Frozen variables in random boolean constraint satisfaction problems

Michael Molloy and Ricardo Restrepo
Department of Computer Science, University of Toronto
10 King's College Road, Toronto, ON

July 3, 2012

Abstract

We determine the exact *freezing threshold*, r^f, for a family of models of random boolean constraint satisfaction problems, including NAE-SAT and hypergraph 2-colouring, when the constraint size is sufficiently large. If the constraint-density of a random CSP, F, in our family is greater than r^f then for almost every solution of F, a linear number of variables are *frozen*, meaning that their colours cannot be changed by a sequence of alterations in which we change $o(n)$ variables at a time, always switching to another solution. If the constraint-density is less than r^f, then almost every solution has $o(n)$ frozen variables.

Freezing is a key part of the clustering phenomenon that is hypothesized by non-rigorous techniques from statistical physics. The understanding of clustering has led to the development of advanced heuristics such as Survey Propogation. It has been suggested that the freezing threshold is a precise algorithmic barrier: There is reason to believe that for densities below r^f the random CSPs can be solved using very simple algorithms, while for densities above r^f one requires more sophisticated techniques in order to deal with frozen clusters.

1 Introduction

The clustering phemonenon is arguably the most important development in the study of random constraint satisfaction problems (CSP's) over the past decade or so. Statistical physicists have discovered that for typical models of random constraint satisfaction problems, the structure of the solution space appears to undergo remarkable changes as the constraint density increases.

At first, all solutions are very similar in that we can change any one solution into any other solution via a sequence of small local changes; i.e. by changing only $o(n)$ variables-at-a-time, always having a satisfying solution. This remains true for almost all solutions until the *clustering threshold* [40], at which point they shatter into an exponential number of clusters. Roughly speaking: one can move from any solution to any other *in the same cluster* making small local changes, but moving from one cluster to another requires changing a linear number of variables in at least one step. As we increase the density further, we reach the *freezing threshold* [50]. Above that point, almost all clusters[1] contain *frozen variables*; that is, variables whose values do not change for any solutions in the cluster. At higher densities, we find other thresholds, such as the *condensation threshold* [36] above which the largest cluster contains a positive proportion of the solutions. Eventually we reach the *satisfiability threshold*, the point at which there are no solutions.

The methods that are used to describe these phenomena and determine the values of the thresholds are mathematically sophisticated, but are typically not rigorous. Nevertheless, they have transformed the rigorous study of random CSP's.

For one thing, this picture explained things that mathematicians had already discovered. For some problems (eg. k-NAE-SAT [10], k-SAT [12] and k-COL [11]) the second moment method had been used to prove the existence of solutions at densities that are close to, but not quite, the hypothesized satisfiability threshold. We now understand that this is because the way that the second moment method was applied cannot work past the condensation threshold. As another example, it had long been observed that at a point where the density is still far below the satisfiability threshold, no algorithms are proven to find solutions for many of the standard random CSP models. We now know that this observed "algorithmic barrier" is asymptotically equal to the clustering threshold as k grows, and so the difficulties are brought on by the onset of clusters ([3] provides some rigorous grounding for this). It has been suggested that this algorithmic barrier may occur precisely at the freezing threshold; i.e. the formation of clusters does not cause substantial algorithmic difficulties until the clusters have frozen variables (see section 1.1 below).

Although the clustering picture is, for the most part, not established rigorously, understanding it has led to substantial new theorems [1, 5, 6, 20–23, 28, 29, 33, 39, 47, 51]. For example, [23] used our understanding of how condensation has foiled previous second moment arguments to modify those arguments and obtain a remarkably tight bound on the satisfiability threshold for k-NAE-SAT. [21] used our understanding of clustering to design an algorithm that provably solves random k-SAT up to densities of $O(2^k \ln k/k)$, which is the asymptotic value of the clustering threshold. A particularly impressive heuristic result is the Survey Propogation algorithm [16, 42], which experimentally has solved random 3-SAT on 10^7 variables at densities far closer to the satisfiability threshold than anyone had previously been able to handle, even on fewer than 1000 variables. This algorithm was designed specifically to take advantage of the clustering picture.

Of course, another thrust has been to try to rigorously establish pieces of the clustering picture [3, 4, 8, 24, 30, 45, 47, 53]. We have been most successful with k-XOR-SAT; i.e. a random system of

[1] By this we mean: all but a vanishing proportion of the clusters, when weighted by their size.

boolean linear equations. The satisfiability threshold was established in [27] for $k = 3$ and in [26] for $k \geq 4$. More recently, [8,30] each established a very precise description of the clustering picture. It should be noted that the solutions of a system of linear equations are very well-understood, and that was of tremendous help in the study of the clustering of the solutions. Other CSP's, for which we do not have nearly as much control over the solutions, have been more resistant to rigorous analysis.

The contribution of this paper is to rigorously determine the precise freezing threshold for a family of CSP models including k-NAE-SAT and hypergraph 2-colouring. The freezing threshold for k-COL was determined by the first author in [45]; prior to this work, k-COL and k-XOR-SAT are the only two common models for which the freezing threshold was determined rigorously.

We follow the approach of [45], but we differ mainly in: (i) Where [45] analyzed the *Kempe-core*, we need to analyze the **-core*, which was introduced in [13] to prove the existence of frozen variables in random k-SAT. (ii) Rather than carrying out the analysis for a single model, we carry it out simultaneously for a family of models.

Our informal description of freezing described it in terms of the clusters. At this point, not enough information about clustering has been established rigorously to permit us to define freezing in those terms. (Eg. we do not know the clustering threshold for any interesting model except k-XOR-SAT.) So our formal definition of a frozen variable avoids the notion of clustering.

Definition 1.1. *An ℓ-path of solutions of a CSP F is a sequence $\sigma_0, \sigma_1, ..., \sigma_t$ of solutions, where for each $0 \leq i \leq t-1$, σ_i and σ_{i+1} differ on at most ℓ variables.*

Definition 1.2. *Given a solution σ of a CSP F, we say that a variable x is* ℓ-frozen *with respect to σ if for every ℓ-path $\sigma = \sigma_0, \sigma_1, ..., \sigma_t$ of solutions of F, we have $\sigma_t(x) = \sigma(x)$.*

In other words, it is not possible to change the value of v by changing at most ℓ vertices at a time. Roughly speaking, the solutions in the same cluster as σ are the solutions that can be reached by a $o(n)$-path. So x is $o(n)$-frozen with respect to σ if x has the same value in every solution in the same cluster as σ. Thus, this definition is essentially equivalent to the informal one if the clustering picture is accurate.

We make critical use of the *planted model* (section 3) introduced in [3]. We prove that one can use the planted model up to a certain density, and so we want the freezing threshold to be below that density. It will be if the constraint size k is sufficiently large; $k \geq 30$ will do.

We analyze CSP-models satisfying certain properties: non-trivial, feasible, symmetric, balance-dominated, and 1-essential (defined in section 2). The first four are needed to permit the planted model; the fifth allows us to focus on the *-core. Given such a CSP model Υ, we define constants $r_f(\Upsilon), r_p(\Upsilon)$ and function $\lambda(\Upsilon, r)$ below. Our main theorem is that $r_f(\Upsilon)$ is the freezing threshold for Υ and that $\lambda(\Upsilon, r)$ is the proportion of frozen vertices. We require the density to be below $r_p(\Upsilon)$ in order to apply the planted model. This is not just a technicality - if the density is significantly above $r_p(\Upsilon)$, then it will be above the condensation threshold and the formulas that we provide will fail to hold.

Given a CSP-model Υ, $C(\Upsilon, n, M)$ is a random instance of Υ on n variables and with M constraints (see Section 2). We say that a property holds w.h.p. (with high probability) if it holds with probability tending to 1 as $n \to \infty$.

Theorem 1.3. *Consider any non-trivial, feasible, symmetric, balance-dominated, and 1-essential CSP-model Υ with $r_f(\Upsilon) < r_p(\Upsilon)$. Let σ be a uniformly random solution of $C(\Upsilon, n, M = rn)$.*

(a) For any $r_f(\Upsilon) < r < r_p(\Upsilon)$, there exists a constant $0 < \beta < 1$ for which:

(i) w.h.p. there are $\lambda(\Upsilon, r)n + o(n)$ variables that are βn-frozen with respect to σ.

(ii) w.h.p. there are $(1 - \lambda(\Upsilon, r))n + o(n)$ variables that are not 1-frozen with respect to σ.

(b) For any $r < r_f(\Upsilon)$, w.h.p. at most $o(n)$ variables are 1-frozen with respect to σ.

In other words: for $r > r_f$, a linear number of variables are αn-frozen, while for $r < r_f$, all but at most $o(n)$ variables are not even 1-frozen. Furthermore, for $r > r_f$ we specify the specific number of αn-frozen vertices, up to an additive $o(n)$ term. All but at most $o(n)$ of the other vertices are not even 1-frozen.

We remark that for k-COL and k-XOR-SAT, we have "$\omega(n)$-frozen" rather than "1-frozen", for some $\omega(n) \to \infty$. The reason that the unfrozen variables are so much more unrestricted in the present models, arises from the fact that they are outside of the *-core. Part (b) probably remains true upon replacing "$o(n)$" with "zero". The $o(n)$ terms arises from a limitation of using the planted model.

For $k \geq 30$ we always have $r_f(\Upsilon) < r_p(\Upsilon)$ and so our theorem applies.

For densities below the freezing threshold, our proof yields that, in fact, almost all variables can be changed via a $o(n)$-path of length 1:

Theorem 1.4. *Consider any non-trivial, feasible, symmetric, balance-dominated, and 1-essential CSP-model Υ with with $r_f(\Upsilon) < r_p(\Upsilon)$ Let σ be a uniformly random solution of $C(\Upsilon, n, M = rn)$ with $r < r_f(\Upsilon)$.*

For any $\omega(n) \to \infty$, w.h.p. for all but at most $o(n)$ variables x, there is a solution σ' such that (i) $\sigma'(x) \neq s(x)$ and (ii) $\sigma'(x), \sigma(x)$ differ on at most $\omega(n)$ variables.

As mentioned above, our theorems apply to k-NAE-SAT and hypergraph 2-colouring, two of the standard benchmark models. k-NAE-SAT is a k-CNF boolean formula which is satisfied if every clause contains at least one true literal and at least one false literal. For hypergraph 2-colouring, we are presented with a k-uniform hypergraph and we need to find a boolean assignment to the vertices so that no hyperedge contains only vertices of one sign. Thus, it is equivalent to an instance of k-NAE-SAT where every literal is signed positively. See Appendix 8 for a discussion of other models to which our theorems apply.

We should emphasize that the clustering picture described above is very rough. The mathematical analysis used by statistical physicists to determine the various thresholds actually studies properties of certain Gibbs distributions on infinite trees rather than solutions of random CSP's. The clustering picture is a common geometric interpretation and it is not exact. Nevertheless, there is very strong evidence that something very close to this picture should hold.

1.1 The algorithmic barrier

A great deal of the interest in random CSP's arises from the long-established observation that as the densities approach the satisfiability threshold, the problems appear to be extremely difficult to solve [18, 43]. Much work has gone into trying to understand what exactly causes dense problems to be so algorithmically challenging (eg. [2, 19, 21, 46]).

It has been suggested (eg. [25, 34, 35, 50, 52, 54]) that, for typical CSP's, the freezing threshold forms an algorithmic barrier. For $r < r_f$ very simple algorithms (eg. greedy algorithms with minor backtracking steps) will w.h.p. find a satisfying solution, but for $r > r_f$ one requires much more sophisticated algorithms (eg. Survey Propogation). It has been proposed that the following simple algorithm should succeed for $r < r^f$:

Suppose that Theorem 1.4 were to hold for *every* solution σ. We build our CSP one random constraint at a time, letting F_i denote the CSP with i constraints. We begin with a solution σ_0 for F_0 (σ_0 can be any assignment). Then we obtain σ_{i+1} from σ_i as follows: If σ_i does not violate the $(i+1)$st constaint added, then we keep $\sigma_{i+1} = \sigma_i$. Otherwise, we modify σ_i into another solution σ' of F_i in which the values of the variables in the $(i+1)$st constraint are changed so that it is satisfied; then we set $\sigma_{i+1} = \sigma'$. If Theorem 1.4 holds for σ_i, then we can change each of the k variables in that constraint by changing only $o(n)$ other variables. Expansion properties of a random CSP imply that these small changes will (usually) not interfere with each other and so we can change each of the k variables to whatever we want. Thus we will eventually end up with a solution σ_M to our random CSP F_M.

However, Theorem 1.4 does not hold for *every* solution, only most of them. This is not just a limit of our proof techniques - it is believed that it does not hold for an exponentially small, but positive, proportion of the solutions. So proving that this algorithm works would require showing that we never encounter one of those solutions.

To see, intuitively, why the onset of freezing may create algorithmic difficulties, consider *near-solutions* - assignments which violate only a small number of constraints, say $o(n)$ of them. The near-solutions will also form clusters (because of *high energy barriers*; see [3]). Furthermore, almost all clusters of near-solutions will not contain any solutions. This is because, above the freezing threshold, almost all solution clusters have a linear number of frozen variables and so after adding only $o(n)$ constraints, we will pick a constraint that violates the frozen variables. This will violate all solutions in that cluster, thus forming a near-solution cluster that contains no actual solutions. Of course, this description is non-rigorous but it provides a good intuition.

Now consider a greedy algorithm with backtracking. As it sets its variables, it will approach a near-solution ρ. At that point, it cannot move to a near-solution in a different cluster than ρ, without employing a backtracking step that changes a linear number of variables. So the algorithm will need to be sophisticated enough to approach one of the rare near-solution clusters that contains solutions.

There is a second freezing threshold, above which *every* cluster has frozen variables. [54] suggests that this is another algorithmic barrier above which even the sophisticated algorithms fail to find solutions. One indication is that, empirically, every solution σ found by Survey Propogation is such that no variables are frozen with respect to σ. So somehow, the algorithm is drawn to those rare unfrozen clusters, and hence may fail when there are no such clusters.

1.2 Related work

The clustering picture for k-NAE-SAT and hypergraph 2-colouring was analyzed non-rigorously in [25]. There are hundreds of other papers from the statistical physics community analyzing clustering and related matters. Some are listed above; rather than listing more, we refer the reader to the book [41].

Achlioptas and Ricci-Tersenghi [13] were the first to rigorously prove that freezing occurs in a random CSP. They studied random k-SAT and showed that for $k \geq 8$, for a wide range of edge-densities below the satisfiability threshold and for *every* satisfying assignment σ, the vast majority of variables are 1-frozen w.r.t σ. They did so by stripping down to the *-core, which inspired us to do the same here. One difference between their approach and ours is that the variables of the *-core are 1-frozen by definition, whereas much of the work in this paper is devoted to proving that, for our models, they are in fact $\Theta(n)$-frozen. We expect that our techniques should be able to prove that the 1-frozen variables established in [13] are, indeed, $\Theta(n)$-frozen.

[3] provides the asymptotic (in k) value for the appearance of what they call *rigid* variables in various random CSP's, including k-NAE-SAT and hypergraph 2-colouring. The definition of rigid is somewhat weaker than frozen, but a simple modification extends their proof to show the same for frozen vertices. So [3] provided the asymptotic, in k, location of the freezing threhold for NAE-SAT and hypergraph 2-colouring.

[3, 4, 47] establish the existence of what they call *cluster-regions* for various CSP's; these are proven to be w.h.p. $\Theta(n)$-separated but are not shown to be w.h.p. well-connected, including k-NAE-SAT and hypergraph 2-colouring. They prove that by the time the density exceeds $(1+o_k(1))$ times the hypothesized clustering threshold the solution space w.h.p. shatters into an exponential number of $\Theta(n)$-separated cluster-regions, each containing an exponential number of solutions. While these cluster-regions are not shown to be well-connected, the well-connected property does not seem to be crucial to the difficulties that clusters pose for algorithms. So [3] was a very big step towards explaining why an algorithmic barrier seems to arise asymptotically close to the clustering threshold.

[9, 10] provided the first asymptotically tight lower bounds on the satisfiability threshold of k-NAE-SAT and hypergraph 2-colouring, achieving a bound that is roughly equal to the condensation threshold. [24] provides an even stronger bound for hypergraph 2-colouring, extending above the condensation threshold. [23] provides a remarkably strong bound for k-NAE-SAT - the difference between their upper and lower bounds decreases exponentially with k.

2 CSP models

A *boolean constraint* of arity k consists of k *ordered* variables $(x_1, \ldots, x_k)$ together with a boolean function $\varphi : \{-1,1\}^k \to \{0,1\}$. This function constrains the set of variables to take values $\sigma = (\sigma_1, \ldots, \sigma_k) \in \{-1,1\}^k$ such that $\varphi(\sigma_1, \ldots, \sigma_k) = 1$. We say that the constraint is *satisfied* by a boolean assignment σ if it evaluates to 1 on σ.

A *constraint satisfaction problem (CSP)* is a set of constraints, where the a^{th} constraint is formed by a boolean function φ_a over the variables $(x_{i_{1,a}}, \ldots, x_{i_{k,a}})$, with $i_{j,a} \in [n]$. A CSP, H, defines a boolean function $F^{(H)} : \{-1,1\}^n \to \{0,1\}$ given by

$$F^{(H)}(\sigma_1, \ldots, \sigma_n) := \prod_a \varphi_a(\sigma_{i_{1,a}}, \ldots, \sigma_{i_{k,a}}).$$

Given $\sigma \in \{-1,1\}^n$, we say that σ is a *satisfying assignment*, or *solution*, of the CSP H if σ satisfies every constraint of H; i.e. if $F^{(H)}(\sigma) = 1$.

A *CSP model* is a set Φ of boolean functions, together with a probability distribution $p : \Phi \to [0,1]$ defined on it (we assume implicitly that the support of p is Φ). Our random CSPs are:

Definition 2.1. *Given a CSP model $\Upsilon = (\Phi, p)$, a* random CSP*, $C(\Upsilon, n, M)$, is a CSP over the variables $\{x_1, \ldots, x_n\}$ consisting of M constraints $\{\varphi^{(a)}(x_{i_{1,a}}, \ldots, x_{i_{k,a}}) : a = 1, \ldots, M\}$ where the* *boolean constraints $\{\varphi^{(a)} : a = 1, \dots, M\}$ are drawn independently from Φ according to the distribution p, and the k-tuples $\{(x_{i_{1,a}}, ..., x_{i_{k,a}}) : a = 1, \dots, m\}$ are drawn uniformly and independently from the set of k-tuples of $\{x_1, \dots, x_n\}$.*

We consider random CSP-models $\Upsilon = (\Phi, p)$ with the following properties.

Definition 2.2.
***Non-trivial**: There is at least one $\varphi \in \Phi$ that is not satisfied by $x_1 = ... = x_k = 1$ and at least one $\varphi \in \Phi$ that is not satisfied by $x_1 = ... = x_k = -1$.*

***Feasible**: For any $\varphi \in \Phi$, and every assignment to any $k-1$ of the variables, at least one of the two possible assignments to the remaining variable will result in φ being satisfied.*

***Symmetric**: For every $\varphi \in \Phi$, and for every assignment x, we have $\varphi(x) = \varphi(-x)$, where $-x$ is the assignment obtained from x by reversing the assignment to each variable.*

***Balance-dominated** Consider a random assignment σ where each variable is independently set to be 1 with probability q and -1 with probability $1-q$, and let φ be a random constraint from Φ with distribution p. The probability that σ satisfies φ is maximized at $q = \frac{1}{2}$.*

Those four properties will allow us to apply the planted model. 'Non-trivial' is a standard property to require. 'Feasible' is also quite natural, although some models do not satisfy it. The other two properties help us to bound the second moment of the number of solutions, which in turn enables us to use the planted model.

Our final property allows us to analyze frozen variables using the *-core.

Definition 2.3. 1-essential: *Given a boolean constraint φ and an assignment σ that satisfies φ, we say that the variable x is* essential *for (φ, σ) if changing the value of x results in φ being unsatisfied. We say that a set Φ of constraints is* 1-essential *if for every $\varphi \in \Phi$, and every σ satisfying φ, at most one variable is essential for (φ, σ). A CSP-model (Φ, p) is* 1-essential *if Φ is 1-essential.*

For example: in hypergraph 2-colouring, x is essential iff its value is different from that of every other variable in ϕ; in k-XOR-SAT, every variable is essential. It is easily confirmed that k-SAT, hypergraph 2-colouring and k-NAE-SAT are 1-essential, but k-XOR-SAT is not.

3 The planted model

Consider any CSP-model $\Upsilon = (\Phi, p)$. Theorem 1.3 concerns a uniformly random satisfying assignment of $C(\Upsilon, n, M)$; i.e. a pair (F, σ) drawn from:

Definition 3.1. *The* uniform model *$U(\Upsilon, n, M)$ is a random pair (F, σ) where F is taken from the $C(\Upsilon, n, M)$ model and σ is a uniformly random satisfying solution of F.*

The uniform model is very difficult to analyze directly. So instead we turn to the much more amenable planted model:

Definition 3.2. *The* planted model *$P(\Upsilon, n, M)$ is a random pair (F, σ) chosen as follows: Take a uniformly random assignment $\sigma \in \{-1, +1\}^n$. Next select a random F drawn from $C(\Upsilon, n, M)$ conditional on σ satisfying F.*

Remark: Note that we can select F by choosing M independent constraints. Each time, we choose a uniformly random k-tuple of k variables, then choose for those variables a constraint $\varphi \in \Phi$ with probability distribution p. If σ does not satisfy the constraint then reject and choose a new one. Equivalently, we can choose the k-tuples non-uniformly where the probability that a particular k-tuple is chosen is proportional to the probability that, upon choosing φ for that set, the constraint will be satisfied by σ. Then we choose $\varphi \in \Phi$ with probability p conditional on σ satisfying φ.

It is not hard to see that the uniform and planted models are not equivalent. In the planted model, a CSP is selected with probability roughly proportional to the number of satisfying assignments. Nevertheless, Achlioptas and Coja-Oghlan [3] proved that, under certain conditions, one can transfer results about the planted model to the uniform model when Υ is k-COL, k-NAE-SAT or hypergraph 2-colouring (also k-SAT, but under stronger conditions). Montanari, Restrepo and Tetali [47] extended this to all Υ in a class of CSP-models, including all models that are non-trivial, feasible, symmetric, and balance-dominated.

For each non-trivial, feasible, symmetric and balance-dominated CSP-model Υ we define (in Appendix 8) a constant $r_p(\Upsilon)$, which is the highest density for which we can use the planted model. The following key tool essentially follows from Theorem B.3 of [47], except that they do not explicitly mention $r_p(\Upsilon)$, instead giving an implicit lower bound under appropriate conditions. It was first proven in [3] for NAE-SAT, hypergraph 2-COL and a few other models.

Lemma 3.3. *Consider any non-trivial, feasible, symmetric, and balance-dominated CSP-model Υ. For every $r < r_p(\Upsilon)$, there is a function $g(n) = o(n)$ such that: Let $\mathcal{E}$ be any property of pairs (F, σ) where σ is a satisfying solution of F. If*

$$\mathbf{Pr}(P(\Upsilon, n, M = rn) \text{ has } \mathcal{E}) > 1 - e^{-g(n)},$$

then

$$\mathbf{Pr}(U(\Upsilon, n, M = rn) \text{ has } \mathcal{E}) > 1 - o(1).$$

In Appendix 8, we prove that if Υ is also 1-essential, then for $k \geq 30$, we have $r_p(\Upsilon) > r_f(\Upsilon)$ and so Theorem 1.3 is non-trivial. In fact, $r_p(\Upsilon) = \Theta(\frac{k}{\ln k}) r_f(\Upsilon)$. The bound $k \geq 30$ can be lowered, and for some specific models Υ it can be lowered significantly. For example, for k-NAE-SAT and hypergraph 2-colouring, one can probably prove that $k \geq 6$ will do.

4 The *-core

The *-core was introduced in [13] to study frozen variables in random k-SAT.

Fix a satisfying assignment σ, and consider a variable x. Suppose that there are no constraints φ such that x is essential for (φ, σ). Then, by the definition of essential, we can change x and still have a satisfying assignment. So x is not frozen. This inspires the following:

Definition 4.1. *Consider a CSP F with a satisfying assignment σ. The* *-core *of (F, σ) is the sub-CSP formed as follows:*
Iteratively remove every variable x such that for every constraint $\varphi \in F$: x is not essential for (φ, σ). When we remove a variable, we also remove all constraints containing that variable.

Note that the order in which variables are deleted will not affect the outcome of the iterative procedure. So the *-core is well-defined, albeit possibly empty.

As described above, it is clear that the first variable removed is not frozen. Expansion properties of a random CSP - in particular the fact that it is locally tree-like - imply that almost every variable removed is not frozen. Furthermore, we will prove that if the model is 1-essential then almost all variables that remain in the *-core are frozen. Having proven those two key results, Theorem 1.3 follows from an analysis of the *-core process.

Now suppose that our CSP-model is 1-essential. A key observation is that the *-core depends only on the constraints that have essential variables. I.e., if we first remove all constraints with no essential variables from the CSP and then apply the *-core process, the set of vertices in the resultant *-core will not change.

Definition 4.2. *Given a 1-essential CSP, F, and a satisfying solution σ, we define* the hypergraph $\Gamma(F, \sigma)$ *as follows: The vertices are the variables of F and the variables of each constraint of F form a hyperedge, if that constraint has an essential variable. That essential variable is called the* essential vertex *of the hyperedge.*

Note that we can find the *-core of (F, σ) by repeatedly deleting from $\Gamma(F, \sigma)$ vertices that are not essential in any hyperedges. The resulting hypergraph is called the **-core of* $\Gamma(F, \sigma)$.

The precise model for the random hypergraph $\Gamma(F, \sigma)$ varies with Υ (see appendix 10). However, the size of the *-core as a function of the number of hyperedges is the same for all such models.

We define:

$$\alpha_k := \inf_{x>0} \frac{x}{(1-e^{-x})^{k-1}}.$$

Also, for $\alpha > \alpha_k$, let $x_k(\alpha)$ be the maximum value of $x \geq 0$ such that $\frac{x}{(1-e^{-x})^{k-1}} = \alpha$ and set

$$\rho_k(\alpha) = 1 - e^{-x_k(\alpha)}.$$

In Appendix 11, we prove

Lemma 4.3. *Consider any 1-essential CSP-model $\Upsilon = (\Phi, p)$ of arity k, and a random CSP, F, drawn from $P(\Upsilon, n, M = rn)$. Suppose $\Gamma(F, \sigma)$ has $\alpha n + o(n)$ hyperedges. For any $g(n) = o(n)$, with probability at least $1 - e^{-g(n)}$:*

*(a) If $\alpha > \alpha_k$ then the *-core of $\Gamma(F, \sigma)$ has $\rho_k(\alpha)n + o(n)$ vertices.*

*(b) If $\alpha < \alpha_k$ then the *-core of $\Gamma(F, \sigma)$ has $o(n)$ vertices.*

This allows us to analyze our family of models simultaneously by working directly with the *-core of $\Gamma(F, \sigma)$. We prove that almost all vertices of the *-core are $\Theta(n)$-frozen variables in F and almost all vertices outside of the *-core are not even 1-frozen in F.

In Appendix 9, we define for any 1-essential CSP-model $\Upsilon = (\Phi, p)$, a constant $\xi(\Upsilon) > 0$ and prove:

Lemma 4.4. *For any $g(n) = o(n)$ and $r > 0$, with probability at least $1 - e^{-g(n)}$, the number of constraints in $P(\Upsilon, n, M = rn)$ that have an essential variable is $\xi(\Upsilon)rn + o(n)$.*

This yields Theorem 1.3 (see appendix 10) with:

$$r_f(\Upsilon) = \alpha_k / \xi(\Upsilon); \qquad \lambda(\Upsilon, r) = \rho_k(\xi(\Upsilon)r).$$

In Appendix 10, we describe the models that we use to analyze $\Gamma(F, \sigma)$ and the *-core of $\Gamma(F, \sigma)$.

5 Unfrozen variables outside of the *-core

Let x be a vertex of $\Gamma(F,\sigma)$ which is not in the *-core of $\Gamma(F,\sigma)$. We will consider how x can be removed during the peeling process used to find the *-core of $\Gamma(F,\sigma)$. More specifically, we consider a sequence of vertices, culminating in x, which could be removed in sequence by the peeling process.

Definition 5.1. *A* peeling chain *for a vertex $x \in \Gamma(F,\sigma)$ is a sequence of vertices $x_1,...,x_\ell = x$ such that each x_i is not essential for any hyperedges in the hypergraph remaining after removing $x_1,...,x_{i-1}$ from $\Gamma(F,\sigma)$. The* depth *of the chain is the maximum distance from one of the vertices to x. The* *-depth *of x is the minimum depth over all peeling chains for x.*

In Appendix 11, we will prove:

Lemma 5.2. *For any $\epsilon > 0$, there exists constant L such that: For all $g(n) = o(n)$, the probability that at least ϵn vertices of $\Gamma(F,\sigma)$ that are not in the *-core of $\Gamma(F,\sigma)$ have *-depth greater than L is less than $e^{-g(n)}$.*

This is enough to prove that all but $o(n)$ variables outside the *-core are 1-frozen as follows:

Consider any $\epsilon > 0$. With probability at least $1 - e^{-g(n)}$, $\Gamma(F,\sigma)$ has fewer than ϵn vertices of *-depth greater than L. Consider any vertex x of *-depth at most L. Consider a peeling chain for x of depth at most L and let W be the set of all hyperedges that contain at least one vertex of the peeling chain.

If no hyperedges of W form a cycle, then it is easy to see that we can change all of the variables in the peeling chain, one-at-a-time and still have a satisfying assignment for F. Indeed, this follows from a straightforward induction on L. Therefore, the variable x is not 1-frozen. The case where W contains a cycle is rare enough to be negligible (see the appendix). So for all $\epsilon > 0$ there are fewer than ϵn variables outside of the *-core that are not 1-frozen, as required.

This argument also leads to:

Proof of Theorem 1.3: This theorem follows as above, by adding the observation that with sufficiently high probability, almost all vertices outside the *-core have a peeling chain of size $O(1)$. We can change the corresponding variable by changing a subset of the entire peeling chain. See Appendix 10 for the short proof. □

6 Frozen variables in the *-core

Most of the work in this paper is in proving that almost all vertices in the *-core of $\Gamma(F,\sigma)$ are $\Theta(n)$-frozen. To do so, we first study the structure of sets of variables that can be changed to obtain a new solution. Note that if changing the value of every variable of S yields a solution, then every constraint whose essential variable is in S must contain at least one other variable in S. This leads us to define:

Definition 6.1. *A* flippable set *of the *-core of $\Gamma(F,\sigma)$ is a set of vertices S such that for every $x \in S$ and every *-core hyperedge f in which x is essential, S contains another vertex of f.*

For every vertex $x \in S$, since x is in the *-core, there will be at least one such hyperedge f.

We prove that for some $\phi'(n) = o(n)$ and constant $\zeta > 0$, with sufficiently high probabilty, there are no flippable sets of size $\phi'(n) < a < \zeta n$. This will be enough to prove that at most $o(n)$ vertices lie in flippable sets, which in turn will be enough to show that almost all of the *-core is frozen.

We apply the first moment method. Unfortunately, we cannot apply it directly to the number of flippable sets because the existence of one flippable set S typically leads to the existence of an exponential number of flippable sets formed by adding to S vertices x such that (i) x is essential in exactly one hyperedge, and (ii) that hyperedge contains a non-essential vertex in S. So instead we focus on something that we call *weakly flippable sets*, which do not contain such vertices x. Roughly speaking: every flippable set can be formed from a weakly flippable set by repeatedly adding vertices x in that manner. We prove that with sufficently high probability:

(a) There are no weakly flippable sets of size $\phi(n) < a < \zeta n$.

(b) There are no weakly flippable sets of size at most $\phi(n)$ which extend to a flippable set of size greater than $\phi'(n)$.

This establishes our bound on the sizes of flippable sets. (This is not quite true - we also need to consider *cyclic sets* - but it provides a good intuition.)

Let H_1 denote the vertices that are essential in exactly one hyperedge. Define a *one-path* to be a sequence of vertices $x_1, ..., x_{t+1}$ such that for each $1 \leq i \leq t$: $x_i \in H_1$ and x_{i+1} is in the hyperedge in which x_i is essential. Note that if x_{t+1} is in a flippable set S, then we can add the entire one-path to S and it will still be flippable. This ends up implying that if we have a proliferation of long one-paths, then we would not be able to prove (b). It turns out that a proliferation of long one-paths would also prevent us from proving (a).

Consider a vertex $x \in H_1$ and the edge f in which x is essential. Intutively, the expected number of other members of H_1 that are in f is $(k-1)|H_1|$ divided by the size of the *-core. We prove (Lemma 10.3) that this ratio is less than 1. This implies that one-paths do not “branch” and so we do not tend to get many long one-paths. So our bound on this ratio plays a key role in establishing both (a) and (b).

This is just an intuition. In fact, *one-paths* are not explicitly mentioned anywhere in the proofs. For all the details, see Appendix 12.

7 Further Challenges

Of course, one ongoing challenge is to continue to rigorously establish parts of the clustering picture. By now, it is clear that in order to establish satisfiability thresholds or understand the algorithmic challenges for problems with densities approaching that threshold, we will need a strong understanding of clustering.

Another challenge is to try to establish whether the freezing threshold is, indeed, an algorithmic barrier. For several CSP-models, we now know the precise location of that threshold, and we have a very good understanding of how it arises and which variables are frozen. Perhaps we can use that understanding to prove that a simple algorithm works for all densities up to that threshold and/or establish that frozen clusters will indeed neccesitate more sophistication.

Another challenge is to determine the freezing threshold for a wider variety of CSP-models. These techniques rely crucially on the planted model; at this point there is no known way to get to the exact threshold without it. This prevents us from extending our results to k-SAT and many other models as the planted model does not work nearly well enough, mainly because the number of solutions is not sufficiently concentrated. A more important challenge would be to devise a better means to analyze random solutions to CSP's drawn from those models.

Acknowledgment

The authors are supported by an NSERC Discovery Grant and an NSERC Accelerator Fund.

References

[1] E. Abbe, A. Montanari. *On the concentration of the number of solutions of random satisfiability formulas.* arXiv:1006.3786v1

[2] D. Achlioptas, P. Beame, and M. Molloy. *A sharp threshold in proof complexity yields lower bounds for satisfiability search.* J. Comput. Syst. Sci., **68** (2), 238–268 (2004).

[3] D. Achlioptas and A. Coja-Oghlan. *Algorithmic Barriers from Phase Transitions.* Proceedings of FOCS (2008), 793 - 802. Longer version available at arXiv:0803.2122

[4] D. Achlioptas, A. Coja-Oghlan and F. Ricci-Tersenghi. *On the solution-space geometry of random constraint satisfaction problems.* Random Structures and Algorithms **38** (2011), 251 - 268.

[5] D. Achlioptas and R. Menchaca-Mendez. *Exponential lower bounds for DPLL algorithms on satisfiable random 3-CNF formulas.* Proceedings of SAT (2012).

[6] D. Achlioptas and R. Menchaca-Mendez. *Unsatisfiability bounds for random CSPs from an energetic interpolation method.* Proceedings of ICALP (2012).

[7] D. Achlioptas and M. Molloy. *The analysis of a list- coloring algorithm on a random graph.* Proceedings of FOCS (1997), 204 212.

[8] D. Achlioptas and M. Molloy. *The solution space geometry of random linear equations.* arXiv:1107.5550v1

[9] D. Achlioptas and C. Moore. *On the 2-colorability of random hypergraphs.* Proceedings of RANDOM (2002).

[10] D. Achlioptas and C. Moore. *Random k-SAT: Two moments suffice to cross a sharp threshold.* SIAM J. Comp., 36, (2006), 740 - 762.

[11] D. Achlioptas and A. Naor. *The two possible values of the chromatic number of a random graph.* Annals of Mathematics, **162** (2005), 1333 1349.

[12] D. Achlioptas and Y. Peres. *The threshold for random k-SAT is* $2^k \log 2 - O(k)$. J.AMS **17** (2004), 947 - 973.

[13] D. Achlioptas and F. Ricci-Tersenghi. *On the solution-space geometry of random constraint satisfaction problems.* Proceedings of STOC (2006), 130 - 139.

[14] N. Alon and J. Spencer. *The Probabilistic Method.* Wiley.

[15] K. Azuma. *weighted sums of certain dependent random variables.* Tokuku Math. J. **19** (1967), 357 - 367.

[16] A. Braunstein, M. Mezard and R. Zecchina. *Survey propagation: an algorithm for satisfiability.* Random Structures and Algorithms **27** (2005), 201 - 226.

[17] S. Chan and M. Molloy. *A dichotomy theorem for the resolution complexity of random constraint satisfaction problems.* Proceedings of FOCS 2008.

[18] P. Cheeseman, B. Kanefsky and W. Taylor. *Where the really hard problems are.* Proceedings of IJCAI (1991), 331 - 337.

[19] V. Chvátal and E. Szemerédi. *Many hard examples for resolution.* J. ACM, **35**(4), 759-768 (1988).

[20] A. Coja-Oghlan. *A better algorithm for random k-SAT.* SIAM Journal on Computing **39** (2010), 2823 - 2864.

[21] A. Coja-Oghlan. *On belief propagation guided decimation for random k-SAT.* Proc. 22nd SODA (2011), 957 - 966.

[22] A. Coja-Oghlan and C. Efthymiou. *On independent sets in random graphs.* Proc. 22nd SODA (2011), 136 - 144.

[23] A. Coja-Oghlan and K. Panagiotou. *Catching the k-NAESAT threshold.* Proceedings of STOC (2012).

[24] A. Coja-Oghlan and L. Zdeborov. *The condensation transition in random hypergraph 2-coloring.* Proceedings of SODA (2012).

[25] L. Dall'Asta, A. Ramezanpour and R. Zecchina. *Entropy landscape and non-Gibbs solutions in constraint satisfaction problems.* Phys. Rev. E 77, 031118 (2008).

[26] M. Dietzfelbinger, A. Goerdt, M. Mitzenmacher, A. Montanari, R. Pagh and M. Rink Tight thresholds for cuckoo hashing via XORSAT. Preprint (2010), arXiv:0912.0287v3

[27] O. Dubois and J. Mandler. *The 3-XORSAT threshold.* In Proc. 43rd FOCS (2002), p 769.

[28] U. Feige, A. Flaxman, and D. Vilenchik. *On the diameter of the set of satisfying assignments in random satisfiable k-CNF formulas.* SIAM J. Disc.Math. **25** (2011), 736 - 749. (2011)

[29] A. Gerschenfeld and A. Montanari. *Reconstruction for models on random graphs.* Proceedings of FOCS 2007.

[30] M. Ibrahimi, Y. Kanoria, M. Kraning and A. Montanari. *The set of solutions of random XORSAT formulae.* Proceedings of SODA 2012. Longer version available at arXiv:1107.5377

[31] S. Janson, T. Łuczak and A. Ruciński. Random Graphs. Wiley, New York (2000).

[32] J.H.Kim. *Poisson cloning model for random graphs.* arXiv:0805.4133v1

[33] M. Krivelevich, B. Sudakov, and D. Vilenchik. *On the random satisfiable process.* Combinatorics, Probability and Computing **18** (2009), 775 - 801.

[34] F. Krzakala and J. Kurchan. *Constraint optimization and landscapes.*

[35] F. Krzakala and J. Kurchan. *A landscape analysis of constraint satisfaction problems.*

[36] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian and L. Zdeborova. *Gibbs States and the Set of Solutions of Random Constraint Satisfaction Problems.* Proc. Natl. Acad. Sci., (2007).

[37] F. Krzakala, A. Pagnani and Martin Weigt. *Threshold values, stability analysis, and high-q asymptotics for the coloring problem on random graphs.* Phys. Rev. E, 70(4):046705, (2004).

[38] S. Kudekar and N. Macris. *Decay of correlations for sparse graph error correcting codes.* SIAM J. Disc. Math. **25** (2011), 956 - 988.

[39] E. Maneva, E. Mossel and M. J. Wainwright. *A new look at Survey Propagation and its generalizations.* JACM **54**, (2007).

[40] M. Mézard, T. Mora, and R. Zecchina *Clustering of Solutions in the Random Satisfiability Problem.* Phys. Rev. Lett., **94** (19), 197205 (2005).

[41] M. Mezard and A. Montanari. *Information, Physics and Computation.* Oxford University Press, (2009).

[42] M. Mezard, R. Zecchina *The random K-satisfiability problem: from an analytic solution to an efficient algorithm.* Phys. Rev. E **66**, (2002).

[43] D. Mitchell, B. Selman and H. Levesque. *Hard and Easy Distributions of SAT Problems.* Proceedings of AAAI 1992, 459 - 465.

[44] M. Molloy. *Cores in random hypergraphs and boolean formulas.* Random Structures and Algorithms **27**, 124 - 135 (2005).

[45] M. Molloy. *The freezing threshold for k-colourings of a random graph.* Proceedings of STOC (2012).

[46] M. Molloy and M. Salavatipour. *The resolution complexity of random constraint satisfaction problems.* SIAM J. Comp. **37**, 895 - 922 (2007).

[47] A. Montanari, R. Restrepo and P. Tetali. *Reconstruction and clustering in random constraint satisfaction problems.* SIAM J. Disc. Math. **25** (2011), 771 - 808.

[48] E. Mossel and Y. Peres. *Information flow on trees.* Ann. Appl. Probab. **13** (2003),817 844.

[49] R. Mulet, A. Pagani, M. Weigt and R. Zecchina. *Coloring random graphs.* Phys. Rev. Lett. **89**, 268701 (2002).

[50] G. Semerjian, *On the freezing of variables in random constraint satisfaction problems.*

[51] A. Sly. *Reconstruction of random colourings.* Commun. Math. Phys. **288** (2009), 943 961.

[52] L. Zdeborová and F. Krzakala. *Phase transitions in the colouring of random graphs.* Phys. Rev. E 76, 031131 (2007).

[53] L. Zdeborová and F. Krzakala. *Quiet planting in the locked constraint satisfaction problems.* SIAM J. Discrete Math. **25** (2011) 750 - 770.

[54] L. Zdeborová. *Statistical physics of hard optimization problems.* Acta Physica Slovaca **59** (2009), 169 - 303.

Appendix

8 The transfer theorem

We recall our properties of a CSP-model $\Upsilon = (\Phi, p)$:

Definition 8.1.
***Non-trivial**: There is at least one $\varphi \in \Phi$ that is not satisfied by $x_1 = ... = x_k = 1$ and at least one $\varphi \in \Phi$ that is not satisfied by $x_1 = ... = x_k = -1$.*

***Feasible**: For any $\varphi \in \Phi$, and every assignment to any $k-1$ of the variables, at least one of the two possible assignments to the kth variable will result in φ being satisfied.*

***Symmetric**: For every $\varphi \in \Phi$, and for every assignment x, we have $\varphi(x) = \varphi(-x)$, where $-x$ is the assignment obtained from x by reversing the assignment to each variable.*

***Balance-dominated** Consider a random assignment σ where each variable is independently set to be 1 with probability q and -1 with probability $1-q$, and let φ be a random constraint from Φ with distribution p. The probability that σ satisfies φ is maximized at $q = \frac{1}{2}$.*

Given a boolean function $\varphi \in \Phi$, denote by $S_\varphi \subset \{-1, +1\}^k$ the set of satisfying assignments of φ and by I_φ its complement. Now, let $\varphi(x) = \sum_{Q \subseteq \{-1,1\}^k} \varphi_Q \prod_{i \in Q} x_i$ be its *Fourier expansion.* Such expansion is unique with $\varphi_Q = \sum_{x \in \{-1,1\}^k} \varphi(x) \prod_{i \in Q} x_i$. Notice, in particular, that $\varphi_\emptyset = \frac{|S_\varphi|}{2^k} = \sum_Q \varphi_Q^2$ and also that $\varphi_{\{i\}} = 0$ if φ is balanced. Now, we define the polynomial $p_\varphi(\theta)$ as follows,

$$p_\varphi(\theta) := \sum_{Q \subseteq \{-1,1\}^k} (\varphi_Q / \varphi_\emptyset)^2 \theta^{|Q|}$$

Also, we define the binary entropy function $H(\theta)$ as

$$H(\theta) := -\frac{1+\theta}{2} \ln(1+\theta) - \frac{1-\theta}{2} \ln(1-\theta)$$

We define

$$r_p(\Upsilon) := \inf_{\theta \in (0,1)} \frac{-H(\theta)}{\mathbf{Exp}_\varphi[\ln(p_\varphi(\theta))]}.$$

We will now prove Lemma 3.3, which we restate:
Lemma 3.3 *Consider any non-trivial, feasible, symmetric, and balance-dominated CSP-model Υ. For every $r < r_p(\Upsilon)$, there is a function $g(n) = o(n)$ such that: Let $\mathcal{E}$ be any property of pairs (F, σ) where σ is a satisfying solution of F. If*

$$\mathbf{Pr}(P(\Upsilon, n, M = rn) \textit{ has } \mathcal{E}) > 1 - e^{-g(n)},$$

then

$$\mathbf{Pr}(U(\Upsilon, n, M = rn) \textit{ has } \mathcal{E}) > 1 - o(1).$$

The proof follows the argument employed in [47] to prove Theorem B.3, which followed the same spirit of similar results in [3].

Proof. In what follows, we will take expectations over a random φ chosen from Φ with distribution p. Thus, for a variable $X(\varphi)$, we have $\mathbf{Exp}(X) = \sum_{\varphi\in\Phi} p(\varphi)X(\varphi)$.

Let ξ_φ be the number of clauses with constraint φ in the random CSP H drawn from Υ. Let $\mathcal{F}$ be the event 'For all $\varphi \in \Phi$, $|\xi_\varphi - \alpha p_\varphi n| < n^{1/2+\gamma}$' ($\gamma$ fixed in $(0, 1/2)$). So, $\mathcal{F}$ holds w.h.p.

We say that a solution σ is *balanced* if the number of variables assigned $+1$ is either $\lceil \frac{n}{2} \rceil$ or $\lfloor \frac{n}{2} \rfloor$. Let Z_b be the number of balanced solutions of H, let Z be the number of solutions of H and let $Z_b(\theta)$ be the number of *pairs* of balanced solutions $x^{(1)}$,$x^{(2)}$ with discrepancy θ, that is, such that $\frac{1}{n}\sum_{i=1}^n x_i^{(1)} x_i^{(2)} = \theta$.

Now,

$$\frac{\mathbf{Exp}[Z_b^2 \mathbf{I}(\mathcal{F})]}{(\mathbf{Exp}[Z_b \mathbf{I}(\mathcal{F})])^2} = \sum_{\theta\in U_n} \frac{\mathbf{Exp}[Z_b(\theta)\mathbf{I}(\mathcal{F})]}{(\mathbf{Exp}[Z_b \mathbf{I}(\mathcal{F})])^2}$$

where $U_n := \{i/n : i = -n, \ldots, n\}$. From lemma A.2 in [47], it is the case that

$$\frac{\mathbf{Exp}[Z_b(\theta)\mathbf{I}(\mathcal{F})]}{(\mathbf{Exp}[Z_b \mathbf{I}(\mathcal{F})])^2} \leq Cn^{-1/2} \exp\left(n(H(\theta) + \alpha \mathbf{Exp}[\ln(p_\varphi(\theta))] + o(1))\right)$$

where C does not depends on θ (neither the $o(1)$ term). Now, if $\alpha < r_p(\Upsilon)$, it is the case that

$$H(\theta) + \alpha \mathbf{Exp}[\ln(p_\varphi(\theta))] < 0 \text{ for all } \theta \in (0, 1). \tag{1}$$

On the other hand, since Υ is symmetric, $\varphi(\theta) = (-\frac{1}{2} + \alpha \mathbf{Exp}\left[\sum_{|Q|=2}(\varphi_Q/\varphi_\emptyset)^2\right])\theta^2 + O(\theta^3)$. Therefore, using the fact that

$$\lim_{\theta\to 0} \frac{-H(\theta)}{\mathbf{Exp}[\ln(p_\varphi(\theta))]} = \frac{1/2}{\mathbf{Exp}\left[\sum_{|Q|=2}(\varphi_Q/\varphi_\emptyset)^2\right]},$$

then it is the case $H(\theta) + \alpha \mathbf{Exp}[\ln(p_\varphi(\theta))] < -c\theta^2$ for some $c > 0$ and θ close enough to 0. Combining this fact with eq. (1) we have that for some $c' > 0$,

$$H(\theta) + \alpha \mathbf{Exp}[\ln(p_\varphi(\theta))] < -c'\theta^2 \text{ for all } \theta \in (0, 1). \tag{2}$$

Now,

$$\begin{aligned} \frac{\mathbf{Exp}[Z_b^2 \mathbf{I}(\mathcal{F})]}{(\mathbf{Exp}[Z_b \mathbf{I}(\mathcal{F})])^2} &\leq \frac{C}{n^{1/2}} \sum_{\theta\in U_n} \exp(-c'n(\theta^2 + o(1))) & (3) \\ &\leq Cn^{1/2} \int_{-\infty}^{\infty} \exp(-c'n(\theta^2 + o(1))) & (4) \end{aligned}$$

And the last quantity is bounded by a constant C_0 (not depending on n). Thus, from Paley-Zygmund, for every $\epsilon > 0$ and all $n \geq n_0$ it is the case that $\Pr(Z_b > e^{-n\epsilon}\mathbf{Exp}[Z_b]) \geq C_0/2$.

Now, because Υ is balance-dominated, we have that $\mathbf{Exp}[Z] \leq n\mathbf{Exp}[Z_b]$. Therefore, for n large enough, we have that

$$\Pr(Z > e^{-n\epsilon}\mathbf{Exp}[Z]) \geq \Pr(Z_b > ne^{-n\epsilon}\mathbf{Exp}[Z_b]) \geq \Pr(Z_b \geq e^{-n(\epsilon/2)}\mathbf{Exp}[Z_b]) \geq c_{\epsilon/2}.$$

Now, it is easy to see that $\mathbf{Exp}[Z]$ is exponential in n for $\alpha < r_p(\Upsilon)$ (Indeed $\mathbf{Exp}[Z]$ is exponential for $\alpha < r_{sat}(\Upsilon) := \frac{\ln 2}{\mathbf{Exp}_\varphi[\ln(1+|I_\varphi|/|S_\varphi|)]} = \frac{-H(1)}{\mathbf{Exp}[\ln(p_\varphi(1))]}$). Therefore, recalling the result from

Appendix C in [47], the event '$Z > e^{-n\epsilon}\mathbf{Exp}[Z]$' has a sharp threshold in the parameter α. (There, it is proved that the event '$Z > B^n$', where $B > 1$, has a sharp threshold for non-trivial feasible CSP models). Therefore, necesarily, $\Pr(Z > e^{-n\epsilon}\mathbf{Exp}[Z]) = 1 - o_n(1)$. This implies therefore, that for some function $g(n)$ of order $o(n)$, it is the case that w.h.p.,

$$\ln(Z) > \ln(\mathbf{Exp}(Z)) - g(n). \tag{5}$$

After this equation is established now the lemma follows. For instance, from Theorem B.3 in [47]. □

Now, recall our other property:

1-essential: *Given a boolean constraint φ and an assignment σ that satisfies φ, we say that the variable x is* essential *for (φ, σ) if changing the value of x results in φ being unsatisfied. We say that a set Φ of constraints is* 1-essential *if for every $\varphi \in \Phi$, and every σ satisfying φ, at most one variable is essential for (φ, σ). A CSP-model (Φ, p) is* 1-essential *if Φ is 1-essential.*

An easy description of a feasible, 1-essential constraint is the following: φ is feasible and 1-essential iff the Hamming distance between any pair of assignments in I_φ is greater than 2. This implies in particular that $|I_\varphi| \leq \frac{2^k}{\binom{k}{2}+1}$. Also, notice that $\varphi_{\{i,j\}} = -\frac{1}{2^k}\sum_{x \in I_\varphi} x_i x_j$. This allows us to prove a more concrete lower bound on the transfer threshold $r_p(\Upsilon)$ that we will use in the next section to establish that $r_p(\Upsilon)$ is above the freezing threshold for large enough k.

Theorem 8.2. *In addition to the above properties, suppose that the CSP model is 1-essential. Define*

$$\Omega_p(\Upsilon) := \mathbf{Exp}_\varphi\left(\frac{|I_\varphi|}{|S_\varphi|}\right),$$

then, it is the case that

$$r_p(\Upsilon) \geq \frac{0.25}{\Omega_p(\Upsilon)}.$$

Proof. Since every constraint $\varphi \in \Phi$ is feasible and 1-essential, we have that

$$\sum_{\{i,j\}}\left(\frac{\varphi_{i,j}}{\varphi_\emptyset}\right)^2 = \sum_{\{i,j\}}\frac{\left(\sum_{x \in I_\varphi} x_i x_j\right)^2}{|S_\varphi|^2} \leq \binom{k}{2}\left(\frac{|I_\varphi|}{|S_\varphi|}\right)^2$$

Therefore, since

$$\sum_{|Q|\geq 3}\varphi_Q^2\theta^{|Q|} \leq \sum_{|Q|\geq 3}\varphi_Q^2\theta^3 \leq \left(\sum_Q \varphi_Q^2 - \varphi_\emptyset^2\right)\theta^3 = \varphi_\emptyset(1-\varphi_\emptyset)\theta^3,$$

we have that

$$p_{\varphi(\theta)} \leq 1 + \binom{k}{2}\left(\frac{|I_\varphi|}{|S_\varphi|}\right)^2\theta^2 + \frac{|I_\varphi|}{|S_\varphi|}\theta^3$$

And, since $|I_\varphi| \leq \frac{2^k}{\binom{k}{2}+1}$, and therefore $\binom{k}{2}\left(\frac{|I_\varphi|}{|S_\varphi|}\right)^2 \leq \frac{|I_\varphi|}{|S_\varphi|}$, we get that

$$p_{\varphi(\theta)} \leq 1 + 2\frac{|I_\varphi|}{|S_\varphi|}\theta^2$$

Thus,

$$\mathbf{Exp}_{\varphi}[\ln(p_{\varphi}(\theta))] \leq 2\theta^2\Omega_p(\Upsilon)$$

Now, we finally conclude that

$$r_p(\Upsilon) = \inf_{\theta\in(0,1)} \frac{-H(\theta)}{\mathbf{Exp}_{\varphi}[\ln(p_{\varphi}(\theta))]} \geq \frac{0.5}{\Omega_p(\Upsilon)} \inf_{\theta\in(0,1)} \frac{-H(\theta)}{\theta^2} = \frac{0.25}{\Omega_p(\Upsilon)}. \tag{6}$$

□

We close this section by discussing the CSP-models that satisfy our five conditions: non-trivial, feasible, symmetric, balance-dominated, and 1-essential.

Our properties are rich enough to permit a large class of CSP-models beyond hypergraph 2-coloring and k-NAE-SAT. For example, we can construct a model in the following way:

Represent the assignments in $\{-1,+1\}^k$ as the k-dimensional hypercube H_k, and so two assignments are adjacent if the differ on exactly one variable. Let L_ϵ denote the vertices $x \in H_k$ with $\sum x_k > \epsilon k$. Consider any subset $I \subseteq L_\epsilon$ containing no two vertices of distance at most two. We use $-I$ to denote the subset formed by switching the sign of every vertex in I, and set $J := I \cup -I$ to be the assignments which violate our constraint φ_J. I.e., $\varphi_J(x) := 1$ iff $x \notin J$.

Now consider any set Φ of constraints of this form in which at least one is non-trivial (i.e. has $(1,1,...,1) \in J$). Let $\Upsilon = (\Phi, p)$ for any p (such that *supp*$(p) = \Phi$). For *any k large enough in terms of* ϵ, Φ satisfies our five properties. For instance, hypergraph 2-coloring is formed in this way with $I := (1,...1)$.

Given a constraint φ and some $s \in \{-1,+1\}^k$, we define the constraint φ^s as $\varphi^s(x_1,...,x_k) = \varphi(s_1x_1,...,s_kx_k)$. We can allow $\epsilon = 0$ and drop the condition that k must be large if (a) no two vertices of J are within distance 2, and (b) for every $\varphi \in \Phi$ and every $s \in \{-1,+1\}^k$, we have $\varphi^s \in \Phi$ and $p(\varphi^s) = p(\varphi)$. For instance, k-NAE-SAT is formed in this way with $I := (1,...,1)$.

9 Essential hyperedges

Consider any nontrivial, feasible, symmetric 1-essential CSP-model $\Upsilon = (\Phi, p)$. We will draw (F, σ) from the planted model $P(\Upsilon, n, M)$. We begin by taking a random assignment σ for the variables $x_1,...,x_n$ and note that $|\Lambda^+|, |\Lambda^-| = \frac{1}{2}n + o(n)$ with probability at least $1 - e^{-g(n)}$, for any $g(n) = o(n)$. So we can assume that this condition holds.

In what follows, we will take expectations over a random φ chosen from Φ with distribution p. Thus, for a variable $X(\varphi)$, we have $\mathbf{Exp}(X) = \sum_{\varphi\in\Phi} p(\varphi)X(\varphi)$.

For every $\varphi \in \Phi$, recall from the previous section that S_φ is the set of assigments in $\{-1,+1\}^k$ that satisfy φ and $I_\varphi = \overline{S_\varphi}$ is the set that do not satisfy φ. We define $S^e_\varphi \subseteq S_\varphi$ to be the set of assignments that satisfy φ and for which φ has an essential variable. Noting that switching the essential variable of an assignment in S^e_φ yields an assignment in I_φ, and using the fact that Υ is feasible, it is easy to see that $|S^e_\varphi| = k|I_\varphi|$.

Since $|\Lambda^+|, |\Lambda^-| = \frac{1}{2}n + o(n)$, it follows that when picking a constraint in the planted model, we choose φ with probability proportional to $p(\varphi)|S_\varphi| + o(1)$. Thus, defining $\Omega_f := \frac{\mathbf{Exp}|I_\varphi|}{\mathbf{Exp}|S_\varphi|}$, the probability that φ has an essential variable is:

$$\xi(\Upsilon) = k\Omega_f + o(1).$$

So the number of constraints that have an essential variable is distributed as the binomial $BIN(M = rn, \xi(\Upsilon))$. Concentration of the binomial variable implies Lemma 4.4.

Now recall the type of φ, as defined in Section 10.1. For a constraint $\varphi \in \Phi$, define $I_\varphi(a, b) := \{x \in I_\varphi : x \text{ has } a\ 1's \text{ and } b\ -1's\}$ then the clause φ has exactly $(b+1)|I_\varphi(a, b+1)|$ assignments of type $(1; a, b)$ and $(a+1)|I_\varphi(a+1, b)|$ assignments of type $(-1; a, b)$. Therefore, when picking a constraint in the planted model, if we condition on the event that it has an essential variable, then the conditional probability that it has type $\tau = (1; a, b)$ is

$$\gamma_\tau = \frac{(b+1)\mathbf{Exp}[|I_\varphi(a, b+1)|]}{k\mathbf{Exp}[|I_\varphi|]} + o(1)$$

and to be of type $\tau = (-1; a, b)$ is

$$\gamma_\tau = \frac{(a+1)\mathbf{Exp}[|I_\varphi(a+1, b)|]}{k\mathbf{Exp}[|I_\varphi|]} + o(1)$$

Since Υ is symmetric, $\varphi(x) = \varphi(-x)$ for every assignment x. It follows that $|I_\varphi(a, b)| = |I_\varphi(b, a)|$ and therefore $\gamma_{\tau=(1;a,b)} = \gamma_{\tau=(-1;b,a)} + o(1)$.So, noting that we can exchange a, b in the following definition:

$$\gamma^+ := \sum_{\tau=(1,a,b)} \gamma_\tau, \qquad \gamma^- := \sum_{\tau=(-1,a,b)} \gamma_\tau,$$

we have $\gamma^+ = \gamma^- = \frac{1}{2} + o(1)$.

In other words:

Lemma 9.1. *When we choose a random clause for the planted model, and condition on it having an essential variable: the probability that the essential variable is in Λ^+ is equal to the probability that it is in Λ^- plus $o(1)$.*

We close this section by showing that $r_f(\Upsilon) < r_p(\Upsilon)$ for sufficiently large k.

Proposition 9.2. *For any nontrivial, symmetric, feasible, balance-dominated, 1 essential CSP model Υ of arity k:*

(a) For every $k \geq 27$, $r_p(\Upsilon) > r_f(\Upsilon)$.

(b) Asymptotically in k, $\frac{r_f(\Upsilon)}{r_p(\Upsilon)} \lesssim \frac{\ln k}{k}$.

Proof. Notice first that

$$\Omega_p = \mathbf{Exp}\left[\frac{|I_\varphi|}{|S_\varphi|}\right] \leq \frac{\mathbf{Exp}[|I_\varphi|]}{2^k(1 - \frac{1}{\binom{k}{2}+1})} \leq \frac{\mathbf{Exp}[|I_\varphi|]}{(1 - \frac{1}{\binom{k}{2}+1})\mathbf{Exp}[|S_\varphi|]} = \frac{\Omega_f}{(1 - \frac{1}{\binom{k}{2}+1})}.$$

Notice also that $\alpha_k \leq \frac{2\ln(k)}{(1-1/k^2)^{k-1}}$. Therefore, since

$$\frac{2\ln(k)}{k(1 - 1/k^2)^{k-1}} \leq (1/4)(1 - \frac{1}{\binom{k}{2} + 1})$$

for $k \geq 27$, then

$$r_f(\Upsilon) \leq \frac{2\ln(k)}{\Omega_f k(1 - 1/k^2)^{k-1}} \leq \frac{(1/4)}{\Omega_p} \leq r_p(\Upsilon),$$

by Theorem 8.2. Then, part (a) follows. To prove part (b) we use the same inequality, that is

$$\frac{r_f(\Upsilon)}{r_p(\Upsilon)} \leq \frac{8\ln(k)}{k(1-\frac{1}{\binom{k}{2}+1})(1-1/k^2)^{k-1}} \sim 8\ln(k)/k$$

□

10 The *-core

Lemma 10.1. *Consider any 1-essential CSP-model $\Upsilon = (\Phi, p)$ of arity k, and a random CSP, F, drawn from $P(\Upsilon, n, M = rn)$. Suppose $\Gamma(F, \sigma)$ has $\alpha n + o(n)$ hyperedges with $\alpha \neq \alpha_k$. For any $g(n) = o(n)$ and constant $\epsilon > 0$, there exist constants $T, Z, \beta > 0$ such that, with probability at least $1 - e^{-g(n)}$:*

*(a) All but $o(n)$ vertices of the *-core of $\Gamma(F, \sigma)$ are βn-frozen variables for (F, σ).*

*(b) All but at most ϵn vertices outside the *-core of $\Gamma(F, \sigma)$ are either (i) not T-frozen variables for (F, σ) or (ii) within distance Z from a cycle of length at most Z.*

This yields Theorem 1.3:

Proof of Theorem 1.3: Consider (F, σ) drawn from the uniform model $U(\Upsilon, n, M = rn)$. A simple first moment calculation shows that the expected number of variables that are within distance Z of a cycle of length at most Z in the underlying hypergraph of F is $O(1)$. Therefore w.h.p. there are $o(n)$ such vertices.

For part (b): If $r > r_f(\Upsilon)$ then $\alpha > \alpha_k$. Consider any $\epsilon > 0$. Lemma 3.3 allows us to transfer Lemmas 4.3, 10.1, 4.4 to (F, σ) to establish that w.h.p. all but at most ϵn variables are either T-frozen with respect to σ or are within distance Z of a cycle of length at most Z. W.h.p. there are $o(n)$ variables of the latter type, and so all but at most $\epsilon n + o(n)$ vertices are T-frozen. By letting T tend to infinity we can take ϵ arbitrarily small thus obtaining part (b).

For part (a): If $r > r_f(\Upsilon)$ then $\alpha < \alpha_k$. Again, we transfer Lemmas 4.3, 10.1, 4.4 to (F, σ). This shows that w.h.p. all but $o(n)$ of the vertices of the *-core are frozen. The same argument as for part (b) shows that w.h.p. all but $o(n)$ of the vertices outside of the *-core are frozen. Part (a) follows since $\lambda(\Upsilon, r) = \rho_k(\xi(\Upsilon)r) = \rho_k(\alpha)$ and w.h.p. the size of the *-core is $\rho_k(\alpha)n + o(n)$. □

Lemma 10.1(a) is proven in Section 12. Lemma 10.1(b) follows from Lemma 5.2 as follows:

Proof of Lemma 10.1(b): Consider any $\epsilon > 0$. With probability at least $1 - e^{-g(n)}$, $\Gamma(F, \sigma)$ has fewer than ϵn vertices of *-depth greater than L, where L comes from Lemma 5.2. Consider any vertex x of *-depth at most L. Consider a peeling chain for x of depth at most L and let W be the set of all hyperedges that contain at least one vertex of the peeling chain.

If some hyperedges of W form a cycle, then there must be a cycle of length at most $2L$ within distance L of x. If no hyperedges of W form a cycle, then it is easy to see that we can change all of the variables in the peeling chain, one-at-a-time and still have a satisfying assignment for F. Indeed, this follows from a straightforward induction on L. Therefore, the variable x is not 1-frozen. □

10.1 Our hypergraph models

Consider any 1-essential CSP, F, and any solution σ.

The vertices of $\Gamma(F, \sigma)$ are partitioned into two sets Λ^+, Λ^- containing those variables which are assigned $+1, -1$ respectively under σ.

Definition 10.2. *For each hyperedge $e \in \Gamma(F, \sigma)$: Let a be the number of non-essential vertices of e in Λ^+ and let b be the number of non-essential vertices of e in Λ^-. The* type *of e is defined to be:*

- *$(1, a, b)$ if the essential vertex vertex of e is in Λ^+;*
- *$(-1, a, b)$, if the essential vertex vertex of e is in Λ^-.*

The type *of a constraint of (F, σ) with an essential vertex, is the type of the corresponding hyperedge in $\Gamma(F, \sigma)$.*

Now consider a nontrivial, feasible, symmetric, balance-dominated, 1-essential CSP-model Υ and choose a random (F, σ) from the planted model $P(\Upsilon, n, M)$. Recalling the Remark following Definition 3.2, we can selected the constraints of F independently. Given the partition Λ^+, Λ^-, and a type τ, we let $w(\tau) = w(\tau, \Lambda^+, \Lambda^-)$ denote the probability that a selected constraint has type τ, conditional on it having an essential vertex. (See Appendix 9 for further discussion.) Note that $w(\tau)$ depends only on $\Upsilon, |\Lambda^+|, |\Lambda^-|$. Note further that, conditional on a hyperedge e having type τ, every choice of the vertices of e which is consistent with τ is equally likely. Thus, when choosing $\Gamma(F, \sigma)$ we can choose the type of a hyperedge first and then its vertices. This leads us to:
Model A:

1. Partition the vertices into Λ^+, Λ^- uniformly at random.
2. For $i = 1$ to M, choose the ith hyperedge e_i as follows:
   (a) Choose the type (s, a, b) of e_i (where $s \in \{+1, -1\}$), where type τ is chosen with probability $w(\tau)$.
   (b) Choose the essential vertex for e_i uniformly from the appropriate set, Λ^+ or Λ^-, according to s.
   (c) Choose a vertices uniformly from Λ^+ and b vertices uniformly from Λ^-. These are the non-essential vertices of e_i.

In some cases, it will be useful to fix the essential vertex of every hyperedge, along with the assignment σ, and then choose our planted hypergraph. In this case, for $s \in \{-1, +1\}$, we use $w^s(\tau) = w(\tau, \Lambda^+, \Lambda^-)$ denote the probability that a selected constraint has type τ, conditional on it having an essential vertex in Λ^s. We can use the following model:
The Essential Model:

1. We are given a partition the vertices into Λ^+, Λ^-.
2. For $i = 1$ to M, we are given the essential vertex of e_i. We choose the rest of e_i as follows:
   (a) Choose the type (s, a, b) of e_i, where s is already determined and type τ is chosen with probability $w^s(\tau)$.

(b) Choose a vertices uniformly from Λ^+ and b vertices uniformly from Λ^-. These are the non-essential vertices of e_i.

The essential model will be useful in analyzing the *-core of $\Gamma(F, \sigma)$.

We let H_1 denote the set of vertices $v \in H^*$ that are essential in exactly one hyperedge. We use H_1^+, H_1^- to denote $H_1 \cap \Lambda^+, H_1 \cap \Lambda^-$, the vertices of H_1 corresponding to variables assigned $+1, -1$ by σ. The following lemma will be key in proving that most of H^* is frozen:

Lemma 10.3. *If Υ is non-trivial, feasible, symmetric, and balance-dominated and if $\alpha > \alpha_k$ then there exists $\gamma = \gamma(\Upsilon, \alpha) > 0$ such that: for any $g(n) = o(n)$, with probability at least $1 - e^{-g(n)}$,*

(a) $|V(H^*) \cap \Lambda^+|, |V(H^*) \cap \Lambda^-| = |V(H^*)|(\frac{1}{2} + o(1))$;

(b) $|H_1^+|, |H_1^-| \leq \frac{\frac{1}{2}-\gamma}{k-1}|V(H^*)|$.

The proof appears in Appendix 11.

We close this section with:

Proof of Theorem 1.3: Since $r < r_f(\Upsilon)$, w.h.p. the *-core is empty. During the proof of Lemma 11.2 in Appendix 11, we prove that for D sufficiently large, with probability at least $1 - e^{-g(n)}$, fewer than ϵn vertices are within distance L of a vertex with degree greater than D. It follows that for all but at most ϵn vertices of depth at most I, the size of their peeling chain is at most $(kD)^I = O(1)$. We can change any such variable by changing a subset of the entire peeling chain in one step. So, applying Lemma 5.2, we see that for all but $2\epsilon n$ variables v, we can change v by changing at most $(kD)^I$ variables.

We use Lemma 3.3 to show that this holds w.h.p. in the uniform model. Then by taking D arbitrarily large and ϵ arbitarily small,we obtain the theorem. □

11 Analysis of the *-core process

Recall that Υ is a non-trivial, feasible, symmetric, balance-dominated, and 1-essential CSP-model, and that we draw (F, σ) from the planted model.

Let H denote the hypergraph $\Gamma(F, \sigma)$. H has $M = \alpha n$ edges. We will analyze the *-core process on H (recall Section 4). We follow the analysis of [44], being careful to obtain a failure probability of at most $e^{-g(n)}$ for any $g(n) = o(n)$; alternatively, we could have followed the analysis of [32].

Recall that Λ^+, Λ^- denotes the sets of vertices corresponding to variables of sign $+1, -1$ in σ. We can assume that $|\Lambda^+|, |\Lambda^-| = \frac{1}{2}n + o(n)$, as this occurs with probability $1 - e^{-g(n)}$ for any $g(n) = o(n)$.

Let $H(0) = H$ and define $H(i + 1)$ to be the hypergraph obtained by removing every vertex in $H(i)$ that is not essential for any hyperedges, along with all hyperedges in which that vertex is non-essential. We call this operation a *parallel round* of the *-core process. We begin by analyzing $H(i)$ for constant i, using Model A from section 10.1.

We let ρ_i^+, ρ_i^- denote the probability that a vertex $v \in \Lambda^+, \Lambda^-$ survives the i parallel rounds; i.e. $\mathbf{Pr}(v \in H(i))$. Initially $\rho_0^+ = \rho_0^- = 1$; it will follow by induction that $\rho_i^+ = \rho_i^- + o(1)$. So we will recursively define ρ_i and show that $\rho_i^+ = \rho_i^- = \rho_i + o(1)$.

Consider any vertex v. Note that $v \in H(i + 1)$ iff there is at least one hyperedge f in which v is the essential vertex and every non-essential vertex is in $H(i)$. Lemma 9.1 implies the following key property:

Property 11.1. *For every vertex v, the expected number of hyperedges in which v is essential is $\alpha + o(1)$.*

Consider any hyperedge e in which v is the essential vertex. Let the other vertices be $u_1, ..., u_{k-1}$. By induction, $\mathbf{Pr}(u_j \in H(i)) = \rho_i + o(1)$ for each $1 \le j \le k-1$. W.h.p. F is locally tree-like; in particular v does not lie within distance i of a cycle of length at most $2i$. From this, it is straightforward to show that these $k-1$ events are nearly independent, and so $\mathbf{Pr}(u_1, ..., u_{k-1} \in H(i)) = \rho_i^{k-1} + o(1)$. Furthermore, Property 11.1 and the fact that w.h.p. v does not lie near a short cycle imply that the expected number of hyperedges in which v is essential and all non-essential vertices are in $H(i)$ is $\lambda_i + o(n)$ where

$$\lambda_i = \alpha \rho_i^{k-1}.$$

A similar straightforward expected number calculation shows that for any $t > 0$, the expected number of t-tuples of such hyperedges is $\lambda_i^t + o(1)$; again, the key point is that if there are no nearby short cycles, then the hyperedges occur nearly independently. So the Method of Moments (see e.g. [31]) implies that the number of such hyperedges is distributed asymptotically as a Poisson. In particular, the probability that there is at least one is $\rho_{i+1} + o(1)$ where

$$\rho_{i+1} = 1 - e^{-\lambda_i} = 1 - e^{-\alpha\rho_i^{k-1}}.$$

In other words $\rho_{i+1}^+, \rho_{i+1}^- = \rho_{i+1} + o(1)$, thus completing the induction. We define

- X_i^+, X_i^- is the number of vertices of Λ^+, Λ^- in $H(i)$;
- Y_i^+, Y_i^- is the number of hyperedges in $H(i)$ whose essential vertex is in Λ^+, Λ^-;
- A_i^+, A_i^- is the number of vertices of Λ^+, Λ^- in $H(i)$ that are not essential in any hyperedges of $H(i)$;
- B_i^+, B_i^- is the number of vertices of Λ^+, Λ^- that are essential in exactly one hyperedge of $H(i)$.

By the above calculations, $\mathbf{Exp}(X_i^+), \mathbf{Exp}(X_i^-) = \frac{1}{2}\rho_i n + o(n)$. Since every hyperedge has exactly one essential variable, those calculations yield $\mathbf{Exp}(Y_i^+), \mathbf{Exp}(Y_i^-) = \frac{1}{2}\lambda_i n + o(n)$. A_i^+, A_i^- count the vertices that are in $H(i)$ but not in $H(i+1)$; so $\mathbf{Exp}(A_i^+), \mathbf{Exp}(A_i^-) = \frac{1}{2}(\rho_i - \rho_{i+1})n + o(n)$. Since the number of edges in which v is essential is asymptotic to a Poisson with mean λ_i, $\mathbf{Exp}(B_i^+), \mathbf{Exp}(B_i^-) = \frac{1}{2}\lambda_i e^{-\lambda_i} n + o(n)$. We will prove below that these variables are all highly concentrated.

Lemma 11.2. *For any fixed $i \ge 0$, and any constant $\epsilon > 0$, there exists $\eta = \eta(\epsilon, \alpha, i, \Upsilon)$:*

(a) $\mathbf{Pr}(|X_i^+ - \frac{1}{2}\rho_i n| > \epsilon n) < e^{-\eta n}$, $\mathbf{Pr}(|X_i^- - \frac{1}{2}\rho_i n| > \epsilon n) < e^{-\eta n}$

(b) $\mathbf{Pr}(|Y_i^+ - \frac{1}{2}\lambda_i n| > \epsilon n) < e^{-\eta n}$, $\mathbf{Pr}(|Y_i^- - \frac{1}{2}\lambda_i n| > \epsilon n) < e^{-\eta n}$;

(c) $\mathbf{Pr}(|A_i^+ - \frac{1}{2}(\rho_i - \rho_{i+1})n| > \epsilon n) < e^{-\eta n}$, $\mathbf{Pr}(|A_i^- - \frac{1}{2}(\rho_i - \rho_{i+1})n| > \epsilon n) < e^{-\eta n}$;

(d) $\mathbf{Pr}(|B_i^+ - \frac{1}{2}\lambda_i e^{-\lambda_i} n| > \epsilon n) < e^{-\eta n}$,$\mathbf{Pr}(|B_i^- - \frac{1}{2}\lambda_i e^{-\lambda_i} n| > \epsilon n) < e^{-\eta n}$.

We defer the proof to the end of this appendix.

We let $\rho = \lim_{i\to\infty} \rho_i$, which exists since ρ_i is positive and decreasing. So ρ must satisfy $\rho = 1 - e^{-\alpha\rho^{k-1}}$. Setting $\lambda = \lim_{i\to\infty} \lambda_i = \alpha\rho^{k-1}$, we obtain:

$$\rho = 1 - e^{-\lambda}; \qquad \text{so } \lambda = \alpha(1-e^{-\lambda})^{k-1}; \qquad \text{so } \alpha = \frac{\lambda}{(1-e^{-\lambda})^{k-1}}.$$

We will prove:

Lemma 11.3. *For any $g(n) = o(n)$, with probability at least $1 - e^{-g(n)}$:*

*(a) If $\alpha < \alpha_k$ then the *-core of $\Gamma(F,\sigma)$ has $o(n)$ vertices.*

*(b) If $\alpha > \alpha_k$ then the *-core of $\Gamma(F,\sigma)$ has*

(i) $\frac{1}{2}\rho n + o(n)$ vertices in each of Λ^+, Λ^-;

(ii) $\frac{1}{2}\lambda n + o(n)$ hyperedges with essential vertices in each of Λ^+, Λ^-;

(iii) $\frac{1}{2}\lambda e^{-\lambda} n + o(n)$ vertices in each of Λ^+, Λ^- that are essential in exactly one hyperedge.

Recalling that $\alpha_k = \inf_{x>0} \frac{x}{(1-e^{-x})^{k-1}}$, we have that for $\alpha < \alpha_k, \rho = 0$. For $\alpha > \alpha(k)$ we define $x_k(\alpha)$ to be the maximum $x > 0$ such that $\alpha = \frac{x}{(1-e^{-x})^{k-1}}$. It is straightforward to show that $x_k(\alpha) < 1$, $\lambda = x_k(\alpha)$ and $\rho = 1 - e^{-\lambda}$. Thus Lemma 11.3 implies Lemma 4.3 and Lemma 10.3(a). Also, Lemmas 11.2, 11.3 imply Lemma 5.2 as follows:

Proof of Lemma 5.2: For any $\epsilon > 0$ we can choose I such that $\rho_I < \rho + \epsilon$. The number of vertices outside of the *-core with *-depth greater than I is $X_I^+ + X_I^-$ minus the size of the *-core, and hence is less than ϵn. This proves the lemma with $L = I$. □

We will require the following bound:

Lemma 11.4. *For any $\alpha > \alpha_k$ there exists $\gamma > 0$ such that $\lambda e^{-\lambda} < (1-\gamma)\rho/(k-1)$.*

Proof. Let x_1 be the value of $x > 0$ that minimizes $\frac{x}{(1-e^{-x})^{k-1}}$. It is straightforward to check that for $\alpha > \alpha_k$ we have $x_k(\alpha) > x_1$. Differentiating, we see that

$$(1-e^{-x_1})^{k-1} - (k-1)x_1 e^{-x_1}(1-e^{-x_1})^{k-2} = 0; \qquad \text{so } 1 - e^{-x_1} = (k-1)x_1 e^{-x_1}.$$

Clearly $\frac{1-e^{-x}}{xe^{-x}} = \frac{1}{x}(e^x - 1) = (1 + \frac{x}{2} + ...)$ is increasing with x. So for $x > x_1$ we have $\frac{1-e^{-x}}{xe^{-x}} > k-1$ which yields the lemma since $\lambda = x_k(\alpha), \rho = 1 - e^{-x_k(\alpha)}$. □

Note that Lemmas 11.3, 11.4 imply Lemma 10.3(b).

Proof of Lemma 11.3: We will choose a small constant $\zeta > 0$. Lemma 11.2 implies that we can choose I sufficiently large that, with probability at least $1 - e^{-g(n)}$:

$$(\frac{1}{2}\rho - \zeta)n < X_I^+, X_I^- < (\frac{1}{2}\rho + \zeta)n; \qquad Y_I^+, Y_I^- < (\frac{1}{2}\lambda + \zeta)n; \qquad A_I^+, A_I^- < \frac{1}{2}\zeta n; \qquad B_I^+, B_I^- < (\frac{1}{2}\lambda e^{-\lambda} + \zeta)n.$$

Recall that the order in which vertices are removed during the *-core process does not affect the outcome. So we can remove them as follows: First, we carry out I parallel rounds. Then we remove vertices that are not essential in any edges one-at-a-time in arbitrary order; eg. we can pick one of the removable vertices uniformly at random, or we can choose the removable vertex with the lowest label.

After the I parallel rounds, we expose the vertices that remain, W, the number of hyperedges that remain, Y_I and for each remaining hyperedge f we expose its essential vertex, ess(f). The following observation allows us to analyze $H(I)$ using the Essential Model (section 10.1).

Observation: Consider any two hypergraphs Ω, Ω' on the same subset of the vertices of H, with edge set $\{e_1, ..., e_\ell\}$ and $\{e'_1, ..., e'_\ell\}$, such that: for each $1 \leq j \leq \ell$, the hyperedges e_j, e'_j have the same type and the same essential vertex. Then Ω, Ω' are equally likely to be $H(I)$.

To see this, let R be any hypergraph such that applying I iterations of the parallel process to R yields Ω. Form R' from R by replacing every edge of R that is in Ω by the corresponding edge from Ω'. Then applying I iterations of the parallel process to R' will yield Ω'. Furthermore, $\mathbf{Pr}(\Gamma(F,\sigma) = R) = \mathbf{Pr}(\Gamma(F,\sigma)) = R'$.

Note that this observation allows us to model $H(I)$ using the Essential Model. So we expose the vertices of $H(I)$, and for each hyperedge $e \in H(I)$ we expose the essential vertex of e. We let H_1 denote the set of vertices that are essential in exactly one hyperedge; so $|H_1| = B_I^+ + B_I^- < (\lambda e^{-\lambda} + 2\zeta)n$.

From here, the analysis is nearly identical to that from the proof of Lemma 12.11.

Our first step will be to expose the type of every hyperedge; recall that we choose these types independently and the probability that a hyperedge with essential vertex in Λ^s has type τ is $w^s(\tau)$. For each vertex $x \in H_1$, if the type of the hyperedge in which x is essential (s, a, b) then we say $a(x) = a, b(x) = b$. We set $A = \sum_{x \in H_1} a(x)$ and $B = \sum_{x \in H_1} b(x)$. As in the proof of Lemma 12.11, with probability at least $1 - e^{-g(n)}$ we have $A, B = |H_1|(\frac{1}{2}(k-1) + o(1))$.

As we remove vertices one-at-a-time from $H(I)$, we let L denote the set of removable vertices that remain. So initially, $|L| = A_I^+ + A_I^- < \zeta n$. At each step, we remove a vertex w from L. For each hyperedge f containing w, if the essential vertex ess(f) is in H_1 then we add ess(f) to L.

We carry out up to $\frac{4\zeta}{\gamma}n$ steps. If we do not reach the *-core before that time, then we must have added a total of at least $\frac{4\zeta}{\gamma}n - \zeta n$ vertices to L during those steps.

To determine which vertices are added to L we expose the following information: For each remaining hyperedge f, we ask whether w is a non-essential vertex of f. If it is, then we delete f and place ess(f) into L if ess$(f) \in H_1$. If w is not in f, then we do not expose the non-essential vertices of f.

Suppose $w \in \Lambda^+$. As in the proof of Lemma 12.11, it is easy to compute that the vertices of H_1 that will be added to L are determined by at most H_1 independent trials of total probability at most

$$\frac{A}{X_I^+ - \frac{8\zeta}{\gamma}n} < \frac{(\frac{1}{2}\lambda e^{-\lambda} + \zeta)n}{(\frac{1}{2}\rho - \zeta)n - \frac{8\zeta}{\gamma}n} < 1 - \frac{1}{2}\gamma,$$

for ζ sufficiently small, by Lemma 11.4. Similarly, if $w \in \Lambda^-$ then we have at most H_1 independent trials of total probability at most $1 - \frac{1}{2}\gamma$.

Summing over the first $\frac{4\zeta}{\gamma}n$ iterations, the total number of vertices added to L is upperbounded in distribution by the sum of $\frac{4\zeta}{\gamma}n \times H_1$ independent trials, each with probability $\Theta(n^{-1})$ and with total expectation $\frac{4\zeta}{\gamma}n(1 - \frac{1}{2}\gamma) = \frac{4\zeta}{\gamma}n - 2\zeta n$. Standard concentration results for binomial variables yield that the probability that they total more than $\frac{4\zeta}{\gamma}n - \zeta n$ is at most $e^{-\delta n}$ for some $\delta > 0$.

So for every $\zeta > 0$, there exists $\delta > 0$ such that with probability at least $e^{-\delta n}$, we halt within $\frac{4\zeta}{\gamma}n$

steps. If we halt within that many steps then the number of vertices of Λ^+ that are in the *-core is between X_I^+ and $X_I^+ - \frac{4\zeta}{\gamma}n$ and hence is within $\frac{5\zeta}{\gamma}n$ of $\frac{1}{2}\rho n$. Since we can take ζ arbitrarily small, this implies that for any $g(n) = o(n)$, the number of such vertices is $\frac{1}{2}\rho n + o(n)$ with probability at least $1 - e^{-g(n)}$. The same argument applies to the other parameters, thus proving Lemma 11.3. □

It only remains to prove our concentration lemma:

Proof of Lemma 11.2: We will apply Azuma's Inequality [15] which implies (see eg. [14]) that for any random variable $Q = Q(H) = O(n)$, if changing the vertices of one of the αn hyperedges in H can change Q by at most an additive constant, then $\mathbf{Pr}(|Q - \mathbf{Exp}(Q)| > \epsilon n) < e^{-\Theta(n)}$.

We start with the concentration of X_i^+. Note that whether v is counted in X_i^+ is determined entirely by the subgraph induced by $N^i(v)$, the set of vertices within distance i of v. In an extreme case, changing a single hyperedge f can affect X_i^+ by a lot, if f is within distance i of many vertices. So we fix a large constant D and define:

$$\begin{aligned} \Psi_D &= \text{ the set of vertices that are within distance } i \text{ of a vertex } u \text{ of degree } > D, \\ X_i^+(D) &= \text{ the number of vertices of } \Lambda^+ \setminus \Psi_D \text{ that are in } H(i). \end{aligned}$$

Changing a single hyperedge can affect $X_i^+(D)$ by at most $2k((k-1)D)^i = O(1)$. Indeed, if changing the vertices of f affects whether $v \in X_i^+$ then v is connected to one of the old or new vertices of f by a path of length at most i. If any vertex on a hyperedge of that path has degree greater than D then $v \in \Psi_D$ and so v will not count towards $X_i^+(D)$. So each of the $2k$ old or new vertices of f can affect at most $((k-1)D^i)$ vertices v. Therefore, there exists $\eta_1 = \eta_1(D, \epsilon, k, i, \Upsilon)$ such that

$$\mathbf{Pr}(|X_i^+(D) - \mathbf{Exp}(X_i^+(D))| > \frac{1}{3}\epsilon n) < e^{-\eta_1 n}.$$

A standard property of random graphs (and indeed an easy calculation) yields that by taking D sufficently large, we can make $\mathbf{Exp}(|\Psi_D|)$ an arbitrarily small multiple of n. (Roughly: the expected number of vertices u of degree greater than D drops exponentially in D while the expected number of vertices within distance i of each such u is linear in D, for fixed i.) So we choose D such that $\mathbf{Exp}(|\Psi_D|) < \frac{1}{3}\epsilon n$.

Next we show that $|\Psi_D|$ is concentrated. A similar argument to that above shows that changing the vertices of a single hyperedge f can affect $|\Psi_D|$ by at most $2k((k-1)D)^i = O(1)$. Indeed, if changing f affects whether $v \in \Psi_D$ then v is connected to one of the old or new vertices of f by a path of length at most i. If any vertex on the hyperedges of that path has degree greater than D then $v \in \Psi_D$ regardless of the choice of f. So each of the $2k$ old or new vertices of f can affect at most $((k-1)D^i)$ vertices v. Therefore, there exists $\eta_2 = \eta_2(D, \epsilon, k, i, \Upsilon)$ such that

$$\mathbf{Pr}(|\Psi_D - \mathbf{Exp}(|\Psi_D|)| > \frac{1}{6}\epsilon n) < e^{-\eta_2 n}.$$

Noting that $X_i^+(D) < X_i^+ < X_i^+(D) + |\Psi_D|$ and applying linearity of expectation, we have

$$\mathbf{Pr}(|X_i^+ - \mathbf{Exp}(X_i^+)| > \epsilon n) < e^{-\eta_1 n} + e^{-\eta_2 n} < e^{-\eta n},$$

for any $\eta < \eta_1, \eta_2$. The proof for the remaining parameters is nearly identical. □

We close this section by noting that by the same reasoning as for the Observation in the proof of Lemma 11.3, we can model the *-core of $\Gamma(F, \sigma)$ using the Essential Model. We do so in section 12.

12 Frozen variables in the *-core

Here, we prove Lemma 10.1(a). Recall that (F, σ) is drawn from $P(\Upsilon, n, M = rn)$ where Υ is symmetric and 1-essential.

H^* is the *-core of $\Gamma(F, \sigma)$, so every edge of H^* has exactly one essential vertex and every vertex is essential for at least one edge. H_1 is the set of vertices that are essential in exactly one hyperedge of H^*. We need to show that, with sufficiently high probability, all but $o(n)$ vertices in H^* are βn-frozen variables of (F, σ).

Definition 12.1. *For each vertex $x \in H_1$, we use $e(x)$ to denote the unique hyperedge of H^* in which x is the essential vertex.*

Definition 12.2. *A* flippable set *of H^* is a set of vertices $S \subset H^*$ such that for every $x \in S$ and for every hyperedge $f \in H^*$ in which x is essential, S contains another vertex of f.*

Note that, since every hyperedge of H^* has exactly one essential variable, that other vertex is not essential for f.

Given two boolean assignments σ, σ' to the variables of F, we let $\sigma \Delta \sigma'$ denote the set of variables x for which $\sigma(x) \neq \sigma'(x)$.

Proposition 12.3. *(a) If σ' is any solution of F, then $(\sigma \Delta \sigma') \cap H^*$ is a flippable set.*

(b) The union of any collection of flippable sets is a flippable set.

Proof. For (a): if $(\sigma \Delta \sigma') \cap H^*$ is not a flippable set, then there is some $x \in (\sigma \Delta \sigma') \cap H^*$ and a hyperedge $f \in H^*$ such that x is essential for f and $\sigma \Delta \sigma'$ contains no other vertices of f. Since $H^* \subset \Gamma(F, \sigma)$, this means that x is essential for the constraint corresponding to f in (F, σ), and that σ' changes the value of x but not of any other variables in f. Therefore σ' violates f and so σ' is not a solution for F.

For (b): this is immediate from the definition of a flippable set. □

To prove Lemma 10.1(a), we will show that there exists $\phi'(n) = o(n)$, $\zeta > 0$ such that, with sufficiently high probability, there are no flippable sets S in H^* of size $\phi'(n) \leq |S| \leq \zeta n$. We will apply a first moment bound. A direct approach does not work, because of a "jackpot phenomena": The existence of a flippable set S typically implies the existence of an exponential number of other flippable sets formed by adding to S variables $x \in H_1$ with the property that $e(x)$ contains a member of S. To overcome this issue, we focus instead on sets with the following property.

Definition 12.4. *We say that a set $A \subseteq H^* \setminus H_1$ is* weakly flippable *if there exists $P \subseteq H_1$ such that $A \cup P$ is flippable. A is ψ-*weakly flippable *if there exists such a P with $|P| \leq \psi$.*

Given a flippable set $S \subseteq H^*$, we consider a directed graph $D(S) \subseteq D$. The vertices of $D(S)$ are the vertices of S; the edges of $D(S)$ are defined as follows:

- For each $x \in S \cap H_1$, we choose one other variable $x' \in e(x)$ that is in S, and we add the edge $x \dashrightarrow x'$ to $D(S)$.

Note that, since S is flippable, there is at least one such x'. It is not important which one we choose, but to be specific we could, eg., choose the lowest indexed variable from amongst all variables of S (other than x) in $e(x)$.

Thus, every vertex in $D(S)$ has outdegree either 0 or 1. We define:

- $A_S = S \setminus H_1$. Note that A_S is the set of vertices with outdegree 0 in $D(S)$.
- C_S is the set of all vertices on directed cycles of $D(S)$. Note that those directed cycles are disjoint since the maximum outdegree is 1.

Since the outdegree of every vertex outside of A_S is one, there is a directed path from every $x \in S \setminus (A_S \cup C_S)$ to $A_S \cup C_S$.

Definition 12.5. *A set of vertices $x_1, \dots, x_l \in H_1$ is* cyclic *if for some permutation $\pi \in \mathcal{S}_l$, $x_{\pi(j)}$ is in $e(x_j)$ for every $1 \leq i \leq a$.*

Definition 12.6. *Given a set $A \subseteq H^*$, the* closure *of A, $cl(A)$ is the set of all vertices x such that, either*

(a) $x \in A$, or

(b) $x \in H_1 \setminus A$ and there is a sequence $x = x_0, x_1, ..., x_\ell$ where (i) $x_\ell \in A$ and (ii) for all $i < \ell$: $x_i \in H_1 \setminus A$ and $x_{i+1} \in e(x_i)$.

Proposition 12.7. *If S is a flippable set, then:*

(a) A_S is weakly flippable.

(b) C_S is cyclic.

(c) $S \subseteq cl(A_S \cup C_S)$.

Proof. (a) follows from the definition of weakly flippable, with $P = S \cap H_1$.

(b) follows from the definition of cyclic, where the directed cycles of $D(S)$ form π.

For (c), if $x \in S \setminus (A_S \cup C_S)$, then $x \in H_1$ and the directed path from x to $A_S \cup C_S$ in $D(S)$ indicates that x satisfies condition (b) of Definition 12.6. □

Lemma 12.8. *Suppose that for some ϕ, ϕ', ψ we have:*

(a) There is no ψ-weakly flippable set $A \subseteq H^* \setminus H_1$ such that $\phi < |A| < \psi$.

(b) There is no cyclic set C such that $\phi < |C| < \psi$.

(c) There is no set A such that $|A| \leq 2\phi$ and $|cl(A)| > \phi'$.

Then there is no flippable set $S \subseteq H^*$ such that $\phi' < |S| < \psi$.

Proof. We apply Proposition 12.7. Let S be a flippable set with $|S| < \psi$. Then A_S is ψ-weakly flippable. Thus, by (a), $|A_S| \leq \phi$. Since C_S is cyclic and $|C_S| \leq |S| < \psi$, (b) implies $|C_S| \leq \phi$. Therefore, $|A_S \cup C_S| \leq 2\phi$, which by (c) implies that $|S| \leq |\mathrm{cl}\,(A_S \cup C_S)| \leq \phi'$. The lemma follows. □

The following lemmas establish that the conditions of Lemma 12.8 hold with sufficiently high probability.

Lemma 12.9. *There exists $\zeta = \zeta(\Upsilon, \alpha) > 0$, and for any $g(n) = o(n)$, there exists $\phi(n)$ satisfying $g(n) << \phi(n) = o(n)$ such that:*
The probability that there is a (ζn)-weakly flippable set A of H^* with $\phi(n) < |A| < \zeta n$ is at most $e^{-g(n)}$.

Lemma 12.10. *There exists $\zeta = \zeta(\Upsilon, \alpha) > 0$, and for any $g(n) = o(n)$, there exists $\phi(n)$ satisfying $g(n) << \phi(n) = o(n)$ such that:*
The probability that there is a cyclic set C in H^* with $\phi(n) < |C| < \zeta n$ is at most $e^{-g(n)}$.

Lemma 12.11. *There exists $\zeta = \zeta(\Upsilon, \alpha) > 0$, and for any $\phi(n) = o(n)$, there exists $\phi'(n) = o(n)$ such that: The probability that there is a set $A \subset H^*$ with $|A| < 2\phi(n)$ and $|cl(A)| > \phi'(n)$ is at most $e^{-\phi'(n)}$.*

These lemmas yield Lemma 10.1(a) as follows:

Proof of Lemma 10.1(a): Note that we can take $\phi'(n) > g(n)$. Lemmas 12.8, 12.9, 12.10, 12.11 imply that for all $g(n) = o(n)$, there exists $\phi'(n) = o(n)$ such that with probability at least $1 - 3e^{-g(n)}$ the *-core H^* of $\Gamma(F, \sigma)$ has no flippable set of size between $\phi'(n)$ and ζn. So suppose that there is no such flippable set in H^*.

Let $S_1, ..., S_t$ be all flippable sets in H^* of size less than ζn. Thus each $|S_i| < \phi'(n)$. Assume by induction that $| \cup_{i=1}^{j} S_i| < \phi'(n)$. Then $| \cup_{i=1}^{j+1} S_i| < 2\phi'(n) < \zeta n$. By Proposition 12.3(b), $\cup_{i=1}^{j+1} S_i$ is a flippable set and hence it must have size less than $\phi'(n)$. Therefore $| \cup_{i=1}^{t} S_i| < \phi'(n)$.

Now consider any sequence of solutions $\sigma = \sigma_0, \sigma_1, ..., \sigma_\ell$ in which the assignment changes for at least one variable in $H^* \setminus (\cup_{i=1}^{t} S_i)$. Let i be the lowest index so that $\sigma_i(x) \neq \sigma(x)$ for some $x \in H^* \setminus (\cup_{i=1}^{t} S_i)$. Therefore $x \in (\sigma_i \Delta \sigma) \cap H^*$ which, by Proposition 12.3(a), is a flippable set. Since $x \notin \cup_{i=1}^{t} S_i$, this implies $|(\sigma_i \Delta \sigma) \cap H^*| \geq \zeta n$. By our choice of i, $|(\sigma_i \Delta \sigma_{i-1}) \cap H^*| \geq |(\sigma_i \Delta \sigma) \cap H^*| - |\cup_{i=1}^{t} S_i| \geq \zeta n - \phi'(n)$. Therefore every variable in $H^* \setminus (\cup_{i=1}^{t} S_i)$ is $(\zeta n - \phi'(n))$-frozen. This yields Lemma 10.1(a) for any $\beta < \zeta$ after rescaling $g(n)$. □

We prove Lemmas 12.9, 12.10, 12.11 in the next three subsections. In each case, we will study H^* using the Essential Model. See the discussion at the end of Appendix 11 explaining why it is valid to do so.

12.1 Weakly-flippable sets: Proof of Lemma 12.9

Suppose that $A \subseteq H^* \setminus H_1$ is a (ζn)-weakly flippable set with $\phi(n) < |A| < \zeta n$.

Set $a := |A|$ and note that there are at least $2a$ hyperedges of H^* whose essential variables are in A, since A contains no variables of H_1. Let $e_1, ..., e_{2a}$ denote exactly $2a$ such hyperedges; to be specific, the $2a$ with the lowest indices. Since A is (ζn)-weakly flippable, for each $1 \leq j \leq 2a$ there exists a sequence of vertices $x_{j,0}, x_{j,1}, ..., x_{j,t_j}$ such that:

(i) $x_{j,0} \in A$ is the essential vertex of e_j;

(ii) $x_{j,t_j} \in A$;

(iii) $x_{j,1} \in e_j$, and if $t_j > 1$ then for each $1 \leq i \leq t_j - 1$: $x_{j,i} \in H_1$ and $x_{j,i+1} \in e(x_{j,i})$.

Note that possibly $t_j = 1$ in which case e_j contains a non-essential member of A.

These sequences are not necessarily disjoint. However, since $e_j \neq e_{j'}$ for all $j \neq j'$, we can take initial portions of them so that the portions in H_1 are disjoint. I.e., there exist $l_1, \ldots, l_{2a} \geq 0$ with $\sum_{j=1}^{2a} l_j \leq \zeta n$ such that

(i) the vertices $x_{j,i} : 1 \leq j \leq 2a, 1 \leq i \leq l_j$ are distinct;

(ii) for $j = 1, \ldots, 2a$: $x_{j,l_j+1} \in A \cup \{x_{j',i} : 1 \leq j' < j, 1 \leq i \leq l_{j'}\}$.

We will bound the expected number of such collections of sequences, when H^* is chosen from the Essential Model. So we expose the vertices of H^*, and for each hyperedge $e_1, ..., e_{\alpha n}$ we expose the essential vertex of e_i. By Lemma 10.3(a) we can assume that $|H^* \cap \Lambda^+|, |H^* \cap \Lambda^+| = \frac{1}{2}|H^*| + o(n)$.

Fix some $0 \leq \ell \leq \zeta n$. First we will choose $l_1, ..., l_{2a} \geq 0$ summing to ℓ. The number of choices is $\binom{\ell+2a-1}{2a-1}$.

Next, we choose A; note that this determines $e_1, ..., e_{2a}$ and their essential vertices $x_{1,0}, ..., x_{2a,0}$. The number of choices is $\binom{|H^*|}{a} \leq \binom{n}{a}$.

Next we choose the remaining vertices. To do so, we first determine their signs; i.e. which are in Λ^+ and which are in Λ^-. So for each j, we choose a pattern θ_j - a sequence of $l_j + 2$ terms from $\{+, -\}$ indicating the signs of $x_{j,0}, ..., x_{j,l_j+1}$. Note that, since $x_{j,0}$ is already determined, the first sign of θ_j is already known. Recall from Lemma 10.3 that we can assume $|H_1^+|, |H_1^-| < |H^*| \times \frac{\frac{1}{2}-\gamma}{k-1}$. Thus, given θ_j, the number of choices for $x_{j,1}, ..., x_{j,l_j}$ is at most $(|H^*| \times \frac{\frac{1}{2}-\gamma}{k-1})^{l_j}$.

Finally, we choose $x_{j,l_j+1} : 1 \leq j \leq 2a$. These are not neccesarily distinct, and they are all members of $A \cup \{x_{j,i} : 1 \leq j < 2a, 1 \leq i \leq l_j\}$. So the number of choices is at most $(a + \ell)^{2a}$.

Having selected these vertices, we bound the probability that all edges are as required.

Consider selecting the type of a hyperedge whose essential vertex is in Λ^+; thus we are select type $\tau = (1, a, b)$ with probability $w^{+1}(\tau)$. We let $g = g(\Upsilon)$ denote the expected value of a, and so the expected value of b is $k - 1 - g$. Because Υ is symmetric and $|H^* \cap \Lambda^+| = |H^* \cap \Lambda^-|(1 + o(1))$, it follows that $w^{+1}(1, a, b) = w^{-1}(-1, b, a) + o(1)$. So when we select the type of a hyperedge whose essential vertex is in Λ^-, the expected value of b is $g + o(1)$.

Now we select the types and then the non-essential vertices for the hyperedges $e_1, ..., e_{2a}$ and $e(x_{j,i}) : 1 \leq j \leq 2a, 1 \leq i \leq l_j$. Recall that we choose those vertices uniformly from $\Lambda^+ \cap H^*$ or $\Lambda^- \cap H^*$ depending on what the type of the hyperedge tells us the sign of the vertex should be. For each j, we require that $x_{j,1}$ is a non-essential vertex of e_j and that $x_{j,i+1}$ is a non-essential vertex of $e(x_{j,i})$. By Lemma 10.3(a), $|\Lambda^+ \cap H^*|, |\Lambda^- \cap H^*| = |H^*|(\frac{1}{2} + o(1))$, and so for each hyperedge this event occurs with probability $\frac{2g+o(1)}{|H^*|}$ if the essential and non-essential vertices have the same sign, and $\frac{2(k-1-g)+o(1)}{|H^*|}$ otherwise. Note also that these events are independent.

We let $y(\theta_j)$ denote the number of terms in θ_j that are the same as the preceding term. So the probability that the required vertices are selected as non-essential vertices in each hyperedge is:

$$\left(\frac{2g + o(1)}{|H^*|}\right)^{\sum_{j=1}^{2a} y(\theta_j)} \left(\frac{2(k-1-g) + o(1)}{|H^*|}\right)^{\sum_{j=1}^{2a} l_j+1-y(\theta_j)}.$$

Putting this all together yields that the expected number of (ζn)-weakly flippable sets A, given

a, ℓ, is at most:

$$\begin{aligned}
&\binom{\ell+2a-1}{2a-1}\binom{n}{a}\left(|H^*| \times \frac{\frac{1}{2}-\gamma}{k-1}\right)^{\sum_j l_j} (a+\ell)^{2a} \\
&\quad \times \sum_{\theta_1,\ldots,\theta_{2a}} \left(\frac{2g+o(1)}{|H^*|}\right)^{\sum_{j=1}^{2a} y(\theta_j)} \left(\frac{2(k-1-g)+o(1)}{|H^*|}\right)^{\sum_{j=1}^{2a} l_j+1-y(\theta_j)} \\
< \; &\binom{\ell+2a}{2a}\binom{n}{a}\left(|H^*| \times \frac{\frac{1}{2}-\gamma}{k-1}\right)^{\ell} (a+\ell)^{2a} \left(\frac{2+o(1)}{|H^*|}\right)^{2a+\ell} \sum_{\theta_1,\ldots,\theta_{2a}} g^{\sum_{j=1}^{2a} y(\theta_j)} (k-1-g)^{\sum_{j=1}^{2a} l_j+1-y(\theta_j)} \\
< \; &\left(\frac{e(\ell+2a)}{2a}\right)^{2a} \left(\frac{en}{a}\right)^{a} \left(\frac{3(a+\ell)}{|H^*|}\right)^{2a} \left(\frac{1-\gamma}{k-1}\right)^{\ell} \prod_{j=1}^{2} a \sum_{\theta_j} g^{y(\theta_j)} (k-1-g)^{l_j+1-y(\theta_j)}. \qquad (7)
\end{aligned}$$

For each value of y, there are $\binom{l_j+1}{y}$ patterns θ_j with $y(\theta_j) = y$, since the first sign in θ_j is already chosen. This implies

$$\prod_{j=1}^{2} a \sum_{\theta_j} g^{y(\theta_j)} (k-1-g)^{l_j+1-y(\theta_j)} = \prod_{j=1}^{2a} \sum_{y=0}^{l_j+1} \binom{l_j+1}{y} g^{y} (k-1-g)^{l_j+1-y} = (k-1)^{\ell}.$$

So (7) is at most

$$\left(\frac{e(\ell+2a)}{2a}\right)^{2a} \left(\frac{en}{a}\right)^{a} \left(\frac{3(a+\ell)}{|H^*|}\right)^{2a} \left(\frac{1-\gamma}{k-1}\right)^{\ell} (k-1)^{\ell} < \left(1+\frac{\ell}{2a}\right)^{2a} \left(1+\frac{\ell}{a}\right)^{2a} \left(\frac{Ca}{n}\right)^{a} (1-\gamma)^{\ell},$$

for some constant $C > 9e^3(\frac{n}{|H^*|})^2$ (see Lemma 4.3). So the total expected number of A with $\phi(n) < |A| < \zeta n$ is at most:

$$\sum_{a=\phi(n)}^{\zeta n} \left(\frac{Ca}{n}\right)^{a} \sum_{\ell \geq 0} \left(1+\frac{\ell}{a}\right)^{4a} (1-\gamma)^{\ell}.$$

To bound this, it is easy to see that $\left(1+\frac{\ell}{a}\right)^{4a}(1-\frac{\gamma}{2})^{\ell}$ is maximized at $\ell = O(a)$ and hence is at most $Y^{4a}(1-\frac{\gamma}{2})^{4a} < Y^{4a}$ for some constant $Y = Y(\gamma)$. Since $1-\gamma < (1-\frac{\gamma}{2})^2$, this yields an upper bound of:

$$\sum_{a=\phi(n)}^{\zeta n} \left(\frac{Ca}{n}\right)^{a} \sum_{\ell \geq 0} Y^{4a}(1-\frac{\gamma}{2})^{\ell} = O(1)\left(\frac{CY^4 a}{n}\right)^{a}.$$

By taking $\zeta < \frac{1}{2CY^4}$, this is less than $\sum_{a=\phi(n)}^{\zeta n} 2^{-a}$ which is less than $e^{-g(n)}$ for any $\phi(n) >> g(n)$. □

12.2 Cyclic sets: Proof of lemma 12.10

Note that the vertices of a cyclic set are partitioned into cycles by the permutation π. We will fix a constant Z, to be named later. A *small-cyclic* set is a cyclic set in which each cycle has length at most Z. A *large-cyclic* set is a cyclic set in which each cycle has length greater than Z.

Lemma 12.12. *For any $g(n) = o(n)$ there exists $\phi(n)$ such that with probability at least $1 - e^{-g(n)}$:*

(a) H^* has no small-cyclic sets of size at least $\frac{1}{2}\phi(n)$.

(b) H^* has no large-cyclic sets of size at least $\frac{1}{2}\phi(n)$.

This clearly proves Lemma 12.10 as any cyclic set of size at least $\phi(n)$ contains either a small-cyclic set or a large-cyclic set of size at least $\frac{1}{2}\phi(n)$. Again, we work in the Essential Model.

Proof of (a): We say that a *cycle* in $\Gamma(F, \sigma)$ is a set of vertices $x_1, ..., x_\ell$ such that x_i, x_{i+1} lie in a common hyperedge of $\Gamma(F, \sigma)$ for each i (addition is mod ℓ). For any x_i, x_j, the probability that x_i, x_j share a hyperedge in $\Gamma(F, \sigma)$ is less than c/n, for some constant $c = c(\Upsilon, \alpha)$.

If H^* has a small-cyclic set of size at least $\frac{1}{2}\phi(n)$, then the hypergraph $\Gamma(F, c)$ must contain at least $\phi(n)/(2Z)$ cycles of size at most Z, and so it must contain at least $\phi(n)/(2Z^2)$ cycles of size exactly z for some $z \leq Z$. Setting $W := \phi(n)/(2Z^2)$, the probability of this occurring for z is less than:

$$\frac{n^{zW}}{W!}\left(\frac{c}{n}\right)^{zW} = \frac{(c^z)^W}{W!} < \frac{1}{2Z}e^{-g(n)},$$

if $\phi(n) >> g(n)$. (Note that the dependency between the events that the zW pairs of vertices each share a hyperedge goes in the right direction for this bound to hold.) Summing over all $z \leq Z$ proves (a). □

Proof of (b): We will bound the expected number of large cyclic sets of size a.

A pattern θ is a sequence of a terms from $\{+, -\}$ indicating the signs of $x_1, ..., x_a$. By Lemma 10.3, we can assume that $|H_1^+|, |H_1^-| < \frac{\frac{1}{2}-\gamma}{k-1}|H^*|$. So for any pattern θ, the number of choices for $x_1, ..., x_a$ is at most $\left(\frac{\frac{1}{2}-\gamma}{k-1}|H^*|\right)^a$.

Given a pattern θ and a permutation π, we let $y(\theta, \pi)$ denote the number of i such that $x_i, x_{\pi(i)}$ have the same sign. Recall g from the proof of Lemma 12.9; the same reasoning as in that proof says that, for any choice of $x_1, ..., x_a$ in agreement with θ, the probability that $x_{\pi(i)}$ is a non-essential vertex in $e(x_i)$ for every i is $\left(\frac{2g+o(1)}{|H^*|}\right)^{y(\theta,\pi)}\left(\frac{2(k-1-g)+o(1)}{|H^*|}\right)^{a-y(\theta,\pi)}$.

We let $c(\pi)$ denote the number of cycles in π; since we are considering large-cyclic sets, we only need to consider permutations π with $c(\pi) < a/Z$. For any π, y, the number of choices of θ with $y(\theta, \pi) = y$ is at most $2^{c(\pi)}\binom{a}{y} < 2^{a/Z}\binom{a}{y}$. Indeed, there are $\binom{a}{y}$ choices of the values of i for which $x_i, x_{\pi(i)}$ have the same sign; given one such choice, the pattern is determined by fixing the sign of one vertex in each of the $c(\pi)$ cycles. Note that this is an upper bound; as for some π, y, parity conditions will imply that there is no such θ.

To bound the expected number of large-cyclic sets of size a, we sum over all ordered choices of

$x_1, ..., x_a$ and all choices of π, and then divide by $a!$, obtaining:

$$\begin{aligned}
& \frac{1}{a!}\sum_\theta\left(\frac{\frac{1}{2}-\gamma}{k-1}|H^*|\right)^a \sum_\pi\left(\frac{2g+o(1)}{|H^*|}\right)^{y(\theta,\pi)}\left(\frac{2(k-1-g)+o(1)}{|H^*|}\right)^{a-y(\theta,\pi)} \\
< \; & \left(\frac{1-\gamma}{k-1}\right)^a \frac{1}{a!}\sum_\pi\sum_{y=0}^{a} 2^{a/Z}\binom{a}{y} g^y(k-1-g)^{a-y} \\
\leq \; & \left(\frac{1-\gamma}{k-1}2^{1/Z}\right)^a (k-1)^a \\
< \; & (1-\frac{1}{2}\gamma)^a
\end{aligned}$$

if Z is chosen large enough that $(1-\gamma)2^{1/Z} < (1-\frac{1}{2}\gamma)$.

So the probability that there is a large-cyclic set of size at least $\frac{1}{2}\phi(n)$ is at most $O(1)(1-\frac{1}{2}\gamma)^{\frac{1}{2}\phi(n)} < \frac{1}{2}e^{-g(n)}$ for any $\phi(n) >> g(n)$. □

12.3 Closure: Proof of lemma 12.11

We will choose $\phi'(n) >> \phi(n)$. Again, we work in the Essential Model.

Consider a set A of size at most $2\phi(n) = o(\phi'(n))$. We can find $\mathrm{cl}\,(A)$ using the following search:

```
1. Initialize $C = \emptyset, L = A$.
2. While $L \neq \emptyset$
    (a) Choose $u \in L$.
    (b) For every $w \in H_1 \setminus (C \cup L)$ such that $u \in e(w)$, add $w$ to $L$.
    (c) Remove $u$ from $L$ and add $u$ to $C$.
```

When this procedure halts, $C = \mathrm{cl}\,(A)$. Note that $|\mathrm{cl}\,(A)\,|$ is the number of times that we execute the loop in Step 2. If $|\mathrm{cl}\,(A)\,| > \phi'(n)$ then during the first $\phi'(n)$ iterations we never reach $L = \emptyset$ and so we must add a total of more than $\phi'(n) - |A| = \phi'(n)(1-o(1)) > \phi'(n)(1-\frac{1}{2}\gamma)$ vertices to L in Step 2(b), where γ comes from Lemma 10.3. We will bound the probability of that occuring.

We analyze H^* using the Essential Model. So we expose the vertices of H^*, and for each hyperedge $e_1, ..., e_{\alpha n}$ we expose the essential vertex of e_i. By Lemma 10.3(a) we can assume that $|H^* \cap \Lambda^+|, |H^* \cap \Lambda^+| = \frac{1}{2}|H^*| + o(n)$.

Our first step will be to expose the type of every hyperedge; recall that we choose these types independently and the probability that a hyperedge with essential vertex in Λ^s has type τ is $w^s(\tau)$. For each vertex $x \in H_1$, if the type of x is chosen to be (s, a, b) then we say $a(x) = a, b(x) = b$. We set $A = \sum_{x \in H_1} a(x)$ and $B = \sum_{x \in H_1} b(x)$.

Because Υ is symmetric and $|H^* \cap \Lambda^+| = |H^* \cap \Lambda^-|(1 + o(1))$, it follows that $w^{+1}(1, a, b) = w^{-1}(-1, b, a) + o(1)$. This implies that for $x \in \Lambda^+, y \in \Lambda^-$, $\mathbf{Exp}(a(x)) = \mathbf{Exp}(b(y)) + o(1)$ and $\mathbf{Exp}(b(x)) = \mathbf{Exp}(a(y)) + o(1)$, and it follows that $\mathbf{Exp}(A), \mathbf{Exp}(B) = |H_1|(\frac{1}{2}(k-1) + o(1))$. The number of hyperedges of each type is a binomial variable and so is easily seen to be highly enough concentrated that with probability at least $1 - e^{-g(n)}$ we have $A, B = |H_1|(\frac{1}{2}(k-1) + o(1))$.

Now we analyze our search. We can choose u arbitrarily in Step 2(a); to be specific, we choose the $u \in L$ with the lowest index. To carry out Step 2(b): for every $x \in H_1 \setminus (C \cup L)$, we expose whether $u \in e(x)$; if $u \notin e(x)$ then we do not expose the non-essential vertices of $e(x)$.

Suppose $u \in \Lambda^+$. To test whether $u \in e(x)$, we ask whether u is one of the $a(x)$ non-essential variables from Λ^+. Initially, the probability is $\frac{a(x)}{|H^* \cap \Lambda^+|}$; as the procedure progresses, this increases as we have exposed that the members of C are not in $e(x)$. But since $|C| \leq \phi'(n)$ it never exceeds $\frac{a(x)}{|H^* \cap \Lambda^+| - \phi(n)}$. We ask this for every $x \in H_1 \setminus (C \cup L)$ resulting in at most H_1 independent trials of total probability at most

$$\frac{A}{|H^* \cap \Lambda^+| - \phi(n)} = \frac{|H_1|(\frac{1}{2}(k-1) + o(1))}{\frac{1}{2}|H^*|(1 + o(1))} < 1 - \gamma,$$

by Lemma 10.3(b). Similarly, if $u \in \Lambda^-$ then we have at most H_1 independent trials of total probability at most $1 - \gamma$.

Summing over the first $\phi'(n)$ iterations, the total number of vertices added to L is upperbounded in distribution by the sum of $\phi'(n)H_1$ independent trials, each with probability $\Theta(n^{-1})$ and with total expectation $\phi'(n)(1-\gamma)$. Standard concentration results for binomial variables yield that the probability that they total more than $\phi'(n)(1 - \frac{1}{2}\gamma)$ is at most $e^{-c\phi'(n)}$ for some $c = c(g, k, \gamma)$.

So the expected number of sets A of size at most $\phi(n)$ for which $|\mathrm{cl}\,(A)\,| \geq \phi'(n)$ is at most

$$\sum_{a=1}^{\phi(n)} \binom{|H^*|}{a} e^{-c\phi'(n)} < \phi(n) \binom{n}{\phi(n)} e^{-c\phi'(n)} < \phi(n) \left(\frac{ene^{-c\phi'(n)/\phi(n)}}{\phi(n)} \right)^{\phi(n)}.$$

By choosing $\phi(n) \log(n/\phi(n)) << \phi'(n) = o(n)$, this probability is less than $e^{-\phi(n)}$, as required. □