sim \mu_{G+e}(\cdot \mid Y_e = 0) = \mu_G$. Generate X'' by first sampling $X_{\sim T(e)}$ according to X, and then sampling $X''_{T(e)}$ from the conditional law of $X^*_{T(e)} \mid X^*_{\sim T(e)} = X_{\sim T(e)}$. Let Y'' be the edge-coverage vector of X''. Then, for $p' = 1/(n\log n)$,*

$$d_{\mathrm{TV}}(Y_{\sim e}, Y''_{\sim e}) = O\left(\frac{\log^{15/2} n}{\sqrt{n}}\right).$$

Proof. We construct auxiliary triangles that reproduce, away from e, the edge effect of the triangles in $X''_{T(e)}$. Let

$$V_v(X) = \{w : \{u,w\} \in E(X),\ \{v,w\} \in G \setminus E(X)\},$$

$$V_u(X) = \{w : \{v,w\} \in E(X),\ \{u,w\} \in G \setminus E(X)\},$$

and

$$V_{uv}(X) = \{w : \{u,w\}, \{v,w\} \in G \setminus E(X)\}.$$

For an edge $\{v,w\}$ newly created by a triangle in $X''_{T(e)}$, choose

$$w' \in W_{E(X)}(v,w) = \{z : \{v,z\}, \{w,z\} \in E(X)\}$$

uniformly and add the triangle $\{v,w,w'\}$. This creates $\{v,w\}$ without creating other new edges outside $E(X)$. The analogous rule handles new edges $\{u,w\}$. If both $\{u,w\}$ and $\{v,w\}$ are new, add one auxiliary triangle through u,w and one through v,w.

This auxiliary construction can be implemented by three kernels K_u, K_v, K_{uv}. The kernels K_u and K_v add one triangle, while K_{uv} adds two triangles. The number of applications of each kernel is stochastically dominated by a binomial random variable with mean $O(np')$, because each triangle in $T(e)$ is present with conditional probability $O(p')$.

We now bound the total variation cost of one application. For K_v, conditioning on all triangles not incident to u or v, the likelihood ratio between $K_v(X)$ and X is

$$L_v(x) = \sum_{\substack{t=\{v,w,w'\}:t\in x \\ w\in V_v(x-t),\ w'\in W_{E(x-t)}(v,w)}} \frac{1}{|V_v(x-t)|\,|W_{E(x-t)}(v,w)|}\frac{\mu_G(x-t)}{\mu_G(x)}.$$

For every summand,

$$\frac{\mu_G(x-t)}{\mu_G(x)} = \Theta(1/p')$$

because adding t increases the triangle count by one and creates one new edge. Uniform two-star density and Lemma B.17 imply that, with high probability,

$$|V_v(X)| = \Theta(n^2p'), \qquad |W_{E(X)}(v,w)| = \Theta(n^3p'^2),$$

and the number of summands has the corresponding concentrated mean. Consequently

$$|L_v(X) - 1| = O\left(\frac{\log n}{n^4 p'^{7/2}}\right)$$

with high probability. Lemma B.4 gives

$$d_{\mathrm{TV}}(K_v(X), X) = O\left(\frac{\log n}{n^4 p'^{7/2}}\right).$$

The same proof gives the same bound for K_u.

For K_{uv}, the likelihood ratio contains two wedge-completion denominators and one V_{uv} denominator. The concentrated scales are

$$|V_{uv}(X)| = \Theta(n), \qquad |W_{E(X)}(u,w)| = |W_{E(X)}(v,w)| = \Theta(n^3 p'^2),$$

and the same likelihood-ratio concentration gives

$$d_{\mathrm{TV}}(K_{uv}(X), X) = O\left(\frac{\log^3 n}{n^6 p'^{11/2}}\right).$$

By Lemma B.5 and the $O(np')$ expected number of applications,

$$d_{\mathrm{TV}}(Y_{\sim e}, Y''_{\sim e}) \le O(np') \left[O\left(\frac{\log n}{n^4 p'^{7/2}}\right) + O\left(\frac{\log^3 n}{n^6 p'^{11/2}}\right)\right].$$

At $p' = 1/(n \log n)$, the second term dominates and equals

$$O\left(\frac{\log^{15/2} n}{\sqrt{n}}\right).$$

□

Lemma B.23 (Reverse perturbation estimate)**.** *Let G be c-uniformly two-star dense and let $p' = 1/(n \log n)$. Then*

$$d_{\mathrm{TV}}(\mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 0), \mathcal{L}_{G+e}(Y_{\sim e} \mid Y_e = 1)) = O\left(\frac{\log^{17/2} n}{\sqrt{n}}\right).$$

Consequently,

$$d_{\mathrm{TV}}(\mathcal{L}_{G+e}(Y), \mathcal{L}_{G+e}(Y_{\sim e}) \times \mathcal{L}_{G+e}(Y_e)) = O\left(\frac{\log^{17/2} n}{\sqrt{n}}\right).$$

Proof. By Lemma B.21, it suffices to compare $Y_{\sim e}$ under $\mu_G = \mu_{G+e}(\cdot \mid Y_e = 0)$ with $Y^*_{\sim e}$. Lemma B.20 controls the non-containing triangle variables, and Lemma B.22 controls the additional edge effect of the triangles in $T(e)$ after reweighting. Combining the two terms using Lemma B.3, and then multiplying by the $O(\log n)$ factor from Lemma B.21, gives

$$O\left(\log n \left[\frac{\log^2 n}{\sqrt{n}} + \frac{\log^{15/2} n}{\sqrt{n}}\right]\right) = O\left(\frac{\log^{17/2} n}{\sqrt{n}}\right).$$

The final product-form statement follows because a joint law whose two conditional laws given $Y_e = 0$ and $Y_e = 1$ are within η total variation is within η of the product of its marginal on Y_e and its marginal on $Y_{\sim e}$. □

B.7 The High-Probability Good Graph Class

For the reverse proof at $p'_\star = 1/(n \log n)$, define

$$\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2,$$

where $\mathcal{G}_1$ is the class of $p^2/3$-uniformly two-star dense graphs, and $\mathcal{G}_2$ is the class of graphs satisfying, for every edge e,

$$\left|\mu_{G+e}(Y_e = 1) - p_e^+\right| \le C_p n^{-5/2} p'^{-2} \sqrt{\log n},$$

with

$$p_e^+ = \mathbb{E}_{G' \sim \mathrm{RGT}(n,p,p')} \mu_{G'+e}(Y_e = 1).$$

Lemma B.24 (Concentration of posterior edge marginals)**.** *If $G \sim \mathrm{RGT}(n,p,p')$ with $p' = 1/(n \log n)$, then*

$$\mathbb{P}(G \in \mathcal{G}_2) = 1 - o(1).$$

Proof. Represent $G \sim \mathrm{RGT}(n,p,p')$ as a function of independent base-edge indicators $x \sim \mathrm{Ber}(p)^{\binom{n}{2}}$ and independent triangle indicators $z \sim \mathrm{Ber}(p')^{\binom{n}{3}}$. For fixed e, set

$$F_e(x,z) = \mu_{\mathrm{RGT}(x,z)+e}(Y_e = 1).$$

On the event that $\mathrm{RGT}(x,z)$ is uniformly two-star dense, Corollary B.16 and Lemma B.15 imply the Lipschitz bounds

$$\left|\Delta_{x_{e'}} F_e\right| = \begin{cases} O(1/(n^3 p'^2)), & e' \cap e \neq \emptyset, \\ O(1/(n^3 p')), & e' \cap e = \emptyset, \end{cases}$$

and the same bounds for changing a triangle indicator, according to whether the triangle shares a vertex with e. The bounded-differences variance proxy is therefore

$$\sigma^2 = O(n^{-5} p'^{-4}).$$

The complement of the two-star dense event has probability $e^{-\Omega(n)}$. McDiarmid's inequality gives

$$\mathbb{P}\Big(|F_e - \mathbb{E} F_e| > C_p n^{-5/2} p'^{-2} \sqrt{\log n}\Big) \le n^{-C}$$

for arbitrarily large fixed C, by choosing C_p sufficiently large. A union bound over $O(n^2)$ edges proves the lemma. □

Lemma B.25 (Good graphs under intermediate planted laws)**.** *Fix $S \subset [n]$ with $|S| = k$, and let $\mathrm{RGT}_i(n,p,p',S,q)$ be the law obtained from $\mathrm{RGT}(n,p,p')$ by resampling the first i edges of $\binom{S}{2}$ from $\mathrm{Ber}(q)$. If*

$$k = o(n^{1/4} \log^{-17/4} n),$$

then, uniformly over $0 \le i \le \binom{k}{2}$,

$$\mathbb{P}_{G \sim \mathrm{RGT}_i(n,p,p',S,q)}(G \in \mathcal{G}) = 1 - o(1/n).$$

Proof. For $\mathcal{G}_1$, Lemma B.8 applies because at most $\binom{k}{2} = o(n)$ edges are resampled.

For $\mathcal{G}_2$, Lemma B.24 gives the result under the null. If G and G' differ in i edges while both are in $\mathcal{G}_1$, Corollary B.16 gives

$$|\mu_{G+e}(Y_e = 1) - \mu_{G'+e}(Y_e = 1)| = \widetilde{O}(i/n).$$

Here $i \le \binom{k}{2} = o(n^{1/2} \log^{-17/2} n)$, so this perturbation is

$$o(n^{-1/2} \log^{5/2} n) = o\Big(n^{-5/2} p'^{-2} \sqrt{\log n}\Big)$$

at $p' = 1/(n \log n)$. Hence $\mathcal{G}_2$ also holds uniformly over the intermediate planted laws. $\square$

B.8 The Reverse Reduction

Theorem B.26 (Reverse map preserves the planted signal)**.** *Let $p' = 1/(n \log n)$, $0 < p < q < 1$, and*

$$k = o(n^{1/4} \log^{-17/4} n).$$

For every fixed $S \subset [n]$ with $|S| = k$,

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\mathrm{rev}}^{p,p'}(\mathrm{RGT}(n, p, p', S, q)), \mathrm{PDS}(n, p, S, qp_e)\Big) = o(1),$$

where

$$p_e = \mathbb{E}_{G \sim \mathrm{RGT}(n,p,p')} \mathbb{P}\big(e \in \mathcal{A}_{\mathrm{rev}}^{p,p'}(G + e)\big) = 1 - O(np').$$

Proof. The null map is exact by Lemma B.11. We apply Theorem 5.2 to the planted edge sequence $\binom{S}{2}$. Let $K = \binom{k}{2}$. By Lemma B.25, the good set $\mathcal{G}$ has probability $1 - o(1/n)$ under each intermediate planted law; this is $1 - o(1/K)$ in the present regime.

For $G - e$, no posterior triangle can cover e, so

$$\mathcal{A}_{\mathrm{rev}}^{p,p'}(G - e) = \mathcal{A}_{\mathrm{rev}}^{p,p'}(G - e)_{\sim e} \times \mathrm{Ber}(0).$$

Thus $p_e^- = 0$, and Condition (C2−) is exact.

For $G + e$, Lemmas B.19 and B.23 imply Condition (C1) for the edge-coverage vector Y. Since the reverse map is obtained from Y by independently resampling covered edges from $\mathrm{Ber}(p)$, Lemma B.1 transfers the same bound to the output graph:

$$\varepsilon = O\left(\frac{\log^{17/2} n}{\sqrt{n}}\right).$$

The product statement in Lemma B.23 gives Condition (C2+).

For Condition (C3), note that

$$\mathbb{P}\big(e \in \mathcal{A}_{\mathrm{rev}}^{p,p'}(G + e)\big) = 1 - (1 - p)\mu_{G+e}(Y_e = 1).$$

The definition of $\mathcal{G}_2$ therefore implies that this marginal is within

$$O(n^{-5/2} p'^{-2} \sqrt{\log n}) = O(n^{-1/2} \log^{5/2} n)$$

of

$$p_e = \mathbb{E}_{G' \sim \mathrm{RGT}(n,p,p')} \mathbb{P}\big(e \in \mathcal{A}_{\mathrm{rev}}^{p,p'}(G' + e)\big).$$

This is smaller than the displayed ε.

The perturbation theorem gives

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\mathrm{rev}}^{p,p'}(\mathrm{RGT}(n,p,p',S,q)), \mathrm{PDS}(n,p,S,qp_e)\Big) = O\left(K \frac{\log^{17/2} n}{\sqrt{n}}\right) + o(1).$$

Since

$$K \frac{\log^{17/2} n}{\sqrt{n}} = O\left(k^2 \frac{\log^{17/2} n}{\sqrt{n}}\right) = o(1),$$

the desired total variation bound follows.

Finally, by Lemma B.13,

$$\mu_{G+e}(Y_e = 1) = O(np')$$

with high probability under $G \sim \mathrm{RGT}(n,p,p')$, and the failure event has negligible contribution. Therefore

$$1 - p_e = (1-p)\,\mathbb{E}\mu_{G+e}(Y_e = 1) = O(np'),$$

which proves $p_e = 1 - O(np')$. □

B.9 Completion of the RGT–ER Equivalence

Proof of Theorem 5.4. The forward direction follows from Theorem B.10 with $p_1' = 0$ and $p_2' = p'$. Since the theorem in Chapter 5 assumes

$$k = o(n^{1/4} \log^{-17/4} n),$$

the forward condition $k = o(n^{1/4} \log^{-5/4} n)$ is automatically satisfied. Thus

$$\mathcal{A}_{\triangle}^{p'}(\mathrm{G}(n,p)) = \mathrm{RGT}(n,p,p')$$

exactly, and

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\triangle}^{p'}(\mathrm{PDS}(n,p,S,q)), \mathrm{RGT}(n,p,p',S,f_{p'}(q))\Big) = o(1),$$

where

$$f_{p'}(q) = q + (1-q)\big(1 - (1-p')^{n-2}\big) = q + O(np') = q + o(1).$$

For the reverse direction at $p' = p'_\star = 1/(n \log n)$, Theorem B.26 gives

$$d_{\mathrm{TV}}\Big(\mathcal{A}_{\mathrm{rev}}^{p,p'_\star}(\mathrm{RGT}(n,p,p'_\star,S,q)), \mathrm{PDS}(n,p,S,qp_e)\Big) = o(1),$$

with

$$p_e = 1 - O(np'_\star) = 1 + o(1).$$

The null mapping is exact under exact posterior sampling by Lemma B.11; using Glauber dynamics with total variation error $o(1)$ gives the stated randomized polynomial-time version.

For $p' < p'_\star$, first apply the forward triangle-addition map with

$$p'_\Delta = \frac{p'_\star - p'}{1 - p'}$$

so that the triangle density becomes $p'_\star$. The planted density becomes

$$q_\Delta = q + (1-q)\big(1 - (1-p'_\Delta)^{n-2}\big).$$

Then apply the reverse map at $p'_\star$, obtaining

$$g_{p'}(q) = q_\Delta p_e = \left[q + (1-q)\big(1 - (1-p'_\Delta)^{n-2}\big)\right] p_e = q + o(1).$$

The maps do not depend on S, and all bounds are uniform over fixed S. Mixing over a uniformly random planted set therefore preserves the same $o(1)$ total variation bounds. The same argument applies to the joint law (S, G).

Finally, both $f_{p'}$ and $g_{p'}$ perturb fixed $q \in (0,1)$ by $o(1)$, uniformly in the stated regime. Hence

$$f_{p'}(g_{p'}(q)) = q + o(1), \qquad g_{p'}(f_{p'}(q)) = q + o(1).$$

This proves Theorem 5.4. □

B.10 Polynomial-Time Implementation

The ideal reverse map uses an exact sample from μ_G. By Corollary B.18, Glauber dynamics mixes in $O(n^3 \log n)$ steps when $p' = o(1/n)$, in particular at $p' = 1/(n \log n)$. Running the chain long enough to obtain total variation error $\eta = o(1)$ perturbs both the null and planted output laws by at most $o(1)$. Taking $\eta = o(K^{-1})$, where $K = \binom{k}{2} \le n^2$, keeps the perturbation negligible in the planted proof and still requires only polynomial time. The forward map is plainly polynomial time by enumerating $\binom{n}{3}$ triangles.

Thus both maps in Theorem 5.4 are randomized polynomial-time maps.

Appendix C

Technical Proofs for Local Recovery from Hypergraph Projections

This appendix proves the results stated in Chapter 6. Throughout, $d \geq 3$ is fixed unless stated otherwise, and

$$H \sim H_d(n,p), \qquad p = (c + o(1))n^{-d+1+\delta}, \qquad c \in (0,\infty).$$

We write

$$N_d = \binom{n}{d}.$$

For a fixed finite d-uniform hypergraph K, write $e(K) = |E(K)|$, $v(K) = |V(K)|$, and

$$m(K) = \max_{K' \subseteq K} \frac{e(K')}{v(K')}.$$

C.1 Preliminary Estimates

Lemma C.1 (One-hit binomial estimate)**.** *Let $X_n \sim \mathrm{Bin}(m_n, p_n)$, where $m_n \to \infty$ and $m_n p_n \to 0$. Then*

$$\mathbb{P}(X_n \geq 1) = (1 + o(1))m_n p_n.$$

Moreover,

$$(1 - o(1))m_n p_n \leq \mathbb{P}(X_n \geq 1) \leq m_n p_n.$$

Proof. The upper bound is the union bound:

$$\mathbb{P}(X_n \geq 1) \leq \mathbb{E}X_n = m_n p_n.$$

For the lower bound,

$$\mathbb{P}(X_n \geq 1) \geq \mathbb{P}(X_n = 1) = m_n p_n (1 - p_n)^{m_n - 1}.$$

Since $m_n p_n \to 0$, we have

$$(1 - p_n)^{m_n - 1} = \exp\{-(1 + o(1))m_n p_n\} = 1 - o(1).$$

Combining the two inequalities gives the claim. $\square$

Lemma C.2 (Fixed subhypergraph threshold)**.** *Let K be a fixed d-uniform hypergraph. Then*

$$\mathbb{P}(K \subseteq H) = \begin{cases} o(1), & p = o(n^{-1/m(K)}), \\ 1 - o(1), & p = \omega(n^{-1/m(K)}), \\ \Omega(1), & p = \Theta(n^{-1/m(K)}). \end{cases}$$

Equivalently, under the parameterization $p = (c + o(1))n^{-d+1+\delta}$, the appearance exponent is

$$\delta(K) = d - 1 - \frac{1}{m(K)}.$$

Proof. Let X_K be the number of labeled embeddings of K into H. Since K is fixed,

$$\mathbb{E}X_K = \Theta(n^{v(K)}p^{e(K)}).$$

More generally, for every nonempty $K' \subseteq K$,

$$\mathbb{E}X_{K'} = \Theta(n^{v(K')}p^{e(K')}).$$

If $p = o(n^{-1/m(K)})$, then for some subgraph $K_* \subseteq K$ attaining $m(K)$,

$$n^{v(K_*)}p^{e(K_*)} = \left(n^{1/m(K)}p\right)^{e(K_*)} = o(1).$$

Thus $\mathbb{E}X_{K_*} = o(1)$, and Markov's inequality gives

$$\mathbb{P}(K_* \subseteq H) = o(1).$$

Since $K \subseteq H$ implies $K_* \subseteq H$, we get $\mathbb{P}(K \subseteq H) = o(1)$.

If $p = \omega(n^{-1/m(K)})$, then for every nonempty $K' \subseteq K$,

$$n^{v(K')}p^{e(K')} = \left(n^{v(K')/e(K')}p\right)^{e(K')} \geq \left(n^{1/m(K)}p\right)^{e(K')} \to \infty.$$

The standard second-moment calculation for fixed subhypergraphs gives

$$\frac{\mathrm{Var}(X_K)}{(\mathbb{E}X_K)^2} = o(1).$$

Indeed, pairs of embeddings overlap in a fixed subhypergraph K', and the contribution of such overlaps is bounded by $O((n^{v(K')}p^{e(K')})^{-1}) = o(1)$. Hence $X_K > 0$ with probability $1 - o(1)$.

If $p = \Theta(n^{-1/m(K)})$, then $\mathbb{E}X_K = \Theta(1)$ for strictly balanced K, and otherwise the usual Poisson or second-moment lower bound applied to an $m(K)$-densest subgraph gives a nonvanishing probability of appearance. This proves the threshold statement. $\square$

C.2 Uniform Partial Recovery

C.2.1 Density of projected d-cliques

Let

$$q_n = \mathbb{P}\left(\binom{[d]}{2} \subseteq E(\mathrm{Proj}(H))\right).$$

Lemma C.3 (Projected clique density)**.** *We have*

$$q_n = (1+o(1))\left[p + \left(\frac{c\,n^{\delta-1}}{(d-2)!}\right)^{\binom{d}{2}}\right].$$

Consequently,

$$q_n = \begin{cases} (1+o(1))p, & \delta < \dfrac{d-1}{d+1}, \\ p + \Theta(p), & \delta = \dfrac{d-1}{d+1}, \\ \omega(p), & \delta > \dfrac{d-1}{d+1}. \end{cases}$$

Proof. First, $[d]$ is a clique in the projection if $[d]$ itself is a hyperedge, which occurs with probability p.

Now condition on $[d] \notin E(H)$. For a pair $e = \{i, j\} \subset [d]$, let A_e be the event that some hyperedge $h \neq [d]$ contains e. The number of candidate hyperedges containing e is

$$\binom{n-2}{d-2} - 1,$$

and therefore, by Lemma C.1,

$$\mathbb{P}(A_e) = (1+o(1))\binom{n-2}{d-2}p = (1+o(1))\frac{c\,n^{\delta-1}}{(d-2)!}.$$

If all $\binom{d}{2}$ pairs are covered by hyperedges whose intersections with $[d]$ are exactly those pairs, the contribution is

$$(1+o(1))\left(\frac{c\,n^{\delta-1}}{(d-2)!}\right)^{\binom{d}{2}},$$

because the relevant candidate hyperedge sets are disjoint up to $O(n^{d-3})$ overlaps, and these overlaps contribute lower-order terms.

It remains to show that no other cover contributes a larger order than the two displayed contributions. Any realization of the clique determines a cover U of $\binom{[d]}{2}$ by subsets $u \subseteq [d]$ with $2 \leq |u| \leq d$: a set u appears if some generating hyperedge intersects $[d]$ in exactly u. If $u = [d]$, the contribution is p. For $u \neq [d]$, the probability of realizing a piece with intersection u is

$$(1+o(1))\binom{n-d}{d-|u|}p = n^{1+\delta-|u|+o(1)}.$$

Thus a cover U contributes at most

$$n^{\sum_{u\in U}(1+\delta-|u|)+o(1)}.$$

For covers not using $[d]$, the following convex relaxation bounds the exponent. Let

$$M = |U|, \qquad y_u = \binom{|u|}{2}.$$

Since the cover covers all $\binom{d}{2}$ edges,

$$\sum_{u\in U} y_u \geq \binom{d}{2}.$$

Also

$$|u| = \frac{1+\sqrt{1+8y_u}}{2}.$$

For fixed M, the exponent is at most

$$M(1+\delta) - \sum_{u\in U} \frac{1+\sqrt{1+8y_u}}{2}.$$

The function $-\sqrt{1+8y}$ is convex, so this maximum over

$$y_u \geq 1, \qquad \sum_{u\in U} y_u \geq \binom{d}{2}$$

is attained at a vertex, namely with $M-1$ of the y_u's equal to 1 and the remaining one equal to $\binom{d}{2} - M + 1$. The resulting upper bound is

$$\phi(M) = M(1+\delta) - (M-1)\cdot 2 - \frac{1+\sqrt{1+8(\binom{d}{2}-M+1)}}{2}.$$

As a function of M, ϕ is convex on the interval $1 \leq M \leq \binom{d}{2}$. Thus its maximum over integer M is attained at an endpoint. The endpoint $M=1$ corresponds to the forbidden full set $[d]$ and gives the true-hyperedge term p. The endpoint $M = \binom{d}{2}$ is the individual-edge cover and gives exponent

$$\binom{d}{2}(\delta-1).$$

Intermediate covers are lower order compared with the larger of p and the individual-edge cover. Hence the displayed asymptotic formula follows.

Finally,

$$\frac{\left(cn^{\delta-1}/(d-2)!\right)^{\binom{d}{2}}}{p} = n^{\binom{d}{2}(\delta-1)+d-1-\delta+o(1)}.$$

The exponent is negative, zero, or positive according as

$$\delta \lesseqgtr \frac{d-1}{d+1}.$$

This proves the three regimes. $\square$

C.2.2 Achievability

Proof of achievability in Theorem 6.3. Let H_c be the output of Algorithm 3. Since every true hyperedge creates a d-clique in Ψ,

$$E(H) \subseteq E(H_c).$$

Therefore

$$|E(H_c) \triangle E(H)| = |E(H_c) \setminus E(H)|.$$

By symmetry,

$$\mathbb{E}|E(H_c)| = \binom{n}{d} q_n, \qquad \mathbb{E}|E(H)| = \binom{n}{d} p.$$

Thus

$$\ell(H_c, H) = \frac{\binom{n}{d}(q_n - p)}{\binom{n}{d} p} = \frac{q_n - p}{p}.$$

If $\delta < (d-1)/(d+1)$, Lemma C.3 gives $q_n = (1 + o(1))p$. Hence

$$\ell(H_c, H) = o(1).$$

The same estimator can be applied after forgetting weights in $\mathrm{Proj}_{\mathrm{W}}(H)$, so the weighted achievability statement follows immediately. □

C.2.3 Bayes risk and posterior overlap

Let B^* be the Bayes-optimal partial recovery estimator from $\mathrm{Proj}(H)$, and let H' be an independent posterior copy of H given $\Psi = \mathrm{Proj}(H)$:

$$H' \sim \mathcal{L}(H \mid \Psi), \qquad H' \perp H \mid \Psi.$$

Lemma C.4 (Posterior-overlap identity)**.** *For the Bayes-optimal partial recovery loss,*

$$\ell(B^*, H) \geq 1 - \frac{\mathbb{E}[|E(H) \cap E(H')|]}{pN_d}.$$

Consequently, if

$$\mathbb{E}[|E(H) \cap E(H')|] = o(pN_d),$$

then

$$\ell(B^*, H) = 1 - o(1).$$

Proof. Condition on $\Psi = G$. For $h \in \binom{[n]}{d}$, define

$$\pi_h(G) = \mathbb{P}(h \in E(H) \mid \Psi = G).$$

The Bayes rule includes h exactly when $\pi_h(G) \geq 1/2$. Its conditional error contribution at h is therefore

$$\min\{\pi_h(G), 1 - \pi_h(G)\}.$$

Hence
$$\mathbb{E}\big[|B^*(G)\triangle E(H)| \mid \Psi = G\big] = \sum_{h\in\binom{[n]}{d}} \min\{\pi_h(G), 1-\pi_h(G)\}.$$
Since $0 \le \pi_h(G) \le 1$,
$$\min\{\pi_h(G), 1-\pi_h(G)\} \ge \pi_h(G) - \pi_h(G)^2.$$
Averaging over G,
$$\mathbb{E}|B^*(\Psi)\triangle E(H)| \ge \sum_h \mathbb{E}\big[\pi_h(\Psi) - \pi_h(\Psi)^2\big].$$
Now
$$\sum_h \mathbb{E}\pi_h(\Psi) = \mathbb{E}|E(H)| = pN_d.$$
Also, conditional independence of H and H' given Ψ gives
$$\mathbb{P}(h \in E(H), h \in E(H') \mid \Psi) = \pi_h(\Psi)^2,$$
so
$$\sum_h \mathbb{E}\pi_h(\Psi)^2 = \mathbb{E}[|E(H)\cap E(H')|].$$
Dividing by pN_d proves the result. $\square$

Lemma C.5 (High-overlap posterior samples are negligible above threshold)**.** *Assume*
$$\delta > \frac{d-1}{d+1}.$$
Then
$$\mathbb{E}[|E(H)\cap E(H')|] = o(pN_d).$$
The same statement holds if H' is sampled from the posterior conditioned on $\mathrm{Proj}_{\mathrm{W}}(H)$*.*

Proof. We give the unweighted proof first. Let
$$q = q_n = \mathbb{P}\left(\binom{[d]}{2} \subseteq E(\mathrm{Proj}(H))\right).$$
By Lemma C.3,
$$\frac{q}{p} = n^{\eta+o(1)}$$
for some $\eta > 0$.

Let H_c be the d-clique hypergraph of Ψ. Since $|E(H)| \sim \mathrm{Bin}(N_d, p)$,
$$|E(H)| = (1+o(1))pN_d$$
with high probability. Similarly, standard bounded-dependency concentration for clique counts in this sparse projected model gives
$$|E(H_c)| = (1+o(1))qN_d$$

with high probability. Fix a deterministic sequence $\epsilon_n \downarrow 0$ for which the event

$$\mathcal{Z} = \{|E(H)| \in [(1-\epsilon_n)pN_d, (1+\epsilon_n)pN_d], \quad |E(H_c)| \in [(1-\epsilon_n)qN_d, (1+\epsilon_n)qN_d]\}$$

has probability $1 - o(1)$.

By Cauchy–Schwarz,

$$\mathbb{E}\big[|E(H) \cap E(H')|\mathbf{1}_{\mathcal{Z}^c}\big] \le \big(\mathbb{E}|E(H)|^2\big)^{1/2}\mathbb{P}(\mathcal{Z}^c)^{1/2} = o(pN_d),$$

after slightly enlarging ϵ_n if necessary.

It remains to control the overlap on $\mathcal{Z}$. Fix $M \in (0,1)$. We show

$$\mathbb{P}\big(H, H' \in \mathcal{Z},\ |E(H) \cap E(H')| \ge MpN_d\big) = o(1).$$

For $I \in [MpN_d, (1+\epsilon_n)pN_d]$, condition on $H \in \mathcal{Z}$. The number of choices for I common hyperedges is at most

$$\binom{(1+\epsilon_n)pN_d}{I}.$$

Given these I common hyperedges, H' must choose its remaining hyperedges among d-cliques of Ψ, of which there are at most $(1+\epsilon_n)qN_d$. Thus the number of choices for the remaining $(1+O(\epsilon_n))pN_d - I$ hyperedges is at most

$$\binom{(1+\epsilon_n)qN_d}{(1+\epsilon_n)pN_d - I}.$$

The probability weight of any such H' is at most

$$\left(\frac{p}{1-p}\right)^{(1-\epsilon_n)pN_d}(1-p)^{N_d}.$$

Using Stirling's bound $\binom{a}{b} \le (ea/b)^b$, summing over I, and using $q/p = n^{\eta+o(1)}$, the contribution of overlap $I \ge MpN_d$ is bounded by

$$\sum_{I \ge MpN_d}\left[C\left(\frac{p}{q}\right)^{I/(pN_d)}p^{-o(1)}\right]^{pN_d} = o(1),$$

because

$$\left(\frac{p}{q}\right)^{I/(pN_d)} \le n^{-\eta M + o(1)}$$

and $pN_d = \Theta(n^{1+\delta}) \to \infty$. Therefore

$$\mathbb{E}\big[|E(H) \cap E(H')|\mathbf{1}_{\mathcal{Z}}\big] \le MpN_d + o(pN_d).$$

Letting $M \downarrow 0$ gives

$$\mathbb{E}|E(H) \cap E(H')| = o(pN_d).$$

For weighted projections, choose K so large that the expected number of graph edges with multiplicity at least K is $o(1)$. This is possible because, for a fixed graph edge,

$$\mathrm{Proj}_{\mathrm{W}}(H)_{ij} \sim \mathrm{Bin}\left(\binom{n-2}{d-2}, p\right)$$

has mean $\Theta(n^{\delta-1}) = o(1)$. On the high-probability event that all multiplicities are less than K, the number of possible weighted observations supported on at most $O(pN_d)$ graph edges is larger than the number of unweighted observations by at most a factor $K^{O(pN_d)}$. This changes the displayed counting bound only by a constant exponential factor $\exp\{O(pN_d)\}$, while the overlap penalty contributes

$$(q/p)^{-MpN_d} = \exp\{-(\eta M + o(1))pN_d \log n\}.$$

The logarithmic factor dominates. The same conclusion follows for the weighted posterior. □

Proof of the converse in Theorem 6.3. The unweighted converse follows by combining Lemmas C.4 and C.5. The weighted converse follows from the weighted version of Lemma C.5. □

C.3 Uniform Exact Recovery

C.3.1 MAP and component decomposition

Lemma C.6 (MAP is a minimum-preimage rule)**.** *For all sufficiently large n, the MAP estimator from* $\mathrm{Proj}(H)$ *returns a minimum-cardinality preimage of the observed projection.*

Proof. For any d-uniform hypergraph K,

$$\mathbb{P}(H = K) = p^{e(K)}(1-p)^{N_d - e(K)}.$$

Conditioned on $\mathrm{Proj}(H) = G$, the posterior mass of K is proportional to this expression over all K satisfying $\mathrm{Proj}(K) = G$. Since $p = o(1)$, for all large n,

$$\frac{p}{1-p} < 1.$$

Thus the posterior mass is maximized exactly by minimizing $e(K)$. This proves the claim. □

Lemma C.7 (Decomposition over two-connected components)**.** *Let G be a projected graph, and let $H_c(G)$ be the d-uniform hypergraph consisting of all d-cliques in G. Let*

$$C_1, \ldots, C_m$$

be the two-connected components of $H_c(G)$. Then every preimage K of G decomposes as

$$K = \bigcup_{r=1}^{m} K_r, \qquad \mathrm{Proj}(K_r) = \mathrm{Proj}(C_r).$$

Consequently, a minimum preimage of G is the union of minimum preimages of the graphs $\mathrm{Proj}(C_r)$*.*

Proof. Every hyperedge of any preimage K is a d-clique in G, hence is a hyperedge of $H_c(G)$. Thus every hyperedge of K belongs to one of the components C_r. Let K_r be the set of hyperedges of K that belong to C_r.

If an edge $\{u, v\} \in E(G)$ is covered by a hyperedge $h \in K_r$, then $\{u, v\} \subset h$, so every d-clique of G containing $\{u, v\}$ intersects h in at least two vertices and is therefore in the same two-connected component as h. Hence graph edges covered by different K_r's are disjoint across components, and

$$\mathrm{Proj}(K_r) = \mathrm{Proj}(C_r).$$

The equality $K = \bigcup_r K_r$ follows by construction. Since the projections decompose edge-disjointly across components, minimizing the total number of hyperedges is equivalent to minimizing inside each component separately. □

Lemma C.8 (Bounded two-connected components below $(d-1)/(d+1)$). *If*

$$\delta < \frac{d-1}{d+1},$$

then, with high probability, every two-connected component of $H_c(\mathrm{Proj}(H))$ *has* $O_{d,\delta}(1)$ *hyperedges.*

Proof. We give the exploration proof. Start with a d-clique $h \in H_c$. A new d-clique h' can join the same two-connected component only if $|h \cap h'| \geq 2$, or if it can be reached by a sequence of such steps. Suppose a partially exposed finite hypergraph K has already been generated, and consider adding one two-neighbor h'. If $h' \notin E(H)$, every graph edge in $\binom{h'}{2} \setminus \mathrm{Proj}(K)$ must be covered by true hyperedges of H. This generates a finite cover U of the missing graph edges by intersections with h'. A piece $u \subseteq h'$ of size $|u| \geq 2$ costs probability

$$n^{1+\delta-|u|+o(1)}.$$

If h' shares k vertices with the previously exposed vertex set, then choosing its new vertices has at most n^{d-k} possibilities. Therefore the expected multiplicative increase associated with a growth step is at most

$$n^{d-k} \max_U n^{\sum_{u\in U}(1+\delta-|u|)+o(1)}.$$

The cover optimization is the following. Let $g_k(\delta)$ be the minimum, over covers of $\binom{[d]}{2} \setminus \binom{[k]}{2}$, of

$$\sum_{u\in U}(|u| - 1 - \delta).$$

Then

$$g_k(\delta) + k - d \geq \frac{d-1}{d+1} - \delta \qquad \text{for all } 2 \leq k \leq d.$$

Indeed, relaxing $|u|$ to real variables and writing $y_u = \binom{|u|}{2}$, the same convex argument used in Lemma C.3 reduces the minimum to endpoint covers. At $\delta = (d-1)/(d+1)$, the endpoint values are nonnegative; since the objective is piecewise linear with slopes at most -1, decreasing δ by $(d-1)/(d+1) - \delta$ increases the minimum by at least that amount.

Thus every growth step decreases the expected number of components of a fixed exploration type by a factor at most

$$n^{-\alpha+o(1)}, \qquad \alpha = \frac{d-1}{d+1} - \delta > 0.$$

The expected number of starting d-cliques is $O(n^{1+\delta})$. After

$$T > \frac{2+\delta}{\alpha}$$

growth steps, the expected number of explored structures of depth T is $o(1)$. Since d is fixed, the number of possible finite growth patterns of depth T is $O_{d,T}(1)$. A union bound shows that no component has more than $O_{d,\delta}(1)$ growth steps, and hence no component has more than $O_{d,\delta}(1)$ hyperedges. $\square$

C.3.2 Ambiguity and exact recovery

Lemma C.9 (No ambiguity implies MAP success)**.** *Assume*

$$\delta < \frac{d-1}{d+1}.$$

If, with high probability, no two-connected component of $H_c(\mathrm{Proj}(H))$ *has an ambiguous projection, then MAP achieves exact recovery with probability* $1 - o(1)$.

Proof. By Lemma C.8, all two-connected components have at most L hyperedges with high probability, where $L = O_{d,\delta}(1)$. On this event, each component has only $O_{d,L}(1)$ possible preimages.

Assume no component is ambiguous. Then each component projection has a unique minimum preimage. If K' is any non-minimum preimage of a component whose true preimage is K, then

$$e(K') \geq e(K) + 1.$$

The posterior ratio satisfies

$$\frac{\mathbb{P}(H = K' \mid \mathrm{Proj}(H))}{\mathbb{P}(H = K \mid \mathrm{Proj}(H))} = \left(\frac{p}{1-p}\right)^{e(K')-e(K)} \leq 2p$$

for all large n. Since each component has $O(1)$ competing preimages, the probability that the MAP preimage of a fixed component differs from the true preimage is $O(p)$.

The number of components is at most $|E(H_c)|$. Below the partial recovery threshold, Lemma C.3 gives

$$\mathbb{E}|E(H_c)| = (1 + o(1))pN_d = O(n^{1+\delta}).$$

Therefore a union bound gives total error probability at most

$$O(n^{1+\delta}p) = O(n^{1+\delta}n^{-d+1+\delta}) = O(n^{-d+2+2\delta}) = o(1),$$

because $d \geq 3$ and $\delta < (d-1)/(d+1) < 1$. Thus MAP succeeds with probability $1 - o(1)$. $\square$

Lemma C.10 (The extremal ambiguous family)**.** *Let $\mathcal{A}_d$ be the hypergraph described in Section 6.3.3. Then* $\mathrm{Proj}(\mathcal{A}_d)$ *is ambiguous, and*

$$e(\mathcal{A}_d) = 2d - 1, \qquad v(\mathcal{A}_d) = 2d^2 - 5d + 5.$$

Thus its appearance exponent is

$$d - 1 - \frac{v(\mathcal{A}_d)}{e(\mathcal{A}_d)} = \frac{2d-4}{2d-1}.$$

Proof. The hyperedges of $\mathcal{A}_d$ are

$$h_u = \{u, v_1, \ldots, v_{d-1}\},$$

$$h_{i,u} = \{u, v_i\} \cup S_i^u, \qquad 1 \le i \le d-1,$$

and

$$h_{i,w} = \{w, v_i\} \cup S_i^w, \qquad 1 \le i \le d-1.$$

This gives

$$1 + (d-1) + (d-1) = 2d - 1$$

hyperedges.

The vertex set is

$$\{u, w\} \cup \{v_1, \ldots, v_{d-1}\} \cup \bigcup_{i=1}^{d-1} S_i^u \cup \bigcup_{i=1}^{d-1} S_i^w.$$

The sets are disjoint, and each S_i^u, S_i^w has size $d - 2$. Therefore

$$v(\mathcal{A}_d) = 2 + (d-1) + 2(d-1)(d-2) = 2d^2 - 5d + 5.$$

The projection has two minimum preimages. One is $\mathcal{A}_d$. The other replaces

$$h_u = \{u, v_1, \ldots, v_{d-1}\}$$

by

$$h_w = \{w, v_1, \ldots, v_{d-1}\},$$

leaving all other hyperedges unchanged. The projected edges among $v_1, \ldots, v_{d-1}$ are covered by either h_u or h_w, and the edges incident to u and w through each v_i are already covered by the corresponding $h_{i,u}$ and $h_{i,w}$. Thus the two hypergraphs have the same projection and the same number of hyperedges.

Minimality follows because each set S_i^u contains vertices that occur only inside the clique generated by $h_{i,u}$, so any preimage must contain a hyperedge covering that clique; similarly for $h_{i,w}$. In addition, the clique on $\{v_1, \ldots, v_{d-1}\}$ must be completed by one additional d-hyperedge containing either u or w. Hence at least $2d - 1$ hyperedges are necessary, and the two displayed preimages are minimum preimages.

Finally,

$$d - 1 - \frac{v(\mathcal{A}_d)}{e(\mathcal{A}_d)} = d - 1 - \frac{2d^2 - 5d + 5}{2d - 1} = \frac{2d-4}{2d-1}.$$

□

Lemma C.11 (Extremal ambiguity lower bound). *Let K be a d-uniform hypergraph whose projection is ambiguous. Then*

$$d - 1 - \frac{1}{m(K)} \geq \frac{2d-4}{2d-1}.$$

Equivalently,

$$m(K) \geq \frac{2d-1}{2d^2 - 5d + 5}.$$

Proof. Let K' be another minimum preimage of $\mathrm{Proj}(K)$, $K' \neq K$, with

$$e(K') = e(K).$$

Remove common hyperedges from K and K'. This cannot increase the maximum density needed below, and it leaves two distinct hypergraphs with the same projection on the edges that remain to be explained. Let $E_0 = E(K) \setminus E(K')$ and $E_1 = E(K') \setminus E(K)$, and let U be the set of vertices incident to hyperedges in $E_0 \cup E_1$.

Because $\mathrm{Proj}(K) = \mathrm{Proj}(K')$, every graph edge generated by E_0 is also generated either by E_1 or by a removed common hyperedge. Let P be the set of graph edges in U on which the two covers differ. Each edge in P is supported by a common hyperedge before removal, and therefore may be charged to a hyperedge outside $E_0 \cup E_1$. The deterministic density certificate is

$$\frac{(d-1)|E_0| - |U| + |P|}{|E_0| + |P|} \geq \frac{2d-4}{2d-1}.$$

To see this, reveal the hyperedges in E_0 one at a time. Since no hyperedge of E_0 is present in K', the d-clique induced by such a hyperedge must be covered by proper intersections with hyperedges of K' and by edges in P. A proper cover of a d-clique has total deficit at least $2d-4$ per $2d-1$ units of hyperedge/edge mass; this is the same convex cover inequality used in Lemma C.3, now applied to the symmetric difference of two covers rather than to a single projected clique. Summing the inequality over all hyperedges of E_0 gives the displayed certificate; equality is attained exactly by the configuration $\mathcal{A}_d$.

Now form the subhypergraph $L \subseteq K$ induced by U together with the common hyperedges that support P. It has at most

$$|U| + \sum_i (d - x_i)$$

vertices, where x_i is the number of vertices of the i-th supporting common hyperedge already inside U, and it has at least $|E_0|$ plus the number of supporting common hyperedges as hyperedges. The preceding certificate implies

$$d - 1 - \frac{1}{e(L)/v(L)} \geq \frac{2d-4}{2d-1}.$$

Since $m(K) \geq e(L)/v(L)$, the result follows. □

Remark C.12. Lemma C.11 is the purely deterministic part of the exact-recovery threshold. The proof above is written in the density-certificate form used by the source proof: one compares two minimum covers, charges their symmetric-difference edges to common supporting hyperedges, and applies the same convex cover inequality that controls fake cliques. The extremal family $\mathcal{A}_d$ shows that the bound is tight.

Lemma C.13 (No ambiguous component below the ambiguity exponent)**.** *If*

$$\delta < \frac{2d-4}{2d-1},$$

then with high probability no ambiguous graph appears as the projection of a two-connected component of $H_c(\mathrm{Proj}(H))$.

Proof. By Lemma C.8, in the regime relevant to achievability every two-connected component has at most $L = O_{d,\delta}(1)$ hyperedges with high probability. There are only finitely many isomorphism types of such components.

For any ambiguous component with a preimage K, Lemma C.11 gives

$$d - 1 - \frac{1}{m(K)} \geq \frac{2d-4}{2d-1}.$$

If $\delta < (2d-4)/(2d-1)$, then

$$p = o(n^{-1/m(K)}).$$

By Lemma C.2, the probability that this fixed K appears is $o(1)$. A union bound over finitely many possible K's gives the claim. □

Lemma C.14 (Ambiguous copies above the ambiguity exponent)**.** *If*

$$\delta > \frac{2d-4}{2d-1},$$

then for every fixed M, with probability $1 - o(1)$, the random hypergraph H contains M vertex-disjoint copies of $\mathcal{A}_d$.

Proof. Let $K = \mathcal{A}_d$. By Lemma C.10,

$$d - 1 - \frac{1}{m(K)} = \frac{2d-4}{2d-1}.$$

Thus, if $\delta > (2d-4)/(2d-1)$, then

$$p = \omega(n^{-1/m(K)}).$$

By Lemma C.2, K appears with probability $1 - o(1)$. The same second-moment argument applied to M disjoint labeled copies gives appearance of M vertex-disjoint copies with probability $1 - o(1)$, because M is fixed and the disjoint union has the same maximum-density threshold. □

C.3.3 Proof of the exact recovery theorem

Proof of Theorem 6.6. First assume

$$\delta < \delta_{\mathrm{ex}}(d).$$

Then

$$\delta < \frac{d-1}{d+1},$$

so by Lemma C.8, all two-connected components of H_c have bounded size with high probability. Also

$$\delta < \frac{2d-4}{2d-1},$$

so Lemma C.13 shows that no ambiguous component appears with high probability. Lemma C.9 proves exact recovery by MAP. The bounded component size and Lemma C.7 imply that MAP is computed by brute force on $O(1)$-size components, hence in polynomial time for fixed d.

Now assume

$$\delta > \delta_{\mathrm{ex}}(d).$$

If

$$\delta > \frac{d-1}{d+1},$$

then Theorem 6.3 implies that partial recovery has loss $1 - o(1)$. Exact recovery would imply partial recovery with zero loss, so exact recovery is impossible. More quantitatively, if H' is an independent posterior copy given Ψ, Lemma C.5 implies

$$\mathbb{E}|E(H) \cap E(H')| = o(pN_d).$$

Since $|E(H)| = (1 + o(1))pN_d$ with high probability,

$$\mathbb{P}(H = H') = o(1).$$

Writing posterior masses as $\pi_\Psi(K) = \mathbb{P}(H = K \mid \Psi)$,

$$\mathbb{P}(H = H' \mid \Psi) = \sum_K \pi_\Psi(K)^2.$$

Thus

$$\mathbb{E} \sum_K \pi_\Psi(K)^2 = o(1).$$

The MAP success probability satisfies

$$\mathbb{E} \max_K \pi_\Psi(K) \le \mathbb{E} \left(\sum_K \pi_\Psi(K)^2 \right)^{1/2} \le \left(\mathbb{E} \sum_K \pi_\Psi(K)^2 \right)^{1/2} = o(1).$$

Therefore no estimator succeeds with nonvanishing probability.

It remains only to consider $d = 3, 4$ with

$$\frac{2d-4}{2d-1} < \delta < \frac{d-1}{d+1}.$$

By Lemma C.14, for every fixed M, with probability $1 - o(1)$, H contains M vertex-disjoint copies of $\mathcal{A}_d$. Each copy gives two minimum preimages with the same projection and the same number of hyperedges. Conditional on the projection, the posterior mass of any single choice over these M independent binary ambiguities is at most 2^{-M}. Hence the MAP success probability is at most $2^{-M} + o(1)$. Since M is arbitrary, the success probability is $o(1)$. This proves impossibility. $\square$

C.4 Weighted Exact Recovery

Definition C.15 (Weighted ambiguous graph)**.** A weighted projected graph is weighted-ambiguous if it has two distinct minimum preimages under $\mathrm{Proj}_{\mathrm{W}}$.

Lemma C.16 (Weighted ambiguity requires density at least $(d-1)/(d+1)$)**.** *Let K be a d-uniform hypergraph whose weighted projection has another minimum preimage. Then*

$$d - 1 - \frac{1}{m(K)} \geq \frac{d}{2} - 1 \geq \frac{d-1}{d+1}.$$

Proof. Let $K' \neq K$ be another minimum preimage with

$$\mathrm{Proj}_{\mathrm{W}}(K') = \mathrm{Proj}_{\mathrm{W}}(K).$$

Remove all common hyperedges from K and K'. The weighted projections remain equal after subtracting the common contribution, and $m(K)$ cannot increase by passing to a subhypergraph of K. Thus we may assume

$$E(K) \cap E(K') = \emptyset.$$

Let V' be the set of vertices that appear in at least one hyperedge of K. We claim every $v \in V'$ has degree at least 2 in K. Suppose, for contradiction, that v belongs to exactly one hyperedge $h \in E(K)$. For every $u \in h \setminus \{v\}$, the weighted projected edge $\{u, v\}$ has multiplicity one in $\mathrm{Proj}_{\mathrm{W}}(K)$. Since $\mathrm{Proj}_{\mathrm{W}}(K') = \mathrm{Proj}_{\mathrm{W}}(K)$, each such pair $\{u, v\}$ must be contained in a hyperedge of K'. If all these pairs are contained in the same hyperedge of K', that hyperedge must equal h, contradicting $E(K) \cap E(K') = \emptyset$. If they are not all contained in the same hyperedge, then v belongs to at least two hyperedges of K'. But equality of weighted projections implies that the total weighted degree of v in K' equals that in K, a contradiction. Therefore every $v \in V'$ has degree at least 2.

Let K_0 be the subhypergraph of K induced by V'; this is just K restricted to its non-isolated vertices. Counting vertex-hyperedge incidences,

$$d\, e(K_0) = \sum_{v \in V'} \deg_K(v) \geq 2|V'| = 2v(K_0).$$

Hence

$$\frac{e(K_0)}{v(K_0)} \geq \frac{2}{d}.$$

Thus

$$m(K) \geq \frac{2}{d},$$

and so

$$d - 1 - \frac{1}{m(K)} \geq d - 1 - \frac{d}{2} = \frac{d}{2} - 1.$$

Finally,

$$\frac{d}{2} - 1 - \frac{d-1}{d+1} = \frac{d(d+1) - 2(d+1) - 2(d-1)}{2(d+1)} = \frac{d^2 - 3d}{2(d+1)} = \frac{d(d-3)}{2(d+1)} \geq 0$$

for $d \geq 3$. This proves the lemma. ∎

Proof of Theorem 6.7. If

$$\delta < \frac{d-1}{d+1},$$

then Lemma C.8 gives bounded two-connected components. By Lemma C.16 and Lemma C.2, no weighted-ambiguous component appears with high probability. The weighted analogue of Lemma C.9 then shows that weighted MAP recovers exactly.

If

$$\delta > \frac{d-1}{d+1},$$

the weighted partial recovery loss is $1 - o(1)$ by Theorem 6.3. Exact weighted recovery would imply zero weighted partial loss, so exact weighted recovery is impossible. Equivalently, the posterior-collision argument in the proof of Theorem 6.6 applies to the weighted posterior and gives MAP success probability $o(1)$. □

C.5 Heterogeneous Partial Recovery

We now prove Theorem 6.8. Recall the heterogeneous model:

$$2 \le d_1 < \cdots < d_k, \qquad p_r = n^{1-d_r+\delta_r+o(1)}, \qquad 0 < \delta_r < 1.$$

Let

$$\delta_* = \max_{r \in [k]} \delta_r.$$

C.5.1 Projected cover probabilities

For $\ell \ge 2$, define

$$\Delta_\ell = \max_{r:\ d_r \ge \ell} \delta_r,$$

with the convention $\Delta_\ell = -\infty$ if no $d_r \ge \ell$.

Let $E \subseteq \binom{V}{2}$ be a fixed graph edge set on a fixed vertex set V. A cover of E is a collection $U \subseteq 2^V$ such that

$$|u| \ge 2 \quad \text{for all } u \in U,$$
$$E \subseteq \bigcup_{u \in U} \binom{u}{2},$$

and no $u \in U$ is redundant.

Lemma C.17 (Heterogeneous cover exponent)**.** *For fixed* $E \subseteq \binom{V}{2}$,

$$\mathbb{P}(E \subseteq E(\psi)) = n^{g(E,\Delta)+o(1)},$$

where

$$g(E, \Delta) = \max_U \sum_{u \in U} \left(1 + \Delta_{|u|} - |u|\right),$$

and the maximum is over all covers U *of* E*. In particular,*

$$g(E, \Delta) \le \max_U \sum_{u \in U} (1 + \delta_* - |u|).$$

Proof. Fix $u \subseteq V$, $|u| = \ell \geq 2$. Let A_u be the event that some hyperedge of any allowed degree contains all vertices of u and contains no other vertex of V. For degree $d_r \geq \ell$, the number of candidate hyperedges is

$$\binom{n-|V|}{d_r-\ell} = n^{d_r-\ell+o(1)}.$$

Each appears with probability

$$p_r = n^{1-d_r+\delta_r+o(1)}.$$

Therefore the probability that at least one such degree-d_r hyperedge exists is

$$n^{1-\ell+\delta_r+o(1)}.$$

Taking the union over finitely many r's and using the largest exponent gives

$$\mathbb{P}(A_u) = n^{1-\ell+\Delta_\ell+o(1)}.$$

For a fixed cover U, the events $\{A_u : u \in U\}$ involve disjoint classes of hyperedges because the intersections with V are prescribed to be distinct. Hence

$$\mathbb{P}\left(\bigcap_{u\in U} A_u\right) = n^{\sum_{u\in U}(1+\Delta_{|u|}-|u|)+o(1)}.$$

If all edges in E appear in ψ, then the collection of intersections with V of all generating hyperedges contains a subcollection that is a nonredundant cover of E. Conversely, any realized cover implies $E \subseteq E(\psi)$. Since V is fixed, there are only finitely many covers. The largest exponent dominates the finite union, proving the formula. Replacing every $\Delta_{|u|}$ by δ_* gives the upper bound. □

Lemma C.18 (Clique cover optimization). *Let K_m be the complete graph on $m \geq 3$ vertices. For covers that do not use the full m-set,*

$$g(K_m, \Delta) \leq \max\left\{-\binom{m}{2} + \binom{m}{2}\delta_*,\ m\delta_* - 2m + 3\right\}.$$

Proof. Let U be a cover of K_m not using the full m-set. Write $M = |U|$, and let

$$x_u = |u|, \qquad y_u = \binom{x_u}{2}.$$

Then

$$2 \leq x_u \leq m-1, \qquad \sum_{u\in U} y_u \geq \binom{m}{2}.$$

The exponent is bounded by

$$\sum_{u\in U}(1+\delta_* - x_u) = M(1+\delta_*) - \sum_{u\in U} x_u.$$

Using

$$x_u = \frac{1+\sqrt{1+8y_u}}{2},$$

and relaxing y_u to real variables, the maximum for fixed M is attained at a vertex of

$$y_u \geq 1, \qquad \sum_{u\in U} y_u \geq \binom{m}{2}.$$

Thus $M-1$ variables equal 1, and the final variable equals $\binom{m}{2} - M + 1$. The resulting bound is convex as a function of M, so the maximum over integer M is attained at an endpoint.

The endpoint $M = \binom{m}{2}$ is the all-edges cover, giving

$$\binom{m}{2}(1+\delta_* - 2) = -\binom{m}{2} + \binom{m}{2}\delta_*.$$

The other relevant endpoint is the cover by one $(m-1)$-clique and $m-1$ additional edges from the remaining vertex. It has $M = m$, total set-size

$$(m-1) + 2(m-1) = 3m-3,$$

and exponent

$$m(1+\delta_*) - (3m-3) = m\delta_* - 2m + 3.$$

This proves the stated upper bound. □

Lemma C.19 (Star cover optimization)**.** *Let* Star_m *be the star with* m *edges. Then*

$$g(\mathrm{Star}_m, \Delta) \leq -m + m\delta_*.$$

Proof. Let the star have center 0 and leaves $1, \dots, m$. Any cover piece u that covers r star edges must contain the center and those r leaves, so $|u| \geq r+1$. Its exponent is at most

$$1 + \delta_* - |u| \leq 1 + \delta_* - (r+1) = \delta_* - r.$$

Replacing this piece by the r individual edges gives total exponent

$$r(1+\delta_* - 2) = -r + r\delta_*.$$

Since $0 < \delta_* < 1$,

$$-r + r\delta_* \geq \delta_* - r$$

whenever $r \geq 1$. Therefore the individual-edge cover is optimal, and its exponent is

$$m(1+\delta_* - 2) = -m + m\delta_*.$$

□

C.5.2 Proof of heterogeneous achievability

Proof of Theorem 6.8. Fix j with $d_j \geq 3$, and write

$$m = d_j.$$

By symmetry, it suffices to analyze the target set

$$S = [m].$$

Let

$$p_j = n^{1-m+\delta_j+o(1)}.$$

Then

$$\mathbb{E}|E(H_j)| = \binom{n}{m} p_j.$$

Let A_{FP} be the event that S is a false positive:

$$A_{\mathrm{FP}} = \{S \notin E(H_j)\} \cap \{S \in E(\widehat{H}_j)\}.$$

Let A_{FN} be the event that S is a false negative:

$$A_{\mathrm{FN}} = \{S \in E(H_j)\} \cap \{S \notin E(\widehat{H}_j)\}.$$

Then

$$\mathbb{E}\Big[|E(H_j) \triangle E(\widehat{H}_j)|\Big] = \binom{n}{m} \left[\mathbb{P}(A_{\mathrm{FP}}) + \mathbb{P}(A_{\mathrm{FN}})\right].$$

It is therefore enough to prove

$$\mathbb{P}(A_{\mathrm{FP}}) = o(p_j), \qquad \mathbb{P}(A_{\mathrm{FN}}) = o(p_j).$$

We first control false positives. If a hyperedge of degree at least m contains all vertices of S, then either S itself is a true degree-m hyperedge, or S is contained in a larger clique and is not maximal. Hence a false positive requires the clique on S to be assembled from a cover not using the full m-set. By Lemmas C.17 and C.18,

$$\mathbb{P}(A_{\mathrm{FP}}) \leq n^{\max\left\{-\binom{m}{2}+\binom{m}{2}\delta_*,\, m\delta_*-2m+3\right\}+o(1)}.$$

The hypothesis

$$\delta_* < \frac{m-2}{m} + \frac{2\delta_j}{m(m-1)}$$

is equivalent to

$$-\binom{m}{2} + \binom{m}{2}\delta_* < 1 - m + \delta_j.$$

It also implies

$$m\delta_* - 2m + 3 < 1 - m + \frac{2\delta_j}{m-1} \leq 1 - m + \delta_j,$$

because $m \geq 3$ and $0 < \delta_j < 1$. Thus

$$\mathbb{P}(A_{\mathrm{FP}}) = o\left(n^{1-m+\delta_j}\right) = o(p_j).$$

Now we control false negatives. If $S \in E(H_j)$, then S is a clique in ψ. The maximal clique estimator fails to output S only if S is not maximal, meaning that there exists an outside vertex $v \notin S$ connected in ψ to every vertex of S. For fixed v, this is the event that a star with m edges appears on $S \cup \{v\}$. By Lemmas C.17 and C.19,

$$\mathbb{P}\big(v \text{ is connected to all vertices of } S\big) \leq n^{-m+m\delta_*+o(1)}.$$

A union bound over $v \notin S$ gives

$$\mathbb{P}\big(\exists v \notin S:\ v \text{ is connected to all vertices of } S\big) \leq n^{1-m+m\delta_*+o(1)}.$$

This event is independent of the occurrence of the degree-m hyperedge S, because the latter uses the full set S while the outside-star event uses hyperedges whose intersection with $S \cup \{v\}$ contains v. Hence

$$\mathbb{P}(A_{\mathrm{FN}}) \leq p_j\, n^{1-m+m\delta_*+o(1)}.$$

The assumed condition implies

$$1 - m + m\delta_* < -1 + \frac{2\delta_j}{m-1} \leq 0,$$

with strict inequality because $m \geq 3$ and $0 < \delta_j < 1$. Therefore

$$n^{1-m+m\delta_*+o(1)} = o(1),$$

and

$$\mathbb{P}(A_{\mathrm{FN}}) = o(p_j).$$

Combining the false-positive and false-negative bounds,

$$\mathbb{E}\Big[|E(H_j) \triangle E(\widehat{H}_j)|\Big] = \binom{n}{m} o(p_j) = o(1)\, \mathbb{E}|E(H_j)|.$$

This proves the theorem. $\square$

References

[1] M. E. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582.

[2] P. W. Holland, K. B. Laskey, and S. Leinhardt. "Stochastic Blockmodels: First Steps". In: *Social Networks* 5.2 (1983), pp. 109–137.

[3] E. Abbe. "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177 (2017), pp. 1–86.

[4] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. "Statistical algorithms and a lower bound for detecting planted cliques". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM. 2013, pp. 655–664.

[5] 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: *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE. 2016, pp. 428–437.

[6] M. Brennan, G. Bresler, and W. Huleihel. "Reducibility and Computational Lower Bounds for Problems with Planted Sparse Structure". In: *COLT*. 2018, pp. 48–166.

[7] M. Brennan and G. Bresler. "Reducibility and statistical-computational gaps from secret leakage". In: *Conference on Learning Theory*. PMLR. 2020, pp. 648–847.

[8] A. Bogdanov and L. Trevisan. "Average-case complexity". In: *Foundations and Trends® in Theoretical Computer Science* 2.1 (2006), pp. 1–106.

[9] O. Goldreich. "Notes on Levin's theory of average-case complexity". In: *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*. Springer, 2011, pp. 233–247.

[10] S. B. Hopkins. "Statistical Inference and the Sum of Squares Method". PhD thesis. Cornell University, 2018.

[11] A. S. Wein. "Average-Case Complexity of Tensor Decomposition for Low-Degree Polynomials". In: *arXiv preprint arXiv:2211.05274* (2022).

[12] S. B. Hopkins, P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer. "The power of sum-of-squares for detecting hidden structures". In: *Proceedings of the fifty-eighth IEEE Foundations of Computer Science* (2017), pp. 720–731.

[13] M. S. Granovetter. "The strength of weak ties". In: *American journal of sociology* 78.6 (1973), pp. 1360–1380.

[14] D. J. Watts and S. H. Strogatz. "Collective Dynamics of "Small-World" Networks". In: *Nature* 393.6684 (1998), pp. 440–442.

[15] M. E. Newman. "Coauthorship networks and patterns of scientific collaboration". In: *Proceedings of the National Academy of Sciences* 101.suppl_1 (2004), pp. 5200–5205.

[16] A.-L. Barabási and Z. N. Oltvai. "Network Biology: Understanding the Cell's Functional Organization". In: *Nature Reviews Genetics* 5.2 (2004), pp. 101–113.

[17] M. Brennan, G. Bresler, and D. M. Nagaraj. "Phase transitions for detecting latent geometry in random graphs". In: *Probability Theory and Related Fields* 178 (Sept. 2020), pp. 1215–1289.

[18] M. Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

[19] G. Bresler, C. Guo, and Y. Polyanskiy. "Algorithmic Decorrelation and Planted Clique in Dependent Random Graphs: The Case of Extra Triangles". In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2023, pp. 2149–2158.

[20] K. Bagachev, G. Bresler, and C. Guo. *Middle-Scale Asymptotics of Random Geometric Graphs: Part I*. Manuscript. 2026.

[21] J.-G. Young, G. Petri, and T. P. Peixoto. "Hypergraph reconstruction from network data". In: *Communications Physics* 4.1 (2021), p. 135.

[22] S. Lizotte, J.-G. Young, and A. Allard. "Hypergraph reconstruction from uncertain pairwise observations". In: *Scientific Reports* 13.1 (2023), p. 21364.

[23] Y. Wang and J. Kleinberg. "From Graphs to Hypergraphs: Hypergraph Projection and its Reconstruction". In: *The Twelfth International Conference on Learning Representations*. 2024.

[24] G. Bresler, C. Guo, and Y. Polyanskiy. "Thresholds for Reconstruction of Random Hypergraphs From Graph Projections". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Vol. 247. Proceedings of Machine Learning Research. PMLR, 2024, pp. 632–647.

[25] G. Bresler, C. Guo, Y. Polyanskiy, and A. Yao. "Partial and exact recovery of a random hypergraph from its graph projection". In: *arXiv preprint arXiv:2502.14988* (2025).

[26] A. Morgan and C. Guo. "Achievability of heterogeneous hypergraph recovery from its graph projection". In: *arXiv preprint arXiv:2603.01268* (2026).

[27] A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige, and A. Vijayaraghavan. "Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph". In: *Proceedings of the forty-second ACM symposium on Theory of computing* (2010), pp. 201–210.

[28] Z. Ma and Y. Wu. "Computational barriers in minimax submatrix detection". In: *The Annals of Statistics* 43.3 (2015), pp. 1089–1116.

[29] B. E. Hajek, Y. Wu, and J. Xu. "Computational Lower Bounds for Community Detection on Random Graphs." In: *COLT*. 2015, pp. 899–928.

[30] Y. Luo and A. R. Zhang. “Tensor clustering with planted structures: Statistical optimality and computational limits”. In: *The Annals of Statistics* 50.1 (2022), pp. 584–613.

[31] A. Zhang and D. Xia. “Tensor SVD: Statistical and Computational Limits”. In: *arXiv preprint arXiv:1703.02724* (2017).

[32] V. Feldman, W. Perkins, and S. Vempala. “On the complexity of random satisfiability problems with planted solutions”. In: (2015), pp. 77–86.

[33] I. Diakonikolas, D. M. Kane, and A. Stewart. “Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures”. In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 73–84.

[34] O. Y. Feng, R. Venkataramanan, C. Rush, and R. J. Samworth. “A Unifying Tutorial on Approximate Message Passing”. In: *Foundations and Trends® in Machine Learning* 15.4 (2022), pp. 335–536.

[35] A. Montanari and A. S. Wein. “Equivalence of Approximate Message Passing and Low-Degree Polynomials in Rank-One Matrix Estimation”. In: *arXiv preprint arXiv:2212.06996* (2022).

[36] M. Celentano, A. Montanari, and Y. Wu. “The estimation error of general first order methods”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 1078–1141.

[37] B. Rossman. “On the constant-depth complexity of k-clique”. In: *Proceedings of the fortieth annual ACM symposium on Theory of computing*. ACM. 2008, pp. 721–730.

[38] B. Rossman. “The monotone complexity of k-clique on random graphs”. In: *SIAM Journal on Computing* 43.1 (2014), pp. 256–279.

[39] D. Gamarnik and M. Sudan. “Limits of local algorithms over sparse random graphs”. In: *Proceedings of the 5th conference on Innovations in theoretical computer science*. ACM. 2014, pp. 369–376.

[40] B. Huang and M. Sellke. “Tight lipschitz hardness for optimizing mean field spin glasses”. In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 312–322.

[41] M. Brennan, G. Bresler, S. B. Hopkins, J. Li, and T. Schramm. “Statistical query algorithms and low-degree tests are almost equivalent”. In: *arXiv preprint arXiv:2009.06107* (2020).

[42] F. Koehler and E. Mossel. “Reconstruction on trees and low-degree polynomials”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 18942–18954.

[43] A. S. Wein, A. El Alaoui, and C. Moore. “The Kikuchi hierarchy and tensor PCA”. In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2019, pp. 1446–1468.

[44] I. Zadik, M. J. Song, A. S. Wein, and J. Bruna. “Lattice-based methods surpass sum-of-squares in clustering”. In: *Conference on Learning Theory*. PMLR. 2022, pp. 1247–1248.

[45] B. Barak. "The Complexity of Public-Key Cryptography". In: *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*. Ed. by Y. Lindell. Cham: Springer International Publishing, 2017, pp. 45–77. ISBN: 978-3-319-57048-8. DOI: 10.1007/978-3-319-57048-8_2. URL: https://doi.org/10.1007/978-3-319-57048-8_2.

[46] Q. Berthet and P. Rigollet. "Complexity Theoretic Lower Bounds for Sparse Principal Component Detection." In: *COLT*. 2013, pp. 1046–1066.

[47] T. Wang, Q. Berthet, and R. J. Samworth. "Statistical and computational trade-offs in estimation of sparse principal components". In: *The Annals of Statistics* 44.5 (2016), pp. 1896–1930.

[48] C. Gao, Z. Ma, and H. H. Zhou. "Sparse CCA: Adaptive estimation and computational barriers". In: *The Annals of Statistics* 45.5 (2017), pp. 2074–2101.

[49] M. Brennan and G. Bresler. "Optimal Average-Case Reductions to Sparse PCA: From Weak Assumptions to Strong Hardness". In: *arXiv preprint arXiv:1902.07380* (2019).

[50] M. Brennan, G. Bresler, and W. Huleihel. "Universality of Computational Lower Bounds for Submatrix Detection". In: *arXiv preprint arXiv:1902.06916* (2019).

[51] J. Bruna, O. Regev, M. J. Song, and Y. Tang. "Continuous lwe". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 694–707.

[52] A. Gupte, N. Vafa, and V. Vaikuntanathan. "Continuous LWE is as Hard as LWE & Applications to Learning Gaussian Mixtures". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. 2022, pp. 1162–1173. DOI: 10.1109/FOCS54457.2022.00112.

[53] M. J. Song, I. Zadik, and J. Bruna. "On the cryptographic hardness of learning single periodic neurons". In: *Advances in neural information processing systems* 34 (2021), pp. 29602–29615.

[54] S. Chen, A. Gollakota, A. Klivans, and R. Meka. "Hardness of noise-free learning for two-hidden-layer neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 10709–10724.

[55] A. Daniely and G. Vardi. "From local pseudorandom generators to hardness of learning". In: *Conference on Learning Theory*. PMLR. 2021, pp. 1358–1394.

[56] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. "Cryptographic hardness of learning halfspaces with massart noise". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 3624–3636.

[57] I. Diakonikolas, D. Kane, and L. Ren. "Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 7922–7938.

[58] S. Tiegel. "Hardness of agnostically learning halfspaces from worst-case lattice problems". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 3029–3064.

[59] M. Brennan, G. Bresler, and D. Nagaraj. "Phase Transitions for Detecting Latent Geometry in Random Graphs". In: *arXiv preprint arXiv:1910.14167* (2019).

[60] B. Klimt and Y. Yang. "Introducing the Enron corpus." In: *CEAS*. Vol. 45. 2004, pp. 92–96.

[61] D. Ghoshdastidar and A. Dukkipati. "Consistency of spectral partitioning of uniform hypergraphs under planted partition model". In: *Advances in Neural Information Processing Systems* 27 (2014).

[62] M. C. Angelini, F. Caltagirone, F. Krzakala, and L. Zdeborová. "Spectral detection on sparse hypergraphs". In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2015, pp. 66–73.

[63] J. Gaudio and N. Joshi. "Community Detection in the Hypergraph SBM: Exact Recovery Given the Similarity Matrix". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Vol. 195. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 469–510.

[64] Y. Gu and Y. Polyanskiy. "Weak Recovery Threshold for the Hypergraph Stochastic Block Model". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Vol. 195. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 885–920.

[65] I. Dumitriu and H. Wang. "Optimal and exact recovery on general non-uniform Hypergraph Stochastic Block Model". In: *arXiv preprint arXiv:2304.13139* (2024).

[66] Y. Chen, C. Fang, Z. Lin, and B. Liu. "Relational learning in pre-trained models: a theory from hypergraph recovery perspective". In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org, 2024.

[67] G. Reeves, J. Xu, and I. Zadik. "The All-or-Nothing Phenomenon in Sparse Linear Regression". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. Proceedings of Machine Learning Research. PMLR, 25–28 Jun 2019, pp. 2652–2663.

[68] Y. Wu, J. Xu, and S. H. Yu. "Settling the Sharp Reconstruction Thresholds of Random Graph Matching". In: *IEEE Transactions on Information Theory* 68.8 (2022), pp. 5391–5417.

[69] E. Abbe and A. Montanari. "Conditional Random Fields, Planted Constraint Satisfaction, and Entropy Concentration". In: *Theory of Computing* 11.17 (2015), pp. 413–443.

[70] F. Krzakala and L. Zdeborová. "Hiding Quiet Solutions in Random Constraint Satisfaction Problems". In: *Phys. Rev. Lett.* 102 (23 June 2009), p. 238701.

[71] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. "Optimal errors and phase transitions in high-dimensional generalized linear models". In: *Proceedings of the National Academy of Sciences* 116.12 (2019), pp. 5451–5460.

[72] D. Achlioptas and A. Coja-Oghlan. "Algorithmic Barriers from Phase Transitions". In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. 2008, pp. 793–802.