The Solution Space Geometry of Random Linear Equations

Dimitris Achlioptas,[1,*] **Michael Molloy**[2,†]
[1] *Department of Informatics & Telecommunications, University of Athens, Athens, Greece; e-mail: optas@di.uoa.gr*
[2] *Department of Computer Science, University of Toronto, Ontario, Canada; e-mail: molloy@cs.toronto.edu*

Received 3 March 2012; accepted 5 October 2012
Published online 12 April 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/rsa.20494

ABSTRACT: We consider random systems of linear equations over GF(2) in which every equation binds k variables. We obtain a precise description of the clustering of solutions in such systems. In particular, we prove that with probability that tends to 1 as the number of variables, n, grows: for every pair of solutions σ, τ, either there exists a sequence of solutions starting at σ and ending at τ such that successive solutions have Hamming distance $O(\log n)$, or every sequence of solutions starting at σ and ending at τ contains a pair of successive solutions with distance $\Omega(n)$. Furthermore, we determine precisely which pairs of solutions are in each category. Key to our results is establishing the following high probability property of cores of random hypergraphs which is of independent interest. Every vertex not in the r-core of a random k-uniform hypergraph can be removed by a sequence of $O(\log n)$ steps, where each step amounts to removing one vertex of degree strictly less than r at the time of removal. © 2013 Wiley Periodicals, Inc. Random Struct. Alg., 46, 197–231, 2015

Keywords: random k-XORSAT; solution clustering; hypergraphs; cores

1. INTRODUCTION

In Random Constraint Satisfaction Problems (CSPs) one has a set of n variables all with the same domain D and a set of independently chosen constraints, each of which binds a randomly selected subset of k variables. In the most common setting, both D and k are $O(1)$, while $m = \Theta(n)$. Two canonical examples are random k-SAT and coloring sparse random graphs. A fundamental quantity in the study of random CSPs is the so-called constraint density, i.e., the ratio of constraints-to-variables m/n.

Correspondence to: M. Molloy
*Supported by ERC IDEAS Starting Grant, NSF CAREER Award, and Sloan Fellowship.
†Supported by NSERC Discovery Grant.

There has been much non-rigorous evidence from statistical physics that for many random CSPs, if the constraint density is higher than a specific value, then all but a vanishing proportion of the solutions can be partitioned into exponentially many sets (clusters) such that each set is: (i) well-separated, i.e., has linear Hamming distance from all others, and (ii) in some sense, well-connected. The solution clustering phenomenon has been a central feature of the statistical physics approach to random CSPs and is central to important algorithmic developments in the area, such as Survey Propagation [23].

The mathematical studying of clustering began in [13, 24] where it was shown that in random k-CNF formulas, above a certain density there exist constants $0 < \alpha_k < \beta_k < 1/2$ such that w.h.p. no pair of satisfying assignments has distance in the range $[\alpha_k n, \beta_k n]$. Let us say that two solutions are adjacent if they have Hamming distance 1 and consider the connected components under this notion of adjacency. In [4] it was shown that above a certain density, there exist exponentially many connected components of solutions and, moreover, in *every one of them* the majority of variables are frozen, i.e., take the same value in all assignments in the connected component.

Defining a cluster-region to be the union of one or more connected components, [3] proved that above a certain density, not only do exponentially many connected components exist, but there are exponentially many cluster-regions separated from one another by linear Hamming distance. Moreover, asymptotic bounds were given on the volume, diameter, and separation of these cluster regions. Later, in [2], it was shown for random k-SAT and random graph colouring that when clustering occurs, the emergent cluster-regions are also separated by large energetic barriers, i.e., that any path connecting solutions in different cluster-regions passes through value assignments violating linearly many constraints. This picture of cluster-regions remains unchanged if one considers two solutions to be adjacent if they have Hamming distance $o(n)$. At the same time, though, it lends little information regarding the internal organization of cluster-regions, e.g., the connectivity of each such region.

Until now, it has not been proven that *any* random CSP model exhibits clustering into sets that are both well-separated and well-connected. The main contribution of this paper is to prove that this phenomenon does indeed occur for random k-XOR-SAT, i.e., for systems of random linear equations over GF(2) where each equation contains precisely k variables. We also obtain a precise description of the clusters. We remark that the cluster structure for k-XOR-SAT is much simpler than what is hypothesized for most CSPs, e.g., the clusters are all isomorphic and have the same set of *frozen variables* (see Section 4). Random k-XOR-SAT has long been recognized as one of the most accessible of the fundamental random CSP models, in that researchers have managed to prove difficult results for it that appear to be far beyond our current reach for, e.g., random k-SAT and random graph coloring. For example, the k-XOR-SAT satisfiability threshold was established by Dubois and Mandler [14] for $k = 3$ and by Dietzfelbinger et al. [12] and independently by Pittel and Sorkin [32] for general k.

1.1. Random Systems of Linear Equations

We consider systems of $m = O(n)$ linear equations over n Boolean variables, where each equation binds a constant number of variables. Clearly, deciding whether such a system has satisfying assignments (solutions) can be done in polynomial time by, say, Gaussian elimination. In fact, the set of solutions forms a subspace, so that the sum of two solutions is also a solution. At the same time, it seems that if one fails to exploit the underlying algebraic

structure everything falls apart. For example, if the system is unsatisfiable, finding a value assignment σ that satisfies as many equations as possible, i.e., MAX XOR-SAT, is NP-complete. Moreover, given a satisfiable system and an arbitrary $\sigma \in \{0,1\}^n$, finding a solution nearest to σ is also NP-complete [5]. Finally, random systems of linear equations appear to be extremely difficult both for generic CSP solvers and for SAT solvers working on a SAT encoding of the instance. Indeed, very recent work strongly suggests that among a wide array of random CSPs, random k-XOR-SAT, defined below, is the *most* difficult for random walk type algorithms such as WalkSat [16].

In random k-XOR-SAT, which we study here, each equation binds exactly $k \geq 3$ variables (the case $k = 2$ is trivial). To form the random system of equations $Ax = b$ we take A to be the adjacency matrix of a random k-uniform hypergraph H with n variables and m edges and $b \in \{0,1\}^m$ to be a uniformly random vector. It is straightforward to see, using e.g., Gaussian elimination, that if two systems have the same matrix A, then their solution spaces are isomorphic as b ranges over vectors for which the solution space is not empty. Since we will only be interested in properties of the set of solutions that are invariant under isomorphism, we will assume throughout that $b = \mathbf{0}$. As a result, throughout the paper we will be able to identify the system of linear equations with its underlying hypergraph. Regarding the choice of random k-uniform hypergraphs we will use both standard models $H_k(n,m)$ and $H_k(n,p)$, which respectively correspond to: including exactly m out of the possible $\binom{n}{k}$ edges uniformly and independently, and including each possible edge independently with probability p. (Results transfer readily between the two models when $m = p\binom{n}{k}$.) Our corresponding models of random k-XOR-SAT are:

Definition 1. $X_k(n,m), X_k(n,p)$ are the systems of linear equations over n boolean variables whose underlying hypergraphs are $H_k(n,m), H_k(n,p)$ and where we set $b = \mathbf{0}$.

The usual model for k-XOR-SAT differs from ours only in that it takes a uniformly random Boolean vector b. As described above, these models are equivalent up to isomorphisms of the solution space, and hence we can use our more convenient definition for the purposes of this paper. We will say that a sequence of events $\mathcal{E}_n$ holds *with high probability (w.h.p.)* for such a system if $\lim_{n\to\infty} \Pr[\mathcal{E}_n] = 1$. We will analyze $X_k(n,p)$. All our theorems translate to $X_k(n,m)$ where $m = p\binom{n}{k}$ using a standard argument.

We are interested in the range $p = \Theta(n^{1-k})$ which is equivalent to $m = \Theta(n)$. We note that as $n \to \infty$, the degrees of the variables in such a random system tend to Poisson random variables with mean $\Theta(1)$. This implies that w.h.p. there will be $\Theta(n)$ variables of degree 0 and 1. Clearly, variables of degree 0 do not affect the satisfiability of the system. Similarly, if a variable v appears in exactly one equation e_i, then we can always satisfy e_i by setting v appropriately for any constant b_i. Therefore, we can safely remove e_i from consideration and only revisit it after we have found a solution to the remaining equations. Crucially, this removal of e_i can cause the degree of other variables to drop to 1. This leads us to the definition of the core of a hypergraph.

Definition 2. The *r-core* of a hypergraph H is the maximum subgraph of H in which every vertex has degree at least r.

It is well known [20,27,33] that for every fixed $r \geq 2$, as p is increased, $H_k(n,p)$ acquires a (massive) non-empty r-core suddenly, around a critical edge probability $p = c^*_{k,r}/n^{k-1}$.

Trivially, removing any vertex of degree less than r and all its incident edges from H does not change its r-core. Therefore, the r-core is the (potentially empty) outcome of the

following procedure: repeatedly remove an arbitrary vertex of degree less than r until no such vertices remain. In the case of linear equations we will be particularly interested in 2-cores, as variables outside the 2-core can always be properly assigned.

Definition 3. The *2-core system* is the subsystem of linear equations induced by the 2-core of the underlying hypergraph, i.e., the set of equations whose variables all lie in the 2-core. A *2-core solution* is a solution to the 2-core system. An *extension* of a 2-core solution, σ, is a solution of the entire system of linear equations that agrees with σ on all 2-core variables.

We will show that in the absence of a 2-core, while the diameter of the set of solutions is linear, it is w.h.p. possible to transform any solution to any other solution by changing $O(\log n)$ variables at a time. So, the set of solutions is not only well-connected but pairs of solutions exist at, essentially, every distance-scale. On the other hand, the emergence of the 2-core signals the onset of clustering, as now every pair of solutions is either very close with respect to the 2-core variables, or very far.

Theorem 1. *For every $k \geq 3$ and $c > c^*_{k,2}$, there exists a constant $\alpha = \alpha(c,k) > 0$ such that in $X_k(n, p = c/n^{k-1})$, w.h.p. every pair of solutions either disagree on at least αn 2-core variables, or on at most $\xi(n)$ 2-core variables, for any function $\xi(n) \to \infty$ arbitrarily slowly.*

We will refine the picture of Theorem 1, to prove that as soon as the 2-core emerges, unless two solutions agree on essentially all 2-core variables, transforming one into another requires the simultaneous change of $\Omega(n)$ variables. To identify the relevant 2-core disagreements, we need to define the following notion which is central to our work.

Definition 4. A *flippable cycle* in a hypergraph H is a set of vertices $S = \{v_1, \ldots, v_t\}$ where the set of edges incident to S can be ordered as $e_1, \ldots e_t$ such that each vertex v_i lies in e_i and in e_{i+1} and in no other edges of H (addition mod t).

Thus, the vertices $v_1, \ldots, v_t$ must have degree exactly two in the hypergraph. The remaining vertices in edges $e_1, \ldots, e_t$ can have arbitrary degree and are *not* part of the flippable cycle. The following observation is a simple exercise:

Observation 5. No vertex can lie in two flippable cycles.

Definition 6. A *core flippable cycle* in a hypergraph H is a flippable cycle in the subhypergraph $H_0 \subseteq H$ induced by the 2-core of H.

Thus, in a core flippable cycle, the vertices $v_1, \ldots, v_t$ have degree exactly two *in the 2-core*, but possibly higher degree in H. Note also that H may contain flippable cycles outside the 2-core.

As discussed above, any 2-core solution can be readily extended to the remaining variables. Indeed, this can typically be done in numerous ways since the equations not in the 2-core are far less constrained, e.g., a constant fraction of the equations outside the 2-core form hypertrees very loosely attached to the 2-core. In order to understand the emergence of the clustering of solutions, we will focus on whether we can change the value of a 2-core variable without changing many other 2-core variables.

If σ is any 2-core solution then flipping the value of all variables in a core flippable cycle readily yields another solution of the 2-core, since every equation contains either zero or two of the flipped variables. It is not hard to show that a random hypergraph often contains a handful of short core flippable cycles, implying that 2-core solutions may have Hamming distance $\Theta(1)$. At the same time, though, we will see (Lemma 35) that, for any $\xi(n) \to \infty$ arbitrarily slowly, the total number of vertices in core flippable cycles w.h.p. does not exceed $\xi(n)$, placing a corresponding upper bound on the distance between core solutions that differ only on flippable cycles.

In contrast, we will prove that w.h.p. every pair of core solutions that differ on *even one* 2-core variable not in a flippable cycle, differ in at least $\Omega(n)$ 2-core variables. In other words, flipping the handful of variables in potential flippable cycles, w.h.p. is the *only* kind of movement between 2-core solutions that does not entail the simultaneous change of a massive number of variables.

The above indicates that the following is the appropriate definition of clusters in random k-XOR-SAT.

Definition 7. Two solutions are *cycle-equivalent* if on the 2-core they differ only on variables in core flippable cycles (while they may differ arbitrarily on variables not in the 2-core).

Definition 8. The *solution clusters* of $X_k(n, p = c/n^{k-1})$ are the cycle-equivalence classes, i.e., two solutions are in the same cluster iff they are cycle-equivalent.

Note that in the absence of a 2-core, this definition states that all solutions are in the same cluster. We can now state our main theorems in terms of connectivity properties of clusters.

Definition 9. Two solutions σ, τ of a CSP are *d-connected* if there exists a sequence of solutions $\sigma, \sigma', \ldots, \tau$ such that the Hamming distance of every two successive elements in the sequence is at most d. A set S of solutions is *d-connected* if every pair $\sigma, \tau \in S$ is d-connected. Two solution sets S, S' are *d-separated* if every pair $\sigma \in S, \tau \in S'$ is *not* d-connected.

It appears that for many random CSP's, there is a constant $\alpha > 0$ and a function $g(n) = o(n)$ such that if the constraint density is sufficiently large, then all but a vanishing proportion of the solutions can be partitioned into *clusters* $S_1, \ldots, S_t$ such that:

- Every S_i is $g(n)$-connected.
- Every pair S_i, S_j is αn-separated.

That is the sense in which we said earlier that each cluster is well-connected and that each pair of clusters is well-separated.

Our main theorems are that for k-XOR-SAT, the clusters we defined in Definition 8 satisfy these conditions with $g(n) = O(\log n)$. Note that for this particular CSP, the clusters contain *all* the solutions, rather than all but a vanishing proportion of them.

Theorem 2. *For any constant $c \neq c^*_{k,2}$ and $k \geq 3$, there exists a constant $\alpha = \alpha(c, k) > 0$ such that in $X_k(n, p = c/n^{k-1})$, w.h.p. every pair of clusters is αn-separated.*

In stark contrast, we prove that clusters are internally very well connected.

Theorem 3. *For any constant $c \neq c^*_{k,2}$ and $k \geq 3$, there exists a constant $Q = Q(c,k) > 0$ such that in $X_k(n, p = c/n^{k-1})$, w.h.p. every cluster is $Q \log n$-connected.*

Theorem 3 is nearly tight due to the following.

Observation 10. W.h.p. every cluster contains a pair of solutions that are not $g(n)$-connected, for some $g(n) = \Omega(\log n / \log\log n)$.

Proof. Consider any solution σ to the 2-core, and consider any two extensions σ_0, σ_1 of σ to the entire system such that, for some non-core variable v, we have $\sigma_0(v) = 0$ but $\sigma_1(v) = 1$. Then σ_0, σ_1 must differ in at least one additional variable in every equation containing v implying that their Hamming distance is at least $\deg(v) + 1$.

If T is an acyclic (tree) component of the underlying hypergraph and v is any vertex in T, then, clearly, σ can be extended so that v takes any desired value. Therefore, the maximum degree of any vertex in a tree component is a lower bound for $g(n)$. A tree component T is a d-star if precisely one vertex in T has degree d and all other vertices have degree 1. Computing the second moment of the number of d-stars in a random hypergraph implies that w.h.p. there exist $g(n)$-stars, where $g(n) = \Omega(\log n / \log\log n)$. ∎

So, in a nutshell, we prove that before the 2-core emerges any solution can be transformed to any other solution along a sequence of successive solutions differing in $O(\log n)$ variables. In contrast, after the 2-core emerges, the set of solutions shatters into clusters defined by complete agreement on the 2-core, except for the handful of variables in core flippable cycles: any two solutions that disagree on *even one* 2-core variable not in a core flippable cycle, must disagree on $\Omega(n)$ variables. At the same time, solutions in the same cluster behave like solutions in the pre-core regime, i.e., one can travel arbitrarily inside each cluster by changing $O(\log n)$ variables at a time.

Our proof of Theorem 3 is algorithmic, giving an efficient method to travel between any pair of solutions in the same cluster. Indeed, to prove Theorem 3, we draw heavily from the linear structure of the constraints to: (1) identify a set B of *free variables* such that the $2^{|B|}$ solutions in any cluster are determined by the $2^{|B|}$ assignments to B, (2) prove that we can change these free variables one-at-a-time, each time obtaining a new solution by changing only $O(\log n)$ other variables.

For $c < c^*_{k,2}$, there is no 2-core, and so all solutions belong to the same cluster. For $c > c^*_{k,2}$, but below the k-XOR-SAT satisfiability threshold, the number of 2-core variables exceeds the number of 2-core equations by $\Theta(n)$ (see [12, 14]), and the number of variables on core flippable cycles has expectation $O(1)$ (Lemma 35); it follows that w.h.p. there are an exponential number of clusters. So Theorems 2, 3 yield:

Corollary 11. *For every $k \geq 3$ and c below the k-XOR-SAT satisfiability threshold:*

- *If $c < c^*_{k,2}$, then w.h.p. the entire solution-set of $X_k(n, p = c/n^{k-1})$ is $O(\log n)$-connected.*
- *If $c > c^*_{k,2}$, then w.h.p. the solution-set of $X_k(n, p = c/n^{k-1})$ consists of an exponential number of $\Theta(n)$-separated, $O(\log n)$-connected clusters.*

For $k \geq 3$, the threshold for the appearance of a non-empty 2-core was determined in [20, 27] (see also [11]) to be:

$$c^*_{k,2} = \min_{\lambda>0} \frac{(k-1)!\lambda}{(1-e^{-\lambda})^{k-1}}.$$

For example, when $k = 3$, an exponential number of clusters emerge at $c = 0.13...$ while the satisfiability threshold [14] is at $c = 0.15....$ (These values correspond to $m/n = 0.818...$ and $m/n = 0.917...$ in the $X_k(n, m)$ model.) It is not clear what happens, in terms of clustering, at density $c = c^*_{k,2}$; see the remarks following Theorem 5.

Our proof of Theorem 2 easily extends to all *uniquely extendable* CSPs.

Definition 12 ([10])**.** A constraint of arity k is *uniquely extendable* if for every set of $k-1$ variables and every value assignment to those variables there is precisely one value for the unassigned variable that satisfies the constraint.

Linear equations over GF(2) and unique games are the two most common examples of uniquely extendable (UE) CSPs, but many others exist (see, eg. [10]). Clearly, any instance of a UE CSP Φ is satisfiable iff its 2-core is satisfiable. Thus, it is natural to define clusters analogously to XOR-SAT, i.e., two solutions are in the same cluster if and only if their 2-core restrictions differ only on core flippable cycles. Our proof of Theorem 2 applies readily to any UE CSP, yielding a corresponding theorem, i.e., that there exists $\alpha > 0$ such that if two solutions are not cycle-equivalent they are not αn-connected (see the remark following Proposition 48). However, we do not know whether the analogue of Theorem 3 holds under this definition of clusters, i.e., whether it is possible to travel between cycle-equivalent solutions in small steps. Also, note that while in XOR-SAT changing all the variables in *any* flippable cycle results in another solution, this is not necessarily the case for every UE CSP Φ.

Finally, we note that Theorem 1 follows immediately from Theorem 2 and the fact that w.h.p. there are fewer than $\xi(n)$ vertices on flippable cycles (Lemma 35). Indeed, if two solutions differ on more than $\xi(n)$ 2-core variables, then they disagree on a variable that is not on a flippable cycle. Thus, they are in different clusters and so disagree on at least αn variables, by Theorem 2. So the paper focuses on proving Theorems 2 and 3.

Remark. Ibrahimi, Kanoria, Kraning and Montanari [17] have, independently, obtained similar results to ours. Their definition of clusters is equivalent to ours, and they prove that the clusters are well-connected and well-separated. Their cluster separation result is equivalent to our Theorem 2, but uses a different technique. Their internal connectivity result differs from our Theorem 3 in that (a) they prove that the clusters are polylog(n)-connected rather than $O(\log n)$-connected, and (b) they additionally prove that the clusters exhibit a form of high conductance. Again, their approach is different from the one used here. To prove high conductance, they show that w.h.p. the solution space contains a basis which is polylog(n)-sparse, meaning that each vector in the basis has Hamming distance at most polylog(n) from $\mathbf{0}$. It is easy to see that the set of free variables B that we choose in Section 4 yields a $O(\log n)$-sparse basis. So our proof, along with Lemma 1.1 of [17] combine to yield a stronger conductance result by replacing "$(\log n)^C$" with "$O(\log n)$" in their Theorem 1.

1.2. Cores of Hypergraphs

The main step in our proof of Theorem 3 is to prove a property of the non-2-core vertices in a random hypergraph. As this property is of independent interest, we prove it for non-r-core vertices for general $r \geq 2$.

For any integers $k \geq 2, r \geq 2$ such that $r + k > 4$, the threshold for the appearance of a non-empty r-core in a k-uniform random hypergraph was determined in [20, 27] to be:

$$c^*_{k,r} = \min_{\lambda>0} \frac{(k-1)!\lambda}{\left[e^{-\lambda} \sum_{i=r-1}^{\infty} \lambda^i/i!\right]^{k-1}}. \tag{1}$$

For $r = k = 2$, i.e., for cycles in graphs, the emergence of a 2-core is trivial as any constant-sized cycle has non-zero probability for all $c > 0$. On the other hand, for $r + k > 4$, any r-core has linear size w.h.p. The threshold for the emergence of a 2-core of linear size in a random graph coincides with the threshold for the emergence of a giant component [31], so we set $c^*_{2,2} = 1$, consistent with the expression above after replacing min with inf.

Recall that we can reach the r-core of a hypergraph by repeatedly removing any one vertex of degree less than r, until no such vertices remain. Consider a vertex v not in the r-core, and consider the goal of repeatedly removing vertices of degree less than r until v is removed. We prove that w.h.p. for *every* non-r-core variable v, this can be achieved by removing only $O(\log n)$ vertices.

Definition 13. An *r-stripping sequence* is a sequence of vertices that can be deleted from a hypergraph, one-at-a-time, along with their incident hyperedges such that at the time of deletion each vertex has degree less than r. A *terminal r-stripping sequence* is one that contains all vertices outside the r-core; i.e., a sequence whose deletion leaves the r-core.

Definition 14. For any vertex v not in the r-core, the *depth* of v is the length of a shortest r-stripping sequence ending with v.

Theorem 4. *For any integers $k \geq 2, r \geq 2$ and any constant $c \neq c^*_{k,r}$, let $H = H_k(n, p = c/n^{k-1})$. There exists a constant $Q = Q(c, k, r) > 0$ such that w.h.p., every vertex v in H has depth at most $Q \log n$.*

It is easy to show using standard facts about r-cores of random hypergraphs that for every constant $\epsilon > 0$, there is a constant $T = T(\epsilon)$ such that w.h.p. all but ϵn of the non-core vertices have depth at most T. The challenge here is to prove that w.h.p. *all* non-core vertices have depth $O(\log n)$.

Remark 15. The case $k = r = 2$, i.e. the 2-core of a random graph, follows easily from previously known work. The conclusion of Theorem 4 does not hold at $c = c^*_{2,2} = 1$. (See the remarks following the statement of Theorem 5 below.)

2. RELATED WORK

To get an upper bound on the random k-XOR-SAT satisfiability threshold, observe that the expected number of solutions in a random instance with n variables and m constraints is bounded by $2^n(1/2)^m \to 0$ if $m/n > 1$. As one can imagine, this condition is not tight

since variables of degree 0 and 1 only contribute fictitious degrees of freedom. Perhaps the next simplest necessary condition for satisfiability is $m_c/n_c = \gamma_c \leq 1$, where n_c, m_c is the number of variables and equations in the 2-core, respectively. In [14] Dubois and Mandler proved that, for $k = 3$, this simple necessary condition for satisfiability is also sufficient by proving that for all $\gamma_c < 1$, the number of core solutions is strongly concentrated around its (exponential) expectation. Thus, they determined the satisfiability threshold for 3-XOR-SAT. Dietzfelbinger et al. [12] modify and extend the approach of [14] to determine the satisfiability threshold for general k. A full version of [14] has not been published, but a proof for all $k \geq 3$ appears in [12].

Mézard et al. [25] were the first to study clustering in random k-XOR-SAT. Specifically, they defined the clusters by saying that two solutions are in the same cluster iff they agree on *all* variables in the 2-core. They proved that there exists a constant $\gamma > 0$ such that for any $\theta \in (0, \gamma)$ and any integer $z = \theta n + o(n)$, w.h.p. no two solutions differ on exactly z variables in the 2-core. Based on this fact, they claimed that the clusters they defined are $\Omega(n)$-separated, i.e., that every pair of solutions in different clusters is not γn-connected. As we have already seen, this is false since it does not account for the effect of core flippable cycles. Performing the analysis of solutions that differ on $o(n)$ variables is what allows us to establish that 2-core solutions which differ on $o(n)$ variables must differ *only* on core flippable cycles. Indeed, this is the most difficult part of our proof of Theorem 2.

Mézard et al. [25] also gave a heuristic argument that if σ is any solution and v is a non-core variable, then there exists a solution σ' in which v takes the opposite value from the one in σ such that the distance between σ and σ' is $O(1)$. From this they concluded that clusters are well-connected. Regarding internal connectivity, the clusters of [25] are, indeed, well-connected, i.e., the analogue of Theorem 3 holds for them, since they are subsets of the clusters defined in this paper. However their proof of this fact is flawed; it implies that their clusters are $O(1)$-connected, which is not true by the same argument used as for Observation 10. Proving that the k-XOR-SAT clusters are well-connected was later listed as an open problem in [26].

Finally, as described above, Ibrahimi, Kanoria, Kraning and Montanari [17] have, independently, obtained similar results to ours.

3. PROOF OUTLINE

3.1. Theorem 2: Cluster Separation

Given a solution σ, a *flippable set* is a set of variables S such that flipping the value of all variables in S yields another solution τ. Proving Theorem 2 boils down to proving that w.h.p., in the subsystem induced by the 2-core, every flippable set other than a flippable cycle has linear size.

A common approach to proving analogous statements is to establish that every flippable set, other than a flippable cycle, must deterministically induce a dense subgraph. In particular, if one can prove that for some constant $\epsilon > 0$, every such set is at least $1 + \epsilon$ times as dense as a flippable cycle, then standard arguments yield the desired conclusion. Here, though, this is not the case, due to the possibility of arbitrarily long paths of degree 2 vertices. Specifically, by replacing the edges of any flippable set (that is not a flippable cycle) by *2-linked paths*, one can easily create flippable sets whose density is arbitrarily close to that of a flippable cycle (for a more more precise statement, see the definition of *2-linked*

paths in Section 9). Thus, controlling the number and interactions of these 2-linked paths, an approach similar to that of [1, 29], is crucial to our argument. In order to work on the 2-core, we carry this analysis out on hypergraphs with a given degree sequence.

The key to controlling 2-linked paths is to bound a parameter governing the degree to which they tend to branch. Lemma 32 shows that this parameter is bounded below 1, so while arbitrarily long 2-linked paths will occur, their frequency decreases exponentially with their length.

We note that if we were working on hypergraphs with minimum degree at least 3, then there would be no 2-linked paths, and the proof would have been very easy. All of the difficulties arise from the problem of degree 2 vertices. We note that our approach applies to general degree sequences of minimum degree 2.

3.2. Theorem 3: Connectivity Inside Clusters

The main step in the proof of Theorem 3 is to prove Theorem 4; i.e. that every vertex outside the r-core can be removed by an r-stripping sequence of length $O(\log n)$.

It is often useful to consider stripping the vertices in several parallel rounds.

Definition 16. The *parallel r-stripping process* consists of iteratively removing *all* vertices of degree less than r at once along with any hyperedges containing any of those vertices, until no vertices of degree less than r remain.

To prove that all non-core vertices can be removed by a stripping sequence of length $O(\log n)$, our approach is significantly different below and above the threshold, $c^*_{k,r}$, for the emergence of an r-core in random k-uniform hypergraphs. In both cases, we begin by stripping down to H_B, the hypergraph remaining after B rounds of the parallel stripping process, for a sufficiently large constant B. A simple argument shows that for any non-core vertex v, the number of vertices removed during this initial phase that are relevant to the removal of v, is bounded. Thus, what remains is to show that any non-core vertex in H_B can be removed from H_B by a stripping sequence of length $O(\log n)$.

For $c < c^*_{k,r}$, we prove that there exists a sufficiently large constant $B = B(c, k, r)$ such that all connected components of H_B have size at most $W = O(\log n)$; therefore, any remaining vertex can be removed with an additional W strips. To do this we establish analytic expansions for the degree sequence of H_B as B grows and then apply a hypergraph extension of the main result of Molloy and Reed [28] regarding the component sizes of a random k-uniform hypergraph with a given degree sequence.

For $c > c^*_{k,r}$, a lot more work is required. Once again, 2-linked paths are a major problem. Indeed, it is not hard to see that a long 2-linked path with one endpoint of degree 1, can create a long stripping sequence leading to the removal of its other endpoint.

We first establish that for any $\epsilon > 0$, there exists a sufficiently large constant $B = B(c, k, r, \epsilon)$ such that H_B is sufficiently close to the r-core for two important properties to hold in H_B: (i) there are at most ϵn vertices of degree less than r, and (ii) the "branching" parameter for 2-linked paths, mentioned above, is bounded below 1. Property (ii) allows us to control long 2-linked paths. However, this does not suffice as we need to control, more generally, for large tree-like stripping sequences. To do so, we note that any large tree must either have many leaves, or long paths of degree 2 vertices. Such long paths will correspond to 2-linked paths in the random hypergraph, and so (ii) allows us to control the latter case.

Leaves of the tree will have degree less than r, and so (i) enables us to control the former case.

4. AN ALGORITHM FOR TRAVELING INSIDE CLUSTERS

In this section, we show how we use Theorem 4 to prove Theorem 3. In fact, we require Theorem 5 below, which is somewhat stronger than Theorem 4.

Given a hypergraph H, we consider any terminal r-stripping sequence, $v_1, \ldots, v_t$, i.e., one that removes every vertex outside of the r-core of H. Let H_i denote the hypergraph remaining after removing $v_1, \ldots, v_{i-1}$; so $H_1 = H$ and H_{t+1} is the r-core of H. Let E_i denote the set of at most $r-1$ hyperedges in H_i that contain v_i. We form a directed graph, D, as follows:

Definition 17. The vertices of D are the non-r-core vertices $v_1, \ldots, v_t$, as well as any r-core vertex that shares a hyperedge with a vertex not in the r-core. For each vertex v_i in the stripping sequence, D contains a directed arc (u, v_i) for every vertex $u \neq v_i$ contained in the hyperedges of E_i. Note that if v_i has degree zero in H_i, then $E_i = \emptyset$, and so v_i will have indegree zero in D.

For every vertex v in D, we define $R^+(v)$ to be the set of vertices that can be reached from v. Note that if v is not in the r-core, then the vertices of $R^+(v)$ can be arranged into a (not necessarily terminal) r-stripping sequence ending with v. So to prove Theorem 4, it suffices to show $|R^+(v)| = O(\log n)$ for every such v.

Theorem 5. *For any integers $k \geq 2, r \geq 2$ and any constant $c > 0, c \neq c^*_{k,r}$, let $H = H_k(n, p = c/n^{k-1})$. There exists a constant $Q = Q(k, r, c) > 0$ such that w.h.p. there is a terminal r-stripping sequence of H for which in the digraph D associated with the sequence:*

(a) For every vertex v, $|R^+(v)| \leq Q \log n$.
(b) For $r = 2$, for every core flippable cycle C,

$$\sum_{v \in C} |R^+(v)| \leq Q \log n.$$

Remark 18. The proof of Theorem 5 can be extended to show that w.h.p. for every vertex $v \in D$, the subgraph induced by $|R^+(v)|$ has at most as many arcs as vertices.

Remark 19. The case $r = k = 2$ follows from previously known work. For $c < c^*_{2,2} = 1$, it follows from the fact that w.h.p. every component of $G_{n,p=c/n}$ has size $O(\log n)$ below the giant component threshold $c^*_{2,2}$. For $c > c^*_{k,r}$, it follows from Lemma 5(b) of [31]. Our proof will work for $k = r = 2$, but it is convenient to assume $(k, r) \neq (2, 2)$.

Remark 20. We think that the conclusion of Theorem 5 does not hold at $c = c^*_{k,r} + o(1)$. This is known to be true for the case $k = r = 2$. Indeed, when $c = 1 - \lambda$, for $\lambda = n^{-1/3+\epsilon}, \epsilon > 0$, w.h.p. the size of the largest component is $\Theta(\lambda^{-2} \log n)$ and no component has more than one cycle [22]. A simple first moment analysis yields that w.h.p. there is no

cycle of length greater than $\log n/\lambda$. Furthermore, w.h.p. no vertex has degree greater than $\log n$. It follows that the largest component must contain an induced subtree, none of whose vertices are in the 2-core, which has size $\Theta(\frac{\lambda^{-2}\log n}{\log^2 n/\lambda}) = \Theta(1/(\lambda \log n))$. It is easy to see that such a subtree will contain vertices with depth $\Theta(1/(\lambda \log n))$, which can be as large as n^α for any $\alpha < 1/3$.

The proof of Theorem 5 occupies Sections 7 and 8, after we set out some basic facts about cores in Section 5 and some basic calculations in Section 6. But first, we show that it yields Theorems 3 and 4:

Proof of Theorem 4. This follows immediately from Theorem 5 because the depth of v is at most $|R^+(v)|$. ∎

We are now ready to give our algorithm for traveling between any two assignments in the same cluster while changing $O(\log n)$ variables at a time.

Proof of Theorem 3. Given an arbitrary system of linear equations consider a terminal 2-stripping sequence $v_1, \ldots, v_t$ of its associated hypergraph and let D be the digraph formed from the sequence. For each core flippable cycle, C, we choose an arbitrary vertex $v_C \in C$. Let B be the set consisting of each vertex v_C and every non-2-core vertex with indegree zero in D.

Consider any 2-core solution σ. Consider the system of equations formed from our system by fixing the value of every 2-core variable that does not belong to a core flippable cycle to its value in σ; we call such vertices *fixed vertices*. Recall from Definition 4 that the edges of a flippable cycle contain vertices that are not considered to be vertices of the flippable cycle; such vertices will be fixed. Note that the solutions of this system form a cluster, and that every cluster can be formed in this way from some σ.

We will perform Gaussian elimination on this system in a manner such that B will be the set of free variables that we obtain. Importantly, this set of free variables does not depend on σ, i.e., it will be the same for every cluster.

For each $v \in B$ and for each fixed vertex v, set $\chi(v) = \{v\}$. For each core flippable cycle C, we process all of the edges (i.e., equations) in C except for one of the edges containing v_C. For each vertex $v \in C$, we obtain the equation $v = v_C + z_v$ where z_v is a constant (0 or 1) depending only on the assignment to the fixed vertices in the edges of C; we set $\chi(v) = \{v_C\}$. By Observation 5, the core flippable cycles are edge-disjoint, and so we can carry this out for each core flippable cycle C.

Next, we process the edges not in the 2-core, in reverse removal order, i.e., $E_t, \ldots, E_1$. Note that, since $r = 2$, each E_i contains at most one edge. When processing E_i, we set $\chi(v_i)$ to the symmetric difference of the sets $\chi(u)$, over all $u \in E_i$ other than v_i. That is, a variable z is in $\chi(v_i)$ iff $z \in \chi(u)$ for an odd number of variables $u \in E_i$ other than v_i. Since E_i is the equation $v_i = \sum_{u \in E_i; u \neq v} u$, this is equivalent (by induction) to $v_i = \sum_{w \in \chi(v_i)} w + z_{v_i}$, where z_{v_i} is the sum of z_u over all vertices $u \in \chi(v_i)$ that belong to core flippable cycles. We now note that every non-2-core vertex $v_i \notin B$ has indegree at least 1 in D and so $|E_i| = 1$ and thus $\chi(v_i)$ is defined. For each vertex $u \neq v_i$ in E_i, either $u \in B$, or u is fixed, or $u = v_j$ for some $j > i$, or u is in a core flippable cycle. Therefore, by induction, $\chi(v_i)$ contains only vertices that are in B or are fixed.

Finally, note that possibly $\chi(v_i) = \emptyset$; in that case, $v_i = \sum_{w \in \chi(v_i)} w + z_{v_i} = z_{v_i}$ in every solution. (It is not hard to adapt the proof of Theorem 5 to show that w.h.p. for every i, $\chi(v_i) \neq \emptyset$. But that is not required for the purposes of this paper.)

At this point, all non-fixed vertices are either in B or have been expressed as the sum of vertices in B and fixed vertices. Therefore, the vertices in B are the free variables for the system obtained by fixing the values of the fixed vertices to σ. Thus, there are exactly $2^{|B|}$ solutions to that system, one for each assignment to B. We can move between any two such solutions by changing the assignments to the vertices of B, one at a time. Each time we change the value of a non-2-core vertex $v \in B$, in order to get to another solution, we only need to change a subset of $R^+(v)$ in the digraph D, because only vertices $u \in R^+(v)$ can have $v \in \chi(u)$. Similarly, each time we change the value of some $v_C \in B$, we only need to change a subset of $\cup_{v \in C} R^+(v)$. Thus, by Theorem 5, we can move between any two such solutions changing at most $Q \log n$ variables at a time. This implies Theorem 3, since each cluster is such a solution set. ∎

We close this section by showing how the preceding proof extends to determine all of the frozen variables. A variable is said to be *frozen* in a cluster, if it takes the same value in all assignments of the cluster. In general random CSPs it is hypothesized that the set of frozen variables can differ from cluster to cluster. In random k-XOR-SAT, though, the set of frozen variables depends only on the underlying hypergraph, i.e., is the same for all clusters.

Theorem 6. *In every cluster, the frozen variables consist of the 2-core vertices not in core flippable cycles, and the non-2-core variables v for which $\chi(v) \cap B = \emptyset$.*

Proof. This follows immediately from the fact that B is the set of free variables in a system of linear equations whose solution set is the cluster. ∎

5. RANDOM HYPERGRAPHS AND THEIR CORES

We will use the configuration model of Bollobás [6] to generate a random k-uniform hypergraph H with a given degree sequence. Suppose we are given the degree $d(v)$ for each vertex v; thus $\sum d(v) = kE$ where E is the number of hyperedges. We take $d(v)$ *copies* of each v, and we take a uniformly random partition of these kE vertex-copies into E sets of size k. This naturally yields a k-uniform hypergraph, by mapping each k-set to a hyperedge on the vertices whose copies are in the k-set. Note that the hypergraph may contain loops (two copies of the same vertex in one hyperedge) and multiple edges (two identical hyperedges). It is well known that the probability that this partition yields a simple hypergraph (i.e., one with no loops or multiple edges) is bounded below by a constant for degree sequences[1] satisfying certain conditions. Specifically:

Definition 21. Say that a degree sequence $\mathcal{S}$ is *nice* if $E = \Theta(n)$, $\sum_v d(v)^2 = O(n)$ and $d(v) = o(n^{1/24})$ for all v.

Every degree sequence we will consider will correspond to some subgraph of $H_k(n,p)$ with a linear expected number of edges. Since, as is well known, the degree sequence of such random hypergraphs is nice w.h.p., all the degree sequences we will consider will be

[1] Clearly, we are referring to a sequence of degree sequences $\mathcal{S}_n$ so that asymptotic statements are meaningful. We suppress this point though, throughout, to streamline exposition.

nice. With this in mind, we will make heavy use of the following standard proposition (see eg. [11]) and corollary, as working in the configuration model is technically much easier than working with uniformly random hypergraphs with a given degree sequence.

Proposition 22. *If $\mathcal{S}$ is a nice degree sequence, then there exists $\epsilon > 0$ such that the probability that a random hypergraph with degree sequence $\mathcal{S}$ drawn from the configuration model is simple is at least ϵ.*

This immediately yields:

Corollary 23. *If $\mathcal{S}$ is a nice degree sequence then:*

(a) If property Q holds w.h.p. for k-uniform hypergraphs with degree sequence $\mathcal{S}$ drawn from the configuration model, then Q holds w.h.p. for uniformly random simple hypergraphs with degree sequence $\mathcal{S}$.
(b) For any random variable X, if $E(X) = O(1)$ for k-uniform hypergraphs with degree sequence $\mathcal{S}$ drawn from the configuration model, then $E(X) = O(1)$ for uniformly random simple hypergraphs with degree sequence $\mathcal{S}$.

The following lemma will be very useful. Its exponential term is not tight, but will suffice for our purposes.

Lemma 24. *Consider a random k-uniform hypergraph drawn from the configuration model with E edges, i.e., with total degree kE. For each $i = 2, \ldots, k$, specify ℓ_i sets of i vertex-copies, and set $L = \sum_{i=2}^{k} \ell_i$. The probability that each of these sets appears in some hyperedge, and no two appear in the same hyperedge is less than*

$$\exp\left(\frac{kL^2}{E-L}\right)\prod_{i=2}^{k}\left(\frac{(k-1)(k-2)\cdots(k-i+1)}{(kE)^{i-1}}\right)^{\ell_i}.$$

Proof. We choose the partition of the vertex-copies by processing the specified sets one-at-a-time. To process one of the ℓ_i sets of size i, we first choose one set member γ arbitrarily and then randomly select the remaining $k-1$ vertex-copies of the part containing γ. Every time we do this there are at least $kE - kL$ yet unselected vertex-copies. Thus, the probability we chose all other $i-1$ members of the specified set is at most

$$\begin{aligned}\frac{(k-1)(k-2)\cdots(k-i+1)}{(kE-kL)^{i-1}} &< \frac{(k-1)(k-2)\cdots(k-i+1)}{(kE)^{i-1}} \times \left(\frac{E}{E-L}\right)^{i-1} \\ &< \frac{(k-1)(k-2)\cdots(k-i+1)}{(kE)^{i-1}} \times e^{kL/(E-L)},\end{aligned}$$

since $i \le k$. So the probability that each of the L tuples is chosen to be in a hyperedge is less than

$$\begin{aligned}&\prod_{i=2}^{k}\left(\frac{(k-1)(k-2)\cdots(k-i+1)}{(kE)^{i-1}}\right)^{\ell_i} \times e^{(kL/(E-L))\ell_i} \\ &\qquad = e^{kL^2/(E-L)} \times \prod_{i=2}^{k}\left(\frac{(k-1)(k-2)\cdots(k-i+1)}{(kE)^{i-1}}\right)^{\ell_i}.\end{aligned}$$

∎

5.1. Cores

Recall from Section 4 that Theorem 5 is already known for $k = r = 2$. So we will assume that $k + r > 4$. It is well known that the r-core of a random k-uniform hypergraph is uniformly random conditional on its degree sequence. See [33] for the case $k = 2$, and [27] for the nearly identical proof for general k. In fact, the same is true of the graph remaining after any number of iterations of the parallel stripping process.

Let $H = H_k(n,p)$ be a random k-uniform hypergraph and let $H = H_0, H_1, \ldots$ be the sequence of hypergraphs produced by the parallel r-stripping process. It is well known how (see e.g., [27]) to show the following propositions.

Proposition 25.

(a) For every $i \geq 0$, H_i is uniformly random with respect to its degree sequence.
(b) There exist functions $\rho_0, \rho_1, \ldots$ such that for any fixed integer i, w.h.p. H_i contains $\rho_j(i)n + o(n)$ vertices of degree j and $\frac{1}{k}(\sum_{j\geq 1} j\rho_j(i))n + o(n)$ edges.

Remark 26. The functions $\rho_j(i)$ have explicit recursive expressions, which we give in Section 8. An approximation is stated in Proposition 31 below.

Proposition 25 allows us to use the configuration model to study H_i. We will begin by showing that we can uniformly approximate the total degree of H_i.

Lemma 27. *For every fixed integer $i \geq 0$,*

$$\sum_{v\in H_i} \deg_{H_i}(v) = \left(\sum_{j\geq 1} j\rho_j(i)\right) n + o(n).$$

Proof. Proposition 25 implies that $\sum_{j\geq 1} j\rho_j(i)$ is convergent, else w.h.p. H_i, and hence H, would have a superlinear number of edges.

Consider any fixed J. By Proposition 25, w.h.p. $\sum_{v:\deg_{H_i}(v)\leq J} \deg_{H_i}(v) = \sum_{j=1}^{J} j\rho_j(i)n + o(n)$. For any $\theta > 0$, the convergence of $\sum_{j\geq 1} j\rho_j(i)$ implies that we can choose $J = J(\theta)$ sufficiently large that $\sum_{j>J} j\rho_j(i) < \frac{1}{2}\theta$. Since $H_i \subseteq H_0 = H$, we have $\sum_{v:\deg_{H_i}(v)>J} \deg_{H_i}(v) \leq \sum_{v:\deg_H(v)>J} \deg_H(v)$. The fact that the latter sum is less than $\frac{1}{2}\theta n$ for J sufficiently large is well known and follows from the facts that (i) for each constant ℓ, the number of vertices of degree ℓ in H is w.h.p. $\lambda_\ell n + o(n)$ for a particular $\lambda_\ell = \lambda_\ell(c)$ and (ii) the number of hyperedges in H is highly concentrated around $\frac{1}{k}\sum_{\ell\geq 1} \ell\lambda_\ell n$. Thus, $|\sum_{v\in H_i} \deg_{H_i}(v) - (\sum_{j\geq 1} j\rho_j(i))n| < \theta n$ for every $\theta > 0$, which establishes the lemma. ∎

The following similar bound will also be useful:

Lemma 28. *For every constant d and fixed integer $i > 0$:*

$$\sum_{v:\deg_{H_i}(v)\geq d} \frac{\deg_{H_i}(v)!}{(\deg_{H_i}(v) - d)!} = \left(\sum_{j\geq d} \frac{j!}{(j-d)!}\rho_j(i)\right) n + o(n).$$

Proof. The proof is almost identical to that of Lemma 27 but exploits the concentration of the number of d-stars in H, rather than of the number of hyperedges. (A d-star is a set of

d hyperedges which contain a common vertex.) The concentration of the number of d-stars in H is easily established, e.g., by the Second Moment Method or Talagrand's Inequality. (Indeed, Lemma 27 and its proof are special cases of this lemma and its proof for $d = 1$.) ∎

For any fixed integers k, r and real number $\lambda > 0$, we write

$$\Psi_r(\lambda) = e^{-\lambda} \sum_{i \geq r-1} \lambda^i / i! \quad \text{and} \quad f_{k,r}(\lambda) = f(\lambda) = \frac{(k-1)!\lambda}{\Psi_r(\lambda)^{k-1}}.$$

Recall that for $k+r > 4$, the threshold for the appearance of an r-core in a random k-uniform hypergraph $H_k(n,p)$ with $p = c/n^{k-1}$ is

$$c^*_{k,r} = \min_{\lambda > 0} f_{k,r}(\lambda).$$

We will see that f' has a unique root and, thus, for $c > c^*_{k,r}$ the equation $f(\lambda) = c$ has two solutions.

Definition 29. For $c > c^*_{k,r}$, let $\mu = \mu(c)$ denote the larger of the two solutions of $f(\lambda) = c$.

The following two propositions are standard; see e.g., [27] for proofs.

Proposition 30. *For every fixed $j \geq r$, w.h.p. the r-core contains $(e^{-\mu}\mu^j/j!)n + o(n)$ vertices of degree j. Furthermore, w.h.p. the r-core contains $(\mu/k)\Psi_r(\mu)n + o(n)$ edges.*

Proposition 31. *For every $c \neq c^*_{k,r}$ and $\theta > 0$, there exists $B = B(\theta)$ such that w.h.p.*

(a) H_B contains fewer than θn vertices not in the r-core;
(b) For each $j \geq r$, $|\rho_j(B) - e^{-\mu}\mu^j/j!| < \theta$.

The following lemma will be critical for our analysis.

Lemma 32. *For every $c > c^*_{k,r}$, there exists $\zeta = \zeta(k,r,c) > 0$ such that*

$$(k-1)\frac{\mu^{r-1}}{(r-2)!} < (1-\zeta) \sum_{i \geq r-1} \frac{\mu^i}{i!}, \tag{2}$$

where μ is the larger of the two roots of the equation $f_{k,r}(\lambda) = c$.

Proof.

$$\begin{aligned} f'(\lambda) = 0 \quad &\iff \quad \Psi_r(\lambda) = \lambda(k-1)\Psi_r(\lambda)^{k-2}\Psi_r'(\lambda) \\ &\iff \quad \sum_{i \geq r-1} \frac{\lambda^i}{i!} = (k-1)\frac{\lambda^{r-1}}{(r-2)!}. \end{aligned} \tag{3}$$

Equation (3) yields $c^*_{k,r} = f(\lambda^*)$ for some λ^* satisfying the last equation in (3). For $c > c^*_{k,r}$, since $\mu = \mu(c)$ is the larger of the two roots of $f(\lambda) = c$, it follows that $\mu > \lambda^*$. The lemma now follows by noting that the RHS of (2) divided by the LHS is proportional to $\sum_{i \geq r-1} \frac{\mu^{i-r+1}}{i!}$, which is clearly increasing with μ. ∎

6. PRELIMINARIES TO THE PROOF OF THEOREM 5

Recall that we assume $k + r > 4$ and let $H = H_k(n, p)$ be a random k-uniform hypergraph with $p = c/n^{k-1}$. Let $H = H_0, H_1, \ldots$ be the sequence of hypergraphs produced by the parallel r-stripping process.

As we said above, we will choose a sufficiently large constant B, strip down to H_B, and then focus on $R^+(u) \cap H_B$, making use of the fact that H_B is very close to the 2-core (by Proposition 31). The following will be used to bound the number of vertices that are removed from $R^+(u)$ when stripping down to H_B. For integer $s \geq 0$, we use $N^s(v)$ to denote the s-th neighborhood of v, i.e., the set of vertices within distance s from v. For any set of vertices A, $N^s(A) = \bigcup_{v \in A} N^s(v)$. We consider a single vertex to be a connected set. A straightforward induction yields the following.

Proposition 33. *For any integer i and vertex $u \in H_i$, $R^+(u) \subseteq N^i(R^+(u) \cap H_i)$.*

Lemma 34. *For any $c, s \geq 0$, there exists $\Gamma = \Gamma(c, s)$ such that in a random graph $G(n, p)$ with $p = c/n$, w.h.p. for every connected subset A of vertices $|N^s(A)| \leq \Gamma(|A| + \log n)$.*

Proof. We prove this for the case $s = 1$, i.e., that there is a constant $\gamma > 1$ such that w.h.p. every connected subset of vertices A satisfies $|N(A)| \leq \gamma(|A| + \log n)$. By iterating, we obtain that for every $s \geq 1$, every connected subset of vertices A satisfies $|N^s(A)| \leq f_s(|A|)$ where

$$f_1(x) = \gamma(x + \log n)$$
$$f_{i+1}(x) = \gamma(f_i(x) + \log n), \text{ for } i \geq 1.$$

A simple induction yields $f_i(x) \leq \gamma^i(x + i \log n)$ and that yields the lemma with $\Gamma = s\gamma^s$.

Given any set A of size a, the probability that A is connected is at most the expected number of spanning trees of A which is $a^{a-2}(c/n)^{a-1}$. After conditioning that A is connected, the number of neighbors outside of A is distributed as $\mathrm{Bin}(a(n-a), c/n)$. The probability that this exceeds z is at most

$$\binom{a(n-a)}{z}\left(\frac{c}{n}\right)^z < \left(\frac{eca}{z}\right)^z < 2^{-z}, \qquad \text{for } z > 2eca.$$

For any $\gamma > 2$, if $|N(A)| > \gamma(|A| + \log n)$, then we must have $|N(A) \backslash A| > \frac{1}{2}\gamma(|A| + \log n)$. Taking $\gamma > 4ec$, the expected number of connected sets A satisfying this last inequality is at most

$$\binom{n}{a} a^{a-2}\left(\frac{c}{n}\right)^{a-1} 2^{-\frac{1}{2}\gamma(a+\log n)} < \frac{en}{a^2}(ec)^{a-1} 2^{-\frac{1}{2}\gamma(a+\log n)} < \frac{en}{a^2}\left(\frac{ec}{2^{\gamma/2}}\right)^{a-1} 2^{-\frac{1}{2}\gamma \log n} = n^{-\Theta(\gamma)},$$

for γ sufficiently large. Multiplying by the n choices for a yields the lemma. ∎

Lemma 35. *Fix $k \geq 3$ and let $H = H_k(n, p)$ be a random k-uniform hypergraph with $p = c/n^{k-1}$, where $c > c^*_{k,2}$. The expected number of vertices in core flippable cycles of H is $O(1)$.*

Proof. Let $\mathcal{D}$ be the degree sequence of the 2-core of H. By Corollary 23, we can work in the configuration model. Recalling Definition 29, Proposition 30 and Lemma 32, w.h.p.

(i) $\mathcal{D}$ has total degree $\gamma n + o(n)$, where $\gamma = \mu\Psi_r(\mu)$,
(ii) $\mathcal{D}$ has $\lambda_2 n + o(n)$ vertices of degree 2, where $\lambda_2 = e^{-\mu}\mu^2/2$,
(iii) there exists $\zeta > 0$ such that $2(k-1)\lambda_2 < (1-\zeta)\gamma$.

We first bound the expected number of core flippable cycles of size a. Let $\Lambda = \gamma n + o(n)$ be the total number of vertex copies, and let $L = \lambda_2 n + o(n)$ be the number of copies of degree 2 vertices.

There are $\binom{L}{a}$ choices for the connecting vertices, $\frac{(a-1)!}{2}$ ways to order them into a cycle, and 2^a ways to align their vertex-copies. This yields a pairs $\{y_1, z_1\}, \ldots, \{y_a, z_a\}$ of vertex copies, each of which must land in a hyperedge. We process these pairs one-at-a-time, halting if we ever find that the pair does not land in a hyperedge. To process pair i, we ask only whether z_i lands in the same hyperedge as y_i; if it does we do *not* expose the other vertex-copies in that hyperedge. Thus, prior to processing pair i, we have exposed exactly $2i-2$ vertex-copies, all of degree 2. There are $k-1$ other copies appearing in the same hyperedge as y_i. Each of the $\Lambda - (2i-1)$ unexposed copies (not including y_i) is equally likely to be one of those copies (and, for $k \geq 3$, the exposed copies also have positive probability). So the probability that z_i is one of them is at most $(k-1)/(\Lambda - 2i + 1)$. So the expected number of core flippable cycles of length a is at most:

$$\binom{L}{a}\frac{(a-1)!}{2}2^a \prod_{i=1}^{a}\frac{k-1}{\Lambda - 2i + 1} < \frac{1}{2a}\prod_{i=1}^{a}\frac{2(k-1)(L-i+1)}{\Lambda - 2i + 1}.$$

By condition (iii) above, $2(k-1)L/(\Lambda-1) < 1 - \frac{1}{2}\zeta$, and so $2(k-1)(L-i+1)/(\Lambda - 2i + 1) < 1 - \frac{1}{2}\zeta$ for each i, since $L \leq \frac{1}{2}(\Lambda - 1)$. So the expected number is at most $\frac{1}{2a}(1-\frac{1}{2}\zeta)^a$, and so the expected total number of vertices on core flippable cycles is at most $\frac{1}{2}\sum_{a\geq 1}(1-\frac{1}{2}\zeta)^a = O(1)$. ∎

7. PROOF OF THEOREM 5 ABOVE THE *r*-CORE THRESHOLD

Recall that we can assume $k + r > 4$. We let $H = H_k(n,p)$ be a random k-uniform hypergraph with $p = c/n^{k-1}$. Let $H = H_0, H_1, \ldots$ be the sequence of hypergraphs produced by the parallel r-stripping process. We will choose a terminal r-stripping sequence that is consistent with the parallel process; i.e., in our stripping sequence: for every $i < j$, the vertices deleted in round i of the parallel process come before the vertices deleted in round j of the parallel process.

Let D be the digraph associated with this terminal r-stripping sequence and recall that $R^+(u)$ denotes the set of vertices reachable from a vertex u in D.

7.1. Bound on the Length of Stripping Sequences

Our main challenge is to prove the following lemma. The idea is that we will take B large enough so that by stripping down to H_B, Proposition 31 gives us control of the degree sequence that remains, and Lemma 32 allows us to prove that a certain branching process involving long paths in a graph constructed from H_B dies out.

Lemma 36. *For every $c > c^*_{k,r}$ there exists $B = B(c,k,r)$ and $Q = Q(c,k,r)$ such that w.h.p. for every vertex u, $|R^+(u) \cap H_B| \leq Q \log n$.*

Proof of Theorem 5(a). Consider any vertex u. If $u \notin H_B$, then by Proposition 33, $R^+(u) \subseteq N^B(u)$ in which case Lemma 34 immediately implies that $|R^+(u)| < \Gamma(1 + \log n)$ for some constant $\Gamma = \Gamma(c, B)$.

If $u \in H_B$, then $R^+(u) \subseteq N^B(R^+(u) \cap H_B)$, by Proposition 33. Since, by Lemma 36, $|R^+(u) \cap H_B| \leq Q \log n$, Lemma 34 now implies that $|R^+(u)| < \Gamma(Q \log n + \log n) = Z \log n$ for $Z = \Gamma Q + 1 = Z(c, B) = Z(c, k, r)$. ∎

Definition 37. For any i, we define D_i to be the subdigraph of D induced by the vertices in H_i.

Consider a particular constant i. Let T^+ be a directed tree in D_i with edges directed away from a root u that spans the vertices of $R^+(u) \cap H_i$; e.g., T^+ could be a Breadth First Search or Depth First Search tree from u. Thus, each vertex has indegree at most 1 in T^+, implying:

Proposition 38. *No two arcs of T^+ were formed during the removal of the same hyperedge.*

Definition 39. A *deletion tree rooted at u* is the undirected tree, T, formed by removing the directions from a tree T^+ rooted at u.

To prove Lemma 36, we will bound the expected number of deletion trees T of size greater than $Q \log n$. The following technical lemma bounds the density of small subgraphs of $H_k(n, p)$. It is of a standard flavour and has a standard proof. Given a subset S of the vertices of $H_k(n, p)$, we let $\ell_j(S)$ denote the number of hyperedges that contain exactly j of the vertices of S, and we let $L(S) = \sum_{j=2}^{k} (j-1)\ell_j$.

Lemma 40. *For every $c, \zeta > 0$, there is $\theta > 0$, such that w.h.p. every $S \subseteq H_k(n, p = c/n^{k-1})$ with $|S| \leq \theta n$ has $L(S) < (1 + \zeta)|S|$.*

Proof. Rather than working in the $H_k(n, p)$ model, it will be convenient to work in the $H_k(n, m)$ model, where exactly $m = (c/k!)n$ edges are selected uniformly, independently and with replacement (note that $m = p\binom{n}{k}$). Standard arguments imply that high probability properties in this model transfer to the $H_k(n, p)$ model.

Let $Y_a = Y_a(\zeta)$ denote the number of sets S with $|S| = a$ and $L(S) = (1 + \zeta)|S|$. We will bound $\mathbb{E}(Y_a)$ as follows. Define

$$\mathcal{L}_a = \left\{ (\ell_2, \ldots, \ell_k) : \sum_{j=2}^{k} (j-1)\ell_j \geq (1+\zeta)a \right\}.$$

Choose a vertices and some $(\ell_2, \ldots, \ell_k) \in \mathcal{L}_a$, pick ℓ_j edges for each j, and then multiply by the probability that each edge chooses (at least) the appropriate number of vertices from S. This yields

$$\begin{aligned} E(Y_a) &\leq \binom{n}{a} \sum_{(\ell_2,\ldots,\ell_k)\in\mathcal{L}_a} \prod_{j=2}^{k} \binom{m}{\ell_j} \left[\binom{k}{j} \left(\frac{a}{n}\right)^j \right]^{\ell_j} \\ &< \left(\frac{en}{a}\right)^a \sum_{(\ell_2,\ldots,\ell_k)\in\mathcal{L}_a} \left(\frac{a}{n}\right)^{\sum_{j=2}^{k}(j-1)\ell_j} \prod_{j=2}^{k} \frac{(Ja)^{\ell_j}}{\ell_j!}, \quad \text{for some constant } J = J(c,k) > 0 \end{aligned}$$

$$< \left(\frac{en}{a}\right)^a \left(\frac{a}{n}\right)^{(1+\zeta)a} \prod_{j=2}^{k} \left(\sum_{\ell_j \geq 0} \frac{(Ja)^{\ell_j}}{\ell_j!} \right)$$

$$< e^a \left(\frac{a}{n}\right)^{\zeta a} e^{(k-1)Ja}$$

$$= \left(\frac{\Delta a}{n}\right)^{\zeta a}, \quad \text{for some constant } \Delta = \Delta(c,k,\zeta) > 0.$$

Choosing $\theta = \frac{1}{2\Delta}$, it is standard and straightforward to show $\mathbb{E}\left(\sum_{a=1}^{\theta n} Y_a\right) = o(1)$. ∎

In order to carry out our first moment calculation, we will bound the difference between the degrees of the vertices of T and their degrees in H_i.

Lemma 41. *For any $\delta > 0$, if i is sufficiently large in terms of δ then w.h.p. : For every vertex $u \in D_i$, if T is a deletion tree rooted at u, then $\deg_{H_i}(v) \leq \deg_T(v) + r - 2$ for all but at most $\delta|T| + 3$ vertices $v \in T$.*

Proof. Define S to be the hypergraph with edge set $\{e \cap R^+(u) : e \in H_i, |e \cap R^+(u)| \geq 2\}$. In other words, for each hyperedge $e \in H_i$ that contains at least two vertices of $R^+(u)$, S contains the edge obtained by removing all vertices outside of $R^+(u)$ from e.

Since $V(T) = R^+(u) \cap H_i$, the r-stripping sequence that yields D contains an r-stripping subsequence which removes from H_i only vertices of T, such that all vertices of T except possibly u are removed. Consider $v \in T, v \neq u$. At the point that v is removed, it has degree at most $r-1$ in what remains of H_i. Every other hyperedge of H_i containing v is removed before v, and thus must contain another member of $R^+(u)$. At least one of those $r-1$ hyperedges contains another vertex of $R^+(u)$, namely the parent of v in T. Therefore:

$$\deg_{H_i}(v) \leq \deg_S(v) + r - 2.$$

For $2 \leq j \leq k$, let ℓ_j denote the number of hyperedges with j vertices in S. All vertices of S, except possibly u, are not in the r-core. So, by Lemma 31(a), we know that for any $\theta > 0$ we can select i sufficiently large in terms of θ so that $|S| < \theta n$. If we pick θ sufficiently small in terms of δ, then Lemma 40 implies that w.h.p., $\sum_{j=2}^{k}(j-1)\ell_j < (1+\delta/2)|S|$. So

$$\sum_{v \in R^+(u)} \deg_S(v) = \sum_{j=2}^{k} j\ell_j \leq 2\sum_{j=2}^{k}(j-1)\ell_j < (2+\delta)|S| = (2+\delta)|T|.$$

Now the total T-degree of the vertices in $R^+(u)$ is $2|T| - 2$, since T is a tree with edges of size 2 that spans $R^+(u)$. So for i sufficiently large in terms of δ,

$$\sum_{v \in R^+(u)} \deg_S(v) - \deg_T(v) \leq (2+\delta)|T| - (2|T| - 2) = \delta|T| + 2.$$

So $\deg_T(v) \neq \deg_S(v)$ for at most $\delta|T| + 2$ vertices $v \in R^+(u)$. Also, $\deg_{H_i}(v) \leq \deg_S(v) + r - 2$ for all but at most one $v \in R^+(u)$ (namely $v = u$). This proves the lemma. ∎

Proof of Lemma 36. We will fix a constant $\delta > 0$ that is sufficiently small for various bounds to hold. We also take B sufficiently large for various bounds to hold, including Lemma 41 for $i \geq B$. Let $X_a = X_a(B)$ be the number of deletion trees T in D_B with a vertices. Our goal is to show that there exists some constant $Q > 0$ such that w.h.p. $X_a = 0$

for $a > Q \log n$, so in the following we may allow ourselves to assume that a is greater than some sufficiently large constant.

To prove Lemma 36 we first observe that, by Proposition 31, we can assume H_B is uniformly random conditional on its degree sequence. Since Lemma 36 asserts a property to hold with high probability, it suffices to establish this property in the configuration model for H_B (by Corollary 23(a)). Moreover, recall that by Proposition 31(b), as B is increased w.h.p. the degree sequence of H_B tends to that of the r-core.

Let $v_1, \ldots, v_a$ be the vertices of T. We first specify $d_i = \deg_T(v_i)$ for each i, noting that these degrees must sum to $2a-2$. The number of ways to arrange these a vertices into a tree with a specified degree sequence is $(a-2)!/\prod(d_i-1)!$ and there are a choices for the root, u, of the tree. So, the number of choices for this step is:

$$\frac{a(a-2)!}{\prod(d_i-1)!}.$$

Next we choose the vertices of T. Then for each edge of T, we choose a vertex-copy of each of its endpoints. To do so, for each v_i, we choose a copy of v_i for each of the d_i edges in T incident with v_i. If $\deg_{H_B}(v_i) = j$, then there are $j!/(j-d_i)!$ choices for the d_i copies of v_i. Since $\deg_{H_B}(v_i) \geq d_i$, the number of choices corresponding to v_i is at most $\sum_{w:\deg_{H_B}(w)\geq d_i} \deg_{H_B}(w)!/(\deg_{H_B}(w)-d_i)!$. By Lemma 28, this number is at most $(Y(d_i)+\frac{1}{2}\delta)n$ where

$$Y(d) = Y_B(d) = \sum_{j\geq d} \frac{j!}{(j-d)!}\rho_j(B).$$

Furthermore, if $d_i \leq \deg_{H_B}(v) \leq d_i + r - 2$, then we can use $Y'(d_i)$ rather than $Y(d_i)$ where

$$Y'(d) = Y'_B(d) = \sum_{j=d}^{d+r-2} \frac{j!}{(j-d)!}\rho_j(B).$$

Using $Y'(d_i)$ instead of $Y(d_i)$ will be particularly useful when $d_i \leq 2$. By Lemma 41, for any $\delta > 0$ we can take $B = B(\delta) > 0$ sufficiently large, so that we must use $Y(d_i)$ for at most $\delta a + 3$ vertices v_i. For convenience, we will assume $a > 3/\delta$ so we can take $\delta a + 3 \leq 2\delta a$.

We will upper bound $\mathbb{E}(X_a)$ by using $Y(d_i)$ for every vertex v_i with $d_i \geq 3$ and for exactly $2\delta a$ vertices of degree $d \leq 2$. Let t_1, t_2, t_3 denote the number of vertices v_i for which $d_i = 1, d_i = 2, d_i \geq 3$, respectively. We note that for sufficiently large d, $Y(d)$ is decreasing and so there is a constant d^* such that for all $d \geq 3$, $Y(d) \leq Y(d^*)$. So, if we were to use $Y'(d)$ for every vertex of degree $d \leq 2$ then the overall contribution of the Y, Y' terms would be at most:

$$[(Y'(1)+\frac{1}{2}\delta)n]^{t_1} \cdot [(Y'(2)+\frac{1}{2}\delta)n]^{t_2} \cdot [(Y(d^*)+\frac{1}{2}\delta)n]^{t_3}.$$

We correct for the $2\delta a$ vertices of degree $d \leq 2$ for which we use $Y(d)$. To do so, we multiply by the $\binom{t_1+t_2}{2\delta a} \leq \binom{a}{2\delta a}$ choices for those vertices, and we multiply by $\Upsilon^{2\delta a}$ where, for δ sufficiently small,

$$\Upsilon = \max\left(\frac{Y(1)+\frac{1}{2}\delta}{Y'(1)+\frac{1}{2}\delta}, \frac{Y(2)+\frac{1}{2}\delta}{Y'(2)+\frac{1}{2}\delta}\right) = O(1) \ .$$

This brings the overall contribution of the Y, Y' terms to at most:

$$\binom{a}{2\delta a}\Upsilon^{2\delta a}[(Y(1)+\frac{1}{2}\delta)n]^{t_1}[(Y(2)+\frac{1}{2}\delta)n]^{t_2}[(Y(d^*)+\frac{1}{2}\delta)n]^{t_3}.$$

Having chosen $d_1,\ldots,d_a$ and the vertices $v_1,\ldots,v_a$, we divide by the number of rearrangements of those vertices; i.e. we multiply by

$$\frac{1}{a!}.$$

Finally, we multiply by the probability that each of the $a-1$ pairs of vertex-copies corresponding to edges of T, lands in a hyperedge of the configuration. By Proposition 38, no two such pairs lie in the same hyperedge of H_B. So, we can apply Lemma 24 to the $a-1$ specified pairs of vertex-copies and multiply by

$$\left(\frac{k-1}{kE}\right)^{a-1} e^{ka^2/(E-a)}$$

to get an overall bound, where E is the number of edges in H_B.

Recall that for $c > c^*_{k,r}$, $\mu = \mu(c)$ denotes the larger of the two solutions of $f(\lambda) = c$. By Proposition 31 and Lemma 27 for any $\delta > 0$, we can take B sufficiently large so that

$$\left| kE - \mu \sum_{j\geq r-1} \frac{e^{-\mu}\mu^j}{j!} n \right| \leq \delta n.$$

Our key Lemma 32 now yields that by taking B sufficiently large, we can have δ sufficiently small in terms of ζ that various bounds below hold, including

$$\left(\frac{e^{-\mu}\mu^r}{(r-2)!}+\delta\right)\frac{k-1}{kE/n} < 1-\frac{\zeta}{2}. \tag{4}$$

By Lemma 31, for any $\delta > 0$, we can take B sufficiently large so that $Y'(1) \leq \delta/2$ and $Y'(2) \leq \frac{e^{-\mu}\mu^r}{(r-2)!} + \delta/2$. So, $Y'(1)+\frac{1}{2}\delta, Y'(2)+\frac{1}{2}\delta$ are bounded above by δ and $\frac{e^{-\mu}\mu^r}{(r-2)!}+\delta$, respectively. We let

$$\Psi = 2Y(d^*) > Y(d^*) + \frac{1}{2}\delta,$$

for δ sufficiently small.

Putting all this together, and recalling that $t_1+t_2+t_3 = a$, yields

$$\begin{aligned} E(X_a) &\leq \binom{a}{2\delta a}\Upsilon^{2\delta a}\left(\frac{k-1}{kE}\right)^{a-1} e^{ka^2/(E-a)} \\ &\quad\times \sum_{d_1+\cdots+d_a=2a-2} (\delta n)^{t_1}\left[\left(\frac{e^{-\mu}\mu^r}{(r-2)!}+\delta\right)n\right]^{t_2}(\Psi n)^{t_3}\frac{a(a-2)!}{a!\prod_{i=1}^a (d_i-1)!} \qquad (5) \\ &\leq O(n/a)e^{ka^2/(E-a)}\left(\frac{\Upsilon^{2\delta}}{(2\delta)^{2\delta}(1-2\delta)^{1-2\delta}}\right)^a\left(\frac{k-1}{kE/n}\right)^a \\ &\quad\times \sum_{d_1+\cdots+d_a=2a-2} \delta^{t_1}\left(\frac{e^{-\mu}\mu^r}{(r-2)!}+\delta\right)^{t_2}\Psi^{t_3}. \end{aligned}$$

Note that in the last line, we dropped the $\prod_{i=1}^{a}(d_i-1)!$ term. We can afford to do so, since this is equal to 1 for $d_i = 1$ or 2, which are the most sensitive values.

For δ sufficiently small in terms of ζ,

$$\frac{\Upsilon^{2\delta}}{(2\delta)^{2\delta}(1-2\delta)^{1-2\delta}} < 1+\frac{\zeta}{10}.$$

Since we are dealing with the degree sequence of a tree, we have $t_1 > t_3$. Since $\delta < 1$, we have $\sqrt{\delta}^{t_1} < \sqrt{\delta}^{t_3}$, yielding:

$$E(X_a) < O(n/a)e^{ka^2/(E-a)}\left(1+\frac{\zeta}{10}\right)^a \times \sum_{d_1+\cdots+d_a=2a-2}\left(\sqrt{\delta}\frac{k-1}{kE/n}\right)^{t_1}\left(\left[\frac{e^{-\mu}\mu^r}{(r-2)!}+\delta\right]\frac{k-1}{kE/n}\right)^{t_2}\left(\sqrt{\delta}\Psi\frac{k-1}{kE/n}\right)^{t_3}.$$

Recalling that $E/n = \Omega(1)$ and $\Psi = O(1)$, we choose δ sufficiently small in terms of ζ so that

$$\sqrt{\delta}\frac{k-1}{kE/n},\ \sqrt{\delta}\Psi\frac{k-1}{kE/n} < \frac{\zeta}{100}.$$

This and (4) yield

$$E(X_a) \le O(n/a)e^{ka^2/(E-a)}\left(1+\frac{\zeta}{10}\right)^a \sum_{d_1+\cdots+d_a=2a-2}\left(1-\frac{\zeta}{2}\right)^{t_2}\left(\frac{\zeta}{100}\right)^{a-t_2}.$$

Now we fix t_2 and count the number of choices for $d_1,\ldots,d_a$. There are $\binom{a}{t_2}$ choices for the values of i with $d_i = 2$. The remaining $a-t_2$ degrees sum to $2a-2-2t_2$. The number of choices for sequences of y non-negative integers that sum to z is $\binom{y+z-1}{y-1}$, so the number of choices for these degrees is bounded by $\binom{2(a-t_2)-3}{a-t_2-1} < 2^{2(a-t_2)-3} < 4^{a-t_2}$. Thus,

$$\begin{aligned}
E(X_a) &\le O(n/a)e^{ka^2/(E-a)}\left(1+\frac{\zeta}{10}\right)^a \sum_{t_2=0}^{a}\binom{a}{t_2}4^{a-t_2}\left(1-\frac{\zeta}{2}\right)^{t_2}\left(\frac{\zeta}{100}\right)^{a-t_2}\\
&= O(n/a)e^{ka^2/(E-a)}\left(1+\frac{\zeta}{10}\right)^a\left(1-\frac{\zeta}{2}+\frac{\zeta}{25}\right)^a\\
&< O(n/a)e^{ka^2/(E-a)}\left(1-\frac{\zeta}{4}\right)^a\\
&< O(n/a)\left(1-\frac{\zeta}{16}\right)^a, \qquad (6)
\end{aligned}$$

where the last inequality holds for all a small enough that $e^{ka/(E-a)} < 1+\frac{\zeta}{4}$. Thus, there are constants $Q,\xi > 0$ such that $\mathbb{E}(\sum_{a=Q\log n}^{\xi n} X_a) = o(1)$ and, therefore, w.h.p. there are no deletion trees of size between $Q\log n$ and ξn. Note now that Q,ξ depend only on ζ, c, k, r and ζ depends only on c, k, r.

Using Proposition 31(a), we chose B large enough that w.h.p. H_B contains fewer than ξn vertices outside of the r-core. Since a deletion tree can have at most one vertex in the r-core, this implies that there are no deletion trees of size at least ξn. Therefore, w.h.p. there

are no deletion trees in H_B of size greater than $Q \log n$. Therefore, w.h.p. for all $u \in D_B$, $|R^+(u) \cap H_B| \leq Q \log n$. ∎

7.2. Summing Over a Core Flippable Cycle for $r = 2$

Recall that for Theorem 5(b), we have $r = 2$; i.e., we consider 2-cores for random k-uniform hypergraphs where $k \geq 3$.

Consider any core flippable cycle C with vertices $u_1, \ldots, u_\ell$. In our directed graph D, add edges from u_j to u_{j+1} for each j (addition mod ℓ). Thus, $R^+(u_1) = \cup_{j=1}^{\ell} R^+(u_j)$. We modify the arguments from the proof of part (a) for this setting.

We define T as in the previous section, this time rooted at u_1.

We follow the proof of Lemma 41. Since $u_1, \ldots, u_\ell$ are the only 2-core vertices in S, we still have $|S| \leq \theta n$. Since each u_i has degree 2 in the 2-core, it is easy to see that $\deg_{H_B}(u_i) = \deg_S(u_i) + 1$. The proof of Lemma 41 still holds, yielding

$$\deg_{H_B}(v) \leq \deg_T(v) + 1 \ , \text{ for all but at most } \delta|T| + 3 \text{ vertices } v \in T.$$

(In fact, this time we actually get $\delta|T| + 2$, but that is inconsequential.)

As in Section 7.1, we bound the expected number of such trees of size a; u_1 is the root and hence plays the role of u from Section 7.1. This time, T has the additional property that there is an edge in D from a vertex of T (i.e. u_ℓ) to u_1. To account for this additional property, we adjust (5) as follows: (i) multiply by the number of choices of one of the $a - 1$ other vertices to be u_ℓ; (ii) choose vertex-copies for the extra edge from u_ℓ to u_1; (iii) adjust the term $\left(\frac{k-1}{kE}\right)^{a-1} e^{ka^2/(E-a)}$ which, by Lemma 24, bounded the probability that the $a-1$ pairs of vertex-copies corresponding to edges of T each landed in a hyperedge of the configuration.

For (ii), we use $Y(d(u_j) + 1)$ instead of $Y(d(u_j))$ or $Y'(d(u_j))$ for $j = 1, \ell$. For $j = 1, \ell$, the adjustment for u_j is an increase of a multiplicative factor of at most $(Y(d(u_j) + 1) + \frac{1}{2}\delta)/(Y'(\deg(u_j)) + \frac{1}{2}\delta) < (Y(d^*) + \frac{1}{2}\delta)/(Y'(1) + \frac{1}{2}\delta) = O(1)$. So the overall effect for (ii) is a multiplicative $O(1)$.

For (iii), the hyperedge containing u_1, u_ℓ is in the 2-core and so is distinct from the other $a - 1$ hyperedges. This results in another multiplicative factor of $\frac{k-1}{kE}$ to account for that edge, when applying Lemma 24.

The net result is to multiply $\mathbb{E}(X_a)$ by $O(a/n)$, and so the bound on $\mathbb{E}(X_a)$ in (6) becomes $O(1)\left(1 - \frac{\zeta}{16}\right)^a$. Summing over all a yields that the expected number of core flippable cycles C such that $|\bigcup_{u \in C} R^+(u) \cap H_B| > \xi(n)$ is $o(1)$ for any $\xi(n) \to \infty$, in particular for $\xi(n) = O(\log n)$. Proposition 33 and Lemma 34 yield Theorem 5(b). ∎

8. PROOF OF THEOREM 5 BELOW THE r-CORE THRESHOLD

Recall that Theorem 5 is already known for $r = k = 2$, so we will assume $r + k > 4$. As in the case for $c > c^*_{k,r}$, we will carry out a large but fixed number, I, of rounds of the parallel r-stripping process, ending up with a hypergraph H_I. Because we are below the r-core threshold, this will delete all but a very small, albeit linear, number of vertices. Proposition 25 asserts that the remaining hypergraph is uniformly random conditional on its degree sequence. We will determine this degree sequence and apply the technique from [28] to show that the maximum component size in the remaining hypergraph has size $O(\log n)$. Thus, for every v, we must have $|R^+(v) \cap H_I| = O(\log n)$. Proposition 33 and Lemma 34 then imply that $|R^+(v)| = O(\log n)$ as required.

Let $\mathrm{Po}(\mu)$ denote a Poisson variable with mean μ. Recursively define the following quantities:

$$\begin{aligned}\phi_0 &= 1\\ \lambda_t &= c\phi_t^{k-1}/(k-1)!\\ \phi_t &= \Pr[\mathrm{Po}(\lambda_{t-1}) \geq r-1].\end{aligned}$$

Write $P(\mu, j) = \Pr[\mathrm{Po}(\mu) = j]$.

Lemma 42. *For any constants d, t, the number of vertices of degree d after t rounds of the parallel r-stripping process, w.h.p. is $\rho_t(d)n + o(n)$, where*

$$\rho_t(d) = \begin{cases} P(\lambda_t, d) & \text{for } d \geq r,\\ P(\lambda_t, d) \cdot \Pr\left[\mathrm{Po}(\lambda_{t-1} - \lambda_t) \geq r - d\right] & \text{for } d < r.\end{cases}$$

Proof. We consider a branching process introduced in [33] and analyze it as in [27]. Consider any hypergraph H and any vertex $v \in H$. For each $0 \leq i \leq t+1$, let $L_i(v)$ be the vertices of distance at most i from v (thus $L_0(v) = \{v\}$). For any $u \in L_i(v)$ with $0 \leq i \leq t$, a *child edge* of u is an edge containing u and $k-1$ members of $L_{i+1}(v)$; thus if the distance $t+1$ neighbourhood of v induces a hypertree, then all but at most one of the edges containing u are child edges of u.

We define the process STRIP(v, t) as follows:

For j from t down to 1 do

Remove all vertices in $L_j(v)$ with fewer than $r-1$ child edges;

Remove all edges that contain a removed vertex.

Let X_t denote the number of child edges of v that survive STRIP(v, t), and let Y_t denote the number of child edges of v that survive STRIP$(v, t-1)$ but not STRIP(v, t). If the hypergraph induced by the vertices in $L_{t+1}(v)$ induces a hypertree, then we see that

(A) For $d \geq r$: v survives the first t rounds of the parallel r-stripping process, and has degree d in what remains iff $X_t = d$.

(B) For $1 \leq d < r$: v survives the first t rounds of the parallel r-stripping process, and has degree d in what remains iff $X_t = d$ and $Y_t \geq r - d$.

To analyze STRIP(v, t) on $H = H_k(n, p = c/n^{k-1})$, we make use of the fact that w.h.p. the distance $t+1$ neighbourhood of v induces a hypertree, and so both (A) and (B) hold.

We will argue by induction on t that the probability a particular child u of v survives STRIP(v, t) is $\phi_t + o(1)$. Suppose $u \in L_1(v)$ and $w \in L_2(v)$ is in a child edge of u. Note that w survives STRIP(v, t) iff w survives STRIP$(u, t-1)$ the probability of which, by induction on t, is easily seen to be $\phi_{t-1} + o(1)$. It follows that the expected number of child edges of u that survive STRIP(v, t) is $\frac{c}{n^{k-1}}\binom{n-O(1)}{k-1}(\phi_{t-1} + o(1))^{k-1} = \lambda_{t-1} + o(1)$. Standard arguments show that for any fixed i the probability that the number of such edges is i is $P(\lambda_{t-1}, i) + o(1)$ (we elaborate more below on similar arguments for X_t, Y_t). Therefore, the probability that u survives STRIP(v, t) is $\phi_t + o(1)$, thus completing the induction. By the same argument, $\mathbb{E}(X_t) = \lambda_t + o(1)$. Noting that $Y_t = X_{t-1} - X_t$, this yields $\mathbb{E}(Y_t) = \lambda_{t-1} - \lambda_t + o(1)$.

Consider any child edge e of v in $L_{t+1}(v)$. Whether e counts towards X_t, Y_t or neither is determined entirely by the subtrees of $L_{t+1}(v)$ rooted at the vertices of e other than v. In other

words, X_t, Y_t are determined by the edges containing v in $H_k(n, p = c/n^{k-1})$, and some local information about each edge where the information for any two edges is w.h.p. disjoint. Also, no edge counts towards both X_t and Y_t. From this, it is straightforward to show, using e.g., the Method of Moments, (see [19]) that the joint distribution of X_t, Y_t is asymptotic to independent Poisson variables; specifically, for any fixed integers x, y, $\mathbf{Pr}(X_t = x \wedge Y_t = y)$ is $o(1)$ plus the probability that two independent Poisson variables with means $\mathbb{E}(X_t), \mathbb{E}(Y_t)$ are equal to x, y.

(A) and (B) now yield that the probability that v survives the first t rounds of the parallel stripping process and has degree d in H_t is $\rho_t(d) + o(1)$, and so the expected number of such vertices is $\rho_t(d)n + o(n)$. The lemma now follows as in [27] from a straightforward concentration argument, e.g., a second moment calculation. We omit the details. ∎

The main result of [28] states: Consider a random graph on a fixed degree sequence where $\Lambda(d) \cdot n + o(n)$ vertices have degree d, and where the degree sequence satisfies certain *well-behaved* conditions. If

$$\sum_{d \geq 1} d(2-d)\Lambda(d) > 0, \tag{7}$$

and then w.h.p. all connected components have size $O(\log n)$. A simple adaptation of the proof in [28] provides a generalization to hypergraphs. Specifically, for $k > 2$ it suffices to replace $d(2-d)$ in (7) with

$$f_k(d) = d[1 - (d-1)(k-1)].$$

Proposition 25 allows us to model H_t as a random hypergraph on degree sequence $\rho_0(t), \rho_1(t), \ldots$. Using Lemma 28, it is straightforward to verify that this degree sequence satisfies the well-behaved conditions from [28], and so deduce that if

$$\sum_{d \geq 1} \rho_t(d) f_k(d) > 0, \tag{8}$$

then w.h.p. all components of H_t have size $O(\log n)$.

Since $\Pr[\mathrm{Po}(\lambda) \geq r-1]$ is a strictly increasing function of λ, the sequences ϕ_t, λ_t are strictly decreasing. If they do not tend to zero, then there must be a positive fixed point to the recursion defining them, i.e., a positive solution to

$$\lambda = c \Pr[\mathrm{Po}(\lambda) \geq r-1]^{k-1}/(k-1)! \ .$$

Rearranging yields

$$c = (k-1)!\lambda / \Pr[\mathrm{Po}(\lambda) \geq r-1]^{k-1} \ .$$

Recall now that $c^*_{k,r}$ was defined in (1) as the smallest value of c for which there is such a solution. Since $c < c^*_{k,r}$, we can conclude that $\phi_t, \lambda_t \to 0$ as $t \to \infty$ and we can develop the following asymptotics in t, using $O_t()$ and $\Theta_t()$ to denote asymptotics are with respect to t:

$$\lambda_t = \frac{c}{(k-1)!}\phi_t^{k-1} = \frac{c}{(k-1)!}\left(\Pr[\mathrm{Po}(\lambda_{t-1}) \geq r-1]\right)^{k-1} = \Theta_t(\lambda_{t-1}^{(k-1)(r-1)}). \tag{9}$$

Let $\lambda := \lambda_t$ and $\theta := \lambda_{t-1} - \lambda_t$. Since $(k-1)(r-1) \geq 2$ for $k + r > 4$, we see that (9) implies $\lambda = o_t(\theta)$.

We apply Lemma 42, noting that as $\theta \to 0$, $\Pr[\mathrm{Po}(\theta) \geq r-d] \to P(\theta, r-d)$. Therefore, as $t \to \infty$, inequality (8) is equivalent to

$$(1 + o_t(1)) \sum_{d=1}^{r-1} P(\lambda, d) P(\theta, r-d) f_k(d) + \sum_{d=r}^{\infty} P(\lambda, d) f_k(d) > 0. \tag{10}$$

Note that $f_k(1) = 1$ and $f_k(d) \leq 0$ for $d \geq 2$. So the first sum in (10) is at least

$$P(\lambda, 1) P(\theta, r-1) - \sum_{d=2}^{r-1} P(\lambda, d) P(\theta, r-d) |f_k(d)|. \tag{11}$$

For $1 \leq d \leq r-1$, we have $f_k(d) = O_t(1)$, so the first term in (11) is $\Theta_t(\lambda\theta^{r-1})$ while the sum in (11) is $\Theta_t(\sum_{d=2}^{r-1} \lambda^d \theta^{r-d}) = \Theta_t(\lambda^2 \theta^{r-2})$, since $\lambda = o_t(\theta)$. Therefore, the first sum in (10) is positive and of order $\Theta_t(\lambda\theta^{r-1})$.

At the same time, since $-f_k(d) = kd^2 - d^2 - kd < kd^2$ we get

$$-\sum_{d=r}^{\infty} P(\lambda, d) f_k(d) \leq k \sum_{d=r}^{\infty} P(\lambda, d) d^2 = O_t(\lambda^r), \quad \text{as } \lambda \to 0.$$

Thus, the first sum in (10) is positive $\Theta_t(\lambda\theta^{r-1})$, whereas the second sum is $O_t(\lambda^r)$. Since $\lambda = o_t(\theta)$ it follows that (8) holds for t sufficiently large and, therefore, for I sufficiently large, every component of H_I has size $O(\log n)$. Theorem 5(b) now follows from Proposition 33 and Lemma 34. ∎

9. PROOF OF THEOREM 2

Given a solution, recall that a set S of variables is *flippable* if changing the assignment of every variable in S results in another solution. Note that flippable sets can be characterized in terms of the underlying hypergraph.

Proposition 43. *S is flippable iff every hyperedge contains an even number of members of S.*

So we define:

Definition 44. A *flippable set* in a hypergraph, H, is a nonempty set of vertices, S, such that every edge in H contains an even number of vertices of S.

Recalling Definition 4, we see that a flippable cycle is a flippable set. A flippable set is *minimal* if it does not contain a flippable proper subset. Note that every flippable set contains a minimal flippable subset.

Lemma 45. *For every $c > c^*_{k,2}$ there exists $\alpha > 0$ such that w.h.p. every minimal flippable set in the hypergraph induced by the 2-core of $H_k(n, p = c/n^{k-1})$ either is a core flippable cycle or has size at least αn.*

Lemma 45 follows immediately from Lemma 51 below, and yields Theorem 2 as follows:

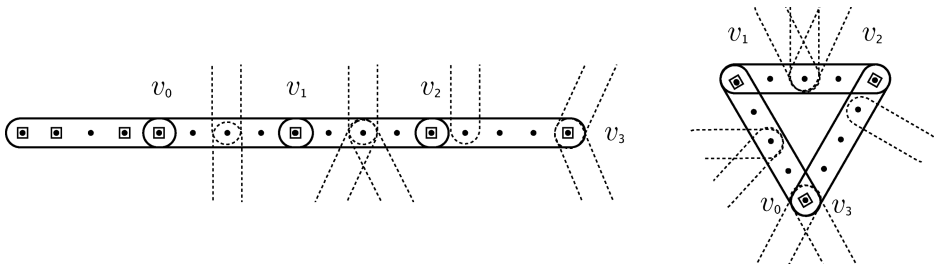

Fig. 1. 2-linked paths with $t = 3$. On the left $v_0 \neq v_3$, while on the right $v_0 = v_3$. Vertices in S are marked with a square.

Proof of Theorem 2. Consider any two solutions σ_1, σ_2 in different clusters. Let S be the variables in the 2-core on which these solutions disagree. Thus, S is a flippable set in the hypergraph induced by the 2-core. Remove all core flippable cycles from S, and let S' be what remains (recall from Definition 4 that a flippable cycle is a set of vertices). Note that S' must also be a flippable set in the hypergraph induced by the 2-core. By the definition of clusters, $S' \neq \emptyset$ as otherwise σ_1, σ_2 would be cycle-equivalent. Let S'' be a mimimal flippable subset of S'. Since S'' contains no core flippable cycles, Lemma 45 implies that $|S''| \geq \alpha n$. Therefore $|S| \geq \alpha n$ and so σ_1, σ_2 differ on at least αn variables.

Any sequence $\sigma, \sigma', \ldots, \tau$ where σ, τ are in different clusters must contain two consecutive solutions that are in different clusters. As argued above, those two solutions differ on at least αn variables. It follows that if σ, τ are in different clusters then σ, t are not αn-connected. ∎

If we could show (deterministically) that the hypergraph induced by any minimal flippable set in a 2-core that is not a core flippable cycle is sufficiently dense, then Lemma 45 would follow by a rather standard argument. Unfortunately, there is no useful lower bound on the density, mainly because of the possibility of very long 2-linked paths in S (defined below). Instead, we follow an approach akin to that of [29], forming a graph $\Gamma(S)$ by contracting those long paths, and making use of the fact that $\Gamma(S)$ is dense (Lemma 50). The main difference from [29] is that here we need to work in the configuration model.

To prove Lemma 45, we first require a few definitions. Note that these concern any hypergraph, not just a 2-core of a random hypergraph.

A hyperedge is *simple* if it is not a loop, i.e., if it does not contain any vertex more than once.

Definition 46 (See Figure 1)**.** Let $\mathcal{H}$ be a k-uniform hypergraph. A *2-linked path* P of a set $S \subseteq V(\mathcal{H})$ is a set of vertices $v_0, \ldots, v_t \in S$ and simple hyperedges $e_1, \ldots, e_t$, where $t \geq 1$, such that

(i) $v_0, \ldots, v_t$ are all distinct except that when $t \geq 2$ we allow $v_0 = v_t$. (Note that if $v_0 = v_t$ then these vertices actually form a cycle and so *2-linked path* is somewhat of a misnomer.)
(ii) Each e_i contains v_{i-1}, v_i and no other vertices of S.
(iii) $v_1, \ldots, v_{t-1}$ all have degree 2 in $\mathcal{H}$; i.e. they do not lie in any edges outside of P.
(iv) If $v_0 = v_t$ then $\deg_{\mathcal{H}}(v_0) > 2$. If $v_0 \neq v_t$ then S is maximal w.r.t. (ii) and (iii); i.e. each of v_0, v_t either has degree $\neq 2$ in $\mathcal{H}$, or lies in a hyperedge $e \notin P$ with $|e \cap S| \neq 2$.

We call v_0, v_t the *endpoints* of the path and $v_1, \ldots, v_{t-1}$ its *connecting vertices*.

Note that if $v_0 = v_t$ then by (iv), $\deg_{\mathcal{H}}(v_0) > 2$ and hence $v_0, \ldots, v_t$ do not form a flippable cycle.

Definition 47. We say that $S \subseteq V(\mathcal{H})$ is a *linked set* if (i) S does not contain a flippable cycle as a subset, (ii) no hyperedge of $\mathcal{H}$ contains exactly one element of S and (iii) every hyperedge e of $\mathcal{H}$ with $|e \cap S| = 2$ is in a 2-linked path of S.

Proposition 48. *Suppose S is a flippable set in a hypergraph where all hyperedges are simple, and S does not contain a flippable cycle as a subset. Then S is a linked set.*

Proof. By Proposition 43, we only need to check condition (iii) of Definition 47. Consider any hyperedge e with $|e \cap S| = 2$. Since e is simple, either e itself forms a 2-linked path in S, or it is easily seen that e can be extended into such a path, unless e lies in a flippable cycle. ∎

Remark. It is easy to see that in any Uniquely Extendible CSP, the set of disagreeing variables of any two solutions must be a flippable set. Since Proposition 48 was derived by only considering the underlying hypergraph (and not the specific constraints), it applies to any UE CSP. Therefore, our Theorem 2 extends readily to every UE CSP since its proof amounts to proving that for some constant $\alpha > 0$, all linked sets are either flippable cycles or contain at least αn variables.

Given a linked set, S, we consider the mixed hypergraph (containing both hyperedges and normal edges) $\Gamma(S)$ formed as follows:

(a) The vertices of $\Gamma(S)$ are the endpoints of the 2-linked paths in S along with all vertices of S that do not lie in any 2-linked paths.
(b) There is an edge in $\Gamma(S)$ between the endpoints of each 2-linked path in S. That edge is a loop if the two endpoints are the same vertex, and so $\Gamma(S)$ is not necessarily simple.
(c) For every hyperedge e of $\mathcal{H}$ with $|e \cap S| > 2$, $e \cap S$ is a hyperedge of $\Gamma(S)$.

Thus $V(\Gamma(S)) \subseteq S$, and since no hyperedge of C contains exactly one element of S, for every $v \in V(\Gamma(S))$ we have $\deg_{\Gamma(S)}(v) = \deg_{\mathcal{H}}(v)$. Any vertex of S that is not in $\Gamma(S)$ is a connecting vertex of a 2-linked path in S.

Proposition 49. *If S is a non-empty linked set, then* $1 \leq |\Gamma(S)| \leq |S|$.

Proof. Any vertex of S that is not in $\Gamma(S)$ is a connecting vertex of a 2-linked path in S. The endpoints of that 2-linked path are in $\Gamma(S)$. Thus $|\Gamma(S)| \geq 1$. The rest follows from the fact that every vertex of $\Gamma(S)$ is a vertex of S. ∎

Note that $\Gamma(S)$ contains hyperedges of size between 2 and k. For each $2 \leq i \leq k$, we define ℓ_i to be the number of i-edges in $\Gamma(S)$.

Lemma 50. *If every vertex in $\mathcal{H}$ has degree at least 2 then* $\sum_{i=2}^{k}(i-1)\ell_i \geq (1 + \frac{1}{2k})|V(\Gamma(S))|$.

Proof. As we said above, every $v \in V(\Gamma(S))$ has the same degree in $\Gamma(S)$ as it does in $\mathcal{H}$. Thus $\Gamma(S)$ has minimum degree at least 2. Consider any v of degree 2 in $\Gamma(S)$. Then v has degree 2 in $\mathcal{H}$ and hence cannot be the endpoint of a 2-linked path in S, unless v lies in at least one hyperedge of $\mathcal{H}$ containing more than 2 members of S. It follows that v lies in at least one hyperedge of $\Gamma(S)$ of size greater than 2. Therefore, at most $\sum_{i=3}^{k} i\ell_i < k\sum_{i=3}^{k} \ell_i$ vertices of $\Gamma(S)$ have degree 2, and so letting Z denote the number of vertices with degree at least 3 in $\Gamma(S)$, we have

$$|V(\Gamma(S))| \le Z + k\sum_{i=3}^{k} \ell_i \le k\left(Z + \sum_{i=3}^{k} \ell_i\right).$$

By the handshaking lemma, $\sum_{i=2}^{k} i\ell_i = \sum_v \deg_{\Gamma(S)}(v)$. Therefore,

$$\begin{aligned}
\sum_{i=2}^{k}(i-1)\ell_i &= \frac{1}{2}\sum_v \deg_{\Gamma(S)}(v) + \sum_{i=2}^{k}(i/2-1)\ell_i \\
&\ge \frac{1}{2}\sum_v \deg_{\Gamma(S)}(v) + \frac{1}{2}\sum_{i=3}^{k}\ell_i \\
&= \sum_v 1 + \sum_v \frac{1}{2}(\deg_{\Gamma(S)}(v) - 2) + \frac{1}{2}\sum_{i=3}^{k}\ell_i \\
&\ge |V(\Gamma(S))| + \frac{1}{2}Z + \frac{1}{2}\sum_{i=3}^{k}\ell_i, \quad \text{since } \deg_{\Gamma(S)(v)} \ge 2 \text{ for all } v \\
&\ge \left(1 + \frac{1}{2k}\right)|V(\Gamma(S))|.
\end{aligned}$$

∎

Let C be the 2-core of $H = H_k(n,p)$. We will apply Lemma 50 with $\mathcal{H} = C$ to prove:

Lemma 51. *There exists $\alpha > 0$ such that w.h.p. C has no non-empty linked set of size less than αn.*

Lemma 45 follows immediately from Lemma 51 and Proposition 48 (since $H_k(n,p)$ contains only simple hyperedges). The proof of Lemma 51 will be reminiscent of the proof of Lemma 40, but significantly more complicated because (i) we are working in the configuration model and (ii) where we had ℓ_2 2-edges in Lemma 51, we have ℓ_2 2-linked paths here. First, we provide a technical lemma.

Lemma 52. *For any integers a, t, given a set of a vertices in $H = H_k(n,p)$, with $p = c/n^{k-1}$ the probability that their total degree exceeds $tkca$ is at most $(e/t)^{act}$.*

Proof. Given a set A of a vertices, let E_A denote the number of hyperedges containing at least one member of A. The total degree in A is at most kE_A. The number of potential edges in E_A is at most $a\binom{n}{k-1} < an^{k-1}$, and so E_A is dominated from above by $\mathrm{Bin}(an^{k-1}, c/n^{k-1})$ and using $\binom{n}{z} \le (ne/z)^z$ we get

$$\mathbf{Pr}\left[\mathrm{Bin}(an^{k-1}, c/n^{k-1}) > act\right] < \binom{an^{k-1}}{act}\left(\frac{c}{n^{k-1}}\right)^{act} < (e/t)^{act}.$$

∎

Proof of Lemma 51. By Corollary 23, we can work in the configuration model. Let $\mathcal{D}$ be the degree sequence of C. Recalling Definition 29, Proposition 30 and our key Lemma 32, we have w.h.p.

(i) $\mathcal{D}$ has total degree $\gamma n + o(n)$, where $\gamma = \mu\Psi_r(\mu)$,
(ii) $\mathcal{D}$ has $\lambda_2 n + o(n)$ vertices of degree 2, where $\lambda_2 = e^{-\mu}\mu^2/2$,
(iii) there exists $\zeta > 0$ such that $2(k-1)\lambda_2 < (1-\zeta)\gamma$.

For each $a \geq 1$, let X_a denote the number of linked sets S in C for which $|\Gamma(S)| = a$ and let $X = \sum_{a=1}^{\alpha n} X_a$. Define

$$\mathcal{L}_a = \left\{(\ell_2, \ldots, \ell_k) : \left(1 + \frac{1}{2k}\right) a \leq \sum_{i=2}^{k} (i-1)\ell_i \leq \left(1 + \frac{1}{2k}\right) a + (k-1)\right\}.$$

By Lemma 50, for any linked set S in C with $|\Gamma(S)| = a$, there is some $(\ell_2, \ldots, \ell_k) \in \mathcal{L}_a$ so that $\Gamma(S)$ contains *at least* ℓ_i i-edges for each i.

To bound $E(X_a)$, we begin by choosing a vertices, $A \subseteq V(C)$ and sum over all $t \geq 0$ of the probability that their total degree in C lies in the range $(tkca, (t+1)kca]$. For each t, we upper bound this last probability by the probability that their total degree in H lies in $(tkca, \infty]$. Moreover, to sum over all subsets $A \subseteq V(C)$ we overcount by summing instead over all $A \subseteq V(H)$, and using Lemma 52. Of course, if such a set is not a subset of C then the probability of it contributing to X_a is zero, and so this provides an upperbound on $\mathbb{E}(X_a)$. This yields:

$$\binom{n}{a} \sum_{t \geq 0} \left(\frac{e}{t}\right)^{tca}.$$

Given A, we sum over all possibilities for the values of $(\ell_2, \ldots, \ell_k) \in \mathcal{L}_a$. For each $2 \leq i \leq k$, we choose ℓ_i i-sets of vertex-copies belonging to vertices of A. If the total degree of A is in $(tkca, (t+1)kca]$ then the number of choices for these ℓ_i i-sets is at most

$$\left(\frac{((t+1)kca)^i}{i!}\right)^{\ell_i} / \ell_i! < \frac{((t+1)kca)^{i\ell_i}}{\ell_i!}.$$

Denote the ℓ_2 2-sets as $\{u_1, w_1\}, \ldots, \{u_{\ell_2}, w_{\ell_2}\}$. For each $i = 1, \ldots, \ell_2$, we select $j_i \geq 0$, the number of connecting variables in the 2-linked path from u_i to w_i in S, we choose the j_i degree two connecting variables for that path, and we choose one of the two possible orientations of the vertex-copies of each of those connecting variables. Let $J = j_1 + \cdots + j_{\ell_2}$, be the number of connecting variables selected. Let $L = \lambda_2 n + o(n)$ be the number of degree 2 vertices in C. Then the total number of choices for the connecting vertices and the orientations of their copies is at most

$$\prod_{i=1}^{J} 2(L - i + 1).$$

Next, we apply Lemma 24 to bound the probability that the $\ell_3 + \cdots + \ell_k$ sets of size at least 3 all land in hyperedges of the configuration and that for each $i = 1, \ldots, \ell_2$, the *first* pair in the 2-linked path, i.e., u_i and the first copy of the first of the j_i connecting variables,

lands in a hyperedge of the configuration. Note that $\ell_2 + \cdots + \ell_k \leq \sum_{i=2}^{k}(i-1)\ell_i < 2a + k - 1 < 2a + o(n)$, by the definition of $\mathcal{L}_a$. By assuming $a < \alpha n$ for some sufficiently small α, we get $\gamma n + o(n) - 2a > \frac{1}{2}\gamma n$. Therefore, Lemma 24 yields that this probability is at most

$$\exp\left(\frac{k(\ell_2 + \cdots \ell_k)^2}{\frac{1}{2}\gamma n}\right)\prod_{i=2}^{k}\left(\frac{(k-1)(k-2)\cdots(k-i+1)}{(\gamma n + o(n))^{i-1}}\right)^{\ell_i}$$
$$< \exp\left(\frac{8ka^2}{\gamma n}\right)\prod_{i=2}^{k}\left(\frac{k}{\gamma n}\right)^{(i-1)\ell_i}.$$

Following the analysis of Lemma 24, we have now exposed $\ell_2 + \cdots + \ell_k$ hyperedges of the configuration. Let Λ be the number of unmatched vertex-copies remaining. Since $\ell_2 + \cdots + \ell_k < 2a + k - 1$, we have $\Lambda \geq \gamma n - 2ka + o(n)$. If the other vertex-copies required for the 2-linked paths are still unmatched, then we continue; else we halt observing that in this case, the set of choices made so far cannot lead to a linked set on the chosen vertices.

There are J pairs of vertex copies that each need to be in a hyperedge of the configuration in order to complete the 2-linked paths. Following the same argument as in Lemma 35, the probability of this happening is at most

$$\prod_{i=1}^{J}\frac{k-1}{\Lambda - k(i-1)}.$$

Applying (iii) above, and taking $a < \alpha n$ for α sufficiently small in terms of γ, λ_2, we obtain:

$$\frac{2(k-1)L}{\Lambda} < \frac{2(k-1)\lambda_2 n + o(n)}{\gamma n - 2ka + o(n)} < 1 - \frac{\zeta}{2}.$$

Thus, since $2(k-1)L \leq \Lambda$ (by the previous line) and $k \leq 2(k-1)$, we have $\frac{2(k-1)(L-(i-1))}{\Lambda - k(i-1)} < 1 - \frac{\zeta}{2}$ for each i, leading to

$$\mathbb{E}(X_a) < \binom{n}{a}\sum_{t\geq 0}(\frac{e}{t})^{tca}\sum_{\ell_2,\ldots,\ell_k\in\mathcal{L}_a}\sum_{j_1,\ldots,j_{\ell_2}\geq 0} e^{8ka^2/(\gamma n)}\left(\prod_{i=2}^{k}\frac{((t+1)kca)^{i\ell_i}}{\ell_i!}\right)$$
$$\times\left(\prod_{i=2}^{k}\left(\frac{k}{\gamma n}\right)^{(i-1)\ell_i}\right)\left(\prod_{i=1}^{J}\frac{2(k-1)(L-(i-1))}{\Lambda - k(i-1)}\right)$$
$$< \left(\frac{en}{a}\right)^{a}\sum_{t\geq 0}(\frac{e}{t})^{tca}\sum_{\ell_2,\ldots,\ell_k\in\mathcal{L}_a} e^{8ka^2/(\gamma n)}\left(\prod_{i=2}^{k}\frac{(kca)^{\ell_i}}{\ell_i!}\left(\frac{k^2 ca}{\gamma n}\right)^{(i-1)\ell_i}(t+1)^{i\ell_i}\right)$$
$$\times\sum_{j_1,\ldots,j_{\ell_2}\geq 0}(1-\zeta/2)^{J}.$$

Since $J = j_1 + \cdots + j_{\ell_2}$, we have $\sum_{j_1,\ldots,j_{\ell_2}\geq 0}(1-\zeta/2)^J = \left(\sum_{j\geq 0}(1-\zeta/2)^j\right)^{\ell_2} = (2/\zeta)^{\ell_2}$, yielding:

$$\mathbb{E}(X_a) < \left(\frac{en}{a}\right)^a e^{8ka^2/\gamma n} \sum_{\ell_2,\ldots,\ell_k\in\mathcal{L}_a} \left(\frac{k^2ca}{\gamma n}\right)^{\sum_{i=2}^k (i-1)\ell_i} \times \left(\prod_{i=2}^k \frac{(kca)^{\ell_i}}{\ell_i!}\right)\left(\frac{2}{\zeta}\right)^{\ell_2} \sum_{t\geq 0}\left(\frac{e}{t}\right)^{tca}(t+1)^{\sum_{i=2}^k i\ell_i}.$$

By our choice of $\mathcal{L}_a$

$$\ell_2 \leq \sum_{i=2}^k (i-1)\ell_i \leq \left(1+\frac{1}{2k}\right)a + k - 1,$$
$$\sum_{i=2}^k i\ell_i \leq 2\sum_{i=2}^k (i-1)\ell_i \leq 3a + 2k.$$

Thus, we obtain $(2/\zeta)^{\ell_2} < Z_1^a$ for constant $Z_1 = Z_1(c)$ and, since $(e/t)^{tc}(t+1)^{3+2k}$ is decreasing for large t, we have

$$\sum_{t\geq 0}(e/t)^{tca}(t+1)^{\sum_{i=2}^k i\ell_i} < \sum_{t\geq 0}(e/t)^{tca}(t+1)^{3a+2k} < \sum_{t\geq 0}\left((e/t)^{tc}(t+1)^{3+2k}\right)^a < Z_2^a,$$

for constant $Z_2 = Z_2(c)$. Also using $a \leq n$ we obtain:

$$\mathbb{E}(X_a) < \left(\frac{en}{a}\right)^a e^{8ka/\gamma}(Z_1Z_2)^a \left(\frac{k^2ca}{\gamma n}\right)^{\left(1+\frac{1}{2k}\right)a+k-1} \sum_{\ell_2,\ldots,\ell_k\geq 0}\prod_{i=2}^k \frac{(kca)^{\ell_i}}{\ell_i!}$$
$$< O(1)\left(eZ_1Z_2e^{8k/\gamma}\left(\frac{k^2c}{\gamma}\right)^{1+\frac{1}{2k}}\right)^a \left(\frac{a}{n}\right)^{a/2k}\left(\sum_{\ell\geq 0}\frac{(kca)^\ell}{\ell!}\right)^{k-1}.$$

Applying $(\sum_{\ell\geq 0}\frac{(kca)^\ell}{\ell!})^{k-1} = e^{kca(k-1)}$ we obtain:

$$\mathbb{E}(X_a) < O(1)\left(eZ_1Z_2e^{8k/\gamma}(k^2c/\gamma)^{1+\frac{1}{2k}}e^{ck(k-1)}\right)^a \left(\frac{a}{n}\right)^{a/2k} < Y^a\left(\frac{a}{n}\right)^{a/2k},$$

for a constant $Y = Y(\gamma,\lambda_2,\zeta,b,\xi)$ that does not depend on a, so long as $a < \alpha n$ for sufficiently small $\alpha > 0$. This yields $\mathbb{E}(\sum_{a=1}^{\sqrt{n}} X_a) = o(1)$. Moreover, for all α sufficiently small, $\mathbb{E}(X_a) < 2^{-a}$. Therefore, $\mathbb{E}(\sum_{a\geq\sqrt{n}} X_a) = o(1)$ and, thus, $\mathbb{E}(X) = o(1)$.

Therefore, w.h.p. there is no 2-linked set S with $1 \leq |\Gamma(S)| \leq \alpha n$. The lemma follows from Proposition 49. ∎

ACKNOWLEDGMENTS

We thank two anonymous referees for their many helpful comments.

REFERENCES

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

[2] D. Achlioptas and A. Coja-Oghlan, Algorithmic barriers from phase transitions, In Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS '08), 2008, pp. 793–802.

[3] D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi, On the solution-space geometry of random constraint satisfaction problems, Random Struct Algorithm 38 (2011), 251–268.

[4] D. Achlioptas and F. Ricci-Tersenghi, Random formulas have frozen variables, SIAM J Comput 39 (2009), 260–280.

[5] E. R. Berlekamp, R. J. McEliece, and H. C. A. van Tilborg, On the inherent intractability of certain coding problems, IEEE Trans Inform Theory 24 (1978), 384–386.

[6] B. Bollobás, A probabilistic proof of an asymptotic formula for the number of labelled regular graphs, Eur J Combin 1 (1980), 311–316.

[7] A. Coja-Oghlan, On belief propagation guided decimation for random k-SAT, In Proceedings of the 22nd SODA, 2011, pp. 957–966.

[8] A. Coja-Oghlan and C. Efthymiou, On independent sets in random graphs, In Proceedings of the 22nd SODA, 2011, pp. 136–144.

[9] V. Chvátal and E. Szemerédi, Many hard examples for resolution, J ACM 35 (1988), 759–768.

[10] H. Connamacher and M. Molloy, The exact satisfiability threshold for a potentially intractable random constraint satisfaction problem, SIAM J Disc Math (2004). (in press): Preliminary version in Proceedings of the 45th FOCS.

[11] C. Cooper, The cores of random hypergraphs with a given degree sequence, Random Struct Algorithms 25 (2004), 353–375.

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

[13] H. Daudé, M. Mézard, T. Mora, and R. Zecchina, Pairs of SAT Assignments and Clustering in Random Boolean Formulae, Theor Comput Sci 393 (2008), 260–279.

[14] O. Dubois and J. Mandler, The 3-XORSAT threshold, In Proceedings of the 43rd FOCS, 2002, pp. 769.

[15] D. Fernholz and V. Ramachandran, Cores and connectivity in sparse random graphs, The University of Texas at Austin, Department of Computer Sciences, Technical report TR 04–13, 2004.

[16] M. Guidetti and A. P. Young, Complexity of several constraint-satisfaction problems using the heuristic classical algorithm WalkSAT, Phys Rev E 84 (2011), 011102.

[17] M. Ibrahimi, Y. Kanoria, M. Kraning, and A. Montanari, The set of solutions of random XORSAT formulae, In Proceedings of SODA, 2012.

[18] S. Janson and M. Luczak, A simple solution to the k-core problem, Random Struct Algorithms 30 (2007), 50–62.

[19] S. Janson, T. Łuczak, and A. Ruciński, Random graphs, Wiley, New York, 2000.

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

[21] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. Spielman, Analysis of low density codes and improved designs using irregular graphs, In Proceedings of STOC, 1998.

[22] T. Łuczak, Component behaviour near the critical point of the random graph process, Random Struct Alg 1 (1990), 287–310.

[23] M. Mézard, G. Parisi, and R. Zecchina, Analytic and algorithmic solution of random satisfiability problems, Science 297 (2002), 812.

[24] M. Mézard, T. Mora, and R. Zecchina, Clustering of solutions in the random satisfiability problem, Phys Rev Lett 94 (2005), 197205.

[25] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina, Alternative solutions to diluted *p*-spin models and XORSAT problems, J Stat Phys 111 (2003), 505.

[26] M. Mézard and A. Montanari, Information, physics, and computation, Oxford University Press, 2009.

[27] M. Molloy, Cores in random hypergraphs and boolean formulas, Random Struct Algorithms 27 (2005), 124–135.

[28] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence, Random Struct Algorithms 6 (1995), 161–180.

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

[30] A. Montanari, R. Restrepo, and P. Tetali, Reconstruction and clustering in random constraint satisfaction problems. 2009.

[31] B. Pittel, On trees census and the giant component in sparse random graphs, Random Struct Algorithms 1 (1998), 311–342.

[32] B. Pittel and G. Sorkin, The satisfiability threshold for *k*-XORSAT., ArXiv:1212.1905 (2012).

[33] B. Pittel, J. Spencer, and N. Wormald, Sudden emergence of a giant *k*-core in a random graph, J Comb Theory B 67 (1996), 111–151.