The freezing threshold for k-colourings of a random graph

Michael Molloy*

March 9, 2012

Abstract

We rigorously determine the exact freezing threshold, r_k^f, for k-colourings of a random graph. We prove that for random graphs with density above r_k^f, almost every colouring is such that a linear number of variables are *frozen*, meaning that their colours cannot be changed by a sequence of alterations whereby we change the colours of $o(n)$ vertices at a time, always obtaining another proper colouring. When the density is below r_k^f, then almost every colouring has at most $o(n)$ frozen variables. This confirms hypotheses made using the non-rigorous cavity method.

It has been hypothesized that the freezing threshold is the cause of the "algorithmic barrier", the long observed phenomenon that once the edge-density of a random graph passes $\frac{1}{2}k \ln k(1 + o_k(1))$, no algorithms are proven to find k-colourings, despite the fact that this density is only half the k-colourability threshold.

We also show that r_k^f is the threshold of a strong form of reconstruction for k-colourings of the Galton-Watson tree, and of the graphical model.

*Dept of Computer Science, University of Toronto, molloy@cs.toronto.edu. Research supported by an NSERC Discovery Grant.

1 Introduction

Over the past decade, some groundbreaking hypotheses arising from statistical physics have driven much of the progress on random constraint satisfaction problems (CSP's). In particular, a common geometric interpretation of the *1-Step Replica Symmetry Breaking* hypothesis (1RSB) (see, eg., [40]) says that, at a certain constraint density called the *clustering threshold*, w.h.p.[1] the solution space shatters into an exponential number of *clusters of solutions*, where each cluster is well-connected and any two clusters are well-separated. Furthermore, at a higher density called the *freezing threshold*, there are a linear number of *frozen variables* in almost every cluster; i.e. variables that are fixed throughout the cluster.

These hypotheses have impacted the study of random CSP's in the theoretical computer science community in (at least) two ways: First, an understanding of these hypotheses has led to substantial new results, eg. [17, 38, 11, 55, 26, 48, 18, 19, 25, 1, 33]. Second, much work has gone towards rigorously proving aspects of these hypotheses, eg. [2, 3, 48, 6, 28, 20, 58]. The main contribution of this paper is of the latter type.

In this paper, we rigorously prove hypotheses concerning frozen variables for k-COL; i.e. k-colourability of $G_{n,M}$. This is one of the two most widely studied random CSP's, the other being k-SAT. We establish the *exact* location of the freezing threshold, for k sufficiently large. The asymptotic (in k) value of this threshold follows from [17] (see Section 3 below). The precise value had previously been estimated non-rigorously using the cavity method[57, 54]. We also determine the number of frozen variables, up to a $o(n)$ term.

Our main tool is the *planted model* which Achlioptas and Coja-Oghlan[17] proved could be used to analyze certain random CSP's (see also [48]). Our approach should apply to determine the freezing threshold of most random CSP's for which we can use the planted model, eg. NAE-SAT and hypergraph 2-colourability. We chose to begin with k-COL, because it is the most well-studied such CSP. Unfortunately, we cannot use the planted model for random k-SAT and so we cannot determine the freezing threshold for that model. However, we believe that we can use the technique from this paper to prove that, at higher densities, random k-SAT exhibits freezing; this has been established in [10, 3] for a weaker notion of freezing (see Section 3 below).

To prove our theorem, we strip the random graph down to what we call a *Kempe core*, and prove that almost all of the vertices in the Kempe core are frozen, while almost all of the vertices outside the Kempe core are not frozen.

The Algorithmic Barrier: It has long been observed that most random CSP's appear to be very difficult to solve for a wide range of constraint densities. This was first observed for k-SAT in [16, 42]. For many CSP's, there appears to be what [17] calls an "algorithmic barrier" substantially lower than the density at which they are w.h.p. unsatisfiable. For example: Random instances of k-SAT are known to pass from being w.h.p. satisfiable to w.h.p. unsatisfiable at constraint density $2^k \ln 2 + O(k)$[9], but no algorithm has been proven to w.h.p. find a satisfying solution for problems of density higher than $O(\frac{2^k \ln k}{k})$[17]. The random graph $G_{n,M}$ is known to pass from being w.h.p. k-colourable to w.h.p. not k-colourable at edge-density $k \ln k + O_k(1)$[8], but no algorithm has been proven to w.h.p. find a k-colouring of a random graph with edge-density higher than $\frac{1}{2} k \ln k(1 + o_k(1))$[5, 27]. On the other hand, greedy algorithms succeed, (w.h.p. or with probability bounded away from zero) on CSP's with densities below these points [17, 5, 27]. So below these (asymptotic in k) barriers, finding a solution tends to be relatively easy, and above the barriers it appears to be much more difficult, perhaps even algorithmically intractible.

These barriers are asymptotically (in k) equal to the hypothesized location of the clustering threshold, and this was given rigorous grounding in [17]. Thus, the clustering of the solution space appears to explain the algorithmic barriers. So understanding this clustering is crucial to making further algorithmic progress, or perhaps more importantly, to understanding why exactly random CSP's are so difficult. Moreover, to gain a thorough understanding of random CSP's near the satisfiability threshold, or to *precisely* determine the satisfiability threshold for, eg. random k-COL or k-SAT, it seems clear that we must understand clustering.

In [59, 57, 34] it is argued that *the algorithmic difficulties are not brought on by the clustering threshold, but rather by the freezing threshold.* In other words, the clusters do not pose significant difficulties until they

[1] We say that a property holds *with high probability (w.h.p.)* if it holds with probability tending to one as the number of variables tends to infinity.

have frozen variables. For example, a simple algorithm is proven to work well on 3-COL[7] at densities above the hypothesized clustering threshold (but below the hypothesized freezing theshold). While the clustering threshold is hypothesized to be strictly less than the freezing threshold, their ratio tends to one as k grows. In particular, the freezing threshold is also asymptotic to the observed algorithmic barrier.

The Cavity Method: We close this section by noting that the *cavity method* has been used to predict many thresholds and other important results concerning random CSP's, including satisfiability thresholds (see eg. [40] for many examples). The quest to "rigorize" applications of the cavity method has been one of the most important trends in the study of random structures over the past decade. Very roughly speaking, the cavity method focuses on analyzing the distance-d neighbourhood of a randomly selected vertex for arbitrarily large, but constant, d, making use of the fact that this neighbourhood is w.h.p. a tree. It then hypothesizes the manner in which the remainder of the graph should affect the analysis; this is typically the point which is very difficult to do rigorously as it concerns the long-range dependencies between vertices in the graph. In this paper, we effectively show that as far as freezing is concerned, *the effect of the long-range dependencies is negligible*; the freezing threshold is exactly what the tree-analysis predicts. It is hoped that our techniques will lead to other results along this line.

2 Clusters and Frozen Variables

We study $G_{n,M}$, the random graph with n vertices and M edges, where each such graph is equally likely. We are interested in the range $M = rn$ where r is constant. This model was introduced by Erdős and Rényi in two seminal papers[22, 23]. In these papers, they posed several natural questions about random graphs. All but one have since been answered; the remaining question is: What is the chromatic number of $G_{n,M=rn}$ for $r > \frac{1}{2}$? It is widely believed that for each $k \geq 3$, there is a constant ϕ_k such that for $r < \phi_k$, $G_{n,M=rn}$ is w.h.p. k-colourable while for $r > \phi_k$, $G_{n,M=rn}$ is w.h.p. not k-colourable. The determination of ϕ_k is one of the most important open problems, and indeed the oldest open problem, in random graph theory. Thus far, we do not even know whether ϕ_k exists. Achlioptas and Friedgut[4] proved something close - a function $\phi_k(n)$. Achlioptas and Naor[8] proved that $\phi_k(n) = k \ln k + O_k(\ln k)$.

The 1-RSB analysis was applied to k-COL in eg. [51, 36, 54, 57]. Amongst other things, these papers non-rigorously determine a *clustering threshold*, $r_k^c \approx \frac{1}{2} k \ln k$, at which the associated Gibbs distribution on partitions into an exponential number of pure states. A common geometric interpretation of this phenomenon[48, 54, 59, 35, 36] poses that the k-colourings group into clusters in the following sense:

Let $\Omega_k(G)$ denote the set of k-colourings of a graph G. It is believed that at some density $r \approx \frac{1}{2} k \ln k$, i.e. roughly half the k-colourability threshold, w.h.p. all but a vanishing proportion of $\Omega_k(G)$ can be partitioned into exponentially many sets $S_1, ..., S_x$ such that one can move within S_i by changing the colours of only $o(n)$ vertices at a time, but to move from S_i to $\overline{S_i}$ requires changing a linear number of vertices. More formally:

Definition 2.1. *An ℓ-path of k-colourings of a graph G is a sequence $\sigma_0, \sigma_1, ..., \sigma_t$ of k-colourings of G, where for each $0 \leq i \leq t-1$, σ_i and σ_{i+1} differ on at most ℓ vertices. We say that two k-colourings σ, σ' are ℓ-connected if they can be joined by an ℓ-path $\sigma = \sigma_0, ..., \sigma_t = \sigma'$ for some $t \geq 0$.*

We emphasize that there is no restriction on the length of the path. So two ℓ-connected colourings might differ on arbitrarily many vertices, and we may require an arbitrarily long ℓ-path to join them.

Definition 2.2. *We define an* (a, b)-cluster *to be a subset of colourings $S \subseteq \Omega_k(G)$, such that:*

(a) no pair of colourings $\sigma \in S_i, \sigma' \notin S_i$ is a-connected; and

(b) every pair of colourings $\sigma, \sigma' \in S_i$ is b-connected.

Condition (a) says that clusters are *well-separated*. Condition (b) says that clusters are *well-connected*.

If $a = b + 1$ then (a, b)-clusters exist trivially in every graph. Remarkably, it appears that in $G_{n,M=cn}$ we have (a, b)-clusters when a is much greater than b: $a = \Theta(n), b = o(n)$.

Hypothesis A: *For r sufficiently large: There exists a constant $\alpha > 0$ and a function $\beta(n) = o(n)$ such that w.h.p. all but a vanishing proportion of $\Omega_k(G_{n,M=rn})$ can be partitioned into an exponential (in n) number of $(\alpha n, \beta(n))$-clusters.*

Hypothesis B: *For $r > r_k^f \approx \frac{1}{2}k \ln k$: W.h.p. almost all*[2] *clusters S_i have a linear number of* frozen vertices *v, with the property that for all $\sigma, \sigma' \in S_i$ we have $\sigma(v) = \sigma'(v)$. This does not happen for $r < r_k^f$*

We note that further details are also hypothesized; eg. the clusters change substantially after the *condensation threshold*[35]. It is easy to see that the clusters must have exponential size (eg. from the fact that w.h.p. there are a linear number of degree zero vertices).

We emphasize that the actual hypotheses studied in the physics literature are in terms of pure states of certain Gibbs distributions on the colourings, and are not equivalent to Hypotheses A and B; these are merely common interpretations of the original hypotheses in terms of the geometry of the solution space. In fact, recent evidence indicates that, for some CSP's, there are values of r greater than the clustering threshold for which the clusters are not as well-separated as Hypothesis A posits. Nevertheless, Hypothesis A appears to hold by the time r reaches r_k^f.

Our main theorem proves Hypothesis B, and determines the freezing threshold r_k^f exactly. However, Hypothesis A is not known to hold for k-COL, nor for any other random CSP model other than k-XOR-SAT. So we restate Hypothesis B in a manner that does not involve clusters.

Definition 2.3. *Given a k-colouring σ of a graph G, we say that a vertex v is* ℓ-frozen *with respect to σ if for every ℓ-path $\sigma = \sigma_0, \sigma_1, ..., \sigma_t$ of k-colourings of G, we have $\sigma_t(v) = \sigma(v)$.*

In other words, it is not possible to change the colour of v by changing at most ℓ vertices at a time. It is important to note:

Observation: *If Hypothesis A holds, then for every $\beta(n) < \ell \leq \alpha n$, the frozen vertices in the cluster containing σ are exactly the vertices that are ℓ-frozen with respect to σ.*

In particular, every vertex that is αn-frozen according to Definition 2.3, is also frozen in the sense of Hypothesis B, assuming Hypothesis A.

We define

$$r_k^f = \min_{x>0} \frac{(k-1)x}{2(1-e^{-x})^{k-1}}.$$

For any $r > r_k^f$ we let $x_k(r)$ denote the largest positive solution to $r = \frac{(k-1)x}{2(1-e^{-x})^{k-1}}$.

Our main theorem is that, for k sufficiently large, r_k^f is the precise threshold for most colourings to have a linear number of ℓ-frozen vertices, where ℓ is linear in n:

Theorem 2.4. *There exists a constant integer k_0 such that for all $k \geq k_0$, and for any $\omega(n)$ tending to ∞ arbitrarily slowly with n: Let σ be a uniformly random k-colouring of $G_{n,M=rn}$.*

(a) For any $r > r_k^f$, there exists a constant $0 < \alpha < 1$ for which:

(i) w.h.p. there are $\frac{(k-1)x_k(r)}{2r}n + o(n)$ vertices that are αn-frozen with respect to σ.

(ii) w.h.p. there are $(1 - \frac{(k-1)x_k(r)}{2r})n + o(n)$ vertices that are not $\omega(n)$-frozen with respect to σ.

(b) For any $r < r_k^f$, w.h.p. there are at most $o(n)$ vertices that are $\omega(n)$-frozen with respect to σ.

In other words: for $r > r_k^f$, a linear number of variables are αn-frozen, while for $r < r_k^f$, all but at most $o(n)$ variables are not even $\omega(n)$-frozen for any $\omega(n)$ growing arbitarily slowly with n. Furthermore, for $r > r_k^f$ we specify the specific number of αn-frozen vertices, up to an additive $o(n)$ term. All but at most $o(n)$ of the other vertices are not even $\omega(n)$-frozen.

In fact, we prove something stronger. In Section 5, we define a subset of the vertices which we call the *Kempe core*. r_k^f is the threshold for the appearance of a Kempe core in the planted model and, for $k \geq k_0$,

[2]Here, "almost all" means for all but a vanishing proportion of the clusters when they are weighted by their size.

also in the uniform model. We will prove that w.h.p. all but $o(n)$ vertices of the Kempe core are frozen and at most $o(n)$ vertices outside of the Kempe core are frozen. Thus, given a uniform k-colouring σ of $G_{n,M=rn}$, we w.h.p. specify precisely which vertices are frozen w.r.t. σ up to an error of $o(n)$ vertices.

We do not know the value of k_0; it comes from Theorem 4.3 below, and its value is estimated in [2] to perhaps be roughly 20. Our theorem likely holds for $k \geq 9$ (see below). One would like to strengthen Theorem 2.4 by (i) replacing $k \geq k_0$ with $k \geq 9$, and (ii) replacing $o(n)$ with zero in part (b). In both cases, the bottleneck is the limitations of Theorem 4.3. Both these improvements are likely to be true, although to replace $o(n)$ by zero in (b), we would have to name a specific $\omega(n)$; $\omega(n) = O(\log n)$ might suffice.

Hypothesized values for r_k^f are provided in [57, 54] for $3 \leq k \leq 10$, using the cavity method to determine an expression for r_k^f and using population dynamics to estimate the value of that expression. They first determine an expression for the freezing threshold on the "tree factor graph", which is hypothesized to be equal to the freezing threshold on $G_{n,M}$ so long as it is below the condensation threshold. For $3 \leq k \leq 8$ the freezing threshold appears to be greater than the condensation threshold, and so is hypothesized to be less than the threshold arising from the tree factor graph. Their expression for the threshold on the tree factor graph is equivalent to ours[3], but is more unwieldy. So for $k \geq 9$ our r_k^f agrees with the hypothesized value of the freezing threshold; eg. for $k = 9, 10$ we have $r_k^f = 17.829..., 20.753...$[4] Thus r_k^f is likely to be the precise freezing threshold for $k \geq 9$, but it is only proven to be correct for $k \geq k_0$. It is likely to be true that Theorem 4.3 can be applied at all densities below the hypothesized condensation threshold, and so would imply that r_k^f is indeed the freezing threshold for $k \geq 9$; but this is not proven.

Asymptotically, we have:

$$r_k^f = \frac{1}{2}k(\ln k + \ln\ln k + 1 + o(1)), \tag{1}$$

which agrees with the asymptotics provided in (44) of [57] and (78) of [54].

To be clear: We do not prove that clusters exist above r_k^f, and so we do not know that vertices are frozen in the sense of Hypothesis B, only that they are frozen as in Definition 2.3. But our results imply that for any $r \geq r_k^f$: if Hypothesis A holds then Hypothesis B holds.

2.1 Reconstruction

In the context of graph colourings, the reconstruction problem is as follows: Consider a tree T of height h, eg. a D-regular tree or a Galton-Watson tree, and consider a uniformly random k-colouring of T. Expose the colours of the leaves at distance h from the root, and consider the conditional distribution that they impose on the colour of the root. We say that the colouring is *reconstructable* if, with probability bounded away from zero as h grows, the distribution of the colour of the root is bounded away from the uniform distribution. There has been extensive focus on the *reconstruction threshold*, the average degree at which the colouring is w.h.p. reconstructable (see eg. [55, 26, 50, 48]). The reconstruction threshold for k-colourings of d-regular trees is bounded between $k(\ln k + \ln\ln k + 1 - \ln 2 + o(1))$[50] and $k(\ln k + \ln\ln k + 1 + o(1))$[55].

In the *graphical model*, introduced in [26], one chooses a Galton-Watson tree T, and the colours of the leaves as follows: Choose a random (G, σ) from $G_{n,p=d/n}$, pick a random vertex v, and expose the distance h neighbourhood of v; call that tree T. Now fix the colour $\sigma(u)$ for each leaf u at distance h from the root, and consider a uniformly random k-colouring of T conditional on each leaf u having colour $\sigma(u)$. [48] applies a theorem from [26] to show that, for k-COL and some other CSP models, the reconstruction threshold for the graphical model is equal to that for the Galton-Watson tree model.

The hypothesized location of the clustering threshold (see eg. [35]) is derived from a non-rigorous determination[39] of the reconstruction threshold for Galton-Watson trees, after halving since the average degree of $G_{n,M=rn}$ is $2c$.

Here, we consider a stronger form of reconstruction. [54] defines the *naive reconstruction threshold* to be the maximum d such that: Let T be a Galton-Watson random tree of height h where each vertex has d

[3] We are grateful to an anonymous referee for pointing out this substitution.

[4] [57, 54] report the threshold in terms of the average degree, rather than edge-density and so their values are exactly twice ours. Also note that what we call the freezing threshold is called the rigidity threshold in [57].

expected children, and take a uniformly random k-colouring of T. Uncolour all vertices except for the leaves at distance h from the root. The probability that the remaining colours force exactly one choice for the colour of the root tends to zero as $h \to \infty$. As with reconstruction, we can also consider choosing the tree and colours as in the graphical model.

The above upper bound on the reconstruction threshold was obtained by computing an upper bound on the naive reconstruction threshold. Note that, after doubling, that bound is asymptotic to the freezing threshold of $G_{n,M}$ - see (1) below. [54] hypothesizes that the naive reconstruction threshold is equal to the freezing threshold (after doubling), for a variety of CSP's. [48] also remarks that it is natural to conjecture these thresholds to be equal. Our proof of Theorem 2.4 easily implies that they are indeed equal for k-colourability, with k sufficiently high:

Theorem 2.5. *In both the Galton-Watson tree model for $k \geq 3$, and the graphical model for $k \geq k_0$, the naive reconstruction threshold is at average degree $d = 2r_k^f$.*

In other words, for a uniform colouring of $G_{n,M=rn}$: The probability that the colour of v is forced by the colours of the vertices of distance h from v stays bounded away from zero as $h \to \infty$, iff the probability that v is $\Theta(n)$-frozen is bounded away from zero. Intuitively, this should be expected. We give the proof in Section 7.

2.2 Minimal Rearrangements

[49, 54] describe a connection between freezing and minimal rearrangements. Given a vertex v in a colouring σ, let σ' be a colouring where $\sigma(v) \neq \sigma'(v)$ such that the number of vertices on which σ, σ' differ is minimum; the set of vertices on which these colourings differ is called a *minimal rearrangement* for v. [54] shows non-rigorously that w.h.p. the average over all vertices v of the size of a minimal rearrangement jumps from $O(1)$ to $\Theta(n)$ at the freezing threshold. A simple corollary of our work shows:

Corollary 2.6. *For $k \geq k_0$ (from Theorem 2.4), let σ be a uniformly random k-colouring of $G_{n,M=rn}$.*

(a) For $r < r_k^f$, w.h.p. the average size of a minimal rearrangement is $o(n)$.

(b) For $r > r_k^f$, w.h.p. the average size of a minimal rearrangement is $\Theta(n)$.

We give the short proof in Section 7.

3 Related work

1-RSB analysis for k-colourings of $G_{n,M}$ was first done in [51] (see also [36]). The freezing threshold was studied in great depth in [57, 54, 59]. These studies were non-rigorous, but mathematically sophisticated. [57] was the first paper to argue that freezing may be the cause of the algorithmic barrier.

Achlioptas and Ricci-Tersenghi[10] were the first to rigorously prove any form of freezing in a random CSP. They studied random k-SAT and showed that for $k \geq 8$, for a wide range of edge-densities below the satisfiability threshold and for *every* satisfying assignment σ, the vast majority of variables are 1-frozen w.r.t σ. Equivalently, such vertices are frozen in the connected components of the graph whose vertices are the satisfying assignments, and where a pair of assignments is adjacent if they have Hamming distance 1. These components are 1-connected by definition, but they are not w.h.p. $\Theta(n)$-separated and hence do not satisfy Hypothesis A. However, it is plausible that they are in some sense close to being the clusters of Hypothesis A. [10] proves the existence of the frozen variables by stripping down to an appropriate core, which inspired us to do the same here. One difference between their approach and ours is that the definition of their core implies that its vertices are 1-frozen, whereas much of the work in this paper is devoted to proving that the vertices of our core are $\Theta(n)$-frozen.

[2] proves the existence of what they call *rigid* variables in various random CSP's, including k-COL. The definition of rigid is equivalent to taking $\ell = \Theta(n)$ in Definition 2.3, but requiring $t = 1$. That is, a vertex v

is rigid w.r.t. a k-colouring σ if every σ' with $\sigma'(v) \neq \sigma(v)$ must differ from σ on $\Theta(n)$ vertices. Achlioptas and Coja-Oghlan[2] prove that for $r < (\frac{1}{2} - \epsilon)k \ln k$ w.h.p. there are *no* rigid vertices (and hence no frozen vertices) w.r.t. almost all colourings of $G_{n,M=rn}$, while for $r > (\frac{1}{2} + \epsilon)k \ln k$ w.h.p. there are a linear number of such rigid vertices. A simple argument (see the remark following Corollary 5.5 below) extends their result to show the same for frozen vertices. So [2] provides the asymptotic, in k, location of the freezing threhold. It also provides the asymptotic location of the freezing threshold for NAE-SAT and hypergraph 2-colouring.

[3, 2, 48] establish the existence of what they call *cluster-regions* for various CSP's; these are proven to be w.h.p. $\Theta(n)$-separated but are not shown to be w.h.p. well-connected. For k-COL, [2] proves that for $r > (\frac{1}{2} + \epsilon)k \ln k$ the solution space w.h.p. shatters into an exponential number of $\Theta(n)$-separated cluster-regions, each containing an exponential number of colourings. While these cluster-regions do not satisfy Hypothesis A, the well-connected property of clusters does not seem to be crucial to the difficulties that they pose for algorithms. So [2] was a very big step towards explaining why $\frac{1}{2}k \ln k$ appears to be, asymptotically, an algorithmic barrier.

The clusters of k-XOR-SAT are very well-understood, independently by [6, 28] (see also [21]). We know the clustering threshold, which in this case is equal to the freezing threshold, and have a very good description of the clusters and the frozen variables. The picture is much simpler here; for example, the same variables are frozen in every cluster. The simple linear algebraic characterization of the solution space was very helpful.

4 The planted model

Definition 4.1. *The* uniform model $U_{n,M}$ *is a random pair* (G, σ) *where* G *is taken from the* $G_{n,M=rn}$ *model and* σ *is a uniformly random* k*-colouring of* G*.*

Until a few years ago, the biggest hurdle to theorems such as Theorem 2.4 has been that there is no representation of the uniform model that lends itself to analysis. This hurdle, along with the corresponding hurdles for random k-SAT, and a few other random CSP's, was overcome by Achlioptas and Coja-Oghlan[2] who proved that, under certain conditions, one can work instead with the much simpler planted model. We will use the $G_{n,p}$ version:

Definition 4.2. *The* planted model $P_{n,p}$ *is a random pair* (G, σ) *chosen as follows: Take a uniformly random partition* σ *of* $\{1, ..., n\}$ *into* k *parts* $A_1, ..., A_k$*. Each pair of vertices in two different parts is joined with an edge with probability* p*, where the edge-choices are independent.*

The following is a derivation of a key tool from [2]. See Appendix 9 for more detail.

Theorem 4.3. *[2] For every* k *at least a particular constant* k_0 *and every* $r < 0.9k \ln k$*, there is a function* $f(n) = o(n)$ *such that: Let* $\mathcal{E}$ *be any property of pairs* (G, σ) *where* σ *is a* k*-colouring of* G*. Set* $c = \frac{2k}{k-1}r$*. If*

$$\mathbf{Pr}(P_{n,p=c/n} \text{ has } \mathcal{E}) > 1 - e^{-f(n)},$$

then

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

We define

$$c_k = \min_{y>0} \frac{ky}{(1 - e^{-y})^{k-1}}.$$

For any $c > c_k$ we let $y_k(c)$ denote the largest solution to $c = \frac{ky}{(1-e^{-y})^{k-1}}$. Note that $c_k = \frac{2k}{k-1}r_k$. We define:

$$\lambda_k(c) = y_k(c)/c.$$

We say that v is an ℓ*-frozen variable of* (G, σ) if v is ℓ-frozen with respect to σ. So, roughly speaking, our goal is to prove that c_k is the threshold for $P_{n,p=c/n}$ to have a linear number of αn-frozen variables, and that the failure probability is $1 - e^{-f(n)}$ where $f(n)$ comes from Theorem 4.3.

5 Kempe cores

Given a k-colouring σ of a graph G, with colour classes $A_1, ..., A_k$, a *Kempe chain* is a component of the subgraph induced by two colour classes. Suppose C is a non-empty Kempe chain on colour classes A_i, A_j. Then exchanging the colours i, j on the vertices of C will result in a new k-colouring of G. Note that a single vertex of colour i will constitute a Kempe chain if it has no neighbours of colour j, for some $j \neq i$. Kempe chains were introduced by Kempe[31] in his work on the Four Colour Problem.

It is clear that a vertex that is in a Kempe chain of size at most ℓ is not ℓ-frozen. This inspires us to remove all "small" Kempe chains from our graph, in order to look for frozen vertices. A bit of thought will make it clear that w.h.p. most vertices in Kempe chains of size at most ℓ in the remaining graph are not ℓ-frozen either. This follows from branching properties of the random graph: if C is a small Kempe chain in the remaining graph, w.h.p. the small Kempe chains that were removed from the original graph each have at most one edge to C. Furthermore none of those chains adjacent to C are adjacent to each other. Thus we can flip the vertices on any subset of those chains without them interfering with each other, thus enabling C to be flipped. This inspires us to remove small Kempe chains iteratively.

Of course, we need to specify what we mean by "small". It turns out that w.h.p. there will be no Kempe chains of size between $O(\log n)$ and $\Theta(n)$; i.e. every Kempe chain will either be small or giant. But to be specific, and to strenghten "w.h.p." enough to apply Theorem 4.3, we will take small to mean: of size at most $g(n)$ for some $g(n) = o(n)$ to be specified later. Thus, we apply the following procedure:

Kempe-Strip
Input: a graph G and a k-colouring $\sigma = A_1, ..., A_k$ of G.
While there are any Kempe chains of size at most $g(n)$
Remove the vertices of one such Kempe chain from G.

The (possibly empty) Kempe core is what remains. Note that, as with most core stripping procedures, the output does not depend on the order in which we choose to remove Kempe chains. So the Kempe core is well-defined.

By definition, every vertex in the Kempe core cannot have its colour changed by changing the vertices of a small Kempe chain. We prove the stronger property that almost every vertex in the Kempe core cannot have its colour changed by changing a small subset of vertices which may involve *more than two* colours.

To gain some intuition as to why this could be the case, note first that almost every very small subgraph, i.e. of size $O(1)$, is a tree. A bit of thought will show that if we can change the colours of a tree to obtain another colouring, then that tree contains a subtree which is a Kempe chain. Thus, (most) changes of $O(1)$ vertices can be simulated by a sequence of Kempe-chain switches.

Of course, we still need to deal with the possibility of changing a non-constant but sublinear sized set of vertices which involve more than two colours. That is the source of much of the difficulty in this paper.

The following lemma is proven in Appendix 13. (See also Lemma 11.3(a)).

Lemma 5.1. *For $k \geq 3$, and any $f(n) = o(n)$:*

(a) If $c < c_k$ then with probability at least $1 - e^{-f(n)}$, the Kempe core of $P_{n,p=c/n}$ has size $o(n)$.

(b) If $c > c_k$ then with probability at least $1 - e^{-f(n)}$, the Kempe core of $P_{n,p=c/n}$ has size $k\lambda_k(c)n + o(n)$.

Remark: In fact, for $c < c_k$, w.h.p. the Kempe core of $P_{n,p=c/n}$ has size 0. But this statement fails with probability $1/\text{poly}(n)$.

The Kempe core can be viewed as a variation of the well-studied k-core[52], which is obtained by iteratively removing vertices of degree less than k. Note that for the Kempe core, we will remove all vertices of degree less than $k-1$, as they will be Kempe chains of size one. In addition, we remove many other small subgraphs. At first, we were quite intimidated at the prospect of extending the $(k-1)$-core analysis to the stripping of these more general subgraphs. But we noticed that if viewed from a different angle, the Kempe core has a very natural description:

Observation 5.2. *Kempe-Strip is equivalent to iteratively removing all small components from the bipartite random graph induced by each pair of parts A_a, A_b.*

The reason for this is that the Kempe-chains are precisely those small components. So we can implement Kempe-Strip by an iterative process where each iteration proceeds as follows: For each $1 \leq a, b \leq k$ we remove all vertices outside of the giant component in the bipartite subgraph induced by what remains of A_a, A_b. This is very fortuitous, and it enabled our analysis of the Kempe core.

In [44], we defined a very natural, and much simpler problem of this nature, and analyzed its core. Our primary motivation was to develop a technique to apply to the analysis of the Kempe core. In Appendix 13, we sketch how to adapt that proof to this setting.

Thus the Kempe core has the property that the subgraph induced by each pair of parts is connected. In [43], Achlioptas and this author asked whether the k-colourability threshold was the same as the threshold for a subgraph with that property to appear. The results of this paper answer that question negatively, for large k.

Having established the Kempe core threshold, we next prove that it has the properties that we require for our main theorem. In Appendix 14, we show:

Lemma 5.3. *For $k \geq 3$, $c \neq c_k$, any $f(n) = o(n)$ and any $\epsilon > 0$: There exist constants T, Z such that with probability at least $1 - e^{-f(n)}$, all but ϵn of the vertices outside of the (possibly empty) Kempe core of $P_{n,p=c/n}$ are either (i) not T-frozen, or (ii) within distance Z of a cycle with length less than Z.*

The proof argues that for all but ϵn of the vertices v removed during the stripping process, if v is not as in (ii) then the sequence of Kempe-chains that led to the removal of v form a tree-like structure. This structure allows the Kempe-chains to be switched without interfering with each other, thus allowing v to be changed. Furthermore, each of those Kempe-chains has size at most T and so v is not T-frozen.

It is straightforward to show that w.h.p. $G_{n,M}$ has $o(n)$ vertices as in Lemma 5.3(ii). So by allowing ϵ to be arbitrarily small, we obtain Theorem 2.4 parts (a.ii) and (b). It only remains to prove part (a.i); i.e to prove that almost all of the Kempe core is αn-frozen.

Recall from Section 3 that a vertex is said to be *rigid* if to change its colour *in one step*, we need to change the colours of $\Theta(n)$ other vertices. Conceptually, it seems much easier to show that a vertex is rigid than to show that it is $\Theta(n)$-frozen. Proving rigidity requires understanding the structure of the symmetric difference between two colourings. Proving frozeness requires understanding sequences of colourings, which can become very complicated.

One of the properties which made this proof feasible is that (most of) the set of frozen variables is, in fact, *internally rigid*, as described in the next lemma, which is the key lemma of this paper and is proven in Appendix 12.

Lemma 5.4. *For $k \geq 3$, $c > c_k$, and any $f(n) = o(n)$, there exists constant $\alpha = \alpha(c, k) > 0$ such that with probability at least $1 - e^{-f(n)}$, the Kempe core K of $P_{n,p=c/n}$ has the following property:*
For all but $o(n)$ vertices $v \in K$, any k-colouring of K which differs from σ on v must differ from σ on at least $2\alpha n$ vertices of K.

The $o(n)$ term depends on $f(n)$. This internal rigidity is enough to imply that almost all vertices of the Kempe core are frozen:

Corollary 5.5. *For $k \geq 3$, $c > c_k$, and any $f(n) = o(n)$, there exists constant $\alpha = \alpha(c, k) > 0$ such that with probability at least $1 - e^{-f(n)}$: all but $o(n)$ vertices of the Kempe core K of $P_{n,p=c/n}$ are αn-frozen.*

Proof Lemma 5.4 says we have $\Theta \subseteq K$ with size $|K| - o(n)$ such that every $v \in \Theta$ has the property that any k-colouring of K which differs from σ on v must differ from σ on at least $2\alpha n$ vertices of K, and thus on at least $2\alpha n - o(n) > \alpha n$ vertices of Θ. The proof now follows by taking any sequence of k-colourings of G, $\sigma = \sigma_0, \sigma_1, ..., \sigma_t$, and considering the first step at which a vertex of Θ changes. □

See Appendix 10 for more details.

Remark: [2] proves that for $r > (\frac{1}{2} + \epsilon)k \ln k$, w.h.p. the vertices in a certain core are internally rigid with respect to that core. So the argument for Corollary 5.5 also implies that their rigid variables are frozen.

Recall that we have planted a k-colouring $\sigma = A_1, ..., A_k$. To prove Lemma 5.4, we need to focus on sets of vertices that can be changed to obtain a new colouring:

Definition 5.6. *A* Δ-set *is the symmetric difference of* σ *and some other* k*-colouring of the Kempe core,* K*. Specifically, given such a colouring* σ'*, the set of vertices* $u \in K$ *with* $\sigma(u) \neq \sigma'(u)$ *is a* Δ*-set, which we sometimes denote by* $\sigma \Delta \sigma'$*.*

We would like to show that there are no Δ-sets of size smaller than $\Theta(n)$. Unfortunately, this is not true - we can have small Δ-sets which induce subgraphs with exactly one cycle. We call these *cyclic* Δ-sets, and they are described in Appendix 12. In expectation, the total number of vertices on cyclic Δ-sets is $O(1)$.

By examining the graph theoretic structure of a Δ-set, we can prove that w.h.p. all other Δ-sets have size $\Theta(n)$. At first glance, this would appear to prove our key lemma. However, this property only holds with probability $1 - \frac{1}{\text{poly}(n)}$ which is not enough to apply Theorem 4.3 and transfer the result from the planted model to the uniform model. So instead we have to prove:

Let $\mathcal{D}$ *be the union of all* Δ*-sets of size less than* $2\alpha n$*. With probability at least* $1 - e^{-f(n)}$*,* $|\mathcal{D}| = o(n)$*.*

To just prove that all non-cyclic Δ-sets have size at least $2\alpha n$, we can use an approach that has been used in [47, 15, 6] to prove similar results: We would like to prove that the 2-core of every non-cyclic Δ-set has high edge-density. If we could do so, then a very fast and common argument based on subgraph densities in $G_{n,M}$ proves that every such subgraph must have linear size. It would follow that the 2-core of the Δ-set must have linear size, and hence so must the Δ-set.

Unfortunately, that is not the case. Δ-sets that are not sufficiently dense can arise from long paths of degree 2 vertices. The proliferation of such paths is determined by two branching factors. We bound these branching factors by analyzing the Kempe core, and show that they are both less than one. This allows us to apply a first moment argument to show that w.h.p. the 2-core of a non-cyclic Δ-set must have size $\Theta(n)$.

In order to adapt this approach to bound $|\mathcal{D}|$, we must complicate things in two ways. (1) we need to extend our analysis to the *unions* of Δ-sets. (2) we cannot restrict our attention to the 2-cores of the Δ-sets. The second complication turned out to be the most difficult.

The details of this argument can be found in Appendix 12.

6 2-paths in Δ-sets

We let $K_i \subset A_i$ be the set of vertices from part A_i that are in the Kempe core. We let $K_{i,j}$ denote the bipartite subgraph of the Kempe core induced by K_i, K_j. Recall that each $K_{i,j}$ is connected.

As described above, a key part of our analysis is to bound the proliferation of long paths of degree 2 vertices in the 2-core of a Δ-set. We prove that such paths are of two types:

Definition 6.1. *A* 2-path *in a* Δ*-set* $\sigma \Delta \sigma'$ *is a path* $u_0, ..., u_x$ *in the 2-core of* $\sigma \Delta \sigma'$ *such that*

(a) $x \geq 1$*;*

(b) each u_i *has degree 2 in the 2-core of* $\sigma \Delta \sigma'$*;*

(c) either
Type A: *every* u_i *is in the 2-core of* $K_{\sigma(u_i),\sigma'(u_i)}$*; or*
Type B: *every* u_i *is not in the 2-core of* $K_{\sigma(u_i),\sigma'(u_i)}$ *and, for* $0 \leq i \leq x-1$*,* u_{i+1} *is its unique neighbour on the path to the 2-core of* $K_{\sigma(u_i),\sigma'(u_i)}$*.*

A key basic property of Δ-sets is:

Proposition 6.2. *Let* u *be any vertex in a* Δ*-set* $\sigma \Delta \sigma'$*. Then every neighbour of* u *in* $K_{\sigma(u),\sigma'(u)}$ *is also in* $\sigma \Delta \sigma'$*.*

Proof Every neighbour w of u in $K_{\sigma(u),\sigma'(u)}$ has $\sigma(w) = \sigma'(u)$. Since σ' is a proper colouring, we cannot have $\sigma'(w) = \sigma'(u)$. Therefore $\sigma'(w) \neq \sigma(w)$. □

Thus, each vertex $u \in \sigma\Delta\sigma'$ yields at least one other neighbour $w \in \sigma\Delta\sigma'$ with $\sigma(w) = \sigma'(u)$. Type B 2-paths are formed when a sequence of vertices each yields either one or two such neighbours in the 2-core of $\sigma\Delta\sigma'$. These paths are the most delicate to deal with.

Given a vertex u that is not in the 2-core of $K_{\sigma(u),\sigma'(u)}$, we wish to bound the expected number of neighbours w of u which could act as the next vertex in a Type B 2-path. If we have exposed $u, \sigma(u), \sigma'(u)$ then this determines w as it must be the unique neighbour of u on the path from u to the 2-core of $K_{\sigma(u),\sigma'(u)}$. We have also determined $\sigma(w) = \sigma'(u)$. We have $k-1$ choices for $\sigma'(w)$. We want to bound the probability that w is not in the 2-core of $K_{\sigma(w),\sigma'(w)}$. By analysing the Kempe core, we show that the proportion of non-2-core vertices in each $K_{i,j}$ is strictly less than $\frac{1}{k-1}$. With some work, this implies that the probability w is not in the 2-core of $K_{\sigma(w),\sigma'(w)}$ is less than $\frac{1}{k-1}$. Multiplying by the $k-1$ choices for $\sigma'(w)$ yields a branching factor of less than 1, as required. This is delicate, because w cannot always be treated as just a uniform member of $K_{\sigma(w)}$, and the Kempe core is tricky to deal with.

For Type A paths, we prove that there is a pair of colours a, b such that every u_i satisfies $\{\sigma(u_i), \sigma'(u_i)\} = \{a, b\}$. It follows that the Type A 2-path is a path of degree 2 vertices in the 2-core of $K_{a,b}$. We prove that the 2-core of $K_{a,b}$ is uniform with respect to its degree sequence, and so we can expose it using the configuration model[12]. The branching factor then is twice the number of degree 2 vertices, divided by the total degree, which we show to be less than 1. This portion of the analysis is very much like the corresponding part of [6].

Further details, including how to use this to prove Lemma 5.4, can be found in the appendix.

7 Naive Reconstruction

Here we note how the results of this paper easily imply Theorem 2.5 and Corollary 2.6.

Proof of Theorem 2.5: The naive reconstruction problem on the Galton-Watson tree model (see section 2.1) is easily seen to be equivalent to: Choose a random (G, σ) from $P_{n,p=c/n}$, pick a random vertex v, and expose the distance h neighbourhood of v; call that tree T. Now fix the colour $\sigma(u)$ for each leaf u of T at distance h from v. Does the probability that the colours on these leaves determine the colour of v tend to zero as $h \to \infty$? Note that the average degree in this tree is $\frac{k-1}{k}c$.

If $c > c_k$, then it is easy to see that the answer is No. With probability bounded away from zero, v will be frozen w.r.t. σ. Thus, changing the colour of v requires changing $\Theta(n)$ other vertices. We can assume that those vertices induce a connected subgraph, as otherwise we could change one component at a time. W.h.p., $|T| = o(n)$, and so least one of those vertices must be of distance h from v. Therefore, with probability bounded away from zero, the colours of the leaves *determine* the colour of v. It follows that the naive reconstruction threshold for k-colourings of Galton-Watson trees is at most $\frac{k-1}{k}c_k = 2r_k^f$.

For $c < c_k$, the proof of Lemma 5.3 implies that for every $\epsilon > 0$, there exist I, s such that with probability at least $1-\epsilon$, v can be removed by the deletion of a sequence of at most I Kempe-trees, each of size at most s. (To be specific, each set of at most s vertices will form a Kempe-tree in what remains at the time that they are deleted.) Thus, the colour of v can be changed by changing only the colours of those Kempe-trees. We can assume that the union of these Kempe-trees is connected, as otherwise the deletion of some would have no effect on whether others may be deleted. By taking $h > Is$, none of those Kempe-trees contain any vertices at distance at least h from v. Thus, the colour of v can be changed without changing the colours of any of the leaves in T at distance h from v. Therefore, for every $\epsilon > 0$, the probability that those leaves force the colour of v is less than ϵ for sufficiently large h. Therefore, the naive reconstruction threshold for k-colourings of Galton-Watson trees is at most $\frac{k-1}{k}c_k = 2r_k^f$.

In the graphical model, T is chosen as above, except that we use the uniform model, rather than the planted model. So applying Theorem 4.3 to transfer the results of the preceding paragraphs to $U_{n,M}$, we again prove that the naive reconstruction threshold is $2r_k^f$, although this time we require $k \geq k_0$. □

Proof of Corollary 2.6: For $r > r_k^f$, Theorem 2.4 implies that w.h.p. $\Theta(n)$ vertices are αn-frozen, for

some $\alpha > 0$, and hence their minimal size rearrangements have size at least αn. This proves part (b).

As in the previous proof, for $r < r_k^f$, for any $\epsilon > 0$ there exists $I, s = O(1)$ so that for all but ϵn choices of v, the colour of v can be changed by changing at most Is other vertices; i.e. the minimal size rearrangement for v has size $O(1)$. Taking ϵ arbitrarily small yields part (a). □

8 Future Work

As mentioned above, we expect that we can apply these techniques to determine the freezing threshold for other random CSP models for which one can use the planted model, eg. NAE-SAT and hypergraph 2-colouring. We also expect that we can prove that all but $o(n)$ of the k-SAT variables shown to be 1-frozen in [10] are in fact $\Theta(n)$-frozen. This is ongoing work with R. Restrepo.

Acknowledgement: I am grateful to Dimitris Achlioptas for explaining clustering to me many times over many years. I am also thankful to Lenka Zdeborová for her valuable advice, and to some anonymous referees for very helpful comments.

References

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

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

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

[4] D. Achlioptas and E. Friedgut. *A sharp threshold for k-colorability.* Random Struct. Algorithms **14** (1999), 63 70.

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

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

[7] D. Achlioptas and C. Moore *Almost all graphs with average degree 4 are 3-colorable.* J. Comp. Sys. Sci., **67** (2003), 441 - 471.

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

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

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

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

[12] B. Bollobás. *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs.* Europ. J. Combinatorics **1** 311-316 (1980).

[13] B. Bollobás, *Random Graphs.* 2nd edition. Cambridge University Press, 2001.

[14] B. Bollobás and O. Riordan, *Asymptotic normality of the size of the giant component via a random walk.* arXiv:1010.4595

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

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

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

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

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

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

[21] O. Dubois and J. Mandler. *The 3-XORSAT threshold.* Proc. 43rd FOCS (2002), 769 - 778.

[22] P. Erdős and A. Rényi. *On random graphs I.* Publ. Math. Debrecen **6** (1959), 290 - 297.

[23] P. Erdős and A. Rényi. *On the evolution of random graphs.* Magayar Tud. Akad. Mat. Kutato Int. Kozl. **5** (1960), 17 - 61.

[24] 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).

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

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

[27] G.R.Grimmett and C.J.H.McDiarmid. *On colouring random graphs.* Math. Proc. Cambridge Philos. Soc., **77** (1975), 313324.

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

[29] S. Janson and M. Luczak. *A simple solution to the k-core problem.* Random Structures Algorithms **30** (2007) 50 - 62 (2007).

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

[31] A. B. Kempe. *On the geographical problem of the four colors.* Amer. J. Math., **2** (1879), 193 - 200.

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

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

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

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

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

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

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

[39] M. Mezard and A. Montanari. *Reconstruction on trees and spin glass transition.* J. Stat. Phys. **124** (2006), 1317 - 1350.

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

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

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

[43] M. Molloy. *Thresholds for colourability and satisfiability in random graphs and boolean formulae.* Proceedings of the British Combinatorial Conference (2001).

[44] M. Molloy. *Sets that are connected in two random graphs.* Preprint available at `www.cs.toronto.edu/~molloy/papers.html`

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

[46] M. Molloy and B. Reed. *A critical point for random graphs with a given degree sequence.* Random Structures and Algorithms **6** 161 - 180 (1995).

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

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

[49] A. Montanari and G. Semerjian. *On the dynamics of the glass transition on Bethe lattices.* J. Stat. Phys. **124** (2006) 103 - 189.

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

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

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

[53] B. Pittel and N. Wormald, *Counting connected graphs inside-out.* J. Comb.Th. B **93** (2005), 127 - 172.

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

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

[56] V.E. Stepanov, *Some features of the structure of a random graph near a critical point.* (Russian) Teor. Veroyatnost. i Primenen. **32** (1987), 633657.

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

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

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

Appendix

9 The planted model

Definition 9.1. *The* uniform model $U_{n,M}$ *is a random pair* (G,σ) *where* G *is taken from the* $G_{n,M=rn}$ *model and* σ *is a uniformly random* k*-colouring of* G*.*

Until a few years ago, the biggest hurdle to theorems such as Theorem 2.4 has been that there is no representation of the uniform model that lends itself to analysis. This hurdle, along with the corresponding hurdles for random k-SAT, and a few other random CSP's, was overcome by Achlioptas and Coja-Oghlan[2] who proved that, under certain conditions, one can work instead with the much simpler planted model:

Definition 9.2. *The* planted model $P_{n,M}$ *is a random pair* (G,σ) *chosen as follows: Take a uniformly random partition* σ *of* $\{1,...,n\}$ *into* k *parts* $A_1,...,A_k$*. Then choose* M *random edges, uniformly and without replacement, from all edges whose endpoints are in two different parts.*

In other words, $P_{n,M}$ is chosen by first choosing a uniformly random k-colouring σ of the vertices $\{1,...,n\}$, and then choosing a graph that is uniform amongst all graphs with that vertex set and with M edges, for which σ is a k-colouring. This clearly has a different distribution than what one obtains by carrying those steps out in the other order; i.e. first choosing $G_{n,M}$ and then taking a uniformly random k-colouring of G. But remarkably, [2] proves that one can sometimes transfer w.h.p. properties from the former model to the latter, so long as the failure probability of those properties is nearly exponentially small. The following is a rephrasing of Theorem 6 from [2]; the original statement is somewhat more general.

Theorem 9.3. *[2] For every* k *at least a particular constant* k_0 *and every* $r < 1.9k\ln k$*, there is a function* $f(n) = o(n)$ *such that: Let* $\mathcal{E}$ *be any property of pairs* (G,σ) *where* σ *is a* k*-colouring of* G*. If*

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

then

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

It will be more convenient to work in the $G_{n,p}$ version of the planted model, which we define as follows:

Definition 9.4. *The* planted model $P_{n,p}$ *is a random pair* (G,σ) *chosen as follows: Take a uniformly random partition* σ *of* $\{1,...,n\}$ *into* k *parts* $A_1,...,A_k$*. Each pair of vertices in two different parts is joined with an edge with probability* p*, where the edge-choices are independent.*

The following lemma permits us to work in $P_{n,p}$ rather than $P_{n,M}$ and still be able to apply Theorem 9.3.

Lemma 9.5. *Consider any* $f(n) >> n^{-1/2}$*, any property* $\mathcal{E}$ *of pairs* (σ,G) *where* σ *is a* k*-colouring of* G*, and any constant* r*. Setting* $c = \frac{2k}{k-1}r$*, we have:*
If $\mathbf{Pr}(P_{n,p=c/n} \text{ has } \mathcal{E}) > 1 - e^{-2f(n)}$ *then* $\mathbf{Pr}(P_{n,M=rn} \text{ has } \mathcal{E}) > 1 - e^{-f(n)}$*.*

Proof Let E be the event that $P_{n,p=c/n}$ has exactly rn edges. Standard and easily derived facts about the binomial distribution imply $\mathbf{Pr}(E) = \Theta(n^{-1/2})$. Now,

$$\mathbf{Pr}(P_{n,M=rn} \text{ has } \overline{\mathcal{E}}) = \mathbf{Pr}(P_{n,p=c/n} \text{ has } \overline{\mathcal{E}}|E) \leq \frac{\mathbf{Pr}(P_{n,p=c/n} \text{ has } \overline{\mathcal{E}})}{\mathbf{Pr}(E)} < O(e^{-2f(n)}/n^{-1/2}) < e^{-f(n)}.$$

□

We define

$$c_k = \min_{y>0} \frac{ky}{(1-e^{-y})^{k-1}}.$$

For any $c > c_k$ we let $y_k(c)$ denote the largest solution to $c = \frac{ky}{(1-e^{-y})^{k-1}}$. Note that $c_k = \frac{2k}{k-1} r_k$. We define:

$$\lambda_k(c) = y_k(c)/c.$$

We say that v is an *ℓ-frozen variable of* (G, σ) if v is ℓ-frozen with respect to σ. So, roughly speaking, our goal is to prove that c_k is the threshold for $P_{n,p=c/n}$ to have a linear number of αn-frozen variables, and that the failure probability is $1 - e^{-f(n)}$ where $f(n)$ comes from Theorem 9.3.

10 Kempe cores

Given a k-colouring σ of a graph G, with colour classes $A_1, ..., A_k$, a *Kempe chain* is a component of the subgraph induced by two colour classes. Suppose C is a non-empty Kempe chain on colour classes A_i, A_j. Then exchanging the colours i, j on the vertices of C will result in a new k-colouring of G. Note that a single vertex of colour i will constitute a Kempe chain if it has no neighbours of colour j, for some $j \neq i$. Kempe chains were introduced by Kempe[31] in his work on the Four Colour Problem.

It is clear that a vertex that is in a Kempe chain of size at most ℓ is not ℓ-frozen. This inspires us to remove all "small" Kempe chains from our graph, in order to look for frozen vertices. A bit of thought will make it clear that w.h.p. most vertices in Kempe chains of size at most ℓ in the remaining graph are not ℓ-frozen either. This follows from branching properties of the random graph: if C is a small Kempe chain in the remaining graph, w.h.p. the small Kempe chains that were removed from the original graph each have at most one edge to C. Furthermore none of those chains adjacent to C are adjacent to each other. Thus we can flip the vertices on some subset of those chains without them interfering with each other, thus enabling C to be flipped. This inspires us to remove small Kempe chains iteratively.

Of course, we need to specify what we mean by "small". It turns out that w.h.p. there will be no Kempe chains of size between $O(\log n)$ and $\Theta(n)$; i.e. every Kempe chain will either be small or giant. But to be specific, and to strenghten "w.h.p." enough to apply Theorem 9.3, we will take small to mean: of size at most $g(n)$ for some $g(n) = o(n)$ to be specified later. Thus, we apply the following procedure:

Kempe-Strip
Input: a graph G and a k-colouring $\sigma = A_1, ..., A_k$ of G.
While there are any Kempe chains of size at most $g(n)$
Remove the vertices of one such Kempe chain from G.

The (possibly empty) Kempe core is what remains. Note that, as with most core stripping procedures, the output does not depend on the order in which we choose to remove Kempe chains. So the Kempe core is well-defined.

It is not surprising that any vertex in the Kempe core cannot have its colour changed by changing the vertices of a small Kempe chain. What is surprising is that, almost every vertex in the Kempe core cannot have its colour changed by changing a small subset of vertices which involve *more than two* colours.

To gain some intuition as to why this may be the case, note first that every very small subgraph, i.e. of size $O(1)$, is a tree. Then note that if we can change the colours of a tree to obtain another colouring, then that tree contains a subtree which is a Kempe chain. Thus, any changes of $O(1)$ vertices can be simulated by a sequence of Kempe-chain switches.

The following lemma is one of the main steps in this paper, and is proven in Section 13. (See also Lemma 11.3(a)).

Lemma 10.1. *For $k \geq 3$, and any $f(n) = o(n)$:*

(a) If $c < c_k$ then with probability at least $1 - e^{-f(n)}$, the Kempe core of $P_{n,p=c/n}$ has size $o(n)$.

(b) If $c > c_k$ then with probability at least $1 - e^{-f(n)}$, the Kempe core of $P_{n,p=c/n}$ has size $k\lambda_k(c) + o(n)$.

Remark: In fact, for $c < c_k$, w.h.p. the Kempe core of $P_{n,p=c/n}$ has size 0. But this statement fails with probability $1/\text{poly}(n)$.

A very standard argument shows that w.h.p. any Kempe core (in either model) must have at least linear size. Applied to the uniform model, we obtain:

Lemma 10.2. *For $k \geq 4$ and any constant $r > 0$, there is a constant $\epsilon > 0$ such that w.h.p. $U_{n,M=rn}$ has no non-empty Kempe core of size less than ϵn.*

The proof is at the end of this section. Lemmas 10.1, 10.2, Lemma 9.5 and Theorem 9.3 immediately yield:

Corollary 10.3. *For $k \geq k_0$:*

(a) If $r < r_k^f$ then w.h.p. the Kempe core of $U_{n,M=rn}$ has size zero.

(b) If $r > r_k^f$ then w.h.p. the Kempe core of $U_{n,M=rn}$ has size $\frac{(k-1)x_k(r)}{2r}n + o(n)$.

Proof Theorem 9.3 and Lemma 9.5 allow us to translate our results from $P_{n,p=c/n}$ to $U_{n,M=rn}$, with $r = \frac{k-1}{2k}c$. Note that $x(r) = y(c)$. So the corollary follows from Lemmas 10.1 and 10.2, and the fact that

$$k\lambda_k(c) = \frac{ky_k(c)}{c} = \frac{kx_k(r)}{2kr/(k-1)} = \frac{(k-1)x_k(r)}{2r}.$$

□

The remaining steps are to show that the frozen variables consist of the Kempe core, plus or minus $o(n)$ vertices. In Section 14, we show:

Lemma 10.4. *For $k \geq 3$, any $f(n) = o(n)$ and any $\epsilon > 0$: There exist constants T, Z such that with probability at least $1 - e^{-f(n)}$, all but ϵn of the vertices outside of the (possibly empty) Kempe core of $P_{n,p=c/n}$ are either (i) not T-frozen, or (ii) are within distance Z of a cycle with length less than Z.*

In Section 12 we show that almost all of the Kempe core is *internally rigid* in the following sense:

Lemma 10.5. *For $k \geq 3$, $c > c_k$, and any $f(n) = o(n)$, there exists constant $\alpha = \alpha(c,k) > 0$ such that with probability at least $1 - e^{-f(n)}$, the Kempe core K of $P_{n,p=c/n}$ has the following property:*
For all but $o(n)$ vertices $v \in K$, any k-colouring of K which differs from σ on v must differ from σ on at least $2\alpha n$ vertices of K.

The $o(n)$ term depends on $f(n)$. This internal rigidity is enough to imply that almost all vertices of the Kempe core are frozen:

Corollary 10.6. *For $k \geq 3$, $c > c_k$, and any $f(n) = o(n)$, there exists constant $\alpha = \alpha(c,k) > 0$ such that with probability at least $1 - e^{-f(n)}$: all but $o(n)$ vertices of the Kempe core K of $P_{n,p=c/n}$ are αn-frozen.*

Proof Lemma 10.5 says we have $\Theta \subseteq K$ with size $|K| - o(n)$ such that every $v \in \Theta$ has the property that any k-colouring of K which differs from σ on v must differ from σ on at least $2\alpha n$ vertices of K, and thus on at least $2\alpha n - o(n) > \alpha n$ vertices of Θ. Consider any sequence of k-colourings of K, $\sigma = \sigma_0, \sigma_1, ..., \sigma_t$ such that

(i) for all $v \in \Theta$ and $0 \leq i \leq t-1$, we have $\sigma_i(v) = \sigma(v)$.

(ii) for some $v \in \Theta$ we have $\sigma_t(v) \neq \sigma(v)$.

In other words, t is the first step where a member of Θ changes colour.

By (ii), σ_t must differ from σ on at least αn vertices of Θ. Thus by (i), σ_t must differ from σ_{t-1} on those same αn vertices. Therefore, $\sigma = \sigma_0, \sigma_1, ..., \sigma_t$ is not a αn-path. But if at least one vertex of Θ is not

αn-frozen, then there must be such a αn-path; consider the vertex $v \in \Theta$ whose colour can be changed by the shortest possible αn-path. So all of the vertices of Θ must be αn-frozen. □

This yields our main theorem:

Proof of Theorem 2.4: Corollary 10.3 establishes the location of the Kempe core threshold and the size of the Kempe core.

Theorem 9.3 and Lemma 9.5 allow us to translate our results from $P_{n,p=c/n}$ to $U_{n,M=rn}$, with $r = \frac{k-1}{2k}c$. Part (a.i) then follows from Corollary 10.6.

To obtain parts (a.ii) and (b), we apply Lemma 10.4. It is standard and straightforward to show that the expected number of vertices in $G_{n,M=cn}$ that are within distance Z of a cycle with length less than Z, is $O(1)$. So w.h.p. the number of such vertices in $U_{n,M=cn}$ is less than $\frac{\epsilon}{2}n$. Therefore, for all $\epsilon > 0$ there exists $T = O(1)$ such that w.h.p. all but $\frac{\epsilon}{2}n$ of the vertices outside the Kempe core of $U_{n,p=c/n}$ are T-frozen. Therefore for any $\omega(n)$ tending to ∞, w.h.p. $o(n)$ vertices outside of the Kempe core of $U_{n,p=c/n}$ are $\omega(n)$-frozen. □

We close this section with the proof of Lemma 10.2.

Proof of Lemma 10.2: Every Kempe core has minimum degree at least $k - 1 \geq 3$. Otherwise there will be a vertex v and colour α such that α does not appear on v, nor on any neighbour of v, and so v is a Kempe chain of size 1. It is well-known (see eg. [52]) that for any constant r, there is some $\epsilon > 0$ such that w.h.p. every subgraph of $G_{n,M=rn}$ with at most ϵn vertices has average degree less than 3. Therefore, a.s. $U_{n,M=rn}$ has no Kempe core of size at most ϵn. □

11 Kempe cores in the planted model

11.1 Properties of the Kempe core

Let K be the Kempe core of $P_{n,p=c/n}$ for some $c > c_k$, and for each $1 \leq i \leq k$, we let $K_i = K \cap A_i$ be the vertices of K with colour i. For each $i \neq j$, we let $K_{i,j}$ denote the bipartite subgraph of K induced by (K_i, K_j).

Lemma 11.1. *Consider any two connected bipartite graphs H, H', each with vertex set (K_i, K_j), and with $|E(H)| = |E(H')|$. Then $\mathbf{Pr}(K_{i,j} = H) = \mathbf{Pr}(K_{i,j} = H')$.*

Proof Consider any (G, σ) for which the procedure Kempe-Strip yields a Kempe core with $K_{i,j} = H$. Form G' by replacing the subgraph H in G with H'. Then applying Kempe-Strip to (G', σ) will yield a Kempe core with $K_{i,j} = H'$. Furthermore, G, G' arise with the same probability in $P_{n,p=c/n}$, since they have the same number of edges. Finally, every such H' arises from exactly one such H. This implies the lemma. □

Definition 11.2. *The* 2-core *of a graph is what remains after iteratively deleting any vertices of degree less than 2.*

Remark: It is easy to see that the order in which we delete vertices does not affect what remains at the end, so the 2-core is well-defined.

Recall from Section 4 that for any $c \geq c_k$ we let $y_k(c)$ denote the largest positive solution y to $c =$

$\frac{ky}{(1-e^{-y})^{k-1}}$, and that c_k is defined to be the minimum c such that $y_k(c)$ exists. We define:

$$\begin{aligned}
\lambda_k(c) &= y_k(c)/c \\
\xi_k(c) &= \frac{y_k(c)(1-e^{-y_k(c)}(1+y_k(c)))}{c(1-e^{-y_k(c)})} \\
\mu_k(c) &= \frac{y_k(c)e^{-y_k(c)}}{c(1-e^{-y_k(c)})}\sum_{i\geq 2}\frac{y_k(c)^i}{(i-1)!} \\
\tau_k(c) &= \frac{y_k(c)e^{-y_k(c)}}{c(1-e^{-y_k(c)})}\frac{y_k(c)^2}{2}
\end{aligned}$$

Lemma 11.3. *For any $c > c_k$ and any $f(n) = o(n)$, with probability at least $1 - e^{-3f(n)}$, we have that for every i, j, the subgraph induced by $K_{i,j}$ is connected and:*

(a) $|K_i| = \lambda_k(c)n + o(n)$;

(b) the 2-core of $K_{i,j}$ has $\xi_k(c)n + o(n)$ vertices in K_i and $\xi_k(c)n + o(n)$ vertices in K_j;

(c) the 2-core of $K_{i,j}$ has $\mu_k(c)n + o(n)$ edges;

(d) the 2-core of $K_{i,j}$ has $\tau_k(c)n + o(n)$ degree 2 vertices in K_i and $\tau_k(c)n + o(n)$ degree 2 vertices in K_j.

Remark: The $o(n)$ terms depends on $f(n)$.

We outline the proof of Lemma 11.3 in Section 13.

The branching parameters described in Section 6 concern (a) the proportion of non-2-core vertices in each K_i, and (b) the degree two vertices in the 2-core of $K_{i,j}$:

Lemma 11.4. *For every $c > c_k$, there is $\zeta = \zeta(c) > 0$ such that:*

(a) $1 - \frac{\xi_k(c)}{\lambda_k^c(c)} < \frac{1}{k-1}(1-\zeta)$;

(b) $\frac{2\tau_k(c)}{\mu_k^c(c)} < 1 - \zeta$.

Proof At $c = c_k$, $y = y_k(c)$ is the point that minimizes $h(y) = \frac{ky}{(1-e^{-y})^{k-1}}$. Setting $\frac{\partial}{\partial y}h(y) = 0$ yields:

$(1-e^{-y})^{k-1} = (k-1)ye^{-y}(1-e^{-y})^{k-2}$, which yields $e^y - 1 = (k-1)y$. Thus $\frac{e^y-1}{y} = k-1 > e-1$ and so $y > 1$. Thus $e^y > k-1$ and so $e^y - 1$ grows faster than $(k-1)y$ for $y \geq y_k(c)$. Since $y_k(c)$ increases with c, we have that for every $c > c_k$:

$$e^{y_k(c)} - 1 > (k-1)y_k(c).$$

It will suffice to prove that the LHS is less than the RHS in (a,b), since they do not change with n.

Part (a):

$$\frac{\xi(c)}{\lambda_k^c(c)} = \frac{1-e^{-y_k(c)}(1+y_k(c))}{1-e^{-y_k(c)}} = \frac{e^{y_k(c)}-1-y_k(c)}{e^{y_k(c)}-1} > \frac{e^{y_k(c)}-1-\frac{1}{k-1}(e^{y_k(c)}-1)}{e^{y_k(c)}-1} = \frac{k-2}{k-1}.$$

This implies that the LHS of (a) is less than the RHS, as required.

Part (b):

$$\frac{2\tau_k(c)}{\mu_k(c)} = \frac{y_k(c)^2}{\sum_{i\geq 2}\frac{y_k(c)^i}{(i-1)!}} = \frac{y_k(c)}{\sum_{i\geq 1}\frac{y_k(c)^i}{i!}} = \frac{y_k(c)}{e^{y_k(c)}-1} < \frac{1}{k-1} < 1,$$

as required. □

12 The Kempe core is mostly frozen

In this section, we prove Lemma 10.5. Recall that we are working in the $P_{n,p}$ model. So we have a uniformly random partition σ of the vertices into $A_1, ..., A_k$, and a graph G formed by selecting each of the potential edges between different parts with probability $p = c/n$. Our focus will be on the Kempe core, K, of (G, σ).

Definition 12.1. *A* Δ-set *is the symmetric difference of σ and some other k-colouring of the Kempe core, K. Specifically, given such a colouring σ', the set of vertices $u \in K$ with $\sigma(u) \neq \sigma'(u)$ is a Δ-set, which we sometimes denote by $\sigma\Delta\sigma'$.*

Note that "$v \in \sigma\Delta\sigma'$" means the same thing as "$\sigma(v) \neq \sigma'(v)$".

To prove Lemma 10.5 we will show that, with sufficiently high probability, the union of all Δ-sets of size at most $2\alpha n$ has size $o(n)$. So we define:

Definition 12.2. *A* D-set *is the union of Δ-sets.*

Lemma 12.3. *For any $f(n) = o(n)$, there exists $g(n) = o(n)$ such that with probability at least $1 - e^{-2f(n)}$, no D-set has size between $g(n)$ and $2\alpha n$.*

This yields Lemma 10.5 as follows:

Proof of Lemma 10.5: Let $S_1, ..., S_t$ be all the Δ-sets of size less than $2\alpha n$. Note that for every $v \notin \cup_{i=1}^t S_i$, any k-colouring of the Kempe core which differs from σ on v must differ from σ on at least $2\alpha n$ vertices. So it suffices to prove that $|\cup_{i=1}^t S_i| = o(n)$.

Lemma 12.3 implies that for each $1 \leq i \leq t$ we have $|S_i| \leq g(n)$. So by induction, we have $|\cup_{i=1}^j S_i| \leq |\cup_{i=1}^{j-1} S_i| + |S_j| \leq 2g(n)$ and hence by Lemma 12.3 must be at most $g(n) = o(n)$ since $\cup_{i=1}^j S_i$ is a D-set. Therefore $|\cup_{i=1}^t S_i| \leq g(n) = o(n)$, as required. □

The remainder of this section is devoted to the proof of Lemma 12.3, which is the main lemma in this paper. That proof appears at the end of Subsection 12.2.

12.1 The structure of Δ-sets

To prove Lemma 12.3 we first study the structure of Δ-sets and D-sets. When we say the 2-core of a Δ-set or a D-set, we mean the 2-core of the subgraph of the Kempe core induced by that set. Similarly, when we say a component of a Δ-set or a D-set, we mean a component of the subgraph of the Kempe core induced by that set.

Recalling Lemma 11.3, we suppose we have a Kempe core K satisfying:

Property 12.4. *Each $K_{i,j}$ is connected with a non-empty 2-core, and that 2-core is not a cycle.*

We start with the key observation about Δ-sets:

Proposition 12.5. *Let u be any vertex in a D-set Φ, and let $\sigma\Delta\sigma' \subseteq \Phi$ be a Δ-set containing u. Then every neighbour of u in $K_{\sigma(u),\sigma'(u)}$ is also in $\sigma\Delta\sigma'$ and hence in Φ.*

Proof Every neighbour w of u in $K_{\sigma(u),\sigma'(u)}$ has $\sigma(w) = \sigma'(u)$. Since σ' is a proper colouring, we cannot have $\sigma'(w) = \sigma'(u)$. Therefore $\sigma'(w) \neq \sigma(w)$. □

Lemma 12.6. *Every component of a Δ-set or D-set has a non-empty 2-core.*

Proof If a component does not have a 2-core, then it is a tree. If a D-set is a tree, then so is any Δ-set that it contains. Consider a Δ-set $\sigma\Delta\sigma'$ that induces a tree. We direct the edges of that tree as follows: each u has an edge directed to every neighbour that it has in $K_{\sigma(u),\sigma'(u)}$; by Proposition 12.5, all such neighbours must be in the tree.

Note that an edge uv will be directed in both directions iff $\sigma(u) = \sigma'(v)$ and $\sigma(v) = \sigma'(u)$; contract all such edges. The contracted tree is a tree, and each edge is directed in exactly one direction. So there

must be a node which has no edges directed out of it. Our contraction rule implies that every vertex u contracted into that node has $(\sigma(u), \sigma'(u)) = (a, b)$ or (b, a) for some pair of colours a, b. Furthermore, u has no neighbours in $K_{\sigma(u),\sigma'(u)}$ that were not contracted into the node, else this would have produced a directed edge out of the node. Therefore, the vertices contracted into that node are a component of $K_{a,b}$. But since they form a tree, this violates Proposition 12.4. □

Note the following simple facts:

Proposition 12.7. *If a graph H is connected, then the 2-core of H is empty or connected.*

Proof Strip to the 2-core by repeatedly removing vertices of degree 1. The removal of a degree 1 vertex cannot disconnect a graph. □

Definition 12.8. *We say that a D-set or Δ-set is* complex *if its 2-core does not have any components that are cycles. We say that a D-set or Δ-set is* cyclic *if its 2-core is a cycle.*

Lemma 12.9. *Every D-set is the union of D-sets where at most one is complex and the rest are cyclic.*

Proof Each component of a Δ-set is a Δ-set, since you can switch the colours of the vertices in a Δ-set one component at a time. It follows that each component of a D-set is a D-set, as it is the union of components of Δ-sets, and hence is the union of Δ-sets. By Proposition 12.7, each component of a D-set has a connected 2-core and hence is either cyclic or complex. The lemma follows, noting that the union of the complex components of a D-set is a single complex D-set. □

We now turn our attention to the structure of the vertices outside the 2-core of a graph:

Definition 12.10. *Consider any graph H such that every component of H has a non-empty 2-core. The edges not in the 2-core of H form a forest. We call a tree of that forest a* pendant tree*. By Proposition 12.7, the 2-core of each component is connected. It follows that each pendant tree T contains exactly one vertex of the 2-core; that vertex is the* vertex of attachment *for T. We consider a pendant tree to be rooted at its vertex of attachment; in particular, the parent of a vertex u not in the 2-core is its unique neighbour on the path from u to the 2-core.*

Lemma 12.6 implies that we can apply Definition 12.10 to the graph induced by any D-set, Φ. So for every vertex v not in the 2-core of Φ, we can talk about its parent in Φ. By Proposition 12.7, we can also talk about the parent of any non-2-core vertex in $K_{i,j}$.

Lemma 12.11. *Consider any connected D-set Φ. Consider any non-2-core vertex $u \in \Phi$ and let w be the parent of u in Φ. Then for every Δ-set, $\sigma\Delta\sigma' \subseteq \Phi$ which contains u, we have:*

(a) $\sigma'(u) = \sigma(w)$;

(b) u is not in the 2-core of $K_{\sigma(u),\sigma'(u)}$;

(c) w is the parent of u in $K_{\sigma(u),\sigma'(u)}$.

Proof Let T be the pendant tree of Φ containing u. We proceed by induction.

Note that since $K_{\sigma(u),\sigma'(u)}$ is connected and has linear size (by Property 12.4), u has at least one neighbour in $K_{\sigma'(u)}$.

Base case: u is a leaf of T. Then w is the only neighbour of u in Φ, and so by Proposition 12.5, w must be the only neighbour of u in $K_{\sigma(u),\sigma'(u)}$. Parts (b,c) now follow.

Now suppose (a,b,c) hold for every child of u in T. Consider any child w' of u with $\sigma(w') = \sigma'(u)$. Then by the inductive hypothesis, $\sigma'(w') = \sigma(u)$ and u is the parent of w' in $K_{\sigma(u),\sigma'(u)}$. By Proposition 12.5, every neighbour of u in $K_{\sigma(u),\sigma'(u)}$ must be in T. So u has at most one non-child neighbour, w, in $K_{\sigma(u),\sigma'(u)}$. Parts (a,b,c) now follow for u. □

Consider the digraph Υ defined as follows: for any $u, v \in K$, we have the edge $u \to v$ iff there is some $K_{i,j}$ in which v is a non-2-core vertex, and u is the parent of v in the pendant tree of $K_{i,j}$ containing v.

Definition 12.12. *For each vertex $u \in K$, $T^+(u)$ is the set of vertices that can be reached from u in Υ; i.e. the set containing u and all vertices $w \in K$ for which, in at least one $K_{i,j}$, w is a non-2-core vertex and u is on the unique path from w to the 2-core.*

Lemma 12.13. *For any D-set Φ with 2-core H, we have $\Phi \subseteq \cup_{u \in H} T^+(u)$.*

Proof Consider any pendant tree T of Φ. Lemma 12.11 implies that if we direct the edges of T away from its vertex of attachment, u, then the directed edges will all be in Υ. Therefore $T \subseteq T^+(u)$. This, along with the fact that every $v \in H$ is in $T^+(v)$, implies the lemma. □

Definition 12.14. *A* 2-path *in a Δ-set $\sigma\Delta\sigma'$ is a path $u_0, ..., u_x$ in the 2-core of $\sigma\Delta\sigma'$ such that*

(a) $x \geq 1$;

(b) each u_i has degree 2 in the 2-core of $\sigma\Delta\sigma'$;

(c) either
Type A: *every u_i is in the 2-core of $K_{\sigma(u_i),\sigma'(u_i)}$; or*
Type B: *every u_i is not in the 2-core of $K_{\sigma(u_i),\sigma'(u_i)}$ and, for $0 \leq i \leq x-1$, its parent in $K_{\sigma(u_i),\sigma'(u_i)}$ is u_{i+1}.*

We call u_0, u_x the endpoints*, and $u_1, ..., u_{x-1}$ the* internal vertices.

A 2-path in a D-set Φ is a path $u_0, ..., u_x$ in the 2-core of Φ such that

(a) each u_i has degree 2 in the 2-core of Φ;

(b) $u_0, ..., u_x$ is a 2-path in some Δ-set contained in Φ.

Let H be the 2-core of a D-set Φ. Consider any path $W, v_0, ..., v_r, Y$, $r \geq 1$, in H where each v_i has degree exactly 2 in H and W, Y each have degree at least 3 in H. It will be convenient to set $v_{-1} = W$ and $v_{r+1} = Y$. The next several lemmas concern this path.

Lemma 12.15. *Consider any $0 \leq i \leq r$ and any Δ-set $\sigma\Delta\sigma' \subset \Phi$ that contains v_i. If v_i is in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$ then:*

(a) $v_{i-1}, v_{i+1} \in \sigma\Delta\sigma'$;

(b) $\sigma(v_{i-1}) = \sigma(v_{i+1}) = \sigma'(v_i)$;

(c) v_{i-1}, v_{i+1} are both in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$.

Proof Since v_i is in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$, v_i has at least two neighbours in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$. We will argue that those neighbours must also be in the 2-core of Φ. Hence, they must be v_{i-1}, v_{i+1}, thus establishing (c). Suppose $w \in K_{\sigma'(v_i)}$ is a neighbour of v_i that is not in the 2-core of Φ. By Proposition 12.5, $w \in \sigma\Delta\sigma'$. Since v_i is in the 2-core of Φ, v_i must be the parent of w in Φ. By Lemma 12.11(c), this implies that $\sigma(v_i) = \sigma'(w)$ and w is not in the 2-core of $K_{\sigma(w),\sigma'(w)} = K_{\sigma(v_i),\sigma'(v_i)}$. Therefore every neighbour of v_i in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$ must also be in the 2-core of Φ. Thus, we have part (c).

Since v_{i-1}, v_{i+1} are both in $K_{\sigma(v_i),\sigma'(v_i)}$, and they cannot have the same colour as v_i in σ, we have part (b). Propostion 12.5 gives part (a). □

Lemma 12.16. *If $u_0, ..., u_r$ is a Type A 2-path in the 2-core of a D-set, then there are colours a, b such that $(\sigma(u_i), \sigma'(u_i)) = (a, b)$ for even i, and $(\sigma(u_i), \sigma'(u_i)) = (b, a)$ for odd i; i.e. the sequences $\sigma(u_i)$ and $\sigma'(u_i)$ both alternate over the same two colours. Furthermore, if v, w are the non-path neighbours of u_0, u_x, respectively, then $\sigma(v) = \sigma'(u_0)$ and $\sigma(w) = \sigma'(u_x)$.*

Proof This follows immediately from Lemma 12.15(b). □

Lemma 12.17. *Consider any $0 \le i \le r$ and any Δ-set $\sigma\Delta\sigma' \subset \Phi$ that contains v_i. If v_i is not in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$ then*

(a) one of the two neighbours of v_i is its parent in $K_{\sigma(v_i),\sigma'(v_i)}$;

(b) if that neighbour is not W or Y, then either $v_i, v_{i+1}, ..., v_r$ or $v_i, v_{i-1}, ..., v_0$ is a Type B 2-path.

Proof $K_{\sigma(v_i),\sigma'(v_i)}$ is connected with a non-empty 2-core (by Property 12.4). Since v_i is not in that 2-core, it must have a parent; let w be the parent of v_i in $K_{\sigma(v_i),\sigma'(v_i)}$. By Proposition 12.5, $w \in \Phi$. If w is not in the 2-core of Φ, then because v is a neighbour of w and v is in the 2-core of Φ, v must be the parent of w in Φ. By Lemma 12.11(c), this implies that $\sigma(v_i) = \sigma'(w)$ and v is the parent of w in $K_{\sigma(w),\sigma'(w)}$. But now $\sigma(w) = \sigma'(v_i)$ and $\sigma'(w) = \sigma(v_i)$ so $K_{\sigma(w),\sigma'(w)} = K_{\sigma(v_i),\sigma'(v_i)}$; thus v is the parent of w and w is the parent of v in the same graph - contradiction. Therefore, w is in the 2-core of Φ, and so w must be one of the only two neighbours of v_i in the 2-core of Φ. This establishes part (a).

WLOG, assume w, the parent of of v_i in $K_{\sigma(v_i),\sigma'(v_i)}$, is v_{i+1}. Thus $v_{i+1} \in \sigma\Delta\sigma'$; we will show that v_{i+1} is not in the 2-core of $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})}$, which will allow us to apply the same argument again to v_{i+1}.

Case 1: $\sigma'(v_{i+1}) \neq \sigma(v_i)$. Then v_{i+1} has at most one neighbour of colour $\sigma'(v_{i+1})$ in the 2-core of Φ. As argued above, Lemma 12.11 implies that any neighbour u of v_{i+1} that is not in the 2-core of Φ cannot be in the 2-core of $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})}$. So v_{i+1} has at most one neighbour in the 2-core of $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})}$, and so v_{i+1} is not in the 2-core of $K_{\sigma(v_{i+1},\sigma'(v_{i+1})}$.

Case 2: $\sigma'(v_{i+1}) = \sigma(v_i)$. Then $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})} = K_{\sigma(v_i),\sigma'(v_i)}$. Since v_i is not in the 2-core of $K_{\sigma(v_i),\sigma'(v_i)}$, it is not in the 2-core of $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})}$. So again, v_{i+1} has at most one neighbour of colour $\sigma'(v_{i+1})$ in the 2-core of Φ, and the argument proceeds as in Case 1.

Thus, we can repeat the above argument to show that part (a) holds for v_{i+1}. It is not possible for v_i to be the parent of v_{i+1} in $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})}$; that would require $\sigma'(v_{i+1}) = \sigma(v_i)$, and so $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})} = K_{\sigma(v_i),\sigma'(v_i)}$, and so v_i would be the parent of v_{i+1} in the same graph in which v_{i+1} is the parent of v_i. Therefore, v_{i+2} must be the parent of v_{i+1} in $K_{\sigma(v_{i+1}),\sigma'(v_{i+1})}$. Continuing inductively down the path establishes part (b). □

Lemma 12.18. *$v_0, ..., v_r$ can be split into at most three pieces, each of which either has exactly one vertex or is a 2-path.*

Proof If $r \le 2$ then it is trivial. So we assume $r \ge 3$.

Case 1: There is some v_i meeting the conditions of Lemma 12.15. Let $a = \sigma(v_i)$ and $b = \sigma'(v_i)$. Let j_1 be the largest $0 \le j_1 < i$ such that $(\sigma(v_{j_1}), \sigma'(v_{j_1}))$ is not either (a, b) or (b, a); if no such j_1 exists then we set $j_1 = -1$. Similarly, let j_2 be the largest $i < j_2 \le r$ such that $(\sigma(v_{j_2}), \sigma'(v_{j_2}))$ is not either (a, b) or (b, a); if no such j_2 exists then we set $j_2 = r + 1$.

For all $j_1 + 1 \le j \le j_2 - 1$, $(\sigma(v_j), \sigma'(v_j))$ is either (a, b) or (b, a). This allows us to apply Lemma 12.15 inductively from v_i to v_{j_1+1} and from v_i to v_{j_2-1} and show that each such v_j is in the 2-core of $K_{a,b}$ and is in $\sigma\Delta\sigma'$. Therefore the subpath $v_{j_1+1}, ..., v_{j_2-1}$ either has exactly one vertex (v_i) or is a Type A 2-path.

If $j_2 \le r$ then Lemma 12.15, applied to v_{j_2-1}, implies that $v_{j_2} \in \sigma\Delta\sigma'$ and $\sigma(v_{j_2}) = \sigma'(v_{j_2-1})$. If v_{j_2} were in the 2-core of $K_{\sigma(v_{j_2}),\sigma'(v_{j_2})}$ then Lemma 12.15 applied to v_{j_2} would imply that $\sigma'(v_{j_2}) = \sigma(v_{j_2-1})$, and so $(\sigma(v_{j_2}), \sigma'(v_{j_2}))$ is either (a, b) or (b, a) thus contradicting our choice of j_2. So we can apply Lemma 12.17 to v_{j_2} to show that the subpath $v_{j_2}, ..., v_r$ either has exactly one vertex (v_r) or is a Type B 2-path.

Similarly, if $j_1 \ge 0$ then the subpath $v_{j_1}, ..., v_0$ either has exactly one vertex or is a Type B 2-path. This provides our split into at most three pieces.

Case 2: No v_i meets the conditions of Lemma 12.15. Recall we assume that $r \ge 3$, and pick some $1 \le \ell \le r - 1$. Let $\sigma\Delta\sigma' \subseteq \Phi$ be a Δ-set containing v_ℓ. Since we are in Case 2, v_ℓ is not in the 2-core of $K_{\sigma(v_\ell),\sigma'(v_\ell)}$. So Lemma 12.17 implies that v_ℓ lies in a Type B 2-path extending to either v_0 or v_r; WLOG assume it is v_r. Let $j \le \ell$ be the smallest value such that $v_j, ..., v_r$ is a Type B 2-path. If $j = 0$ then we have one piece. If $j \ge 1$ then, since we are in Case 2, v_{j-1} is not in the 2-core of $K_{\sigma(v_{j-1}),\sigma'(v_{j-1})}$. By Lemma 12.17, $v_{j-1}, v_{j-2}, ..., v_0$ either has exactly one vertex ($j - 1 = 0$) or is a Type B 2-path. Thus we can split into two pieces. □

Definition 12.19. *We define $\mathcal{P}(\Phi)$ to be a vertex-disjoint collection of 2-paths in the 2-core of Φ such that: For every path $W, v_0, ..., v_r, Y$ in the 2-core of Φ, where each v_i has degree exactly 2 in the 2-core of Φ and W, Y each have degree at least 3 in the 2-core of Φ, we can split $v_0, ..., v_r$ into at most three pieces, each of which either has exactly one vertex or is a member of $\mathcal{P}(\Phi)$.*

Lemma 12.18 implies that $\mathcal{P}(\Phi)$ exists. $\mathcal{P}(\Phi)$ might not be uniquely defined. It is possible that there are two different ways to partition some $v_0, ..., v_r$ as in Lemma 12.18, thus yielding different choices for $\mathcal{P}(\Phi)$. If there are multiple choices for $\mathcal{P}(\Phi)$ then we arbitrarily specify one of them.

We partition the vertices of the 2-core of any D-set Φ as follows:

- $V_1(\Phi)$ - the internal vertices of the 2-paths in $\mathcal{P}(\Phi)$;
- $V_2(\Phi)$ - the vertices of the 2-core of Φ that are not in $V_1(\Phi)$.

Lemma 12.20. *For any complex D-set Φ with $|\mathcal{P}(\Phi)| = t$, the 2-core of Φ has at least $\frac{101}{100}|V_2(\Phi)| - t$ edges with both endpoints in $V_2(\Phi)$.*

Proof Let H be the 2-core of Φ. Form H' by contracting every 2-path $u_0, ..., u_x$ in $\mathcal{P}(\Phi)$ into a single edge (u_0, u_x). Since no component of H is a cycle (as Φ is complex), every degree 2 vertex of H lies in a path $W, v, ..., v_r, Y$ as in Definition 12.19. Every such path is contracted into a path with at most 4 degree two vertices. Therefore, H' does not contain any path v_0, v_1, v_2, v_3, v_4 of five degree 2 vertices. From that, it is easy to argue that H' has at least $\frac{101}{100}|V(H_1)|$ edges (this also follows from Lemma 11 of [47]). The lemma now follows since $V(H') = V_2(\Phi)$, there are exactly t contracted edges in H', and each of the $\frac{101}{100}|V_2(\Phi)| - t$ non-contracted edges is an edge of H. □

We close this section by determining the structure of cyclic D-sets.

Lemma 12.21. *Every cyclic D-set is a cyclic Δ-set.*

Proof Let Φ be any cyclic D-set, and let Φ' be any Δ-set contained in Φ. Because the 2-core of Φ is a cycle, it follows that Φ' must contain all of that cycle, otherwise the 2-core of Φ' would be empty, contradicting Lemma 12.6. So Φ is equal to Φ' plus, at most, some subtrees of the pendant trees. It is straightforward to argue, using Lemma 12.11, that adding those pendant trees to Φ' will create another Δ-set. □

Lemma 12.22. *If $u_1, ..., u_r$ is the cycle forming the 2-core of a cyclic Δ-set, then (after possibly reversing the order of the labels): Every u_i is not in the 2-core of $K_{\sigma(u_i),\sigma'(u_i)}$ and its parent is u_{i+1} (addition mod r).*

Thus, we can view this 2-core as the cycle analogue of a Type B 2-path.

Proof If at least one u_i is not in the 2-core of $K_{\sigma(u_i),\sigma'(u_i)}$, then the same reasoning as in the proof of Lemma 12.17(a) implies that its parent in $K_{\sigma(u_i),\sigma'(u_i)}$ is either u_{i-1} or u_{i+1}; WLOG assume it is u_{i+1}. The same reasoning as in the proof of Lemma 12.17(b) implies that u_{i+1} cannot be in the 2-core of $K_{\sigma(u_{i+1}),\sigma'(u_{i+1})}$. So we can repeat the argument inductively around the cycle to prove that the lemma holds.

The only other case is if every u_i is in the 2-core of $K_{\sigma(u_i),\sigma'(u_i)}$. The same reasoning as in the proof of Lemma 12.16 implies that every vertex u_j has $(\sigma(u_j), \sigma'(u_j)) = (a, b)$ or (b, a) for the same two colours (b, a). Furthermore we find that each u_j has no other neighbours in the 2-core of $K_{a,b}$, other than its neighbours in the cycle. It follows that this cycle is the 2-core of $K_{a,b}$, contradicting Property 12.4. □

Note that Lemmas 12.21, 12.22, and 12.11 give a very good description of any cyclic D-set.

12.2 A first moment bound for D-sets

We will bound the expected number of D-sets in terms of various size-parameters. We will focus on the 2-cores of the D-sets.

Let Φ be a complex D-set, and set:

- $a = |V_2(\Phi)|$
- $t = |\mathcal{P}(\Phi)|$
- $j_1, ..., j_t$ are the number of internal vertices in the 2-paths of $\mathcal{P}(\Phi)$
- $J = j_1 + ... + j_t$

Let $X_{a,J}$ denote the number of 2-cores of D-sets with parameters a, J. We bound $E(X_{a,J})$ for all $a + J < 2\alpha n$ as follows:

First we choose the a vertices of V_2. We will overcount by choosing from all of $\{1, ..., n\}$ rather than from just the Kempe core. So there are $\binom{n}{a}$ choices. We also choose a set $\mathcal{E}$ of $\frac{101}{100}a - t$ edges within V_2; there are $\binom{\binom{a}{2}}{\frac{101}{100}a-t}$ choices for $\mathcal{E}$.

Next we choose the value of t. Since the 2-paths of $\mathcal{P}(\Phi)$ are vertex-disjoint and each has two endpoints, we have $t \le \frac{a}{2}$.

Then we choose, from amongst the vertices of V_2, the endpoints (v_i, w_i) of each of the 2-paths $P_1, ..., P_t \in \mathcal{P}(\Phi)$. The number of choices is at most $\binom{a}{t,t,a-2t}t!$.

We define the following events:

- E_1 - the event that the statements of Lemma 11.3(a,b,c,d) hold.
- E_2 - the event that all the edges of $\mathcal{E}$ are present in $P_{n,p}$.
- E_3 - the event that each pair (v_i, w_i) is joined by a 2-path.

For a random variable X and an event E, we use $X \wedge E$ to denote the variable that is equal to X if E holds and 0 if E does not hold. We will actually bound $E(X_{a,J} \wedge E_1)$, recalling from Lemma 11.3 that $\mathbf{Pr}(E_1) \ge 1 - e^{-3f(n)}$.

We begin by noting that, since $t \le \frac{a}{2}$, we have $\frac{101}{100}a - t > \frac{a}{2}$. This yields:

$$\mathbf{Pr}(E_2) \times \binom{\binom{a}{2}}{\frac{101}{100}a - t} \le \binom{\binom{a}{2}}{\frac{101}{100}a - t}\left(\frac{c}{n}\right)^{\frac{101}{100}a-t} \le \left(\frac{e\frac{a^2}{2}}{\frac{101}{100}a - t}\frac{c}{n}\right)^{\frac{101}{100}a-t} < \left(\frac{eca}{n}\right)^{\frac{101}{100}a-t}. \tag{2}$$

In Section 12.3, we will prove:

Lemma 12.23. *There is a constant $R = R(c, k)$ such that if $a + J < 2\alpha n$ then*

$$\mathbf{Pr}(E_3 \wedge E_1 | E_2) < R^a(1 - \frac{\zeta}{2})^J \left(\frac{1}{n}\right)^t.$$

This yields that for $a + J < 2\alpha n$:

$$\begin{aligned}
E(X_{a,J} \wedge E_1) &\le \sum_{t, j_1+...+j_t=J} \binom{n}{a}\binom{a}{t, t, a-2t} t! \left(\frac{eca}{n}\right)^{\frac{101}{100}a-t} R^a(1-\zeta)^J \left(\frac{1}{n}\right)^t \\
&< \sum_{t \ge 0} \left(\frac{en}{a}\right)^a \frac{a!}{t!(a-2t)!}\left(\frac{a}{n}\right)^{\frac{101}{100}a-t} \left(R(ec)^{1.01}\right)^a \left(\frac{1}{n}\right)^t \sum_{j_1+...+j_t=J}(1-\zeta)^J \\
&< \left(Z_1\frac{a}{n}\right)^{\frac{a}{100}} \sum_{t \ge 0} \frac{a^t}{t!} \sum_{j_1+...+j_t=J}(1-\zeta)^J \qquad \text{for some constant } Z_1 = Z_1(c,k) > 0.
\end{aligned}$$

The number of choices for $j_1, ..., j_t \ge 0$ that sum to J is $\binom{J+t-1}{t-1}$. It is straightforward to verify that there is a constant $Z_2 = Z_2(\zeta) = Z_2(c, k) > 1$ such that for any t and $J \ge Z_2 t$, $\binom{J+t-1}{t-1}(1-\frac{z}{2})^J$ is monotone decreasing as J increases. Thus, for $J \ge Z_2 t$, we have $\binom{J+t-1}{t-1}(1-\frac{z}{2})^J < \binom{Z_2 t+t-1}{t-1}(1-\frac{\zeta}{2})^{Z_2 t} < \binom{Z_2 t+t}{t}$, and

for $J < Z_2 t$, we have $\binom{J+t-1}{t-1}(1-\frac{z}{2})^J < \binom{J+t-1}{t-1} < \binom{Z_2 t+t}{t}$. This, along with the bound $1-\zeta < (1-\frac{\zeta}{2})^2$, implies:

$$\sum_{j_1+\ldots+j_t=J} (1-\zeta)^J < \binom{J+t-1}{t-1}(1-\frac{z}{2})^{2J} < \binom{Z_2 t+t}{t}(1-\frac{\zeta}{2})^J < (e(Z_2+1))^t(1-\frac{\zeta}{2})^J. \tag{3}$$

Thus, for $a + J < 2\alpha n$, we have:

$$\begin{aligned} E(X_{a,J} \wedge E_1) &< \left(Z_1 \frac{a}{n}\right)^{\frac{a}{100}} (1-\frac{\zeta}{2})^J \sum_{t\geq 0} \frac{(ea(Z_2+1))^t}{t!} \\ &= \left(Z_1 \frac{a}{n}\right)^{\frac{a}{100}} (1-\frac{\zeta}{2})^J e^{ea(Z_2+1)} \\ &< \left(Z \frac{a}{n}\right)^{\frac{a}{100}} (1-\frac{\zeta}{2})^J \qquad \text{for some constant } Z = Z(c,k) > 0 \\ &< (1-\frac{\zeta}{2})^{a+J}, \end{aligned} \tag{4}$$

by applying $a < 2\alpha n$ and taking $\alpha = \alpha(c,k)$ to be sufficiently small that $2Z\alpha < (1-\frac{\zeta}{2})^{100}$.

A standard straightforward argument uses the second last line above to obtain $\mathbf{Exp}(\sum_{a=1}^{\infty}\sum_{J\geq 0} X_{a,J} \wedge E_1) = o(1)$. Since E_1 holds a.s., and since every non-cyclic Δ-set contains a non-empty 2-core (and hence corresponds to $a > 0$) this yields

Lemma 12.24. *W.h.p. the Kempe core of $P_{n,p=c/n}$ has no complex Δ-sets of size less than αn.*

However, the failure probability that we obtain is $1/\mathrm{poly}(n)$ and so we cannot apply Theorem 9.3 to transfer Lemma 12.24 to the $U_{n,M}$ model. But applying Markov's Inequality to our bound on $E(X_{a,J} \wedge E_1)$ does yield that for any $g(n) = o(n)$, the probability that the Kempe core of $P_{n,p=c/n}$ has a complex D-set with $g(n) \leq a + J < 2\alpha n$ is at most $O(g(n)) \times (1-\frac{\zeta}{2})^{g(n)} + \mathbf{Pr}(\overline{E_1})$. (See the proof of Lemma 12.3 below.)

In order to strengthen this statement to obtain Lemma 12.3, we must account for the non-2-core vertices. To do so, we recall Lemma 12.13, let $u_1, ..., u_{a+J}$ be the vertices of the 2-core of Φ, and define:

- $L = |\cup_{i=1}^{a+J} T^+(u_i)|$.

Let $X_{a,J,L}$ denote the number of 2-cores of D-sets with parameters a, J, L. We will extend the analysis above to bound $E(X_{a,J,L} \wedge E_1)$.

In addition to all the counting described above, we choose L, and we define the event:

- E_4 - the event that $|\cup_{i=1}^{a+J} T^+(u_i)| = L$.

In Section 12.3, we extend Lemma 12.23 to prove:

Lemma 12.25. *There are constants $R = R(c,k)$ and $\rho = \rho(c,k) > 0$ such that for any $2\alpha n > L > \frac{4}{\zeta}(a+J)$:*

$$\mathbf{Pr}(E_4 \wedge E_3 \wedge E_1 | E_2) < (1-\rho)^L \times R^a (1-\frac{\zeta}{2})^J \left(\frac{1}{n}\right)^t.$$

Adding these ingredients to our derivation of (4) yields that for $a + J + L < 2\alpha n$:

For $L > \frac{4}{\zeta}(a+J)$:

$$\begin{aligned} E(X_{a,J,L} \wedge E_1) &\leq \sum_{t,j_1+\ldots+j_t=J} \binom{n}{a}\binom{a}{t,t,a-2t} t! \left(\frac{eca}{n}\right)^{\frac{101}{100}a-t} R^a (1-\zeta)^J \left(\frac{1}{n}\right)^t (1-\rho)^L \\ &< (1-\frac{\zeta}{2})^{a+J}(1-\rho)^L. \end{aligned} \tag{5}$$

This will be sufficient to bound the complex D-sets. For the others, we require:

Lemma 12.26. *For any $f(n) = o(n)$, there exists $g(n) = o(n)$ such that with probability at least $1 - e^{-2f(n)}$, the total number of vertices on all cyclic D-sets is at most $g(n)$.*

We prove this lemma in Section 12.3. We close this section with:

Proof of Lemma 12.3: Note that $L \geq |\Phi|$ by Lemma 12.13. For any Q: By (4), $\mathbf{Exp}(\sum_{a+J\geq Q} X_{a,J} \wedge E_1) < O(Q) \times (1 - \frac{\zeta}{2})^Q$. Therefore, the probability that there is a complex D-set with $a + J \geq Q$ is at most $O(Q) \times (1 - \frac{\zeta}{2})^Q + \mathbf{Pr}(\overline{E_1})$. This is also a bound on the probability that there is a complex D-set with $L \leq \frac{4}{\zeta}(a + J)$ and $L \geq \frac{\zeta}{4}Q$.

Similarly, (5) implies that the probability that there is a complex D-set with $L > \frac{4}{\zeta}(a + J)$ and $L \geq Q$ is at most $(1 - \rho)^Q$. These two bounds, along with the bound $\mathbf{Pr}(\overline{E_1}) < e^{-3f(n)}$ from Lemma 11.3 now imply Lemma 12.3 for an appropriate choice of $g(n)$. □

12.3 Some deferred proofs

Lemma 12.23 *There is a constant $R = R(c, k)$ such that if $a + J < 2\alpha n$ then $\mathbf{Pr}(E_3 \wedge E_1 | E_2) < R^a(1 - \frac{\zeta}{2})^J \left(\frac{1}{n}\right)^t$.*

Proof At this point, we have exposed that the fewer than $2a$ edges of $\mathcal{E}$ are present amongst the a vertices of V_2. Next, we will expose the values of the parameters bounded by Lemma 11.3. If E_1 holds then for each i, j:

- $|K_i| = \lambda_k^c(c)n + o(n)$;
- the 2-core of $K_{i,j}$ has $\xi_k(c)n + o(n)$ vertices in K_i and $\xi_k(c)n + o(n)$ vertices in K_j;
- the 2-core of $K_{i,j}$ has $\mu_k n + o(n)$ edges;
- the 2-core of $K_{i,j}$ has $\tau_k n + o(n)$ degree 2 vertices in K_i and $\tau_k n + o(n)$ degree 2 vertices in K_j.

Let $P_1, ..., P_t$ be the 2-paths of $\mathcal{P}(\Phi)$, where P_i has j_i internal vertices plus endpoints v_i, w_i. We expose each P_i one-at-a-time, and for each P_i, we expose the vertices one-at-a-time beginning with the vertex after v_i.

So consider some P_i that we are exposing, and suppose its vertices are $u_0 = v_i, ..., u_{j_i+1} = w_i$; the first and last of these have already been exposed when we selected the vertices of $V_2(\Phi)$ and the endpoints of the paths. We begin with the case where P_i is a Type B 2-path in some Δ-set $\sigma\Delta\sigma'$ contained in Φ.

First, we choose the colour $\sigma(u_x)$ for each of the j_i internal vertices of P_i; note that $\sigma(u_0), \sigma(u_{j_i+1})$ were determined when we chose $u_0 = v_i, u_{j_i+1} = w_i$. By the definition of a Type B 2-path, $\sigma(u_1), ..., \sigma(_{j_i+1})$ determines $\sigma'(u_0), ..., \sigma'(u_{j_i})$ because $\sigma'(u_x) = \sigma(u_{x+1})$. Since we must have $\sigma(u_{x+1}) \neq \sigma(u_x)$, as those two vertices are adjacent and hence cannot have the same colour, there are $(k-1)^{j_i}$ choices for these colours.

Suppose that we have exposed vertex u_x and are now exposing u_{x+1}, for some $x < j_i$; the case $x = j_i$ is a special case, since $u_{j_i+1} = w_i$ has already been exposed. Prior to exposing u_{x+1}, we have exposed only edges incident with fewer than $a + J < 2\alpha n$ vertices; let Ψ be that set of vertices.

By the definition of a Type B 2-path, u_x is not in the 2-core of $K_{\sigma(u_x),\sigma'(u_x)}$, and u_{x+1} is the parent of u_x in $K_{\sigma(u_x),\sigma'(u_x)}$. We will bound the probability that u_{x+1} is not in the 2-core of $K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$.

Note that if u_{x+1} is one of the exposed vertices Ψ, then we have failed to construct a D-set Φ subject to the specified parameters. So to upper bound the probability that our choices yield such a Φ, we can assume $u_{x+1} \notin \Psi$.

Case 1: $\sigma'(u_{x+1}) \neq \sigma(u_x)$. When we expose the parent of u_x in $K_{\sigma(u_x),\sigma'(u_x)}$, and set it to be u_{x+1}, we expose nothing about u_{x+1} in $K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$, as that is a different graph since we are in Case 1 and since $\sigma'(u_x) = \sigma(u_{x+1})$. So in the random graph $K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$, we have exposed nothing about the edges involving any vertices outside of Ψ.

Consider any graph H that, subject to what has already been exposed, could be $K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$. Consider any H' formed from H by permuting the vertices in $K_{\sigma(u_{x+1})}\backslash\Psi$. It follows from Lemma 11.1, and the fact that we have exposed nothing about edges incident to vertices outside of Ψ, that

$$\mathbf{Pr}(K_{\sigma(u_{x+1}),\sigma'(u_{x+1})} = H) = \mathbf{Pr}(K_{\sigma(u_{x+1}),\sigma'(u_{x+1})} = H').$$

Therefore, the probability that u_{x+1} is not in the 2-core of $K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$ is at most the number of non-2-core vertices in $K_{\sigma(u_{x+1})}$ divided by $|K_{\sigma(u_{x+1})}\backslash\Psi|$. Using the fact that E_1 holds, applying Lemma 11.4(a), and taking α sufficiently small in terms of ζ, this ratio is at most:

$$\frac{\lambda_k^c(c) - \xi_k(c)}{\lambda_k^c(c) - 2\alpha} + o(1) < \frac{1}{k-1}(1 - \frac{\zeta}{2}).$$

Case 2: $\sigma'(u_{x+1}) = \sigma(u_x)$. We argue as in Case 1, except this case is more delicate since $K_{\sigma(u_x),\sigma'(u_x)} = K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$. When we expose the parent of u_x in this graph, we need to bound the probability of that parent being outside of the 2-core.

Consider any graph H that, subject to what has already been exposed, could be $K_{\sigma(u_x),\sigma'(u_x)}$. On the previous step (while considering u_{x-1}) we exposed that u_x is not in the 2-core of this graph. Let H' be the graph obtained from H by removing the edge from u_x to its parent. Condition on the event that H' is the graph obtained from $K_{\sigma(u_x),\sigma'(u_x)}$ by removing the edge from u_x to its parent. Consider adding to H' an edge from u_x to any vertex of $K_{\sigma'(u_x)}$ that is not in the same component of H' as u_x. The choice of that vertex does not affect the 2-core of the resulting graph, nor does it affect the number of edges in the resulting graph. It follows that, by Lemma 11.1, every such vertex is equally likely to be the parent of u_x, under this conditioning. Note that every 2-core vertex of $K_{\sigma'(u_x)}\backslash\Psi$ is eligible to be the parent of u_x. Therefore, the conditional probability that the parent of x is in the 2-core is at least the number of 2-core vertices in $K_{\sigma'(u_x)}\backslash\Psi$ divided by $|K_{\sigma'(u_x)}|$. Since this is true for any choice of H', and since E_1 holds and $|\Psi| < 2\alpha n$, and applying Lemma 11.4(a), it follows that the probability that the parent of u_x is *not* in the 2-core of $K_{\sigma(u_x),\sigma'(u_x)}$) is at most

$$1 - \frac{\xi_k(c) - 2\alpha}{\lambda_k^c(c)} < \frac{1}{k-1}(1 - \frac{\zeta}{2}), \tag{6}$$

if α is sufficiently small in terms of ζ.

So in both cases, we find that the probability that u_{x+1} is not in the 2-core of $K_{\sigma(u_{x+1}),\sigma'(u_{x+1})}$, conditional on what has been exposed thus far, is at most $\frac{1}{k-1}(1 - \frac{\zeta}{2})$.

Finally, we turn our attention to the edge $(u_{j_i}, u_{j_i+1} = w_i)$. Here, we must bound the probability that w_i is the parent of u_{j_i} in $K_{\sigma(u_{j_i}),\sigma'(u_{j_i})}$. Note that $w_i \in \Psi$. We follow similar reasoning as in Case 2 above and argue that every 2-core vertex in $K_{\sigma'(u_{j_i})}$ is at least as likely to be the parent of u_{j_i} as w_i is. So the probability that w_i is the parent is at most the inverse of the number of 2-core vertices in $K_{\sigma'(u_{j_i})}$. Using the fact that E_1 holds, this is at most:

$$\frac{1 + o(1)}{\lambda_k^c(c)n}.$$

Putting this all together, each Type B 2-path P_i contributes to $\mathbf{Pr}(E_3 \wedge E_1 | E_2)$ a factor of at most

$$(k-1)^{j_i} \times (\frac{1}{k-1}(1 - \frac{\zeta}{2}))^{j_i} \times \frac{1 + o(1)}{\lambda_k^c(c)} \times \frac{1}{n}. \tag{7}$$

Next, we consider the case where P_i is a Type A 2-path in some Δ-set $\sigma\Delta\sigma'$ contained in Φ.

At this point, we have exposed all the edges in the $K_{i,j}$'s corresponding to the Type B 2-paths. All vertices on those paths are now in Ψ. We have also exposed information on whether some of the vertices in Ψ are in the 2-cores of the $K_{i,j}$'s.

Now, we will expose the entire 2-core of every $K_{i,j}$. Recall that the edge-sets of each $K_{i,j}$ are independent of each other. Our first step is to expose the vertices of each 2-core, and the degree that each vertex has in

the 2-core; note that part of this information has already been determined. Then we will choose the 2-core using the configuration model[13]. Note that some of the edges have already been determined - specifically, the edges of $\mathcal{E}$ and the edges on the Type B 2-paths.

For each Type A 2-path P_i, following Lemma 12.16, we select the two colours on that path; i.e. the colours a, b such that for each $u \in P_i$ we have $(\sigma(u), \sigma'(u)) = (a, b)$ or (b, a). There are $k(k-1)$ choices for each path. The same reasoning as in the proof of Lemma 12.16 shows that every vertex of P_i must have degree exactly two in the 2-core of $K_{a,b}$.

For each a, b, we let $t_{a,b}$ denote the number of Type A 2-paths for which we selected the colours a, b in the preceding paragraph, and we let $J_{a,b}$ denote the total number of internal vertices on those $t_{a,b}$ paths. We set $J^a_{a,b}, J^b_{a,b}$ be the number of such vertices u for which we determined that $\sigma(u) = a, \sigma(u) = b$ resp.

To upper bound the probability of these paths being formed, we will assume that every endpoint of a Type A 2-path is a vertex of the appropriate colour that has degree 2 in the 2-core of the appropriate $K_{i,j}$.

We now choose the interior vertices for each Type A 2-path P_i. For each $K_{a,b}$, we must select $J^a_{a,b}, J^b_{a,b}$ vertices of colour a, b that have degree two in the 2-core of $K_{a,b}$. Let $L^a_{a,b}, L^b_{a,b}$ be the number of such vertices to choose from in $K_{a,b}$. Since E_1 holds, we have $L^a_{a,b}, L^b_{a,b} = \tau_k(c)n + o(n)$. Since the 2-paths are disjoint (by Definition 12.19), the number of choices is at most:

$$L^a_{a,b}(L^a_{a,b}-1)...(L^a_{a,b}-J^a_{a,b}+1)L^b_{a,b}(L^b_{a,b}-1)...(L^b_{a,b}-J^b_{a,b}+1).$$

Now we choose which vertex-copies of the vertices of each path, including the endpoints, will be matched with each other. The number of choices is at most $2^{J_{a,b}+2t_{a,b}}$. Finally, we bound the probability that these copies will be paired up. Because every edge in the configuration contains a vertex of each colour, the total number of vertex-copies from K_a in the 2-core of $K_{a,b}$ which do not lie in edges of $\mathcal{E}$ or edges of the Type B 2-paths, is the same as the total number from K_b; let $X_{a,b}$ be that number.

We proceed along the paths one-vertex-at-a-time, each time exposing whether the selected copy of that vertex is paired with the selected copy of the next vertex on the path. Every success removes a vertex copy of each colour from the 2-core of $K_{a,b}$. So the probability that all of these $J_{a,b} + t_{a,b}$ pairings occur is:

$$\frac{1}{X_{a,b}(X_{a,b}-1)...(X_{a,b}-(J_{a,b}+t_{a,b})+1)}.$$

This leads to the following bound on the probability that the $t_{a,b}$ Type A 2-paths that use edges from $K_{a,b}$ are formed:

$$\begin{aligned}
&\frac{2^{J_{a,b}+2t_{a,b}}L^a_{a,b}...(L^a_{a,b}-J^a_{a,b}+1)L^b_{a,b}...(L^b_{a,b}-J^b_{a,b}+1)}{X_{a,b}...(X_{a,b}-(J_{a,b}+t_{a,b})+1)}\\
< &\left(\frac{4}{X_{a,b}-J_{a,b}-t_{a,b}}\right)^{t_{a_b}}\frac{2^{J_{a,b}}L^a_{a,b}...(L^a_{a,b}-J^a_{a,b}+1)L^b_{a,b}...(L^b_{a,b}-J^b_{a,b}+1)}{X_{a,b}...(X_{a,b}-J_{a,b}+1)}. \qquad (8)
\end{aligned}$$

Since E_1 holds, the 2-core of $K_{a,b}$ has a total of $\mu_k(c)n + o(n)$ vertex-copies in K_a and $\mu_k(c)n + o(n)$ vertex-copies in K_b. $\mathcal{E}$ contains at most $2an$ edges and the Type B 2-paths contain at most $J + t < 2\alpha n$ edges. So

$$X_{a,b} \geq \mu_k(c)n - 4an + o(n).$$

Therefore, if we choose α to be sufficiently small in terms of ζ, then by Lemma 11.4(b), we have $\frac{2L^a_{a,b}}{X_{a,b}-1}, \frac{2L^a_{a,b}}{X_{a,b}-1} < 1 - \frac{\zeta}{2}$. It follows that for every $x > 0$ we have:

$$\frac{2(L^a_{a,b}-x)}{X_{a,b}-2x-1}, \frac{2(L^a_{a,b}-x)}{X_{a,b}-2x-1} < 1 - \frac{\zeta}{2}.$$

If $L^a_{a,b} = L^b_{a,b}$ then this would yield that the bound of (8) is at most

$$\left(\frac{4}{X_{a,b}-J_{a,b}-t_{a,b}}\right)^{t_{a_b}}(1-\frac{\zeta}{2})^{J_{a,b}}.$$

However, we must multiply by a corrective factor if $J^a_{a,b} \neq J^b_{a,b}$. Noting that $|J^a_{a,b} - J^b_{a,b}| < t_{a,b}$, and that $X_{a,b} - J_{a,b} > X_{a,b} - 2\alpha n > \frac{1}{2}X_{a,b}$ for α sufficiently small, we find that the corrective factor is at most $2^{t_{a,b}}$. Similarly, we have $X_{a,b} - J_{a,b} - t_{a,b} > \frac{1}{2}\mu_k(c)n$, and so our bound is at most:

$$\left(\frac{16}{\mu_k(c)n}\right)^{t_{a_b}}(1-\frac{\zeta}{2})^{J_{a,b}}.$$

We multiply this bound over all a, b. Then we multiply by the contribution from (7) for each Type B 2-path. We also multiply by the 2 choices for whether each P_i is Type A or Type B, and if it is Type B, the $k(k-1)$ choices for its colours - a total of $k(k-1)+1 < k^2$ choices for each path.

Setting $R = k^2 \times \max(\frac{1}{\lambda^c_k(c)}, \frac{16}{\mu_k(c)n})$, and recalling that $t \leq a$, this yields:

$$\mathbf{Pr}(E_3 \wedge E_1 | E_2) \leq (1-\frac{\zeta}{2})^J R^a \left(\frac{1}{n}\right)^t,$$

as required. □

Lemma 12.25 *There are constants $R = R(c,k)$ and $\rho = \rho(c,k) > 0$ such that for any $2\alpha n > L > \frac{4}{\zeta}(a+J)$:*

$$\mathbf{Pr}(E_4 \wedge E_3 \wedge E_1 | E_2) < (1-\rho)^L \times R^a (1-\frac{\zeta}{2})^J \left(\frac{1}{n}\right)^t.$$

Proof We continue the proof from Lemma 12.23. At this point, we have selected the $a+J$ vertices of the 2-core of Φ and exposed information about the edges amongst them that are present. We have exposed the values of the parameters bounded by Lemma 11.3. If E_1 holds then for each i, j:

- $|K_i| = \lambda^c_k(c)n + o(n)$;
- the 2-core of $K_{i,j}$ has $\xi(c)n + o(n)$ vertices in K_i and $\xi(c)n + o(n)$ vertices in K_j;
- the 2-core of $K_{i,j}$ has $\mu_k n + o(n)$ edges;
- the 2-core of $K_{i,j}$ has $\tau_k n + o(n)$ degree 2 vertices in K_i and $\tau_k n + o(n)$ degree 2 vertices in K_j.

Following what was already proven in Lemma 12.23, we now must prove that, condtional on all tha has been exposed thus far,

$$\mathbf{Pr}(E_4) < (1-\rho)^L.$$

As in the proof of Lemma 12.23, at any point of the argument we will use Ψ to denote the set of vertices that have been exposed thus far. Initially, we set Ψ to be the $a+2J$ vertices in the 2-core of Φ. Note that we have only exposed information about edges amongst those vertices, and about whether some of those vertices are in the 2-cores of various $K_{i,j}$'s.

We will expose $\cup_{i=1}^{a+J} T^+(u_i)$ using a branching process. For $i = 1$ to $a+J$, we explore $T^+(u_i)$ via a breadth-first search through all previously unexposed vertices. Initially, u_i is unexplored. At each step, we choose an unexplored vertex w. We expose all children of w in $D \backslash \Psi$; each of those children is unexplored, and we label w as explored. We place all of those children into Ψ. Note that, since we only branch amongst previously unexposed vertices, we might not generate all of $T^+(u_i)$. However, if we miss some $v \in T^+(u_i)$, then either $v \in \Psi$ or v is the descendent of some $v' \in \Psi$. Either way, v was (or will be) encountered during the branching from some other u_j, It follows that this process will indeed expose all of $\cup_{i=1}^{a+J} T^+(u_i)$.

Below, we will prove that at each of the first L steps of this process:

$$\text{The expected number of children of } w \text{ is at most } 1 - \frac{\zeta}{2}, \tag{9}$$

where $\zeta = \zeta(c)$ comes from Lemma 11.4. So we run a sequence of $a+J$ branching processes, each with a branching factor of at most $1 - \frac{\zeta}{2}$. Straightforward facts about branching processes imply that there is a constant $\rho = \rho(\zeta) > 0$ such that for $L > \frac{4}{\zeta}(a+J)$:

The probability that these branching processes yield a total of at least L vertices is at most $(1-\rho)^L$.

This establishes the lemma. It remains to prove (9).

We know $w \in K_i$ where $i = \sigma(w)$. Consider any $j \neq i$ and consider any vertex $x \in K_j \backslash \Psi$ that is not in the 2-core of $K_{i,j}$. For any non-2-core vertex $u \in K_{i,j}$, we use $p(u)$ to denote the parent of u in $K_{i,j}$. We will show that in the pendant tree of $K_{i,j}$ containing x,

$$\mathbf{Pr}(p(x) = w) \leq \frac{1+o(1)}{|K_i \backslash \Psi|}. \tag{10}$$

There are $k-1$ choices for j, $(\lambda_k^c(c) - \xi_k(c))n + o(n)$ choices for x and $|K_i \backslash \Psi| \geq (\lambda_k^c(c) - 2\alpha)n + o(n)$. So for α sufficiently small in terms of ζ, the expected number of children of w is at most

$$(k-1) \times \frac{\lambda_k^c(c) - \xi_k(c)}{\lambda_k^c(c) - 2\alpha} + o(1) < 1 - \frac{\zeta}{2},$$

by Lemma 11.4(a). This establishes (9).

To prove (10), consider any vertex $y \in K_i \backslash \Psi$. If y is in the 2-core of $K_{i,j}$, then the same argument as in Case 2 of the proof of Lemma 12.23 proves that $\mathbf{Pr}(p(x) = w) \leq \mathbf{Pr}(p(x) = y)$.

So suppose y is not in the 2-core of $K_{i,j}$. Let E^* be the event that y is a descendant of x in a pendant tree of $K_{i,j}$.

Claim 1: $\mathbf{Pr}(E^*) = O(n^{-1})$.

Proof: We can expose the unique path from y to the 2-core of $K_{i,j}$ as follows: Set $z := y$. While z is not in the 2-core, set $z := p(z)$. A similar argument to that used in the proof of Lemma 12.23 when considering the edge (u_{j_i}, w_i), yields that at each step: Every 2-core vertex in the opposite part from z is at least as likely to be the parent of z as x is. So at each step, the probability of reaching x is $O(\frac{1}{n})$. Furthermore, at each step, the probability of reaching the 2-core is $\Theta(1)$. It follows that the probability that we reach x before the 2-core, i.e. that x is on the path from y to the 2-core, is $O(n^{-1})$. □

Claim 2: $\mathbf{Pr}(p(x) = w | \overline{E^*}) \leq \mathbf{Pr}(p(x) = y | \overline{E^*})$.

Proof: This is equivalent to showing that $\mathbf{Pr}(p(x) = w \wedge \overline{E^*}) \leq \mathbf{Pr}(p(x) = y \wedge \overline{E^*})$. Consider any graph H for which it is possible, under what has been exposed thus far, for $K_{i,j} = H$ and for which the event $p(x) = w \wedge \overline{E^*}$ holds. Let H' be the graph obtained by replacing the edge (x, w) with (x, y). Since $\overline{E^*}$ holds for H, we have that H' is connected and y is the parent of x in H'. Furthermore, H and H' have the same 2-core and the same number of edges. So H' is possibly $K_{i,j}$, given what has been exposed thus far, and H, H' are both equally likely to be $K_{i,j}$. Since each such H' can arise from at most one such H, the Claim follows. □

Therefore, $\mathbf{Pr}(p(x) = w) \leq \mathbf{Pr}(p(x) = y) + \mathbf{Pr}(p(x) = w \wedge E^*)$. A straightforward extension of the proof of Claim 1 shows that $\mathbf{Pr}(p(x) = w \wedge E^*) = O(n^{-2})$ (we omit the details). Therefore $\mathbf{Pr}(p(x) = w) \leq \mathbf{Pr}(p(x) = y) + O(n^{-2})$. Summing over all $y \in K_i \backslash \Psi$ yields

$$|K_i \backslash \Psi| \times \mathbf{Pr}(p(x) = w) \leq 1 + |K_i| \times O(n^{-2}) = 1 + o(1).$$

This yields (10). □

Lemma 12.26 *For any $f(n) = o(n)$, there exists $g(n) = o(n)$ such that with probability at least $1 - e^{-2f(n)}$, the total number of vertices on all cyclic D-sets is at most $g(n)$.*

Proof sketch: Recall from Lemma 12.21 that every cyclic D-set is a cyclic Δ-set. Recall from Lemma 12.22 that the 2-core of a cyclic Δ-set $\sigma\Delta\sigma'$ is a cycle $u_1, ..., u_r$ such that every u_i is not in the 2-core of $K_{\sigma(u_i),\sigma'(u_i)}$ and its parent is u_{i+1} (addition mod r). We refer to such a cycle as a *Type B cycle*. We start by bounding the expected number of Type B cycles of length r.

First we choose the colours $\sigma(u_1), ..., \sigma(u_r)$; there are fewer than $k(k-1)^r$ choices. Next we choose u_1; there are fewer than n choices. Then we proceed around the cycle: after choosing u_i, we expose it's parent in $K_{\sigma(u_i),\sigma'(u_i)}$ and set that vertex to be u_{i+1}. The same argument as in the proof of Lemma 12.23 shows

that the probability that u_{i+1} is not in the 2-core of $K_{\sigma(u_{i+1},\sigma'(u_{i+1})}$ is less than $\frac{1}{k-1}(1-\frac{\zeta}{2})$. Finally, we bound the probability that u_1 is the parent of u_r in $K_{\sigma(u_r),\sigma'(u_r)}$. The argument used in the proof of Lemma 12.23 for the edge (u_{j_i}, w_i) yields a bound of $\frac{1+o(1)}{\lambda_k^c(c)n}$. Putting this all together, the expected number of Type B cycles of length r is less than $Q(1-\frac{\zeta}{2})^r$, for a constant Q.

Letting Y_t be the number of collections of Type B cycles of total size t, these calculations extend to yield a constant $\gamma = \gamma(c,k)$ such that for any $t < 2\alpha n$,

$$\mathbf{Exp}(Y_t \wedge E_1) < (1-\gamma)^t.$$

Recall Definition 12.12 and observe that if $u_1, ..., u_r$ is the 2-core of a Δ-set, then that Δ-set is contained in $\cup_{i=1}^r T^+(u_i)$ (by an argument very similar to the proof of Lemma 12.13). We let $Y_{t,L}$ denote the number of collections of Type B cycles of total size t for which, denoting the vertices as $u_1, ..., u_t$, we have $|\cup_{i=1}^t T^+(u_i)| \geq L$. The same argument as in the proof of Lemma 12.25 yields that there is are constants $B = B(c,k)$ and $\rho' = \rho'(c,k) > 0$ such that for $2\alpha n > L > Bt$:

$$\mathbf{Exp}(Y_t \wedge E_1) < (1-\gamma)^t(1-\rho')^L.$$

This, and an argument much like the proof of Lemma 12.3, is enough to prove the lemma.

The details will appear in a full version of this paper. □

13 The Kempe core threshold

We adapt the argument from [44], where we analyzed a very similar core, but in a simpler setting. In fact, the main motivation for [44] was to develop a technique that we could use here to analyze the Kempe core. The reader might prefer to read [44] before reading this section.

We let $G_{n_1,n_2,p}$ denote the random bipartite graph whose parts have size n_1, n_2 and where each of the $n_1 n_2$ possible edges is present independently with probability p. We will need the following concentration bound on the size of the giant component of $G_{n_1,n_2,p}$.

Lemma 13.1. *Consider any constant $\beta > 0$ and any constant $c > \beta^{-1}$ and any $n_1, n_2 = \rho n + o(n)$. Let β be the unique solution to $\beta = \rho(1-e^{-\beta c})$. For every $f(n) = o(n)$ there exists $g(n) = o(n)$ such that with probability at least $1 - e^{-4f(n)}$, the largest component of $G_{n_1,n_2,p=c/n}$ has size $\beta n \pm g(n)$ and every other component has size less than $g(n)$.*

Proof sketch: It is straightforward to determine that the expected size of the giant component is $\beta n + o(n)$, using eg. a branching process analysis. The required concentration on the size of that giant component can also be obtained by analysing the branching process, as can the probability that there is at least one other component of size at least $g(n)$. We give the details in a full version of this paper.

We remark that much more is known about the giant component of $G_{n,p}$. In particular, its size has a normal distribution[56, 53, 14]. Presumably one could mimic the fairly short proof from [14] to obtain a normal distribution for the size of the giant component of a random bipartite graph. But what we require is weaker than that, and so the simpler proof described above will suffice. □

We will find the Kempe core using a parallel version of Kempe-Strip from Section 5. At each iteration, we remove *all* Kempe-chains of size at most $g(n)$. The key fact that permits our analysis is:

Observation 13.2. *This procedure is equivalent to repeatedly removing all small components from the bipartite random graph induced by each pair of parts A_a, A_b.*

At the beginning of iteration i, V_a^i will be the vertices remaining from part A_a, for each $1 \leq a \leq k$. For each $a \neq b$, we examine the bipartite subgraph induced by the remaining vertices from A_a, A_b; all vertices from that subgraph that are not in the giant component are deleted. The procedure halts when there are no vertices to delete. A formal description of this procedure is:

STRIP
Set $V_1^1 = A_1, ..., V_k^1 = A_k$.
For $i \geq 1$
for all $a \neq b$, $K_{a,b}^i$ is the vertex-set of the largest component
of the bipartite subgraph induced by (V_a^i, V_b^i).
for every $1 \leq a \leq k$,
set $V_a^{i+1} = \cap_{b \neq a}(K_{a,b} \cap V_a)$.
if $V_a^{i+1} = V_a^i$ for all $1 \leq a \leq k$ then HALT and return $V_1^i, ..., V_k^i$.
if $V_1^{i+1} = ... = V_k^{i+1} = \emptyset$ then HALT and return $\emptyset, ..., \emptyset$.

Note that if every $K_{a,b}^i$ contains at most one component of size greater than $g(n)$, then at each step we are indeed removing all Kempe chains of size less than $g(n)$; i.e. this is equivalent to Kempe-Strip from Section 5. Lemma 13.1 will imply that this is the case.

The following key observation is crucial to our analysis:

Observation 13.3. *Given V_a^i, V_b^i and $x_a = |K_{a,b}^i \cap V_a^i|, x_b = |K_{a,b}^i \cap V_b^i|$, the vertices of $K_{a,b}^i \cap V_a^i$ and $K_{a,b}^i \cap V_b^i$ can be treated as uniformly random subsets of V_a^i, V_b^i of sizes x_a, x_b, respectively.*

Proof We can carry out STRIP by exposing, at each step, the vertex set of $K_{a,b}^i$ without actually exposing the edges of the giant component that $K_{a,b}^i$ induces. Thus, at step $i+1$, any subsets of the appropriate size are equally likely to form the vertex sets of the giant component of $K_{a,b}^{i+1}$. □

We will focus mainly on V_1^i, V_2^i; by symmetry, the other sets V_a^i evolve in a similar manner. It will be useful to focus, in particular, on $K_{1,2}^i$; i.e. the giant component of the bipartite subgraph induced by the remaining vertices from V_1^i, V_2^i. It will be convenient to consider sets U^i, W^i, where we will have $V_1^i \subseteq U^i \subseteq A_1$ and $V_2^i \subseteq W^i \subseteq A_2$. Initially, $U^i = A_1$, $W^i = A_2$; throughout the procedure, vertices are removed from U^i and W^i at the same rate that vertices which lie in small components of any bipartite subgraphs *except for* the one induced by (A_1, A_2) are removed from V_1^i and V_i^2. The sets U^i, W^i will be very close to uniformly chosen from A_1, A_2.

To form U^{i+1}, we remove vertices from V_1^i as follows: (1) Expose the number that should be removed because they are in small components of the subgraphs induced by $(V_1^i, V_b^i), 3 \leq b \leq k$. (2) Select that many vertices uniformly at random from V_1^i; Observation 13.3 permits a coupling by which this is legal. To carry out (2), we actually remove vertices uniformly from U^i until the appropriate number of vertices have been removed from V_1^i. We form W^{i+1} in the analogous manner.

More formally, U^i, W^i are defined by the following modified procedure, which captures STRIP from the viewpoint of the bipartite subgraph on A_1, A_2.

STRIP1
Set $V_1^1 = A_1, ..., V_k^1 = A_k$.
Set $U^1 = A_1, W^1 = A_2$.
For $i \geq 1$
Expose the vertex-set of $K_{1,2}^i$.
For every $3 \leq b \leq k$,
Expose $\ell_b^i = |V_1^i \backslash K_{1,b}^i|$, the number of vertices removed from V_1^i
because they are not in $K_{1,b}^i$.
Repeat ℓ_b^i times
Pick a sequence of vertices chosen uniformly from U^i without replacement
until ℓ_b^i of them are chosen from V_1^i.
This sequence is L_b^i.
Do not remove these vertices from U_i yet; they are still eligible
to be chosen for another value of b.
Set $K_{1,b}^i \cap V_1^i = V_1^i \backslash L_b^i$.

Expose $q_b^i = |V_2^i \backslash K_{2,b}^i|$, the number of vertices removed from V_2^i
because they are not in $K_{2,b}^i$.
Repeat q_b^i times
Pick a sequence of vertices chosen uniformly from W^i without replacement
until q_b^i of them are chosen from V_2^i.
This sequence is Q_b^i.
Do not remove these vertices from W_i yet; they are still eligible
to be chosen for another value of b.
Set $K_{2,b}^i \cap V_2^i = V_2^i \backslash Q_b^i$.
Set $U^{i+1} = U^i \backslash \cup_{3 \le b \le k} L_b^i$.
Set $V_1^{i+1} = (K_{1,2}^i \cap V_1^i) \backslash \cup_{3 \le b \le k} L_b^i$.
Set $W^{i+1} = W^i \backslash \cup_{3 \le b \le k} Q_b^i$.
Set $V_2^{i+1} = (K_{1,2}^i \cap V_2^i) \backslash \cup_{3 \le b \le k} Q_b^i$.
For every $1 \le a, b \le k$,
Expose the remainder of the vertex sets of $K_{a,b}$; i.e. all portions of such vertex sets that do not lie in V_1^i, V_2^i.
Update $V_3^{i+1}, ..., V_k^{i+1}$ as in STRIP
if $V_a^{i+1} = V_a^i$ for all $1 \le a \le k$ then HALT and return $V_1^i, ..., V_k^i$.
if $V_1^{i+1} = ... = V_k^{i+1} = \emptyset$ then HALT and return $\emptyset, ..., \emptyset$.

Note that, for each i, b, the ℓ_b^i vertices of L^i that are in V_1^i are uniform members of V_1^i. So by Observation 13.3, we can couple STRIP1 with STRIP so that they produce the same sets $V_1^i, ..., V_k^i$.

A key observation is that for each iteration j, and every $b \ge 3$, all vertices in the small components of $K_{1,b}^j$ are removed from U^j. So all vertices in $U^i \backslash V_1^i$ are in small components of the subgraph induced by U^i, W^i. The same is true of all vertices in $W^i \backslash V_2^i$. Thus, if the largest components of the bipartite subgraphs induced by (V_1^i, V_2^i) and (U^i, W^i) have linear size, then they must be the same components. It will be convenient to focus on the latter subgraph.

The advantage of dealing with the bipartite subgraph induced by (U^i, W^i) is that U^i, W^i are *nearly* uniform subsets of A_1, A_2. They are not quite uniform, as there is some dependency between U^i and the size of U^{i+1}. However, we inductively sandwich each U^i between two uniformly random subsets of A_1 and each W^i between two uniformly random subsets of A_2. This will allow us to treat $K_{1,2}^i$ as being approximately the giant component of $G_{n_1,n_2,p=c/n}$ where $n_1 = |U^i| + o(n), n_2 = |W^i| + o(n)$, and thus apply Lemma 13.1.

We define recursively:

$$\begin{aligned}
\rho_1 &= \nu_1 = \frac{1}{k} \\
\beta_i &= \rho_i(1 - e^{-\beta_i c}), \qquad \text{for } i \ge 1 \\
\nu_{i+1} &= \nu_i \left(\frac{\beta_i}{\nu_i}\right)^{k-1}, \qquad \text{for } i \ge 2 \\
\rho_{i+1} &= \rho_i \left(\frac{\beta_i}{\nu_i}\right)^{k-2}, \qquad \text{for } i \ge 2
\end{aligned}$$

Recalling the definition of c_k from Section 11, if $c > c_k$ then let $\beta = \beta_k(c)$ be the unique positive solution to $\beta = \frac{1}{k}(1 - e^{-\beta c})^{k-1}$. Set $\rho = \rho_k(c) = (\beta^{k-2}/k)^{\frac{1}{k-1}}$.

Lemma 13.4. *(a) If $c < c_k$ then $\lim_{i\to\infty} \beta_i = 0$, $\lim_{i\to\infty} \nu_i = 0$, $\lim_{i\to\infty} \rho_i = 0$.*

(b) If $c > c_k$ then $\lim_{i\to\infty} \beta_i = \lim_{i\to\infty} \nu_i = \beta$, $\lim_{i\to\infty} \rho_i = \rho$.

Proof $\nu_i = \frac{1}{k}\left(\prod_{j=1}^{i-1} \frac{\beta_i}{\nu_i}\right)^{k-1}$ and $\rho_i = \frac{1}{k}\left(\prod_{j=1}^{i-1} \frac{\beta_i}{\nu_i}\right)^{k-2}$. Therefore $(k\nu_i)^{k-2} = (k\rho_i)^{k-1}$. Taking the

fixed point of the recursive equations, we obtain $\nu = \beta$ and $\beta = \rho(1 - e^{-\beta c})$. This yields:

$$(k\beta)^{k-1} = (k\rho)^{k-1}(1 - e^{-\beta c})^{k-1}; \qquad (k\beta)^{k-1} = (k\beta)^{k-2}(1 - e^{-\beta c})^{k-1}; \qquad \beta = \frac{1}{k}(1 - e^{-\beta c})^{k-1}.$$

The rest of the proof is straightforward, after noting that there is a positive solution to $\beta = \frac{1}{k}(1 - e^{-\beta c})^{k-1}$ iff $c \geq c_k$. □

Lemma 13.5. *For any constant I, and for any $f(n) = o(n)$, there exists $h(n) = o(n)$ such that with probability at least $1 - e^{-3f(n)}$ we have for each $1 \leq i \leq I$:*

(a) $|K^i_{1,2} \cap V^i_1|, |K^i_{1,2} \cap V^i_2| = \beta_i n \pm h(n)$;

(b) $|U^i|, |W^i| = \rho_i n \pm h(n)$;

(c) $|V^i_1|, |V^i_2| = \nu_i n \pm h(n)$.

Proof sketch We proceed by induction. At each iteration i, we will actually bound the set sizes within error terms $h_i(n), h'_i(n)$ and obtain a failure probability of $O(i)e^{-4f(n)}$. So, taking $h(n) = \max(h_I(n), h'_I(n))$, the overall failure probability is $O(I^2)e^{-4f(n)} < e^{-3f(n)}$ since $I = O(1)$.

Note that (b,c) both hold at $i = 1$ with sufficiently high probability, since $A_1, ..., A_k$ is a uniformly random partition of $\{1, ..., n\}$.

Suppose that (b) holds for i. We let U^-, U^+ be two uniformly random sets of vertices from A_1 of sizes $\rho_i n - 2h_i(n), \rho_i n + 2h_i(n)$ respectively, and we couple them so that $U^- \subset U^+$. Furthermore, we couple these sets with the steps of STRIP1 where vertices are removed from U so that, if (b) holds for i then $U^- \subset U^i \subset U^+$. We define W^-, W^+ similarly.

Since U^-, W^- are uniform subsets of A_1, A_2, we can choose U^-, W^- *before* any edges are exposed. So the subgraph induced by (U^-, W^-) is identical in distribution to $G_{|U^-|,|W^-|,p=c/n}$. Similarly for (U^+, W^+).

Let $(X^-, Y^-), (X^+, Y^+)$ be the vertex sets of the giant components of the subgraph induced by (U^-, W^-), (U^+, W^+). Thus, we have $X^- \subset K^i_{1,2} \cap V^i_1 \subset X^+$ and $Y^- \subset K^i_{1,2} \cap V^i_2 \subset Y^+$. Thus, applying Lemma 13.1 to both the subgraph induced by (U^-, W^-) and the subgraph induced by (U^+, W^+), we obtain that with probability at least $1 - e^{-4f(n)}$ we have

$$|K^i_{1,2} \cap V^i_1|, |K^i_{1,2} \cap V^i_2| = \beta_i n \pm h'_i(n),$$

where $h'_i(n) = o(n)$ depends on $h_i(n)$ and $g(n)$. This is (a). By symmetry, both parts of each $K^i_{a,b}$ have size $\beta_i n \pm h'_i(n)$ with probability at least $1 - e^{-4f(n)}$. This implies that for each $3 \leq b \leq k$, $\ell^1_b = |V^i_1| - \beta_i n \pm h'_i(n)$. It follows that for any $u \in U^i$, $\mathbf{Pr}(u \notin L^i_b) = \frac{\beta_i n}{|V^i_1|} + o(1)$ where the $o(1)$ term depends on $h'_i(n)$. Using Observation 13.3 and the fact that the graphs $K_{a,b}$ can be chosen independently,

$$\mathbf{Pr}(u \notin \cup_{b \geq 3} L^i_b) = \left(\frac{\beta_i n}{|V^i_1|}\right)^{k-2} + o(1) = \left(\frac{\beta_i}{\nu_i}\right)^{k-2} + o(1),$$

if (c) holds for V^i_1. It follows that

$$\mathbf{Exp}(|U^{i+1}|) = (\rho_i n \pm h_i(n)) \times \left(\frac{\beta_i}{\nu_i}\right)^{k-2} + o(1) = \rho_{i+1} n + o(n),$$

$$\mathbf{Exp}(|V^i_1|) = |K^i_{1,2} \cap V^i_1| \times \left(\frac{\beta_i}{\nu_i}\right)^{k-2} + o(1) = \nu_i n \left(\frac{\beta_i}{\nu_i}\right)^{k-2} + o(1) = \nu_{i+1} n + o(n),$$

if (a) holds for $K^i_{1,2} \cap V^i_1$. The $o(n)$ terms depend on $h_i(n), h'_i(n)$. Because these sets are determined by the choices of $\Theta(n)$ vertices from U_i, W_i, it is easy to show that they are concentrated with sufficiently high probability. The same argument applies to $W_{i+1}, V^i_2, K^i_{1,2} \cap V^i_2$. We are implicitly carrying out these bounds

$O(k^2) = O(1)$ other times, and we are requiring that there were no failures in the earlier rounds. It follows that the failure probability on this inductive step is $O(i)e^{-4f(n)}$.

Being careful about the order of induction and the accumulation of the $o(1)$ terms completes the lemma. Further details will appear in a full version of this paper. □

By symmetry, the bound in Lemma 13.5(c) applies to V_a^i for all $1 \leq a \leq k$. For any $\epsilon > 0$, Lemma 13.4 implies that we can take I large enough to run STRIP until each V_a^I has size $(\beta \pm \epsilon)n$, and U^I, W^I each have size $(\rho \pm \epsilon)n$.

Next we adapt of the proof of Lemma 5.1 in [44], showing that w.h.p. STRIP will terminate and return sets $V_1, ..., V_k$ where each V_a has size $\beta n \pm o(n)$. These are the sets $K_1, ..., K_k$ of the Kempe core. Furthermore this yields:

Lemma 13.6. *For each a, b, the subgraph $K_{a,b}$ induced by V_a, V_b is sandwiched between the giant components of $G_{n_1,n_2,p=c/n}$ and $G_{n_1',n_2',p=c/n}$ for some $n_1 < n_1', n_2 < n_2'$ and $n_1, n_2, n_1', n_2' = \rho n + o(n)$.*

The failure probability in this final step comes from a single application of Lemma 13.1, along with some simple bounds on binomial variables. So again, we obtain a failure probability of $e^{-3f(n)}$, as required. This proves Lemma 10.1.

We close this section with a proof sketch of Lemma 11.3.

Lemma 11.3 *For any $f(n) = o(n)$, with probability at least $1 - e^{-3f(n)}$, we have that for every i, j:*

(a) $|K_i| = \lambda_k(c)n + o(n)$;

(b) the 2-core of $K_{i,j}$ has $\xi_k(c)n + o(n)$ vertices in K_i and $\xi_k(c)n + o(n)$ vertices in K_j;

(c) the 2-core of $K_{i,j}$ has $\mu_k(c)n + o(n)$ edges;

(d) the 2-core of $K_{i,j}$ has $\tau_k(c)n + o(n)$ degree 2 vertices in K_i and $\tau_k(c)n + o(n)$ degree 2 vertices in K_j.

Proof Sketch: Part (a) follows by observing that $y_k(c) = \beta c$ and so $\lambda_k(c) = \beta$.

Standard analysis of the k-cores of random graphs (see eg. [52, 45, 32, 24, 29]) shows that the bounds of parts (b,c,d) hold w.h.p. for the 2-core of $G_{\rho n,\rho n,p=c/n}$. Furthermore, the technique from [45] (and presumably the techniques from some of the other papers as well) can be used to obtain the necessary concentration; i.e. that there is some $\gamma(n) = o(n)$ (defined in terms of $f(n)$) such that with probability at least $1 - e^{-3f(n)}$ those three parameters differ from $\xi_k(c)n, \mu_k(c)n, \tau_k(c)n$ by at most $\gamma(n)$. Lemma 13.6 then yields Lemma 11.3.

The details will appear in a full version of the paper. □

14 The number of unfrozen variables

We close this paper by sketching the proof of Theorem 2.4 parts (a.i) and (b). I.e., we bound the number of vertices that are not $\omega(n)$-frozen for $\omega(n) = o(n)$.

Fix some constant T, and define T-Kempe-Strip to be the process that we get by replacing $g(n)$ by T in Kempe-Strip and by running the parallel version; i.e.:

T-Kempe-Strip
Input: a graph G and a k-colouring $\sigma = S_1, ..., S_k$ of G.
While there are any Kempe chains of size at most T
Remove the vertices of *every* such Kempe chain from G.

Lemma 14.1. *If v is deleted during the first I iterations of T-Kempe-Strip, and if v is not T-frozen, then v is within distance IT of a cycle with length at most $2IT$ in the original graph.*

Proof Let Γ_i denote the set of Kempe chains that are removed during iteration i. Form a graph Λ on the removed Kempe chains as follows. Each Kempe chain $C \in \Gamma_i$ has an edge pointing to every Kempe chain $C' \in \Gamma_j, j < i$ for which there is an edge joining C, C' in G. Let $R(C)$ be the set of Kempe chains that can be reached from C in Γ. If $R(C)$ induces a tree in Λ, then it is clear that by switching (some of) the chains of $R(C)$, one-at-a-time, starting at the leaves, we can eventually switch C. Note that each such switch changes at most T vertices, and so the vertices of C are not T-frozen.

Consider some C containing v; so $C \in \Gamma_i, i \leq I$. If v is T-frozen, then $R(C)$ does not induce a tree in Λ. It follows that C is within distance I of a cycle of length at most $2I$ in Λ. Since each vertex of Λ corresponds to a connected subgraph of G of size at most T, then v is within distance IT of a cycle with length at most $2IT$ in G. □

Recall that for each colour a, the number of vertices of A_a in the Kempe core of $P_{n,p=c/n}$ is $\beta n + o(n)$, where $\beta = \beta(c)$ is defined recursively in Section 13.

Lemma 14.2. *For any $c > c_k$, any $f(n) = o(n)$ and any $\psi > 0$, there exist $T, I = O(1)$ and $g(n) = o(n)$ such that with probability at least $1 - e^{-2f(n)}$: The number of vertices remaining in A_a after applying I rounds of T-Kempe-Strip to $P_{n,p=c/n}$ is at most $(\beta + \psi)n$.*

Proof sketch: The difference between T-Kempe-Strip and the parallel version of Kempe-Strip (i.e. STRIP from Section 13) is that T-Kempe-Strip does not remove non-giant components of size greater than T. For any $\delta > 0$ there exists T such that, at every iteration, the expected number of vertices in such components is at most $\frac{\delta}{2}n$. We adapt the standard concentration bounds on the number of such vertices in $G_{n,p}$ (see eg. [13]) to the $G_{n,n,p}$ setting to show that, at each iteration, the number of vertices in each component is at most δn with probability at least $1 - e^{-3f(n)}$. It follows that the analogue of Lemma 13.5 holds for T-Kempe-Strip upon adjusting the recursive equations by an additive term of at most δ.

Choose I such that $\beta < \beta_I < \beta + \frac{\psi}{2}$. And choose δ so that the accumulation of those adjustments over I iterations is an additive term of at most $\frac{\psi}{2}$. This yields the lemma. □

We now prove Lemma 10.4.

Lemma 10.4 *For $k \geq 3$, any $f(n) = o(n)$ and any $\epsilon > 0$: There exists constants T, Z such that with probability at least $1 - e^{-f(n)}$, all but ϵn of the vertices outside of the (possibly empty) Kempe core of $P_{n,p=c/n}$ are either (i) not T-frozen, or (ii) are within distance Z of a cycle with length less than Z.*

Proof Given ϵ, we set $\psi = \epsilon/k$, choose T, I as in Lemma 14.2, and set $Z = 2TI$.

Recall that the order in which we remove the Kempe chains in Kempe-Strip does not affect the output. So we could run Kempe-Strip as follows:

Phase I: Run T-Kempe-Strip for I iterations.
Phase II: Run Kempe-Strip until it halts.

Lemma 14.1 implies that every T-frozen vertex removed in Phase I is within distance Z of a cycle with length at most Z.

Suppose $c > c_k$. Lemma 14.2 and Lemma 11.3(a) (recalling that $\beta = \lambda_k(c)/k$) implies that with probability at least $1 - e^{-3f(n)}$, at most $k\psi n = \epsilon n$ vertices are removed during Phase II. This is in $P_{n,p=c/n}$. Lemma 9.3 allows us to transfer to $U_{n,p=c/n}$.

The case $c < c_k$ follows from the same argument, after replacing β by zero, recalling from Lemma 13.4(a) that for $c < c_k$ we have $\lim_{i\to\infty} \beta_i = 0$, and considering the analogue of Lemma 14.2 for $c < c_k$. □