LABEL PROPAGATION ON BINOMIAL RANDOM GRAPHS

Marcos Kiwi[*1], Lyuben Lichev[†2], Dieter Mitsche[‡3], and Paweł Prałat[§4]

[1]Univ. Chile, Santiago, Chile
[2]TU Wien, Vienna, Austria
[3]IMC, Pontifícia Univ. Católica, Santiago, Chile
[4]Toronto Metropolitan University, Toronto, Canada

May 22, 2025

Abstract

We study the behavior of a label propagation algorithm (LPA) on the Erdős-Rényi random graph $\mathcal{G}(n,p)$. Initially, given a network, each vertex starts with a random label in the interval $[0,1]$. Then, in each round of LPA, every vertex switches its label to the majority label in its neighborhood (including its own label). At the first round, ties are broken towards smaller labels, while at each of the next rounds, ties are broken uniformly at random. The algorithm terminates once all labels stay the same in two consecutive iterations. LPA is successfully used in practice for detecting communities in networks (corresponding to vertex sets with the same label after termination of the algorithm). Perhaps surprisingly, LPA's performance on dense random graphs is hard to analyze, and so far convergence to consensus was known only when $np \geq n^{3/4+\varepsilon}$, where LPA converges in three rounds. By defining an alternative label attribution procedure which converges to the label propagation algorithm after three rounds, a careful multi-stage exposure of the edges allows us to break the $n^{3/4+\varepsilon}$ barrier and show that, when $np \geq n^{5/8+\varepsilon}$, a.a.s. the algorithm terminates with a single label. Moreover, we show that, if $np \gg n^{2/3}$, a.a.s. this label is the smallest one, whereas if $n^{5/8+\varepsilon} \leq np \ll n^{2/3}$, the surviving label is a.a.s. not the smallest one. En passant, we show a presumably new monotonicity lemma for Binomial random variables that might be of independent interest.

Keywords: label propagation algorithm, binomial random graph, majority rule, voter model, threshold
MSC Class: 05C80, 60C05, 05D40

1 Introduction

In this paper, we analyze a version of a popular unsupervised learning algorithm for finding communities in complex networks called *label propagation algorithm*. In the specific instance we consider, henceforth referred to as LPA, each vertex starts with a random label in the interval $[0,1]$. The algorithm is completely determined by the relative order of the labels. Thus, as long as they are all different from each other, the exact label values are not relevant. Since this assumption is satisfied with probability 1 for every finite graph, we may (and do) assume for convenience that the initial labels coincide with the indices of the vertices, that is, for all $i \in [n] = \{1, \ldots, n\}$, vertex $v_i \in V$ starts with label i. Then, in each round of

the algorithm, every vertex switches its label to the majority label in its neighborhood (including its own label). Moreover, at the first round, ties are broken towards smaller labels, while at each of the next rounds, ties are broken uniformly at random. (Note that the first round has a special role since, in the beginning, every label is represented only once.) The algorithm ends once the process converges (that is, once no more changes are made at some round) or some predefined maximum number of iterations is reached. Intuitively, the algorithm exploits the fact that a single label can quickly become dominant in a densely connected collection of vertices, but will not rapidly propagate through a sparsely connected region. Hence, labels will likely get trapped inside densely connected vertex classes. Vertices that end up with the same label when the algorithm stops are considered part of the same community. Among the advantages of LPA, compared to other algorithms, is the scant amount of a priori information it needs about the network structure (no parameter is required to be known beforehand), its efficient distributed implementation, simplicity, and success in practice.

Label propagation algorithms have often been used to detect communities [26, 13]; for more background, see the surveys [33, 14, 2] or any book on mining complex networks such as [17] or [24]. Despite their popularity and the fact that their theoretical analyses were identified as an important research question [21, 8, 1], there are only a few theoretical papers in this area published so far. As observed in [7], a mathematical analysis is challenging because of "the lack of techniques for handling the interplay between the non-linearity of the local update rules and the topology of the graph".

Mathematically, the class of label propagation algorithms has many parallels with models of opinion exchange dynamics. These models have been proposed in order to improve our understanding of different social, political and economical processes and found applications in the fields of distributed computing and network analysis. Typically, opinion exchange dynamics assume that individual agents learn by observing each other's actions (the clearest example being perhaps learning on financial markets). One interesting question within this framework is whether consensus (that is, agreement of all agents) is eventually reached.

Bayesian type models are among the most sophisticated opinion exchange models. There, the action of each individual is based on maximizing the expectation of some utility function depending on the information available at some point, see the nice survey [23]. Amid the most famous and mathematically interesting models are the deGroot model (see [9] for more details), where the basic idea is that individuals either have opinion 0 or 1, and constantly update their opinion according to the (possibly weighted) average of their neighbors; in the voter model, individuals again have binary opinions, and at each step, everyone chooses one neighbor (according to possibly non-uniform rules) and adopts the opinion of this neighbor (see [5, 15]); in majority dynamics, individuals have binary or non-binary opinions, and at each step, everyone adapts to the majority opinion of its neighbors (with different tie-breaking mechanisms), see for example [12]). Label propagation algorithms are a special case of non-binary majority dynamics.

In [20], Kothapalli, Pemmaraju, and Sardeshmukh initiated the mathematically formal analysis of a specific variant of label propagation algorithms. More precisely, they proposed to study the performance of the procedure on the *stochastic block model (SBM)*, a random graph model that, in its simplest form, partitions the vertices of a graph into k classes and connects vertices between and within different classes independently according to different probabilities – typically with higher density within vertex classes. When $k = 1$, the stochastic block model corresponds to the *binomial random graph* which, formally speaking, is the distribution $\mathcal{G}(n, p)$ over the class of graphs G on n vertices with vertex set V in which every pair $\binom{V}{2}$ appears independently as an edge in G with probability p. Note that most results for the binomial random graph are asymptotic in nature, and $p = p(n)$ may (and usually does) tend to zero as n tends to infinity. We say that $\mathcal{G}(n, p)$ has some property *asymptotically almost surely* or *a.a.s.* if the probability that $\mathcal{G}(n, p)$ has this property tends to 1 as n goes to infinity. For a detailed treatment of this model, see for example [16, 18, 4]. As we shall see, for the range of the parameter p we investigate in this paper, the binomial random graph has the property that LPA converges quickly a.a.s. Clearly, proving fast convergence of LPA on $\mathcal{G}(n, p)$ to a configuration with a single label for a wide range of values of p would be a strong indication of the strength of the procedure.

The authors of [20] considered the variant of label propagation algorithms where ties are always broken

towards smaller labels. They gave a rigorous analysis of this variant and claimed that for arbitrarily small $\varepsilon > 0$ and $np \geq n^{3/4+\varepsilon}$, a.a.s. after only two iterations, all vertices in $\mathcal{G}(n, p)$ receive label 1. (In fact, a careful checking of their proof shows that three iterations are required.) They also conjectured that there is a constant $c > 0$ such that for all $np \geq c \log n$, their version of the algorithm a.a.s. terminates on $\mathcal{G}(n, p)$, and when it does, all vertices carry the same label. This conjecture was then proved wrong in [19] (see [25] for slightly more details) where the authors showed that there exists $\varepsilon > 0$ such that, for any $np \leq n^\varepsilon$, the procedure a.a.s. terminates on $\mathcal{G}(n, p)$ in a configuration where more than one label is present. Simulations reported in [25, 19] suggest that the behavior of the process changes around $np = n^{1/5}$.

Coincidentally, several recent articles have studied the binary case of majority dynamics when the underlying graph is $\mathcal{G}(n, p)$ and initially every vertex chooses one of two labels with equal probability (in contrast to labels from an interval as in LPA). The question considered is whether all vertices converge to the same label and, if so, how many rounds it takes. Benjamini, Chan, O'Donnell, Tamuz and Tan [3] showed that if $p = \Omega(1/\sqrt{n})$, then $\mathcal{G}(n, p)$ is such that with probability at least 0.4 (over the choice of the random graph and the initial choice of vertex labels), convergence to the most popular label happens after four rounds. In fact, they conjectured that this holds with high probability. The conjecture was recently positively resolved by Fountoulakis, Kang and Makai [11] (see also [28, 34, 31]). Also, see [6] for recent work on majority dynamics in $G(n, p)$ in the case where vertices are initially assigned one of $k = O(1)$ labels uniformly at random.

As it is already implicit in [20], the gist of the analysis of LPA on SBM is to understand the circumstances under which LPA identifies each block. We are not aware of any direct translation of results for LPA over $\mathcal{G}(n, p)$ (that is, LPA over SBM with one block) to results for LPA on SBM. However, we believe that most of the ideas and arguments we rely on carry over to the SBM setting, albeit at the cost of complicating notation, lengthier calculations, and negatively impacting clarity of exposition. Moreover, SBM has community structure and so algorithms typically have no problem with finding communities unless one is close to the detectability threshold. In contrast, our goal is to better understand communities that are formed by pure randomness and how LPA variants react to them. This, we believe, is more challenging and of interest in its own right. Thus, we restrict our study to the case where the underlying network is $\mathcal{G}(n, p)$ (equivalently SBM with $k = 1$ block).

We enlarge the range of values of p for which (a.a.s.) LPA converges to a configuration containing a single label. Specifically, we show that a.a.s. LPA identifies $\mathcal{G}(n, p)$ as a single community. To achieve this, we need to overcome significant technical obstacles. Before discussing them, we first formally state our main contribution and provide an overview of its proof.

1.1 Main results

The following theorem formally states the main result of our paper.

Theorem 1.1. *Suppose that $\varepsilon \in (0, 1/24)$ and $n^{5/8+\varepsilon} \leq np \ll n$. Then, a.a.s. after five rounds of the process, all vertices carry the label that was most represented after the first round. Moreover,*

- *if $n^{2/3} \ll np \ll n$, then a.a.s. this label is 1,*

- *if $n^{5/8+\varepsilon} \leq np \ll n^{2/3}$, then a.a.s. this label is different from 1.*

We note that the first point of the theorem (that is, when $n^{2/3} \ll np \ll n$) is valid also if ties are always broken towards the smaller label, as in [20].

Remark 1.2. In parallel with the proof of the first part of Theorem 1.1, we show that, when $np = \Theta(n^{2/3})$, the winning label is 1 with probability bounded away from 0 and 1 simultaneously.

Interestingly, in the influential paper introducing label propagation as a procedure for community detection, Raghavan, Albert and Kumara [26] state that "although one can observe [from simulations on

real-world networks] the algorithm beginning to converge significantly after about five iterations, the mathematical convergence is hard to prove". Our contribution rigorously establishes, for an expanded range of values of p, that the empirically determined at most five iterations observed by Raghavan, Albert and Kumara are sufficient for a specific variant of LPA to identify $\mathcal{G}(n,p)$ as a single community.[1] We believe that the insight gained by our analysis might be useful in the study of label propagation algorithms as well as opinion dynamic models.

1.2 Outline of the proof

On a high level, the main technical contribution of our paper is an in-depth analysis of an exploration process done in several stages. We will explore only a subset of possible edges at each step, thereby leaving independence for subsequent steps, while at the same time taking the same decisions as the original LPA. More specifically, in both regimes of p considered in Theorem 1.1, we first ensure that a.a.s. only at most $k = \lceil 15 p^{-2}(n^{-1}\log n)^{1/2}\rceil$ labels are carried by more than one vertex after the second round. We partition the set of vertices into three *levels*: A, consisting of the vertices $v_1,\ldots,v_{2k}$ that initially carry labels $1,\ldots,2k$, respectively, B, consisting of all neighbors of vertices in A outside A, and C, consisting of all other vertices. Then, for every label $\ell \in [2k]$, we call *basin* of v_ℓ the set of vertices $B_1(\ell) \subseteq B$ connecting to v_ℓ but not connecting to any of $v_1,\ldots,v_{\ell-1}$.

When $n^{2/3} \ll np \ll n$, we show that a.a.s. the basin of vertex v_1 is the largest one, and we estimate the difference between its size and the size of the ℓ-th basin for all $\ell \in [2,2k]$. Then, at the second round, we design a vertex labeling procedure for the vertices in B and in C based only on the edges incident to $A \cup B$, which (thanks to the fact that essentially, a.a.s. only the labels in $[k]$ matter after two rounds) a.a.s. attributes the same labels as the algorithm. This procedure has the advantage of leaving all edges between vertices in C unexposed, which is then used in the third round. We show that the difference between $|B_1(1)|$ and the remaining basin sizes is amplified in C after the second round, that is, the difference between the number of vertices in C with label 1 and those with any other given label is of larger order than $|B_1(1)| - \max_{\ell \in [2,2k]} |B_1(\ell)|$. In fact, we ensure that a.a.s. this difference becomes so large that after the third round, all vertices in C carry label 1. Note that the conclusion of this last point is made possible by the (crucial) fact that edges between vertices in C were not exposed before, and therefore the graph induced by C remains distributed as $\mathcal{G}(|C|,p)$. Finally, since a.a.s. $|C| = (1-o(1))n$, it is easy to conclude that two more rounds are sufficient to end up in a configuration with all vertices carrying label 1. In the case when $np = \Theta(n^{2/3})$, a similar analysis (conducted in parallel with the proof for the regime $n^{2/3} \ll np \ll n$) shows that a.a.s. we end up in a configuration with all vertices carrying a label following some non-trivial distribution on the positive integers.

The regime $n^{5/8+\varepsilon} \le np \ll n^{2/3}$ is more complicated to analyze. Although the global strategy remains the same, there are several additional technical difficulties.

Firstly, the largest basin now is that of v_{ℓ_1} for some $\ell_1 \in [2k]$ that is a.a.s. different from 1. To analyze the size of $B_1(\ell_1)$ and the difference with the sizes of the remaining basins, we do a careful stochastic approximation of all basin sizes with independent binomial random variables. This step additionally ensures that a.a.s. $\ell_1 = o(k)$.

Moreover, differences between basin sizes are typically smaller than before. As a result, the analysis of the vertex labeling procedure in B similar to the one in the first regime is less direct. Roughly speaking, it is divided into two parts: for any fixed $\ell \in [2k]\setminus\{\ell_1\}$, we first count the number of vertices in $B\setminus(B_1(\ell)\cup B_1(\ell_1))$ that get a label among $\{\ell,\ell_1\}$ at the second round. We show that a.a.s. for every choice of ℓ, the majority of these vertices get label ℓ_1. Then, we prove that a.a.s. for every ℓ as above, the number of vertices in

---

[1]Note however that the family of underlying networks we consider, that is, $\mathcal{G}(n,p)$ for $n^{5/8+\epsilon} \ll np \ll n$, are not directly comparable to specific instances of real world networks as those considered in [26]. Moreover, in the latter, it is reported that simulations converge in 5 rounds but detect many communities (whose size is empirically observed to follow a power law distribution). From the report on the simulations, it is impossible to ascertain what the intra-block edge-density among the detected communities is.

$B_1(\ell)$ that do not change their label at the second round is small. Thus, despite the fact that this allows for more vertices of label ℓ than those with label ℓ_1 in B after the second round, the surplus of vertices with label ℓ_1 in C after the second round remains of larger order, and therefore this allows the spread of label ℓ_1 among all vertices in C after the third round. The proof is then completed as before.

1.3 Technical contributions

As mentioned above, it has been recognized that the analysis of label propagation algorithms involves some non-trivial mathematical challenges. The first and foremost, technical complications arise from the deterministic evolution (except for the eventual tie breaking rules) of the process once the graph and the initial label assignment are fixed (the former being much more challenging to deal with than the latter). One way of bypassing these obstacles is to analyze a process in which the supporting graph is resampled anew at the start of each round (see for example [30]). This significantly simplifies the analysis but is unrelated to our underlying motivation, which is to contribute to the rigorous understanding of when label propagation type algorithms succeed in correctly and efficiently identifying communities.

An important but technically demanding result which, we believe, could have further applications, is Lemma 2.6. It roughly says that, for two independent binomial random variables $X \in \mathrm{Bin}(m, p)$ and $Y \in \mathrm{Bin}(n, p)$ with $m > n$, the ratio of the probabilities that $X > M$ and $Y > M$ for some threshold $M \in [-1, n]$ is an increasing function of M (the formal statement is a bit more complicated, as it needs to keep track of our tie-breaking conventions for the attribution of labels). Furthermore, we propose several couplings that facilitate dealing with intrinsic dependencies inherent to the analysis of label propagation variants. For instance, in Lemma 3.22, the random variables $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$ (that represent, roughly speaking, the number of vertices that get label $\ell \in [2k]$ after the first round) are coupled with a sequence of independent binomial random variables $\mathrm{Bin}(z_\ell, p)$ whose mean is a second order approximation of the expectation of $\mathfrak{B}_1(\ell)$ (and a decreasing function of ℓ). Again, in order to deal with dependencies, in Lemma 3.15 we introduce a decoupling technique that conditions on whether a specific edge uv is present or not in $\mathcal{G}(n, p)$ in order to derive (via the second moment method) a.a.s. bounds for the difference between two particular random variables (both measurable with respect to the edges of $\mathcal{G}(n, p)$).

The determination (via coupling) of the number of vertices that get label $\ell \in [2k]$ after the first round leads to questions concerning the asymptotic distribution of the maximum of independent binomials whose mean has a negative drift. Unfortunately, we could not find, among prior results concerning order statistics of independent but not identically distributed random variables, a result useful to us. In contrast, the analogous question for i.i.d. random variables is significantly simpler and extensively studied but not adapted to our setting. To address these questions, we develop an approach that first determines the asymptotic behavior of the maximum (when ℓ varies over an interval of integers I) of the binomial random variables $\mathrm{Bin}(z_\ell, p)$ again with mean a decreasing function of ℓ (see Lemma 3.24 and Remark 3.25). By comparing the obtained asymptotic distributions for different choices of the interval I, we can identify specific intervals for ℓ where the first and second maximum of the collection of binomials is attained (see Corollary 3.26), and estimate the gap between them (see Lemma 3.28).

Finally, an arguably less significant technical contribution but still worth mentioning, is the derivation of several inequalities concerning the density and distribution function of the difference between two sums (over different number) of i.i.d. Bernoulli random variables (see Lemma 2.3). The novelty here is the use of Berry-Esseen's and Slud's inequalities.

2 Preliminaries and Notation

2.1 Notation

We use mostly standard asymptotic notation. Apart from the classical O, Ω, Θ and o, for any two functions $f : \mathbb{N} \to (0, \infty)$ and $g : \mathbb{N} \to (0, \infty)$, we write $f(n) \gg g(n)$ or $g(n) \ll f(n)$ if $g(n) = o(f(n))$ and $f(n) \sim g(n)$

if $f(n) = (1 + o(1))g(n)$.

We use $\log n$ to denote the natural logarithm of n. We use the following extension of the notation $[n] = \{1, \ldots, n\}$: For given $a, b \in \mathbb{N}$, $a \leq b$, we let $[a, b] = \{a, \ldots, b\}$. For $a \in \mathbb{R}$ and $\epsilon > 0$, we let $a \pm \epsilon$ denote the interval $[a - \epsilon, a + \epsilon]$.

For a vertex $v \in V$, we write $N(v)$ for the set of neighbors of v in $\mathcal{G}(n, p)$, and $N[v] = N(v) \cup \{v\}$ for the closed set of neighbors of v. For any $Z \subseteq V$, let also $N(Z) = \bigcup_{v \in Z} N(v)$ and $N[Z] = N(Z) \cup Z$. Finally, as typical in the field of random graphs, for expressions that clearly have to be integers, we round up or down without specifying when this choice does not affect the argument.

2.2 Preliminaries

The first lemma that we need is a specific instance of Chernoff's bound that we will often find useful (see e.g. Theorem 2.1 in [16]).

Lemma 2.1. *Let $X \in \mathrm{Bin}(n, p)$ be a random variable with binomial distribution with parameters n and p and $\varphi : [-1, \infty) \to \mathbb{R}$ be such that $\varphi(t) = (t + 1) \log(t + 1) - t$. Then, for all $t \geq 0$,*

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \exp\left(-\varphi\left(\frac{t}{\mathbb{E}X}\right) \cdot \mathbb{E}X\right) \leq \exp\left(-\frac{t^2}{2(\mathbb{E}X + t/3)}\right),$$

$$\mathbb{P}(X - \mathbb{E}X \leq -t) \leq \exp\left(-\varphi\left(-\frac{t}{\mathbb{E}X}\right) \cdot \mathbb{E}X\right) \leq \exp\left(-\frac{t^2}{2\mathbb{E}X}\right).$$

The following result is a partial converse of Chernoff's bound, stated in terms of the standard normal distribution. To this end, set

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \qquad \text{for all } t \in \mathbb{R}.$$

To avoid over-cluttering formulas, henceforth we adopt the typical convention of denoting $1 - p$ by q.

Lemma 2.2 (see Lemma 2.1 in [29])**.** *Let $X \in \mathrm{Bin}(n, p)$ be a random variable with binomial distribution with parameters n and $p \leq 1/4$. Then, for every $t \in [0, n - 2np]$,*

$$\mathbb{P}(X \geq \mathbb{E}X + t) \geq 1 - \Phi\left(\frac{t}{\sqrt{npq}}\right).$$

The next lemma analyzes the difference of two independent binomial random variables with the same parameter p but slightly different means.

Lemma 2.3. *Fix $a_1 = a_1(n), a_2 = a_2(n) \in \mathbb{N}$ and $p = p(n) = o(1)$ such that $1 \ll a_2 \leq a_1$ and $\min\{a_2, a_1 - a_2\}p \to \infty$ as $n \to \infty$. Let $X_1 \in \mathrm{Bin}(a_1, p)$ and $X_2 \in \mathrm{Bin}(a_2, p)$ be two independent random variables. Then, there exists a constant $\zeta \in (0, 1/2)$ such that for any fixed constant $M \in \mathbb{R}$,*

$$\mathbb{P}(X_1 - X_2 \geq M) \geq \Phi\left(\frac{(a_1 - a_2)p - M}{\sqrt{(a_1 + a_2)pq}}\right) - \frac{2\zeta}{\sqrt{a_2 p}}. \tag{1}$$

In particular, for n large enough,

$$\mathbb{P}(X_1 - X_2 \geq M) \geq \frac{1}{2} + \frac{1}{5} \min\left\{1, \frac{(a_1 - a_2)p}{\sqrt{(a_1 + a_2)p}}\right\}. \tag{2}$$

Moreover, for any fixed constant $m \in \mathbb{Z}$,

$$\mathbb{P}(X_1 - X_2 = m) = o\Big(\mathbb{P}(X_1 - X_2 \geq m) - \mathbb{P}(X_1 - X_2 < m)\Big).$$

Proof. By the normal approximation of the binomial distribution (see Berry-Esseen's inequality [10]), if $X \in \text{Bin}(a, p)$, then for all $x \in \mathbb{R}$,

$$\left| \mathbb{P}\left(\frac{X - ap}{\sqrt{apq}} \leq x \right) - \Phi(x) \right| \leq \frac{\zeta(p^2 + q^2)}{\sqrt{apq}} \leq \frac{\zeta}{\sqrt{ap}},$$

where $0 < \zeta < 1/2$ is an explicit constant. Let Z_1 and Z_2 be two independent and standard normally distributed random variables. Then,

$$\mathbb{P}(X_1 - X_2 \geq M) = \mathbb{P}\left(\frac{X_1 - a_1 p}{\sqrt{a_1 pq}} \geq \frac{M + X_2 - a_1 p}{\sqrt{a_1 pq}} \right) \geq -\frac{\zeta}{\sqrt{a_1 p}} + \mathbb{P}\left(Z_1 \geq \frac{M + X_2 - a_1 p}{\sqrt{a_1 pq}} \right).$$

Since $Z_1 \geq \frac{M + X_2 - a_1 p}{\sqrt{a_1 pq}}$ if and only if $Z_1\sqrt{a_1 pq} - M + a_1 p \geq X_2$, we get

$$\mathbb{P}(X_1 - X_2 \geq M) \geq -\frac{\zeta}{\sqrt{a_1 p}} + \mathbb{P}\left(\frac{Z_1\sqrt{a_1 pq} - M + (a_1 - a_2)p}{\sqrt{a_2 pq}} \geq \frac{X_2 - a_2 p}{\sqrt{a_2 pq}} \right)$$

$$\geq -\frac{\zeta}{\sqrt{a_1 p}} - \frac{\zeta}{\sqrt{a_2 p}} + \mathbb{P}\left(\frac{Z_1\sqrt{a_1 pq} - M + (a_1 - a_2)p}{\sqrt{a_2 pq}} \geq Z_2 \right).$$

Using the fact that $a_1 \geq a_2$ and that $Z_1\sqrt{a_1 pq} - Z_2\sqrt{a_2 pq}$ is a normally distributed random variable with mean 0 and variance $(a_1 + a_2)pq$, we conclude that

$$\mathbb{P}(X_1 - X_2 \geq M) \geq -\frac{2\zeta}{\sqrt{a_2 p}} + 1 - \Phi\left(\frac{M - (a_1 - a_2)p}{\sqrt{(a_1 + a_2)pq}} \right) = \Phi\left(\frac{(a_1 - a_2)p - M}{\sqrt{(a_1 + a_2)pq}} \right) - \frac{2\zeta}{\sqrt{a_2 p}},$$

and so inequality (1) holds.

Let $\xi > 0$ be a fixed constant. Then, if $\frac{M - (a_1 - a_2)p}{\sqrt{(a_1 + a_2)pq}} \leq -\xi$, recalling that $a_2 p \to \infty$ as $n \to \infty$, we get that $\mathbb{P}(X_1 - X_2 \geq M) \geq 1 - 1.01 \cdot \Phi(-\xi)$. On the other hand, if $\frac{M - (a_1 - a_2)p}{\sqrt{(a_1 + a_2)pq}} > -\xi$, recalling that $(a_1 - a_2)p \to \infty$ as $n \to \infty$, we get

$$1 - \Phi\left(\frac{M - (a_1 - a_2)p}{\sqrt{(a_1 + a_2)pq}} \right) \geq \frac{1}{2} + \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} \frac{|M - (a_1 - a_2)p|}{\sqrt{(a_1 + a_2)pq}} \geq \frac{1}{2} + 0.99 \cdot \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} \frac{(a_1 - a_2)p}{\sqrt{(a_1 + a_2)p}}.$$

Observing that

$$\frac{1}{\sqrt{a_2 p}} \Bigg/ \frac{(a_1 - a_2)p}{\sqrt{(a_1 + a_2)p}} = \frac{1}{\sqrt{(a_1 - a_2)p}} \sqrt{\frac{2}{(a_1 - a_2)p} + \frac{1}{a_2 p}} \to 0 \quad \text{when } n \to \infty,$$

we obtain that

$$\mathbb{P}(X_1 - X_2 \geq M) \geq \frac{1}{2} + 0.98 \cdot \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} \frac{(a_1 - a_2)p}{\sqrt{(a_1 + a_2)p}}.$$

Summarizing,

$$\mathbb{P}(X_1 - X_2 \geq M) \geq \min\left\{ 1 - 1.01 \cdot \Phi(-\xi), \tfrac{1}{2} + 0.98 \cdot \frac{e^{-\xi^2/2}}{\sqrt{2\pi}} \frac{(a_1 - a_2)p}{\sqrt{(a_1 + a_2)p}} \right\}.$$

Taking $\xi = 1$, observing that $0.98 \cdot e^{-\xi^2/2}/\sqrt{2\pi} \approx 0.2371 > \frac{1}{5}$ and verifying in a table of values for the cdf of the standard normal distribution that $\Phi(-1) \approx 0.1587$ (hence, $1 - 1.01 \cdot \Phi(-1) \approx 0.8397 > \frac{1}{2} + \frac{1}{5}$) establishes the second part of the lemma.

Now, for the last part of the lemma, note that from the first part we get that

$$\mathbb{P}(X_1 - X_2 \geq m) - \mathbb{P}(X_1 - X_2 < m) = 2\mathbb{P}(X_1 - X_2 \geq m) - 1 \geq \frac{2}{5} \min\left\{ 1, \frac{(a_1 - a_2)p}{\sqrt{(a_1 + a_2)p}} \right\}.$$

7

On the other hand, since the mode m_* of X_1 is either $\lfloor (a_1 + 1)p \rfloor$ or $\lceil (a_1 + 1)p \rceil - 1$, from Stirling's approximation one can deduce that

$$\mathbb{P}(X_1 - X_2 = m) \leq \mathbb{P}(X_1 = m_*) = (1 + o(1)) \frac{1}{\sqrt{2\pi a_1 pq}} = O\Big(\frac{1}{\sqrt{(a_1 + a_2)p}} \Big).$$

Since by hypothesis $(a_1 - a_2)p \to \infty$ as $n \to \infty$, the last two displayed inequalities yield the last part of the lemma. $\qquad\square$

Remark 2.4. Set $t = a_1 - a_2$. Assuming that $tp = O(1)$ and $a_2 p \to \infty$ in Lemma 2.3, it is still possible to establish that

$$\mathbb{P}(X_1 > X_2) + \frac{1}{2} \cdot \mathbb{P}(X_1 = X_2) = \frac{1}{2} + \Omega\Big(\frac{tp}{\sqrt{a_2 p}} \Big).$$

Indeed, let $X' \in \mathrm{Bin}(a_2, p)$ and $Y \in \mathrm{Bin}(t, p)$ be independent random variables. Since $X' + Y \in \mathrm{Bin}(a_1, p)$ and $X_1 \in \mathrm{Bin}(a_1, p)$, we have

$$\mathbb{P}(X_1 > X_2) + \frac{1}{2} \cdot \mathbb{P}(X_1 = X_2) \geq \mathbb{P}(X' > X_2) + \frac{1}{2} \cdot \mathbb{P}(X' = X_2) + \frac{1}{2} \cdot \mathbb{P}(X' = X_2 - 1)\mathbb{P}(Y \geq 1).$$

The first two terms on the right-hand side above sum up to $\frac{1}{2}$. Next, observe that since $tp = O(1)$, we have $\mathbb{P}(Y \geq 1) = 1 - q^t = \Theta(tp)$. To conclude, recall that for $m \in a_2 p \pm \sqrt{a_2 pq}$ one has that $\mathbb{P}(X' = m)$ and $\mathbb{P}(X_2 = m + 1)$ are $\Omega(1/\sqrt{a_2 p})$, so

$$\mathbb{P}(X' = X_2 - 1) \geq \sum_{m \in a_2 p \pm \sqrt{a_2 pq}} \mathbb{P}(X' = m)\mathbb{P}(X_2 = m + 1) = \Omega\Big(\frac{1}{\sqrt{a_2 p}} \Big).$$

Remark 2.5. The proof of Lemma 2.3 also implies that for every $\varepsilon > 0$ there is a positive integer $N = N(\varepsilon)$ such that as long as $\min\{a_2, a_1 - a_2\}p \geq N$ (the other assumptions therein remain), for every integer M, (1) and (2) are still satisfied, and moreover

$$\mathbb{P}(X_1 - X_2 = M) \leq \varepsilon |\mathbb{P}(X_1 - X_2 \geq M) - \mathbb{P}(X_1 - X_2 < M)|.$$

We finish this preliminary section with the already mentioned important technical lemma, which may be of independent interest thanks to its general formulation.

Lemma 2.6. *Fix integers $n > n' \geq 0$ and a real number $\rho \geq 0$. Consider two independent random variables $X \in \mathrm{Bin}(n, p)$ and $Y \in \mathrm{Bin}(n', p)$. For every integer $M \geq -1$, let*

$$p_X(M, \rho) = \mathbb{P}(X > M, X > Y) + \tfrac{1}{2}\mathbb{P}(X = Y > M)$$
$$+ \tfrac{1}{\rho+1}\mathbb{P}(X = M > Y) + \tfrac{1}{\rho+2}\mathbb{P}(X = Y = M),$$

and define $p_Y(M, \rho)$ similarly by exchanging X and Y. Then, for every integer $M \in [0, n]$,

$$\frac{p_X(M, \rho)}{p_Y(M, \rho)} \geq \frac{p_X(-1, \rho)}{p_Y(-1, \rho)} = \frac{\mathbb{P}(X > Y) + \tfrac{1}{2}\mathbb{P}(X = Y)}{\mathbb{P}(Y > X) + \tfrac{1}{2}\mathbb{P}(Y = X)}.$$

Roughly speaking, in the following sections, we apply Lemma 2.6 with $X \in \mathrm{Bin}(n_1, p)$ (where n_1 is the size of the largest basin after one round, say $B_1(1)$), $Y \in \mathrm{Bin}(n_2, p)$ (where $n_2 < n_1$ is the size of some of the remaining basins, say $B_1(2)$), M being the maximum number of edges between u and $B_1(i)$ over $i \in [\ell] \setminus \{1, 2\}$, and ρ equal to the number of times this maximum is attained. Then, the ratio of the probability that u obtains label 1 and the probability that u obtains label 2 is minimal when $\ell = 2$, or equivalently when $M = 0$.

The proof of Lemma 2.6 requires a significant technical effort and it will given in Section 4.

3 Proof of Theorem 1.1

In this section, we fix $\varepsilon \in (0, 1/24)$ and assume that $n^{5/8+\varepsilon} \leq np \ll n$. To start with, recall from the proof outline that we partition the vertex set V into *levels*. In more detail, Level 1 consists of the vertices in the set $A = \{v_1, \ldots, v_{2k}\}$, where

$$k = k(n) = \left\lceil 15p^{-2}(n^{-1}\log n)^{1/2} \right\rceil, \tag{3}$$

Level 2 consists of their neighbors (that is, $B = N(A) \setminus A$), and Level 3 consists of all remaining vertices (that is, $C = V \setminus (A \cup B)$).

Now, we adopt an important notational convention. Whenever considering a set of vertices $S \subseteq V$, the subset of vertices of S that have label ℓ after round t will be denoted by $S_t(\ell)$. Also, the sizes of the sets A, B, C and V will be denoted $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{C}$ and $\mathfrak{V}$, respectively. Furthermore, the number of elements in $A_t(\ell)$, $B_t(\ell)$ and $C_t(\ell)$ will be denoted $\mathfrak{A}_t(\ell)$, $\mathfrak{B}_t(\ell)$ and $\mathfrak{C}_t(\ell)$, respectively. For a set of labels $W \subseteq [n]$ and a subset of vertices of $S \subseteq V$, we let $S_t(W)$ be the subset of vertices of S that after round t have a label that belongs to W, that is, $S_t(W) = \bigcup_{\ell \in W} S_t(\ell)$. In particular, for a set of labels $W \subseteq [n]$, we will use $V(W) = V_0(W)$ for the set of vertices that initially have a label from W. We adopt the same aforementioned convention when referring to sizes of such sets; for instance, $\mathfrak{B}_1([k])$ equals the number of vertices in Level 2 that after round 1 have a label that belongs to set $[k]$.

3.1 First Two Rounds

In this section, we study what happens during the first two rounds of the label propagation algorithm. Our goal is to show that after round two, a.a.s. almost all vertices carry a label in $[k]$.

To begin with, note that, after the first round, every vertex keeps its own label or switches to a smaller label. So, in particular, all vertices in Level 1 get a label from $[2k]$ after the first round. More importantly, after the first round, every vertex in Level 2 also gets a label from $[2k]$: indeed, while initially every vertex $v \in B$ is assigned a label in $[2k+1, n]$, v has a neighbor in Level 1 with label in $[2k]$. Recall that the set of vertices in Level 2 that get label $\ell \in [2k]$ after the first round (that is, $B_1(\ell)$) is referred to as the *basin* of vertex v_ℓ. Observe that a vertex $v \in B$ belongs to the basin of v_ℓ if and only if v is a neighbor of v_ℓ and is not a neighbor of vertices $v_1, \ldots, v_{\ell-1}$. Thus,

$$B_1(\ell) = N(v_\ell) \setminus \left(A \cup \bigcup_{i=1}^{\ell-1} N(v_i)\right).$$

We now formally state the main result of this section.

Lemma 3.1. *Fix $\varepsilon > 0$ and suppose that $np \geq n^{1/2+\varepsilon}$. Fix $K = K(n) = \lceil 2(\log n)/p \rceil$. Then, a.a.s., after the second round, each vertex v_j with $j \in [n] \setminus [k+1, K-1]$ carries a label in $[k]$. Moreover, a.a.s. every label j between $k+1$ and K may be carried only by the vertex v_j after the second round.*

Before we move to the proof of the above lemma, note that, on the one hand, $k = 1$ if $p \geq \sqrt{15}(\log n/n)^{1/4}$. Also, observe that, when $p = o(1)$, one may easily show that a.a.s. the vertex v_2 does not change its label during the first two rounds (indeed, in this case, v_2 a.a.s. is not adjacent to v_1, and the total number of common neighbors of v_1 and v_2 is a.a.s. of smaller order than the total number of neighbors of v_2). We will then use Lemma 3.1 to correct the main result from [20]. On the other hand, for sparser graphs, $k = k(n)$ tends to infinity as $n \to \infty$, and in this case we obtain improved results at the price of several further technical steps.

Proof of Lemma 3.1. Note that all vertices in $V([k]) \subseteq A$, as well as their neighbors, get a label from $[k]$ after the first round. So, even if some vertices in $V([k])$ switch their labels during the second round, these labels remain in $[k]$. Hence, the desired property holds for all vertices in $V([k])$. It is important to notice that we do not need to expose *any* edges of $\mathcal{G}(n, p)$ to conclude this.

Next, we show that for every two vertices $u, v_j \in V \setminus V([k])$ (so, in particular, $j > k$), with probability $1 - o(n^{-2})$, the number of common neighbors of v_1 and u is larger than the number of vertices in $N[u] \cap N[v_j] \setminus B_1([k])$. This implies that u obtains label j after the second round with probability $o(n^{-2})$. On the one hand, the number of common neighbors of u and v_1 is a random variable Y_1 with distribution $\text{Bin}(n - 2, p^2)$. On the other hand, the number of vertices in $N[u] \cap N[v_j] \setminus B_1([k])$ is dominated by $Y_j + 2$ where Y_j is a binomial random variable with distribution $\text{Bin}(n - 2, q^k p^2)$: indeed, except for vertices u and v_j that could carry label j after the first round, the remaining relevant vertices must be neighbors of u (to influence its label), neighbors of v_j (to obtain label j after the first round) but not adjacent to any of $v_1, \ldots, v_k$, which happens with probability $q^k p^2$.

Denote $r = \mathbb{E}[Y_1 - Y_j] = p^2(n - 2)(1 - q^k) = (1 - o(1))np^3 k$, where for the last equality we used that $1 - q^k = 1 - (1 - p)^k = (1 - o(1))pk$. Clearly,

$$\mathbb{P}(Y_1 \leq Y_j + 2) \leq \mathbb{P}(Y_1 - \mathbb{E}Y_1 \leq -r/2 + 1) + \mathbb{P}(Y_j - \mathbb{E}Y_j \geq r/2 - 1). \tag{4}$$

Thus, by Chernoff's bound, the probabilities on the right-hand side above are bounded from above by $\exp\left(-\frac{(r-2)^2/4}{2(\mathbb{E}Y_1 + (r-2)/6)}\right)$ and $\exp\left(-\frac{(r-2)^2/4}{2(\mathbb{E}Y_j + (r-2)/6)}\right)$, respectively. As a result,

$$\mathbb{P}(Y_1 \leq Y_j + 2) \leq 2\exp\left(-\frac{(r-2)^2/4}{2(\mathbb{E}Y_1 + (r-2)/6)}\right) = 2\exp\left(-(1 + o(1))\frac{n^2 p^6 k^2}{8np^2}\right).$$

By using that $np^4 k^2 \geq 225 \log n$, we conclude by a union bound that, for all pairs (u, v_j) with $j > k$ and $u \neq v_j$, u does not get label j after round 2.

Finally, fix $j \geq K$. It follows that $|N(v_1) \cap N(v_j)|$ dominates a random variable Z_j with distribution $\text{Bin}(n - 2, p^2)$, while $|N[v_j] \setminus N[V([j-1])]|$ is dominated by a random variable $W_j + 1$ where W_j is distributed according to $\text{Bin}(n - j, pq^{j-1})$. Hence, using that for $j > K$,

$$\mathbb{E}Z_j = (n - 2)p^2 \gg npq^{j-1} + 1 \geq \mathbb{E}[W_j + 1],$$

an argument similar to (4) with $\frac{1}{3}\mathbb{E}Z_j$ instead of r shows that

$$\mathbb{P}(Z_j \leq W_j + 1) \leq 2\exp\left(-\frac{(\frac{1}{3}\mathbb{E}Z_j)^2}{2(\mathbb{E}Z_j + \frac{1}{9}\mathbb{E}Z_j)}\right) = \exp(-\Omega(np^2)).$$

Thus, a union bound implies that, for every $j \in [K, n]$, v_j does not carry label j after the second round, as desired. □

At this point, we are able to recover fast convergence of the label propagation algorithm when $k = 1$. We need the following more general lemma.

Lemma 3.2. *Fix $\varepsilon > 0$ and suppose that $np \geq n^{1/2+\varepsilon}$. Also, suppose that at some round R of LPA, at least $0.9n$ of the vertices in $\mathcal{G}(n, p)$ have the same label ℓ. Then, after two more rounds, a.a.s. every vertex carries label ℓ.*

Proof. Let $S, T \subseteq V$ be not necessarily disjoint sets such that $s = |S| = 300p^{-1}$ and $t = |T| = 0.9n$ (one can think of T as vertices having label ℓ at round R, and of S as vertices "trying to avoid" label ℓ at round $R + 1$). We will show that, with probability $1 - o(\binom{n}{s}^{-1}\binom{n}{t}^{-1})$, there is at least one vertex in S which has strictly more neighbors in T than in $V \setminus T$.

On the one hand, observe that the number of potential edges between S and $V \setminus T$ is at most $|S| \cdot |V \setminus T| \leq 0.1sn$, and that the number of potential edges between S and T is at least $s(t - s) \geq 0.8sn$. Thus, the number of edges in $\mathcal{G}(n, p)$ between S and $V \setminus T$ is dominated by $X \in \text{Bin}(0.1sn, p)$ and the number of edges between S and T dominates $Y \in \text{Bin}(0.8sn, p)$. Hence, the probability that, for every vertex $u \in S$,

$|N[u] \cap S| \geq |N(u) \cap T|$ (which is necessary for the vertex u to obtain label different from ℓ at round $R+1$), is at most

$$\mathbb{P}(2X + s \geq Y) \leq \left(\mathbb{P}(X \geq 0.2snp) + \mathbb{P}(Y \leq 0.5snp)\right)$$

$$\leq 2\exp\left(-\frac{snp}{100}\right) = o\left(\binom{n}{s}^{-1}\binom{n}{t}^{-1}\right).$$

Hence, a union bound over the (at most) $\binom{n}{s}\binom{n}{t}$ pairs of sets S, T with $|S| = s$ and $|T| = t$, allows us to deduce that, for every set S as above, at least one vertex in S obtains label ℓ at the next round. In particular, a.a.s. all but at most $300p^{-1}$ vertices carry label ℓ at the next round. Finally, since a.a.s. every vertex has degree $\Omega(np) \gg p^{-1}$, most of the neighbors of every vertex carry label ℓ after round $R+1$, which implies that a.a.s. all vertices get the same label at round $R+2$, as desired. $\qquad\square$

The next corollary is a direct consequence of the fact that $k = 1$ in the range $np \geq \sqrt{15}(\log n)^{1/4}n^{3/4}$, Lemmas 3.1 and 3.2, and the fact that $n - \frac{2}{p}\log n \geq 0.9n$ for all sufficiently large n.

Corollary 3.3. *Suppose that $np \geq \sqrt{15}(\log n)^{1/4}n^{3/4}$. Then, a.a.s. the label propagation algorithm attributes label 1 to all vertices after three rounds.*

Thus, in the sequel, we assume that $np \leq \sqrt{15}(\log n)^{1/4}n^{3/4}$.

3.2 The regime $\Omega(n^{2/3}) = np \leq \sqrt{15}(\log n)^{1/4}n^{3/4}$

In this section, we specify a suitable range of p in the statement of each result since these ranges may sometimes differ. Note that while some results may be shown in more generality, the range is often restricted to the one we need in the sequel.

From Section 3.1 we know that after round 2, a.a.s. all but $O((\log n)/p)$ vertices receive a label in $[k]$. In this section, we establish that after round 5, a.a.s. every vertex has label 1. Most of our effort concentrates on showing that after round 3, a.a.s. every vertex in Level 3 has label 1. This and the observation that a.a.s. the number of vertices in Levels 1 and 2 is $o(n)$ suggests that soon after round 3, all vertices should get label 1.

3.2.1 Properties of the basins

To begin with, we establish Lemma 3.4 (showing that for suitably chosen p, the first basin is close to its mean and the gap between the first and the other basins is sufficiently big), and the stronger Lemma 3.8 (showing that for slightly larger values of p, the sizes of all basins are close to their means). For this, recall that $\mathfrak{B}_1(\ell)$ denotes the size of the basin of v_ℓ, that is, $\mathfrak{B}_1(\ell) = |B_1(\ell)|$. Observe that $\mathfrak{B}_1(\ell) \in \text{Bin}(n - 2k, q^{\ell-1}p)$, so for $\ell \in [2k]$, since $\ell p \leq kp = o(1)$ and $k = o(n)$, we have

$$\mathbb{E}\mathfrak{B}_1(\ell) = (n - 2k)q^{\ell-1}p \sim np.$$

Now, fix

$$\omega = \omega(n) = ((np)^{1/2} \cdot np^2)^{1/2} = n^{3/4}p^{5/4}.$$

In particular, this choice guarantees that $(np)^{1/2} \ll \omega \ll np^2$ as long as $np \gg n^{2/3}$.

Lemma 3.4. *Suppose that $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, the event*

$$\mathcal{E} = \{\text{for every } \ell \in [2, 2k], \ \mathfrak{B}_1(1) - \mathfrak{B}_1(\ell) \geq \tfrac{1}{1.4}(\ell - 1)np^2\} \cap \{|\mathfrak{B}_1(1) - \mathbb{E}\mathfrak{B}_1(1)| \leq \omega\}$$

holds a.a.s.

Proof. Recall that $\mathfrak{B}_1(1)$ and $\mathfrak{B}_1(\ell)$ are binomial random variables with means $(n-2k)p$ and $(n-2k)pq^{\ell-1}$, respectively. Hence, since $\ell p \leq 2kp = o(1)$ and $k = o(n)$,

$$\mathbb{E}\mathfrak{B}_1(1) - \mathbb{E}\mathfrak{B}_1(\ell) = (n-2k)p(1 - (1-p)^{\ell-1}) = (1+o(1))(\ell-1)np^2.$$

Moreover, using that $1 - \frac{1}{1.4} > \frac{1}{5}$, by Chernoff's bound,

$$\begin{aligned}
&\mathbb{P}(\mathfrak{B}_1(1) - \mathfrak{B}_1(\ell) \leq \tfrac{1}{1.4}(\ell-1)np^2) \\
&\leq \mathbb{P}(\mathfrak{B}_1(1) - \mathbb{E}\mathfrak{B}_1(1) \leq -\tfrac{1}{10}(\ell-1)np^2) + \mathbb{P}(\mathfrak{B}_1(\ell) - \mathbb{E}\mathfrak{B}_1(\ell) \geq \tfrac{1}{10}(\ell-1)np^2) \\
&\leq \exp\left(-\frac{((\ell-1)np^2/10)^2}{2\mathbb{E}\mathfrak{B}_1(1)}\right) + \exp\left(-\frac{((\ell-1)np^2/10)^2}{2(\mathbb{E}\mathfrak{B}_1(\ell) + (\ell-1)np^2/30)}\right) \\
&\leq 2\exp\left(-\frac{((\ell-1)np^2)^2}{300np}\right) \\
&= 2\exp\left(-\frac{(\ell-1)^2}{300}np^3\right).
\end{aligned}$$

Since $np^3 \gg 1$, summing over the range $\ell \in [2, 2k]$ shows that the first event in the intersection that determines $\mathcal{E}$ holds a.a.s. For the second event therein, using Chernoff's bound and the fact that $\omega^2 = n^{3/2}p^{5/2} \gg \mathbb{E}\mathfrak{B}_1(1)$ shows that it also holds a.a.s. and finishes the proof. $\square$

Remark 3.5. In fact, the proof of Lemma 3.4 also implies the following result. Suppose that $np = cn^{2/3}$ for some constant $c > 0$ (in particular, $np^3 = c^3$ and $\omega^2 = \Theta(\mathbb{E}\mathfrak{B}_1(1))$). Then, for every $\varepsilon \in (0,1)$ and every sufficiently large positive integer $L = L(\varepsilon, c)$, the event

$$\begin{aligned}
\mathcal{E}_L = &\{\text{for every } \ell \in [L+1, 2k], \, \mathfrak{B}_1(1) - \mathfrak{B}_1(\ell) \geq \tfrac{1}{1.4}(\ell-1)np^2\} \\
&\cap \{|\mathfrak{B}_1(1) - \mathbb{E}\mathfrak{B}_1(1)| \leq Ln^{1/3}\}
\end{aligned}$$

holds with probability at least $1 - \varepsilon$.

Although Remark 3.5 extends Lemma 3.4 in the case $np = \Theta(n^{2/3})$, it fails to provide any information for the few largest basins. To fill this gap, Lemma 3.7 shows that their sizes are sufficiently far from each other with probability close to 1. Before going to the proof itself, we show a technical lemma that may itself be of independent interest.

Lemma 3.6. *Fix $L \in \mathbb{N}$, a set of $L+1$ colors, and suppose that $\widehat{p} = \widehat{p}(n)$ is a real number satisfying $n\widehat{p} = cn^{2/3}$ for some constant $c \in (0, \infty)$. Color the elements in $[n]$ independently so that for every $j \in [n]$, j obtains color $i \in [L]$ with probability $\widehat{p}$, and color $L+1$ with probability $1 - L\widehat{p}$. Denote by X_i the number of vertices in color i, and set $Y_i = \frac{X_i - \widehat{p}n}{\sqrt{\widehat{p}(1-\widehat{p})n}}$. Then,*

$$(Y_1, \ldots, Y_L) \xrightarrow[n\to\infty]{d} (N_1, \ldots, N_L),$$

where $(N_i)_{i=1}^{L}$ are i.i.d. normal random variables with mean 0 and variance 1.

Proof. Define $(\widehat{\mathcal{X}}_i)_{i=1}^{L}$ as L independent subsets of $[n]$ where every number $j \in [n]$ belongs to $\widehat{\mathcal{X}}_i$ with probability

$$\widehat{q} = 1 - (1 - L\widehat{p})^{1/L} = \widehat{p} + \tfrac{1}{2}(L-1)\widehat{p}^2 + O(\widehat{p}^3),$$

where we used that for every $\alpha > 0$, $(1+x)^\alpha = 1 + \alpha x + \tfrac{1}{2}\alpha(\alpha-1)x^2 + O(x^3)$ as $x \to 0$. Set $\widehat{X}_i = |\widehat{\mathcal{X}}_i|$ and $\widehat{Y}_i = \frac{\widehat{X}_i - \widehat{q}n}{\sqrt{\widehat{q}(1-\widehat{q})n}}$. Then, by the central limit theorem for independent binomial random variables,

$$(\widehat{Y}_1, \ldots, \widehat{Y}_L) \xrightarrow[n\to\infty]{d} (\widehat{N}_1, \ldots, \widehat{N}_L), \tag{5}$$

12

where $(\widehat{N}_i)_{i=1}^L$ are i.i.d. normal random variables with mean 0 and variance 1.

We construct the random variables $(X_i)_{i=1}^L$ from the sets $(\widehat{\mathcal{X}}_i)_{i=1}^L$. For every integer j belonging to at least one of the sets $(\widehat{\mathcal{X}}_i)_{i=1}^L$, associate a random variable U_j that is uniformly distributed over $\{i:\ j \in \widehat{\mathcal{X}}_i\}$ so that $(U_j)_{j=1}^n$ are independent. Then, for every $i \in [L]$, denote $\mathcal{X}_i = \{j \in [n]\colon U_j \text{ exists and } U_j = i\}$ and $X_i = |\mathcal{X}_i|$. Note that the probability to belong to $\mathcal{X}_i$ is the same for all $i \in [L]$, and it is exactly $\frac{1}{L}(1 - (1c\widehat{q})^L) = \widehat{p}$, so $(X_i)_{i=1}^L$ have the desired distribution.

Now, on the one hand, an element $j \in [n]$ belongs to at least three sets among $(\widehat{\mathcal{X}}_i)_{i=1}^L$ with probability $O(n^{-1})$. Thus, Markov's inequality shows that a.a.s. the number of these elements is no more than $n^{1/6} = o(n^{1/3})$. On the other hand, for every pair of distinct $i_1, i_2 \in [L]$, Chernoff's bound implies that a.a.s. the number of elements j belonging to $\widehat{\mathcal{X}}_{i_1}$ and $\widehat{\mathcal{X}}_{i_2}$ and to no other set among $(\widehat{\mathcal{X}}_i)_{i \in [L]\setminus\{i_1,i_2\}}$, and satisfying that $U_j = i_1$, is equal to $\frac{1}{2}\widehat{q}^2 n + o(\widehat{q}^2 n) = \frac{1}{2}c^2 n^{1/3} + o(n^{1/3})$. We conclude that a.a.s. for every $i \in [L]$,

$$\widehat{X}_i - \widehat{q}n = \left(X_i - \widehat{p}n - \tfrac{1}{2}(L-1)\widehat{p}^2 n + O(\widehat{p}^3 n)\right) + \tfrac{1}{2}(L-1)c^2 n^{1/3} + o(n^{1/3}) = (X_i - \widehat{p}n) + o(n^{1/3}).$$

Combining this with (5) and the fact that $\sqrt{\widehat{q}(1-\widehat{q})n} = (1+o(1))\sqrt{\widehat{p}(1-\widehat{p})n}$ finishes the proof. $\square$

Lemma 3.7. *Suppose that $np = cn^{2/3}$ for some constant $c > 0$. For every $\varepsilon \in (0,1)$ and every positive integer L, there is $\delta = \delta(\varepsilon, L, c) > 0$ such that the event*

$$\mathcal{G}_{L,\varepsilon} = \left\{ \min_{i,j \in [L]: i \neq j} |\mathfrak{B}_1(i) - \mathfrak{B}_1(j)| \geq \delta n p^2 \right\}$$

holds with probability at least $1 - \varepsilon$.

Proof. For every set $S \subseteq [L]$, denote by X_S the number of vertices in $V \setminus A$ that connect to all vertices in S and do not connect to the vertices in $[L] \setminus S$. If $S = \{i\}$ or $S = \{i,j\}$ for some $i,j \in [L]$, for simplicity of notation we denote X_i and $X_{i,j}$ instead of $X_{\{i\}}$ and $X_{\{i,j\}}$, respectively.

Similarly to the proof of Lemma 3.6, we have that a.a.s.

$$\max_{S \subseteq [L]: |S| \geq 3} X_S = O(n^{1/6}), \tag{6}$$

and

$$X_{i,j} = (1 + o(1))\mathbb{E}X_{i,j} = (1 + o(1))c^2 n^{1/3}. \tag{7}$$

Moreover, for every $i \in [L]$, denote

$$Y_i = \frac{X_i - (n-2k)pq^{L-1}}{\sqrt{(n-2k)pq^{L-1}(1 - pq^{L-1})}}.$$

Thus, by applying Lemma 3.6 with $\widehat{p} = pq^{L-1}$ and $n - 2k$ instead of n,

$$(Y_1, \ldots, Y_L) \xrightarrow[n\to\infty]{d} (N_1, \ldots, N_L), \tag{8}$$

where $(N_i)_{i=1}^L$ are i.i.d. normal random variables with mean 0 and variance 1.

Now, it remains to notice that for every $i \in [L]$, $\mathfrak{B}_1(i) = \sum_{S \subseteq [L]: \min S = i} X_S$. For every $i \in [L]$, denote

$$Z_i = \frac{\mathfrak{B}_1(i) - (n-2k)pq^{L-1}}{\sqrt{(n-2k)pq^{L-1}(1 - pq^{L-1})}} = Y_i + \sum_{S \subseteq [L]: \min S = i, |S| \geq 2} \frac{X_S}{\sqrt{(n-2k)pq^{L-1}(1 - pq^{L-1})}}.$$

Then, combining (6), (7), (8) and the fact that $\sqrt{(n-2k)pq^{L-1}(1 - pq^{L-1})} = (1 + o(1))\sqrt{c}n^{1/3}$ implies that

$$(Z_1, \ldots, Z_L) \xrightarrow[n\to\infty]{d} (N_i + (L-i)c^{3/2})_{i=1}^L. \tag{9}$$

13

In particular, there is a $\delta = \delta(\varepsilon, L, c) > 0$ such that

$$\min_{i,j \in [L]: i \neq j} |N_i - N_j + (j - i)c^{3/2}| \geq 2\delta c^{3/2}$$

with probability at least $1 - \frac{\varepsilon}{2}$, which combined with (9) implies that for all sufficiently large n,

$$\min_{i,j \in [L]: i \neq j} |Z_i - Z_j| \geq \delta c^{3/2}$$

holds with probability at least $1 - \varepsilon$. (Note that the factor 2 disappeared to take into account the error coming from the convergence in distribution.) Coming back to $(\mathfrak{B}_1(i))_{i=1}^L$ finishes the proof. □

We will also need the following lemma for larger values of p.

Lemma 3.8. *Suppose that* $np \in [n^{3/4}(\log n)^{-1/2}, \sqrt{15}n^{3/4}(\log n)^{1/4}]$. *Then, the event*

$$\mathcal{E}' = \{\text{for each } \ell \in [2k], \mathfrak{B}_1(\ell) \in \mathbb{E}\mathfrak{B}_1(\ell) \pm \ell\omega\}$$

holds a.a.s.

Proof. Since $p \gg n^{-1/3}$,

$$(np^{-1})^{1/4}k^{-1} = \Omega((n^3 p^7)^{1/4}(\log n)^{-1/2}) = \Omega(n^{1/6}(\log n)^{-1/2}) \gg 1.$$

Hence, $\mathbb{E}\mathfrak{B}_1(\ell) \sim np = (np^{-1})^{1/4}\omega \gg k\omega \geq \ell\omega$. Recalling that $\omega/(np)^{1/2} = (np^3)^{1/4}$ and applying Chernoff's bound we obtain, for n large enough and uniformly in $\ell \in [2k]$,

$$\mathbb{P}(\mathfrak{B}_1(\ell) \notin \mathbb{E}\mathfrak{B}_1(\ell) \pm \ell\omega) \leq 2\exp\left(-\frac{(\ell\omega)^2}{2(\mathbb{E}\mathfrak{B}_1(\ell) + \ell\omega)}\right) \leq \exp\left(-\frac{1}{3}\ell^2(np^3)^{1/2}\right).$$

In particular, a union bound yields

$$\mathbb{P}(\exists \ell \in [2k], \mathfrak{B}_1(\ell) \notin \mathbb{E}\mathfrak{B}_1(\ell) \pm \ell\omega) \leq \sum_{\ell=1}^{2k} \exp\left(-\frac{1}{3}\ell^2(np^3)^{1/2}\right)$$

$$\leq \exp\left(-\frac{1}{4}(np^3)^{1/2}\right) = o(1),$$

which proves the lemma. □

Remark 3.9. When $np \in [n^{3/4}(\log n)^{-1/2}, \sqrt{15}n^{3/4}(\log n)^{1/4}]$, on the event $\mathcal{E}'$ we have, for n large and uniformly in $\ell \in [2k - 1]$,

$$\begin{aligned}
\mathfrak{B}_1(\ell) - \mathfrak{B}_1(\ell + 1) &\geq (q^{\ell-1}p(n - 2k) - \omega\ell) - (q^\ell p(n - 2k) + \omega(\ell + 1)) \\
&= q^{\ell-1}p^2(n - 2k) - \omega(2\ell + 1) \geq \frac{1}{1.4}np^2.
\end{aligned} \tag{10}$$

In particular, the conclusion of Lemma 3.4 still holds.

Lemma 3.10. *Suppose that* $(n \log n)^{1/2} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. *The event*

$$\mathcal{F} = \left\{\frac{4}{3}knp \leq \mathfrak{B} \leq \frac{8}{3}knp\right\}$$

holds a.a.s.

Proof. Since $(\log n/n)^{1/2} \ll p \leq \sqrt{15}(\log n/n)^{1/4}$, we get that $kp = \Theta(p^{-1}(\log n/n)^{1/2}) = o(1)$. As a result, since $\mathfrak{B} \in \text{Bin}(n - 2k, 1 - q^{2k})$ and $k = o(n)$, we have $\mathbb{E}\mathfrak{B} = (n - 2k)(1 - q^{2k}) = (2 - o(1))knp$ and the lemma follows directly from Chernoff's bound. □

Note that Lemma 3.10 holds for a wider range of np than needed in this section, and it will be used in the proof of both the first and the second point of Theorem 1.1.

3.2.2 Consequences of the basin sizes: the second round

In this section, we mostly concentrate on the regime $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. Modifications for the regime $np = \Theta(n^{2/3})$ are minor and mostly consist in the fact that the event $\mathcal{E}_L$ (defined as in Remark 3.5) does not concern the basins of $v_2, \ldots, v_L$. We discuss these modifications in remarks after the corresponding lemmas for the first regime.

Let us denote

$$\Lambda = \Lambda(n,p) = \frac{1}{2} + \frac{1}{5}\min\left\{1, \frac{\sqrt{np^4}}{2}\right\}.$$

Next, we show that conditionally on $\mathcal{E} \cap \mathcal{F}$, a.a.s. after the second round, the number of vertices of label ℓ in Level 3 decreases as a function of ℓ. Here, Lemma 3.1 is crucial since it establishes that the label of almost all vertices v in Level 3 can a.a.s. be attributed based only on the edges between v and vertices in Level 2. In fact, the vertices for which the above does not hold are so few that it will be convenient to think of them as very rare exceptions that, as we shall see, do not have any significant influence on LPA when n is large. In order to formalize this, we now define an alternative label propagation procedure.

The Alternative Label Attribution Procedure (ALAP). At round 1, for every $i \in [2k]$, expose consecutively the edges from v_i to $V \setminus (N(V([i-1])) \cup A)$. Also, expose all edges between vertices in A. On the basis of these edges, attribute labels to the vertices in $A \cup B$ only (which coincide with the labels these vertices receive after round 1 in LPA).

At round 2, given a vertex $u \in B \cup C$, expose the edges from u to B and denote by $U \subseteq [2k]$ the set of labels such that u has the same number of neighbors in $B_1(i)$ for every $i \in U$, and strictly less in $B_1(i)$ for every $i \in [2k] \setminus U$. Then, we pick one label from U uniformly at random and attribute it to u.

Finally, at round 3, we expose the edges between vertices in C and let ALAP evolve according to the same rules as LPA on the set of labels present after round 2.

Two important remarks are due at this point. First, note that by Lemma 3.1 LPA and ALAP on $G(n,p)$ can be coupled so that a.a.s. only the vertices in $V([K] \setminus [2k])$ may receive different labels at the second round. We call such incorrect label attributions (from the point of view of LPA) *mistakes* of ALAP and show that, essentially, their presence does not affect the re-partition of the remaining labels. Second, note that the edges exposed at the first round of ALAP allow to define the basins $(B_1(i))_{i=1}^{2k}$ (which coincide with the basins for LPA). In particular, all events defined in Section 3.2.1 are measurable with respect to the edges exposed at round 1 of ALAP and the results from that section still hold. Thus, from this point on, by abuse we use the notation and terminology introduced for LPA for ALAP instead.

Lemma 3.11. *Suppose that* $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. *The event "after the second round of ALAP, there are at least* $\frac{1}{2}(2\Lambda - 1)n$ *vertices in Level 3 with label 1, and the number of vertices with label* $\ell \in [2, 2k]$ *is at least by* $\frac{1}{2}\mathfrak{C}_2(1)\left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)$ *smaller than the number of vertices with label 1", that is,*

$$\left\{\mathfrak{C}_2(1) \geq \tfrac{1}{2}(2\Lambda - 1)n \text{ and for every } \ell \in [2, 2k], \mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \geq \tfrac{1}{2}\left(1 - \left(\tfrac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\mathfrak{C}_2(1)\right\},$$

holds a.a.s.

Proof. Fix $t_n = np^2/\sqrt{2}$. Condition on $\mathcal{E}$ (as in Lemma 3.4), on $\mathcal{F}$ (as in Lemma 3.10), and the edges (and non-edges) incident to all vertices in Level 1. Moreover, if

$$np \in [n^{3/4}(\log n)^{-1/2}, \sqrt{15}n^{3/4}(\log n)^{1/4}],$$

condition on $\mathcal{E}'$ as well (as in Lemma 3.8). Note that, by definition of ALAP, $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$ are all measurable in terms of the edges between Levels 1 and 2. Since we are conditioning on $\mathcal{E}$ and $\mathcal{F}$ which hold a.a.s. (by Lemma 3.4 and Lemma 3.10), if $np \leq n^{3/4}(\log n)^{-1/2}$, it is sufficient to show the conclusion of the lemma

conditionally on $\mathcal{E} \cap \mathcal{F}$, while if $np \in [n^{3/4}(\log n)^{-1/2}, \sqrt{15}n^{3/4}(\log n)^{1/4}]$, we condition on $\mathcal{E} \cap \mathcal{E}' \cap \mathcal{F}$ instead (we may do so since $\mathcal{E}'$ holds a.a.s.).

Fix a vertex $u \in C$ and, for every $\ell \in [2k]$, define p_ℓ as the probability that u received label ℓ in ALAP after round 2. Our goal will be to compare p_1 and p_ℓ. First, recall that, conditionally under the event $\mathcal{E} \cap \mathcal{F}$ (or $\mathcal{E} \cap \mathcal{E}' \cap \mathcal{F}$, respectively), we have $\mathfrak{B}_1(1) \geq \mathfrak{B}_1(\ell) + (\ell - 1)t_n$. Then, we apply Lemma 2.6 with $n = \mathfrak{B}_1(1)$, $n' = \mathfrak{B}_1(\ell)$, M equal to the maximum number of edges between u and v_i for $i \in [2k] \setminus \{1, \ell\}$, and ρ being the number of times that this maximum is attained over the set $i \in [2k] \setminus \{1, \ell\}$. This shows that

$$\frac{p_1}{p_\ell} \geq \frac{\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(1), p) > \mathrm{Bin}(\mathfrak{B}_1(\ell), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(1), p) = \mathrm{Bin}(\mathfrak{B}_1(\ell), p))}{\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) > \mathrm{Bin}(\mathfrak{B}_1(1), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) = \mathrm{Bin}(\mathfrak{B}_1(1), p))}. \tag{11}$$

Using the inequality $\mathfrak{B}_1(1) \geq \mathfrak{B}_1(\ell) + (\ell - 1)t_n$, the latter further implies that

$$\frac{p_1}{p_1 + p_\ell} \geq \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(1), p) > \mathrm{Bin}(\mathfrak{B}_1(\ell), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(1), p) = \mathrm{Bin}(\mathfrak{B}_1(\ell), p)) \tag{12}$$
$$\geq \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(1), p) \geq \mathrm{Bin}(\mathfrak{B}_1(1) - (\ell - 1)t_n, p) + 1).$$

Further, we apply Lemma 2.3 with $a_1 = \mathfrak{B}_1(1)$, $a_2 = \mathfrak{B}_1(1) - (\ell - 1)t_n - 1/p = (1 - o(1))np$ (where we use that the difference of $(\ell - 1)np^2/1.4$ guaranteed between the basins by the event $\mathcal{E}$ dominates $(\ell - 1)t_n + 1/p$), $M = 1$ and p; note that its assumptions are satisfied as, under the event $\mathcal{E}$, we have that $1 \ll a_2 \leq a_1$ and $\min\{a_2, a_1 - a_2\}p \geq t_n p = np^3/\sqrt{2} \gg 1$. Since

$$\frac{(a_1 - a_2)p - M}{\sqrt{(a_1 + a_2)pq}} \geq \frac{(\ell - 1)t_n p}{\sqrt{2pa_1}} = \frac{(\ell - 1)np^3}{2\sqrt{pa_1}}$$

and $\frac{np}{4} \leq a_2 \leq a_1 \leq (n - 2k)p + \omega \leq np(1 + p)$, we deduce that

$$\frac{p_1}{p_1 + p_\ell} \geq \Phi\left(\frac{(\ell - 1)np^3}{2\sqrt{np^2(1 + p)}}\right) - \frac{4\zeta}{\sqrt{np^2}}, \tag{13}$$

which leads to

$$p_\ell \leq \frac{1 - \Phi\left(\frac{(\ell-1)np^3}{2\sqrt{np^2(1+p)}}\right) + \frac{4\zeta}{\sqrt{np^2}}}{\Phi\left(\frac{(\ell-1)np^3}{2\sqrt{np^2(1+p)}}\right) - \frac{4\zeta}{\sqrt{np^2}}}p_1 = \frac{1 - \Phi\left(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\right) + \frac{4\zeta}{\sqrt{np^2}}}{\Phi\left(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\right) - \frac{4\zeta}{\sqrt{np^2}}}p_1. \tag{14}$$

Now, we show that the expression on the right-hand side is at most $\left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}p_1$. We do this in two steps. First, suppose that $np^4 = o((\log n)^{-1})$. Note that

$$\left(\frac{1 - \Lambda}{\Lambda}\right)^{\ell-1} = \left(\frac{1 - \sqrt{np^4}/5}{1 + \sqrt{np^4}/5}\right)^{\ell-1} = \left(1 - \frac{2}{5}\sqrt{np^4} + O(np^4)\right)^{\ell-1}$$
$$= \exp\left(-\frac{2}{5}(\ell - 1)\sqrt{np^4} + O((\ell - 1)np^4)\right), \tag{15}$$

and if $(\ell - 1)\sqrt{np^4} \leq \varepsilon$ for some sufficiently small $\varepsilon > 0$, then using that $(1 + p)^{-1/2} = 1 + O(p)$, we get

$$\Phi\left(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\right) - \frac{4\zeta}{\sqrt{np^2}} = \frac{1}{2} + \frac{(\ell - 1)\sqrt{np^4}}{2\sqrt{2\pi}} + O\left((\ell - 1)^2 np^4 + (\ell - 1)\sqrt{np^4} \cdot p + \frac{1}{\sqrt{np^2}}\right).$$

Consequently, using that $(\ell - 1)\sqrt{np^4} \cdot p = o((\ell - 1)^2 np^4)$,

$$\frac{1 - \Phi\left(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\right) + \frac{4\zeta}{\sqrt{np^2}}}{\Phi\left(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\right) - \frac{4\zeta}{\sqrt{np^2}}} = 1 - \frac{2(\ell - 1)\sqrt{np^4}}{\sqrt{2\pi}} + O\left((\ell - 1)^2 np^4 + \frac{1}{\sqrt{np^2}}\right), \tag{16}$$

16

and using that $\frac{2}{5} < \frac{2}{\sqrt{2\pi}}$ shows the desired inequality when ε is sufficiently small. On the other hand, when $\varepsilon^{-1} \geq (\ell-1)\sqrt{np^4} \geq \varepsilon$, the inequalities $np^2 \gg 1$ and $\ell np^4 \leq k np^4 = O(\sqrt{np^4 \log n}) = o(1)$ ensure that it is sufficient to prove that for every $x > 0$, $\frac{1-\Phi(x)}{\Phi(x)} < \exp(-\frac{4}{5}x)$, or equivalently

$$\Phi(x)(1 + \exp(-\tfrac{4}{5}x)) - 1 > 0,$$

and then use this inequality for $x = \frac{1}{2}(\ell-1)\sqrt{np^4}$. The latter could be checked via tedious analysis; we provide a link[2] with a numerical justification instead (using that $\Phi(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}(\frac{x}{\sqrt{2}})$, where erf is the error function). Finally, if $(\ell-1)\sqrt{np^4} \geq \varepsilon^{-1}$ for some sufficiently small ε, using that $\frac{x}{1-x} \leq 2x$ when x is small together with (15), the left-hand side of (16) is at most

$$2\left(\left(1 - \Phi\left(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\right)\right) + \frac{4\zeta}{\sqrt{np^2}}\right) \leq 2\exp\left(-\frac{(\ell-1)^2 np^4}{8(1+p)}\right) + \frac{8\zeta}{\sqrt{np^2}} \ll \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1},$$

where to show that $1/\sqrt{np^2} \ll \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}$, we used that $(\ell-1)\sqrt{np^4} \leq k\sqrt{np^4} \leq \sqrt{15 \log n} \ll \log(np^2)$.

Now, suppose that $(\log n)^{-2} \leq np^4 \leq 225 \log n$ or equivalently

$$np \in [n^{3/4}(\log n)^{-1/2}, \sqrt{15} n^{3/4}(\log n)^{1/4}].$$

Then, on the event $\mathcal{E}'$, Remark 3.9 ensures that $\mathfrak{B}_1(\ell) - \mathfrak{B}_1(\ell+1) \geq t_n = np^2/\sqrt{2}$ for every $\ell \in [2k-1]$. By an application of Lemma 2.6 similar to (11) and (12), we obtain that

$$\frac{p_\ell}{p_\ell + p_{\ell+1}} \geq \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) > \mathrm{Bin}(\mathfrak{B}_1(\ell+1), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) = \mathrm{Bin}(\mathfrak{B}_1(\ell+1), p))$$

$$\geq \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) \geq \mathrm{Bin}(\mathfrak{B}_1(\ell) - t_n, p) + 1),$$

implying that

$$p_{\ell+1} \leq \frac{1 - \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) - \mathrm{Bin}(\mathfrak{B}_1(\ell) - t_n, p) \geq 1)}{\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) - \mathrm{Bin}(\mathfrak{B}_1(\ell) - t_n, p) \geq 1)} p_\ell.$$

Applying Lemma 2.3 for $a_1 = \mathfrak{B}_1(\ell)$, $a_2 = \mathfrak{B}_1(\ell) - t_n$, $M = 1$ and p leads to $p_{\ell+1} \leq \frac{1-\Lambda}{\Lambda} p_\ell$, which by an immediate induction leads to $p_{\ell+1} \leq (\frac{1-\Lambda}{\Lambda})^\ell p_1$.

Thus, recalling that $(p_\ell)_{\ell=1}^{2k}$ adds up to 1,

$$1 = \sum_{\ell=1}^{2k} p_\ell \leq p_1 \sum_{\ell=0}^{\infty} \left(\frac{1-\Lambda}{\Lambda}\right)^\ell = p_1 \frac{\Lambda}{2\Lambda - 1}. \tag{17}$$

Now, for all $\ell \in [2k]$, recall that $\mathfrak{C}_2(\ell)$ equals the number of vertices in Level 3 that get label ℓ at the second round. Since our vertex labeling procedure and the original algorithm may be coupled so that a.a.s. all vertices receive the same labels in both at the second round, we abuse notation and identify $\mathfrak{C}_2(\ell)$ with the number of vertices in Level 3 that get label ℓ at the second round in the procedure.

If $np^4 \geq 4$, then $\Lambda = 1/2 + 1/5$ and $k \leq \lceil 15(\log n)^{1/2}/2 \rceil$. Combining this with (17) shows that $\mathbb{E}\mathfrak{C}_2(1) \geq p_1 \mathfrak{C} \geq \frac{n}{2} \gg \log n$. Hence, by Chernoff's bound,

$$\mathbb{P}\big(\mathfrak{C}_2(1) \in p_1\mathfrak{C} \pm 2(p_1\mathfrak{C}\log n)^{1/2}\big) = 1 - o(n^{-1}),$$
$$\mathbb{P}\big(\mathfrak{C}_2(\ell) \leq p_\ell\mathfrak{C} + \max\{2(p_\ell\mathfrak{C}\log n)^{1/2}, (\log n)^2\}\big) = 1 - o(n^{-1}). \tag{18}$$

In particular, with probability $1 - o(n^{-1})$, we get that $\mathfrak{C}_2(1) \geq \frac{3}{4}p_1\mathfrak{C} \geq \frac{2}{3}p_1 n \geq \frac{1}{2}(2\Lambda - 1)n$, which proves the first part of the lemma. On the other hand, (18) implies that for every $\ell \in [2, 2k]$, with probability $1 - o(n^{-1})$,

$$\mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \geq \big(p_1\mathfrak{C} - 2(p_1\mathfrak{C}\log n)^{1/2}\big) - \big(p_\ell\mathfrak{C} + \max\{2(p_\ell\mathfrak{C}\log n)^{1/2}, (\log n)^2\}\big)$$

$$\geq (p_1 - p_\ell)\mathfrak{C} - 4(p_1\mathfrak{C}\log n)^{1/2}$$
$$\geq \frac{2}{3}\left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)p_1\mathfrak{C}$$
$$\geq \frac{1}{2}\left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\mathfrak{C}_2(1),$$

and by a union bound the statement of the lemma follows for the case $np^4 \geq 4$.

Now, consider the case $np^4 < 4$. By Chernoff's bound we have that for every $\varepsilon > 0$,

$$\mathbb{P}(|\mathfrak{C}_2(1) - \mathbb{E}\mathfrak{C}_2(1)| \geq (np_1)^{1/2+\varepsilon}) = O(n^{-2}).$$

Recalling that $1 \leq p_1\frac{\Lambda}{2\Lambda-1}$, we get $\mathbb{E}\mathfrak{C}_2(1) = (1 - o(1))np_1 \geq (2\Lambda - 1)n = \Omega(\sqrt{n^3p^4}) \geq (\log n)^2$. Hence, with probability $1 - O(n^{-2})$, we have $\mathfrak{C}_2(1) \geq \frac{1}{2}(2\Lambda - 1)n$, which proves the first part of the lemma. On the other hand,

$$\mathbb{P}\left(\mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \leq \left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\frac{\mathfrak{C}_2(1)}{2}\right) \leq \mathbb{P}\left(\mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \leq \left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\frac{2np_1}{3}\right)$$
$$+ \mathbb{P}\left(\mathfrak{C}_2(1) \geq \frac{4np_1}{3}\right).$$

Since $\mathbb{E}\mathfrak{C}_2(1) = (1-o(1))np_1$, by Chernoff's bound, the second term on the right-hand side above is $O(n^{-2})$, while the first is bounded from above by

$$\mathbb{P}\left(\mathfrak{C}_2(1) \leq np_1 - \left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\frac{np_1}{6}\right) + \mathbb{P}\left(\mathfrak{C}_2(\ell) \geq \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}np_1 + \left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\frac{np_1}{6}\right).$$

Using Chernoff's bound again (and recalling that $\mathbb{E}\mathfrak{C}_2(\ell) \leq np_\ell \leq (\frac{1-\Lambda}{\Lambda})^{\ell-1}np_1$), both probabilities above can be bounded from above by

$$\exp\left(-\left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)^2\frac{np_1}{100}\right) \leq \exp\left(-\left(\frac{2\Lambda-1}{\Lambda}\right)^2\frac{np_1}{100}\right) \leq \exp\left(-\frac{(2\Lambda-1)^3n}{100}\right) = O(n^{-2}),$$

where in the last inequality we used that $\frac{2\Lambda-1}{\Lambda} \leq p_1$, and for the equality we used that $(2\Lambda - 1)^3 = \Omega((np^4)^{3/2}) = \Omega(n^{-1/2})$. Summarizing, uniformly in $\ell \in [2, 2k]$,

$$\mathbb{P}\left(\mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \leq \left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\frac{\mathfrak{C}_2(1)}{2}\right) = O(n^{-2}).$$

The lemma follows by a union bound over all $\ell \in [2, 2k]$. $\qquad\square$

Remark 3.12. For $np = cn^{2/3}$ for some constant $c > 0$, Lemma 3.11 holds by replacing the original statement with "Given $\varepsilon \in (0, 1)$ and $L = L(\varepsilon, c)$ (provided by Remark 3.5), conditionally on $\mathcal{E}_L$, the event obtained by intersecting $\{\max_{i\in[L]}\mathfrak{C}_2(i) \geq \frac{1}{2}(2\Lambda - 1)n\}$ and

$$\left\{\text{for every } \ell \in [L+1, 2k], \max_{i\in[L]}\mathfrak{C}_2(i) - \mathfrak{C}_2(\ell) \geq \frac{1}{2}\left(1 - \left(\frac{1-\Lambda}{\Lambda}\right)^{\ell-1}\right)\max_{i\in[L]}\mathfrak{C}_2(i)\right\}$$

holds a.a.s.". The necessary modifications are as follows. First, at the beginning of the proof, we replace $\mathcal{E}$ by $\mathcal{E}_L$, and the applications of Lemma 2.3 become applications of Remark 2.5 instead. Define $p_* = \max_{i\in[L]} p_i$. Then, (13) and the consequent analysis holds for all $\ell \in [L+1, 2k]$ and p_* instead of p_1. Moreover, (17) must be replaced by

$$p_*\left(L + \frac{\Lambda}{2\Lambda - 1}\right) \geq 1.$$

Remark 3.13. Fix $np = cn^{2/3}$ for some constant $c > 0$, and define

$$\widehat{\ell}_1 = \min\left\{\ell \in [L] : \mathfrak{C}_2(\ell) = \max_{i \in [L]} \mathfrak{C}_2(i)\right\}. \tag{19}$$

Then, replacing Lemma 3.4 by Lemma 3.7, and Lemma 2.3 by Remark 2.4 in the proof of Lemma 3.11 implies that conditionally on the event of Lemma 3.7, a.a.s. for every $\ell \in [L] \setminus \{\widehat{\ell}_1\}$,

$$\mathfrak{C}_2(\widehat{\ell}_1) - \mathfrak{C}_2(\ell) = \Omega((2\Lambda - 1)\mathfrak{C}_2(\widehat{\ell}_1)). \tag{20}$$

Except replacing p_1 by p_* (as defined in Remark 3.12), no additional modifications are needed.

Our next goal will be to bound from above the number of vertices in Level 2 which carry the most frequent label in this level after the second round. The following observation will be a technical tool in the proof of this bound.

Observation 3.14. *Suppose that $\Omega(n^{5/8}) = np \le \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, every vertex in Level 1 and Level 2 is connected to at most 10 vertices in Level 1 a.a.s.*

Proof. For any $j \in [n]$, the number of neighbors of v_j in Level 1 is stochastically dominated by $\mathrm{Bin}(2k, p)$. Since $kp = p^{-1}n^{-1/2+o(1)} \le n^{-1/9}$, by a union bound over all vertices we conclude that

$$\mathbb{P}(\exists j \in [n], |N(v_j) \cap A| \ge 10) \le n\binom{2k}{10}p^{10} = O(n(kp)^{10}) = o(1),$$

as desired. $\qquad\square$

The next result is an analogue of Lemma 3.11 but concerning vertices at Level 2. However, unlike in Lemma 3.11 where $\mathfrak{C}_2(1)$ was approximated by a binomial random variable, the lower bound on $\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell)$ in Lemma 3.15 is given in terms of $\mathbb{E}\mathfrak{B}_2(1)$ and not of $\mathfrak{B}_2(1)$ itself due to the lack of an appropriate upper bound on $\mathfrak{B}_2(1)$.

Lemma 3.15. *Suppose that $n^{2/3} \ll np \le \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, a.a.s. for every $\ell \in [2, 2k]$, the difference between $\mathfrak{B}_2(\ell)$ and $\mathfrak{B}_2(1)$ is bounded from above by $44np$.*

Proof. Fix $t_n = np^2/\sqrt{2}$. As in the proof of Lemma 3.11, we condition on the events $\mathcal{E}$ (from Lemma 3.4), $\mathcal{F}$ (from Lemma 3.10), and the statement of Observation 3.14, which all hold a.a.s. Moreover, if $np \in [n^{3/4}(\log n)^{-1/2}, \sqrt{15}n^{3/4}(\log n)^{1/4}]$, we also condition on $\mathcal{E}'$ (see Lemma 3.8).

Step 1. Fix a vertex $u \in B$, its associated (random) set U (as defined in the second round of ALAP) and, for every $\ell \in [2k]$, define $\widehat{p}_\ell = \widehat{p}_\ell(u)$ as the probability that the uniformly chosen label from U given to u at the second round of ALAP is ℓ. In particular, the sum of $(\widehat{p}_\ell)_{\ell=1}^{2k}$ is 1. Despite the fact that $(\widehat{p}_\ell)_{\ell=1}^{2k}$ depend on the choice of a vertex in Level 2 (due to the label of the vertex itself as well as its edges towards Level 1), we will show that, for every $\ell \in [2, 2k]$ and almost all vertices $u \in B$, $\widehat{p}_1(u) - \widehat{p}_{\ell+1}(u)$ is uniformly bounded from below.

Recall that, at the end of round 1, only the edges (and non-edges) with two endpoints in A and the edges (and non-edges) $\{v_i\}_{i=1}^\ell \times B_1(\ell)$, $\ell \in [2k]$ have been exposed. Fix $\ell \in [2, 2k]$ and say a vertex $u \in B$ is ℓ-*good* if $u \in B_1(i)$ for some $i \notin \{1, \ell\}$ such that v_i is not incident to any of v_1 and v_ℓ. In particular, Observation 3.14 implies that, for every $\ell \in [2, 2k]$, the vertices in all but at most 22 basins in B are ℓ-good. Moreover, for every ℓ-good vertex u, conditionally on the event $u \in B_2(\{1, \ell\}) \cap B_1(i)$ for some $i \in [2k]$, the choice of a label depends only on currently unexposed edges towards $B_1(1) \cup (N(v_1) \cap \{v_j\}_{j=i+1}^{2k})$ and $B_1(\ell) \cup (N(v_\ell) \cap \{v_j\}_{j=i+1}^{2k} \setminus N(v_1))$. We denote by $\widehat{B}_1(1)$ and $\widehat{B}_1(\ell)$ the latter sets, by $\widehat{\mathfrak{B}}_1(1)$ and $\widehat{\mathfrak{B}}_1(\ell)$ their respective sizes, and by $\widehat{p}_1'$ and $\widehat{p}_\ell'$ the probability that u obtains label 1 and label ℓ, respectively.

Recall that, by our conditioning on the event $\mathcal{E}$, $\mathfrak{B}_1(1) - \mathfrak{B}_1(\ell) \geq (\ell - 1)t_n$ and, by Observation 3.14, $\max\{\mathfrak{B}_1(1) - \widehat{\mathfrak{B}}_1(1), \mathfrak{B}_1(\ell) - \widehat{\mathfrak{B}}_1(\ell)\} \leq 10$. Thus, for $\ell \in [2, 2k]$ and an ℓ-good vertex u, Lemma 2.6 applies with $n = \widehat{\mathfrak{B}}_1(1)$, $n' = \widehat{\mathfrak{B}}_1(\ell)$, M equal to the maximum number of vertices in $A \cup B$ with label $i \in [2k] \setminus \{1, \ell\}$ and at distance at most 1 from u, and ρ equal to the number of times this maximum is attained. As a result,

$$\frac{\widehat{p}_1'}{\widehat{p}_\ell'} \geq \frac{\mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) > \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) = \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p))}{\mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p) > \mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) = \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p))}.$$

By transforming this inequality further, we obtain that

$$\frac{\widehat{p}_1'}{\widehat{p}_1' + \widehat{p}_\ell'} \geq \mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) > \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) = \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p))$$

$$\geq \mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) \geq \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p) + 1),$$

or equivalently

$$\widehat{p}_\ell' \leq \frac{1 - \mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) - \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p) \geq 1)}{\mathbb{P}(\mathrm{Bin}(\widehat{\mathfrak{B}}_1(1), p) - \mathrm{Bin}(\widehat{\mathfrak{B}}_1(\ell), p) \geq 1)} \widehat{p}_1'.$$

Now, note that Lemma 2.3 may be applied for $a_1 = \widehat{\mathfrak{B}}_1(1)$, $a_2 = \widehat{\mathfrak{B}}_1(\ell)$, $M = 1$ and p; indeed, under the event $\mathcal{E}$ and the statement of Observation 3.14, we have that $1 \ll a_2 \leq a_1$ and $\min\{a_2, a_1 - a_2\}p \geq t_n p = np^3/\sqrt{2} \gg 1$. As in the proof of Lemma 3.11, this yields

$$\forall \ell \in [2, 2k], \ \widehat{p}_\ell' \leq \left(\frac{1 - \Lambda}{\Lambda}\right)^{\ell-1} \widehat{p}_1' \ \text{and, in particular,} \ \widehat{p}_1' \geq \frac{2\Lambda - 1}{\Lambda}. \tag{21}$$

It immediately follows, using Lemma 3.10 and Chernoff's bound, that a.a.s. $\mathfrak{B}_2(1) \geq \frac{1}{2}\frac{2\Lambda-1}{\Lambda} \cdot \frac{4knp}{3} \geq (2\Lambda - 1)\frac{knp}{3}$.

Since Chernoff's inequality and a union bound imply that a.a.s. every basin has size at most $2np$ and, as a result, for every $\ell \in [2, 2k]$, there are at most $44np$ non-ℓ-good vertices in B. To finish the proof, it suffices to show that a.a.s., for every $\ell \in [2, 2k]$, the number of ℓ-good vertices ending up with label 1 after two rounds of ALAP is larger than the number of ℓ-good vertices ending up with label ℓ.

Step 2. We fix $\ell \in [2, 2k]$, denote by B' the set of ℓ-good vertices in B, and denote by $\mathfrak{B}_2'(1)$ and $\mathfrak{B}_2'(\ell)$ the number of ℓ-good vertices obtaining label 1 and ℓ after round 2, respectively. We concentrate on bounding from above the variance of $\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell) = \sum_{v \in B'}(\mathbb{1}_{v \in B_2(1)} - \mathbb{1}_{v \in B_2(\ell)})$. Note that,

$$\mathbb{E}[(\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell))^2] = \sum_{u \in B'} \mathbb{E}[(\mathbb{1}_{u \in B_2(1)} - \mathbb{1}_{u \in B_2(\ell)})^2]$$
$$+ \sum_{u,v \in B': u \neq v} \mathbb{E}[(\mathbb{1}_{u \in B_2(1)} - \mathbb{1}_{u \in B_2(\ell)})(\mathbb{1}_{v \in B_2(1)} - \mathbb{1}_{v \in B_2(\ell)})] \tag{22}$$

and

$$(\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)])^2 = \sum_{u,v \in B'} \mathbb{E}[\mathbb{1}_{u \in B_2(1)} - \mathbb{1}_{u \in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v \in B_2(1)} - \mathbb{1}_{v \in B_2(\ell)}]. \tag{23}$$

To bound the first summation in (22), observe that, since $B_2(1)$ and $B_2(\ell)$ are disjoint, $(\mathbb{1}_{u \in B_2(1)} - \mathbb{1}_{u \in B_2(\ell)})^2 = \mathbb{1}_{u \in B_2(1)} + \mathbb{1}_{u \in B_2(\ell)}$, so

$$\sum_{u \in B'} \mathbb{E}[(\mathbb{1}_{u \in B_2(1)} - \mathbb{1}_{u \in B_2(\ell)})^2] = \sum_{u \in B'} \mathbb{E}[\mathbb{1}_{u \in B_2(1)} + \mathbb{1}_{u \in B_2(\ell)}]$$
$$= \mathbb{E}[\mathfrak{B}_2'(1) + \mathfrak{B}_2'(\ell)] \leq 2\mathbb{E}[\mathfrak{B}_2'(1)]. \tag{24}$$

20

Next, recall that $\widehat{p}'_\ell \leq (\frac{1-\Lambda}{\Lambda})^{\ell-1}\widehat{p}'_1$ for all $\ell \in [2, 2k]$. Thus, by definition of $\mathfrak{B}'_2(\ell)$, for $\ell \in [2, 2k]$

$$\mathbb{E}[\mathfrak{B}'_2(1) - \mathfrak{B}'_2(\ell)] = \Omega((\widehat{p}'_1 - \widehat{p}'_\ell)knp) = \Omega\Big(\widehat{p}'_1\Big(1 - \Big(\tfrac{1-\Lambda}{\Lambda}\Big)^{\ell-1}\Big)knp\Big).$$

Since $\frac{1-\Lambda}{\Lambda} = 1 - \frac{2\Lambda-1}{\Lambda} < 1$, the right-hand side expression above is minimized at $\ell = 2$ and

$$\mathbb{E}[\mathfrak{B}'_2(1) - \mathfrak{B}'_2(\ell)] = \Omega\Big(\widehat{p}'_1\Big(\tfrac{2\Lambda-1}{\Lambda}\Big)knp\Big) = \Omega\Big(\widehat{p}'_1(2\Lambda - 1)knp\Big).$$

In particular, since $\mathbb{E}[\mathfrak{B}'_2(1)] = \Theta(\widehat{p}'_1 knp)$ and $\widehat{p}'_1 \geq \frac{2\Lambda-1}{\Lambda} \geq 2\Lambda - 1$, we get

$$p\big(\mathbb{E}[\mathfrak{B}'_2(1) - \mathfrak{B}'_2(\ell)]\big)^2 = \Omega\big((\widehat{p}'_1)^2(2\Lambda - 1)^2 p(knp)^2\big) = \Omega\big((2\Lambda - 1)^3 knp^2\,\mathbb{E}[\mathfrak{B}'_2(1)]\big) \gg \mathbb{E}[\mathfrak{B}'_2(1)],$$

where the last inequality comes from the fact that when $2\Lambda - 1 = \Omega(1)$, then $(2\Lambda - 1)^3 knp^2 = \Omega(\sqrt{n\log n}) \gg 1$, and when $2\Lambda - 1 = \Omega(\sqrt{np^4})$, then $(2\Lambda - 1)^3 knp^2 = \Omega((np^4)^{3/2}\sqrt{n\log n}) \gg (np^3)^2 \gg 1$. By (24), we conclude that

$$\sum_{u\in B'} \mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})^2] = o(p(\mathbb{E}[\mathfrak{B}'_2(1) - \mathfrak{B}'_2(\ell)])^2).$$

This, together with (22) and (23) yields

$$\begin{aligned}
\mathbb{V}[\mathfrak{B}'_2(1) - \mathfrak{B}'_2(\ell)] = {}& o\big(p(\mathbb{E}[\mathfrak{B}'_2(1) + \mathfrak{B}'_2(\ell)])^2\big) \\
&+ \sum_{u,v\in B':u\neq v} \Big(\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)})] \\
&\hspace{3cm} - \mathbb{E}[\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}]\Big).
\end{aligned} \tag{25}$$

To bound the summation above, note that by conditioning on whether the edge uv is in G_n, we get

$$\begin{aligned}
&\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)})] \\
&= q\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}) \mid uv \notin G_n] \\
&+ p\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}) \mid uv \in G_n].
\end{aligned} \tag{26}$$

The random variables $\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)}$ and $\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}$ are independent conditionally on the event $uv \notin G_n$, and also on the event $uv \in G_n$; indeed, in both cases, the two variables are measurable in terms of disjoint sets of edges of G_n. Hence,

$$\begin{aligned}
&\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}) \mid uv \notin G_n] \\
&= \mathbb{E}[\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)} \mid uv \notin G_n]\mathbb{E}[\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n],
\end{aligned} \tag{27}$$

and

$$\begin{aligned}
&\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}) \mid uv \in G_n] \\
&= \mathbb{E}[\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)} \mid uv \in G_n]\mathbb{E}[\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n].
\end{aligned} \tag{28}$$

Moreover, if $w \in \{u, v\}$, then

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)}] \\
&= p\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \in G_n] + q\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \notin G_n].
\end{aligned} \tag{29}$$

Now, consider the general term of summation in (25). Replacing the conditional expectations in (26) by their equivalent in (27)-(28), using (29) twice, and some arithmetic, the general term can be rewritten as

$$pq \cdot \prod_{w\in\{u,v\}} \big(\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \in G_n]\big). \tag{30}$$

21

We now claim that for $w \in \{u, v\}$, uniformly in u, v and ℓ,

$$|\mathbb{E}[\mathbb{1}_{w \in B_2(1)} - \mathbb{1}_{w \in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{w \in B_2(1)} - \mathbb{1}_{w \in B_2(\ell)} \mid uv \in G_n]|$$
$$= o(\mathbb{E}[\mathbb{1}_{w \in B_2(1)} - \mathbb{1}_{w \in B_2(\ell)}]), \tag{31}$$

so the general term in the summation in (25) equals

$$o(p\mathbb{E}[\mathbb{1}_{u \in B_2(1)} - \mathbb{1}_{u \in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v \in B_2(1)} - \mathbb{1}_{v \in B_2(\ell)}]),$$

and thus $\mathbb{V}(\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)) = o(p(\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)])^2)$.

Now, we show (31). Suppose $w = u$, as the other case is similar. Let $X_1 = |N[u] \cap \widehat{B}_1(1)|$ and $X_\ell = |N[u] \cap \widehat{B}_1(\ell)|$. Observe that, thanks to Observation 3.14, in each of the unconditional probability space, the space conditioned on $uv \in G_n$ and the space conditioned on $uv \notin G_n$, there are two independent random variables $\widehat{X}_1 \in \text{Bin}(\mathfrak{B}_1(1), p)$ and $\widehat{X}_\ell \in \text{Bin}(\mathfrak{B}_1(\ell), p)$ and a constant m such that $|X_1 - \widehat{X}_1| \leq m$ and $|X_\ell - \widehat{X}_\ell| \leq m$.

In any case, since opening or closing the edge uv leads to a difference of one edge, and by Lemma 2.3 we have that

$$\mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell = m) = o(\mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell \geq m) - \mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell < m)) \tag{32}$$

for any fixed constant m (showing that any constant number of edges does not change the probability of receiving a concrete label significantly), (31) is satisfied for $w = u$.

Since $\mathbb{V}(\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)) = o(p(\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)])^2)$, by Chebyshev's inequality

$$\mathbb{P}\Big(\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell) \leq \tfrac{1}{2}\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)]\Big)$$
$$\leq \mathbb{P}\Big(|\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell) - \mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)]| \geq \tfrac{1}{2}\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)]\Big)$$
$$\leq \frac{4\mathbb{V}(\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell))}{\big(\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)]\big)^2} = o(p).$$

A union bound leads to

$$\mathbb{P}\Big(\exists \ell \in [2, 2k], \mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell) \leq \tfrac{1}{2}\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)]\Big) = o(kp) = o(1),$$

which proves the lemma since $\mathbb{E}[\mathfrak{B}_2'(1) - \mathfrak{B}_2'(\ell)] = \Omega\big(\big(1 - \big(\frac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)\mathbb{E}\mathfrak{B}_2'(1)\big)$. $\square$

Remark 3.16. For $np = cn^{2/3}$, the following strengthening of Lemma 3.15 holds: Given $\varepsilon \in (0, 1)$ and suitably large $L = L(\varepsilon, c)$ (recall Remark 3.5), conditionally on $\mathcal{E}_L$, there is a constant $c_1 > 0$ such that a.a.s., for every $\ell \in [2k]$ different from the index of the largest basin, the difference between $\mathfrak{B}_2(\ell)$ and the size of the largest basin is at most εnp.

Most of the necessary modifications are the same as in Remark 3.12. We emphasize that choosing N in Remark 2.5 sufficiently large allows us to replace (32) by

$$\mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell = m) \leq \varepsilon|\mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell \geq m) - \mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell < m)|$$

for an appropriately small ε, which in turn allows us to replace (31) with

$$|\mathbb{E}[\mathbb{1}_{w \in B_2(1)} - \mathbb{1}_{w \in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{w \in B_2(1)} - \mathbb{1}_{w \in B_2(\ell)} \mid uv \in G_n]|$$
$$\leq \varepsilon\mathbb{E}[\mathbb{1}_{w \in B_2(1)} - \mathbb{1}_{w \in B_2(\ell)}].$$

As a consequence, the last two displays in the proof of Lemma 3.15 are equal to $O(p)$ and $O(kp)$ instead of $o(p)$ and $o(kp)$, which is sufficient for our purposes.

Most importantly, the current statement and proof of Lemma 3.15 work with "the worst-case scenario" where all non-ℓ-good vertices are simply assumed to obtain label ℓ after two rounds of ALAP. To show the

analogous statement when $np = cn^{2/3}$, one can use the following fact (see also Section 3.3): in the said regime, for every sufficiently large $L = L(\varepsilon, c)$, with probability at least $1 - \varepsilon$, the largest basin is among the first L basins and, moreover, it (as well as each of the first L basins) has size of order $np + \Theta(\sqrt{np})$. Thus, for every $\ell \in [2k]$ and $u \in B$, any constant number of already revealed edges incident to u and leading to a vertex with label ℓ is negligible compared to the standard deviation of the random variables $|N(u) \cap B_1(i)|$ for $i \in [L]$. As a result, by choosing L sufficiently large, for every $\ell \in [L]$, the probability that a non-ℓ-good vertex obtains label ℓ is at most $1/L + o(1) \leq \varepsilon$. A second moment argument similar to the one from the proof of Lemma 3.15 shows the statement of the remark.

At this stage, the analysis of the label distribution after the second round is complete. The next lemma is a technical tool in our analysis of the third round.

Lemma 3.17. *Suppose that $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, after the second round of ALAP, a.a.s. $n - o(n)$ of the vertices in Level 3 have the following property: for every $\ell \in [2, 2k]$, their number of neighbors in the second level with label ℓ is at most by $90np^2$ larger than their number of neighbors in the second level with label 1.*

Proof. Condition on the event $\mathfrak{B}_2(1) \leq (\log n)^{1/3} \mathbb{E}\mathfrak{B}_2(1)$ (which holds a.a.s. as well by Markov's inequality), the event $\mathcal{F}$ from Lemma 3.10, the variables $(\mathfrak{B}_2(\ell))_{\ell=1}^{2k}$ and the a.a.s. statement of Lemma 3.15. For every $\ell \in [2, 2k]$, denote by $B_2'(\ell)$ an arbitrary subset of $B_2(\ell)$ of size $\lceil 44np \rceil$, if $\mathfrak{B}_2(\ell) \geq 44np$, and $B_2'(\ell) = B_2(\ell)$ otherwise. Then, Chernoff's inequality and a union bound show that a.a.s., for every $\ell \in [2, 2k]$ and every vertex $u \in C$, there are at most $88np^2$ neighbors of u in $B_2'(\ell)$.

Now, fix a vertex $u \in C$. Also, denote by Y_1 the number of neighbors of u in $B_2(1)$ and, for every $\ell \in [2, 2k]$, denote by Y_ℓ the number of neighbors of u in $B_2(\ell) \setminus B_2'(\ell)$. In particular, $Y_1 \in \mathrm{Bin}(\mathfrak{B}_2(1), p)$ and $Y_\ell \in \mathrm{Bin}(\max\{\mathfrak{B}_2(\ell) - 44np, 0\}, p)$ for $\ell \in [2, 2k]$. Then, Chernoff's bound implies

$$\mathbb{P}(Y_1 + 2np^2 \leq Y_\ell) \leq \mathbb{P}\big(Y_1 \leq \tfrac{1}{2}\mathbb{E}[Y_1 + Y_\ell] - np^2\big) + \mathbb{P}\big(Y_\ell \geq \tfrac{1}{2}\mathbb{E}[Y_1 + Y_\ell] + np^2\big)$$
$$\leq \mathbb{P}\big(Y_1 - \mathbb{E}Y_1 \leq -np^2\big) + \mathbb{P}\big(Y_\ell - \mathbb{E}Y_\ell \geq np^2\big)$$
$$\leq 2\exp\Big(-\tfrac{n^2p^4}{2\mathbb{E}[Y_1] + 2np^2}\Big)$$
$$\leq 2\exp\Big(-\tfrac{n^2p^4}{8np^2k}\Big),$$

where for the last inequality follows from the fact that the event $\mathcal{F}$ holds, so $p\mathfrak{B}_2(1) \leq 3np^2k = O(\sqrt{n\log n})$. However, $n^2p^4 = \Omega(n^{2/3})$, so the latter bound allows us to conclude by a union bound over $\ell \in [2, 2k]$. $\square$

Remark 3.18. When $np = cn^{2/3}$ for some constant $c > 0$, by replacing Lemma 3.15 with Remark 3.16, we may deduce that, conditionally on the event $\mathcal{E}_L \cap \mathcal{G}_{L,\varepsilon}$ (see Lemmas 3.5 and 3.7), a.a.s. almost all vertices in Level 3 satisfy the following property: for every $\ell \in [2, 2k]$, their number of neighbors in the second level with label ℓ is at most by $3\varepsilon np^2$ larger than their number of neighbors in the second level with label 1.

Next, we observe that a.a.s. all vertices in Level 3 have far more neighbors in $C_2(1)$ than in $C_2(\ell)$ for all $\ell \in [2, 2k]$. Fix

$$\mathfrak{N} := \tfrac{1}{10}(2\Lambda - 1)\mathfrak{C}_2(1).$$

Lemma 3.19. *Suppose $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, a.a.s. $\mathfrak{N} \geq \tfrac{1}{20}(2\Lambda - 1)^2 n$ and every vertex u in Level 3 satisfies $|N(u) \cap C_2(1)| \geq |N(u) \cap C_2(\ell)| + \mathfrak{N}p$ for all $\ell \in [2, 2k]$.*

Proof. We condition on the event that in ALAP,

$$\Big\{\mathfrak{C}_2(1) \geq \tfrac{1}{2}(2\Lambda - 1)n \text{ and for every } \ell \in [2, 2k], \mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \geq \tfrac{1}{2}\big(1 - \big(\tfrac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)\mathfrak{C}_2(1)\Big\}, \qquad (33)$$

which happens a.a.s. by Lemma 3.11 and therefore implies the first part of the lemma.

Fix a vertex $u \in C$. Then, for every $\ell \in [2, 2k]$, (33) implies that $\mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \geq 3\mathfrak{N}$. Now, fix $a_1 = \mathfrak{C}_2(1) - 1$ and $a_2 = \frac{1}{2}(1 + (\frac{1-\Lambda}{\Lambda})^{\ell-1})\mathfrak{C}_2(1)$, and define $X_1 = |N(u) \cap C_2(1)|$ and $X_2 = |N(u) \cap C_2(\ell)|$. Notice that X_1 dominates $\mathrm{Bin}(a_1, p)$ while X_2 is dominated by $\mathrm{Bin}(a_2, p)$. Thus, using Chernoff's bound, we get that

$$\mathbb{P}(X_1 - X_2 \leq 3\mathfrak{N}p) \leq \mathbb{P}(X_1 - a_1 p \leq \mathfrak{N}p) + \mathbb{P}(X_2 - a_2 p \geq \mathfrak{N}p)$$

$$\leq \exp\left(-\frac{\mathfrak{N}^2 p^2}{2(a_1 p + \mathfrak{N}p)}\right) + \exp\left(-\frac{\mathfrak{N}^2 p^2}{2(a_2 p + \mathfrak{N}p)}\right).$$

Since each of the right-hand side terms above are bounded by $\exp(-\Omega((2\Lambda - 1)^2 \mathfrak{C}_2(1)p)) = \exp(-\Omega((2\Lambda - 1)^3 np))= o(n^{-2})$, a union bound over all $\ell \in [2, 2k]$ implies that $X_1 \leq X_2 + \mathfrak{N}p$ with probability at most $o(kn^{-2}) = o(n^{-1})$. Finally, a union bound over all vertices in C finishes the proof. $\qquad\square$

Remark 3.20. For every $\varepsilon, c > 0$, there is $\delta = \delta(\varepsilon, c) > 0$ such that the following holds: for $np = cn^{2/3}$, $\mathfrak{N}$ as before the statement of Lemma 3.19 and label $\widehat{\ell}_1$ defined as in (19), the event "$\mathfrak{N} \geq \delta(2\Lambda - 1)^2 n$ and every vertex u in Level 3 satisfies $|N(u) \cap C_2(\widehat{\ell}_1)| \geq |N(u) \cap C_2(\ell)| + \mathfrak{N}p$ for all $\ell \in [2, 2k]$" holds with probability $1 - \varepsilon$. The proof requires only minor modifications.

Lemma 3.21. *Suppose that $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, after round 5 of LPA, a.a.s. all vertices have label 1.*

Proof. Recall $K = \lceil 2(\log n)/p \rceil$. On the one hand, Chernoff's bound and a union bound imply that all vertices $u \in V$ satisfy $|N(u) \cap V([K])| = O(Kp)$. On the other hand, the labels attributed to the vertices by LPA and by ALAP after the second round coincide for all vertices outside $V([K])$. Thus, combining Lemmas 3.17, 3.19, the inequality $\mathfrak{N}p \gg np^2$ and the fact that a.a.s. $\mathfrak{N}p = \Omega((2\Lambda - 1)^2 np) \gg Kp$ (to ensure that a.a.s. $n - o(n)$ vertices get label 1 in the third round not only in ALAP but also in LPA) and Lemma 3.2 yields the result. $\qquad\square$

This proves Theorem 1.1 in the regime $n^{2/3} \ll np \ll n$. For the regime $np = \Theta(n^{2/3})$, replacing Lemma 3.17 with Remark 3.18 and Lemma 3.19 with Remark 3.20 in the proof of Lemma 3.21 shows that conditionally on the event $\mathcal{E}_L \cap \mathcal{G}_{L,\varepsilon}$ and the event from Remark 3.20, a.a.s. only one label survives. However, since these events hold jointly with probability at least $1 - 3\varepsilon$, and ε could be chosen arbitrarily small, thus implying Remark 1.2.

3.3 The regime $n^{5/8+\varepsilon} \leq np \ll n^{2/3}$

Now, we fix $\varepsilon \in (0, 1/24)$ and concentrate on the regime $n^{5/8+\varepsilon} \leq np \ll n^{2/3}$ in ALAP. Note that all results in this section will be proved in this regime only, so we omit it from the statements of the lemmas.

Recall that $\mathfrak{A} = |A| = 2k$ with k defined in (3). Our first task is to analyze the maximum of $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$, which we denote by $\mathfrak{B}^{(1)}$. For every $\ell \in [2k]$, define

$$z_\ell = n - (\ell - 1)np + \tfrac{1}{2}(\ell - 1)(\ell - 2)np^2.$$

Lemma 3.22. *One may couple the sequence $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$ with a sequence of independent random variables $(Z_\ell)_{\ell=1}^{2k}$ such that a.a.s. for all $\ell \in [2k]$ we have:*

- $|Z_\ell - \mathfrak{B}_1(\ell)| \leq (np)^{2/5}$,

- $Z_\ell \in \mathrm{Bin}(z_\ell, p)$.

Proof. For every $\ell \in [2k]$, set $x_\ell = \ell^2 np^3$. First, we show that for every $\ell \in [2k]$,

$$|\mathbb{E}\mathfrak{B}_1(\ell) - np + (\ell - 1)np^2| \leq x_\ell. \tag{34}$$

Recall that $\mathfrak{B}_1(\ell) \in \mathrm{Bin}(n - 2k, q^{\ell-1}p)$, so $\mathbb{E}\mathfrak{B}_1(\ell) = (n - 2k)q^{\ell-1}p$. Since $q^{\ell-1} \leq 1 - (\ell - 1)p + \ell^2 p^2$, we get

$$\mathbb{E}\mathfrak{B}_1(\ell) - np + (\ell - 1)np^2 \leq nq^{\ell-1}p - np + (\ell - 1)np^2 \leq \ell^2 np^3 = x_\ell.$$

Moreover, since $q^{\ell-1} \geq 1 - (\ell - 1)p$, we have

$$\mathbb{E}\mathfrak{B}_1(\ell) - np + (\ell - 1)np^2 = np(q^{\ell-1} - 1 + (\ell - 1)p) - 2kq^{\ell-1}p \geq -2kp.$$

To conclude the proof of (34), observe that $2kp \leq 32\sqrt{\frac{\log n}{np^2}} \leq 32n^{1-4\epsilon}p^3\sqrt{\log n} \leq x_\ell$ for n large enough.

Now, define the event

$$\mathcal{A}_\ell = \left\{ \left| \sum_{j=1}^{\ell-1} \left(\mathfrak{B}_1(j) - np + (j-1)np^2 \right) \right| \leq \sum_{j=1}^{\ell-1} x_j + 2\left(2\log n \sum_{j=1}^{\ell-1} \mathbb{E}\mathfrak{B}_1(j) \right)^{1/2} \right\}.$$

Then, (34) and the triangle inequality imply that

$$\overline{\mathcal{A}_\ell} \subseteq \left\{ \left| \sum_{j=1}^{\ell-1} (\mathfrak{B}_1(j) - \mathbb{E}\mathfrak{B}_1(j)) \right| \geq 2\left(2\log n \sum_{j=1}^{\ell-1} \mathbb{E}\mathfrak{B}_1(j) \right)^{1/2} \right\}.$$

Since $\sum_{j=1}^{\ell-1} \mathfrak{B}_1(j) \in \mathrm{Bin}(n - 2k, 1 - q^{\ell-1})$, by Chernoff's bound

$$\mathbb{P}(\overline{\mathcal{A}_\ell}) \leq \mathbb{P}\left(\left| \sum_{j=1}^{\ell-1} (\mathfrak{B}_1(j) - \mathbb{E}\mathfrak{B}_1(j)) \right| \geq 2\left(2\log n \sum_{j=1}^{\ell-1} \mathbb{E}\mathfrak{B}_1(j) \right)^{1/2} \right)$$

$$\leq 2\exp\left(-(2 + o(1))\log n \right) \leq \frac{1}{n}.$$

Note that $\mathfrak{B}_1(\ell)$ equals the number of edges from v_ℓ to $V \setminus (A \cup B_1([\ell - 1]))$. We define the random variable Z_ℓ as the number of edges from v_ℓ to $V \setminus (A \cup U_\ell)$ where U_ℓ is a set of vertices of size $n - z_\ell - 2k$ defined as follows:

- if $\mathfrak{B}_1([\ell - 1]) \geq n - z_\ell - 2k$, then U_ℓ consists of $n - z_\ell - 2k$ arbitrary vertices in $B_1([\ell - 1])$,

- otherwise, construct U_ℓ by adding a set of arbitrary $n - z_\ell - 2k - \mathfrak{B}_1([\ell - 1])$ vertices from $V \setminus (A \cup B_1([\ell - 1]))$ to $B_1([\ell - 1])$.

Note that $(Z_\ell)_{\ell=1}^{2k}$ are independent random variables such that $Z_\ell \in \mathrm{Bin}(z_\ell, p)$ for all $\ell \in [2k]$. Moreover, using that by the triangle inequality

$$|n - z_\ell - 2k - \mathfrak{B}_1([\ell - 1])| \leq 2k + \left| \sum_{j=1}^{\ell-1} (np - (j-1)np^2 - \mathfrak{B}_1(j)) \right|,$$

one can deduce that conditionally on $\mathcal{A}_\ell$, $|\mathfrak{B}_1(\ell) - Z_\ell|$ is stochastically dominated by a random variable

$$Y_\ell \sim \mathrm{Bin}\left(2k + \sum_{j=1}^{\ell-1} x_j + 2\left(2\log n \sum_{j=1}^{\ell-1} \mathbb{E}\mathfrak{B}_1(j) \right)^{1/2}, p \right).$$

Using that $\mathbb{E}Y_\ell \leq p\left(2k + \sum_{j=1}^{\ell-1} x_j + 10\sqrt{\ell np \log n} \right) = O(kp + k^3 np^4 + \sqrt{knp^3 \log n}) = o((np)^{2/5})$, Chernoff's bound implies that

$$\mathbb{P}\left(Y_\ell \geq (np)^{2/5} \right) \leq \frac{1}{n}.$$

Hence, for all $\ell \in [2k]$, conditionally on the event $\mathcal{A}_\ell$, we have $|\mathfrak{B}_1(\ell) - Z_\ell| \geq (np)^{2/5}$ with probability $O(n^{-1})$, so the desired result follows by a union bound over all $\ell \in [2k]$. $\square$

Remark 3.23. Note that the previous proof can be extended for $np \in [n^{1/2+\varepsilon}, n^{2/3}]$ with any $\varepsilon \in (0, 1/6]$ by further expansion of $\mathfrak{B}_1(\ell)$. However, since this lemma is not the true bottleneck in our argument, we do not pursue this here.

Our next aim is to give a lower bound on the gap between the maximum and the second maximum of $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$. We do this by estimating this gap for the sequence $(Z_\ell)_{\ell=1}^{2k}$ instead, and transfer our conclusion to $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$ using Lemma 3.22. Define $Z^{(1)} = \max\{Z_\ell \colon \ell \in [2k]\}$ and $Z^{(2)} = \max\{Z_\ell \colon \ell \in [2k], Z_\ell < Z^{(1)}\}$. To begin with, we estimate $Z^{(1)}$.

Lemma 3.24.
$$\frac{Z^{(1)} - np}{\sqrt{np \log(1/(np^3))}} \xrightarrow{\mathbb{P}} 1 \text{ as } n \to \infty.$$

Proof. For any $\zeta > 0$, define $T_n = T_n(\zeta) = np + \sqrt{\zeta np \log(1/(np^3))}$. Then, by independence of $(Z_\ell)_{\ell=1}^{2k}$, we have $\mathbb{P}\big(\bigcap_{\ell=1}^{2k}\{Z_\ell \leq T_n\}\big) = \prod_{\ell=1}^{2k} \mathbb{P}(Z_\ell \leq T_n)$. We now provide upper and lower bounds for $\mathbb{P}(Z_\ell \leq T_n)$. On the one hand, the fact that $Z_\ell \in \text{Bin}(z_\ell, p)$ implies that $\mathbb{E}Z_\ell = z_\ell p = (1 + o(1))np \gg T_n + 1 - z_\ell p$ (the $o(1)$ terms here and below may be chosen independently of ℓ). Combining this with Chernoff's bound yields

$$\mathbb{P}(Z_\ell \geq T_n + 1) = \mathbb{P}(Z_\ell - \mathbb{E}Z_\ell \geq T_n + 1 - z_\ell p) \leq \exp\left(-(1 + o(1))\frac{(T_n + 1 - z_\ell p)^2}{2np}\right).$$

On the other hand, recalling that $q = 1 - p$, Slud's inequality (Lemma 2.2) yields

$$\mathbb{P}(Z_\ell \geq T_n + 1) \geq 1 - \Phi\left(\frac{T_n + 1 - z_\ell p}{\sqrt{z_\ell pq}}\right),$$

where we recall that Φ is the cumulative density function of a standard normal random variable. Using that, as $x \to \infty$,

$$\int_x^\infty e^{-u^2/2}du \geq \int_x^{x+1/x^2} e^{-u^2/2}du \geq \frac{1}{x^2}e^{-x^2/2+o_x(1)} = e^{-(1+o_x(1))x^2/2},$$

we obtain that

$$\mathbb{P}(Z_\ell \geq T_n + 1) \geq \exp\left(-(1 + o(1))\frac{(T_n + 1 - z_\ell p)^2}{2np}\right).$$

By our choice of T_n, expanding the square and canceling the factor np, we have

$$\frac{1}{2np}(T_n + 1 - z_\ell p)^2 = \frac{1}{2np}\left(\sqrt{\zeta np \log(1/(np^3))} + (1 + o(1))(\ell - 1)np^2\right)^2 \tag{35}$$

$$= \frac{1}{2}\zeta \log(1/(np^3))$$

$$+ (1 + o(1))\left((\ell - 1)\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}(\ell - 1)^2 np^3\right).$$

Hence, since $1 - x = e^{-(1+o(1))x}$ as $x \to 0$ and $\mathbb{P}(Z_\ell \leq T_n) = 1 - \mathbb{P}(Z_\ell \geq T_n + 1) = 1 - o(1)$,

$$\prod_{\ell=1}^{2k} \mathbb{P}(Z_\ell \leq T_n) = \prod_{\ell=1}^{2k}(1 - \mathbb{P}(Z_\ell \geq T_n + 1)) = \exp\left(-(1 + o(1))\sum_{\ell=1}^{2k} \mathbb{P}(Z_\ell \geq T_n + 1)\right)$$

$$= \exp\left(-(np^3)^{\zeta/2+o(1)}\sum_{\ell=1}^{2k}\exp\left(-(1 + o(1))\left((\ell - 1)\sqrt{\zeta np^3 \log(1/(np^3))}\right.\right.\right. \tag{36}$$

$$\left.\left.\left. + \frac{1}{2}(\ell - 1)^2 np^3\right)\right)\right).$$

26

Let us estimate the last sum. For a lower bound, note that summing up to $\ell_* = (np^3 \log(1/(np^3)))^{-1/2}$ shows that the sum is bounded from below by

$$\left(\frac{1}{\sqrt{np^3 \log(1/(np^3))}}\right)^{1+o(1)} = \frac{1}{(np^3)^{1/2+o(1)}}. \tag{37}$$

On the other hand, for an upper bound, note that

$$
\begin{aligned}
&\sum_{\ell=1}^{2k} \exp\left(-(1+o(1))\left((\ell-1)\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}(\ell-1)^2 np^3\right)\right) \\
&\leq \sum_{\ell=1}^{2k} \exp\left(-(1+o(1))(\ell-1)\sqrt{\zeta np^3 \log(1/(np^3))}\right) \\
&\leq \frac{1}{1 - \exp\left(-(1+o(1))\sqrt{\zeta np^3 \log(1/(np^3))}\right)} \\
&\leq \frac{1}{(np^3)^{1/2+o(1)}}.
\end{aligned}
\tag{38}
$$

We conclude that if $\zeta > 1$, then $\mathbb{P}\left(\bigcap_{\ell=1}^{2k}\{Z_\ell \leq T_n\}\right)$ converges to 1 as $n \to \infty$, and if $\zeta < 1$, it converges to 0, and the proof of the lemma is finished. $\qquad\square$

Now, define $k_* = \frac{1}{2}\sqrt{\frac{\log(1/(np^3))}{2np^3}} \ll 2k$, $Z_*^{(1)} = \max_{1 \leq \ell \leq k_*} Z_\ell$ and $Z_*^{(2)} = \max\{Z_\ell : \ell \in [k_*], Z_\ell < Z_*^{(1)}\}$.

Remark 3.25. Note that (37) and (38) may be adapted to analyze $\max_{\ell \in [\lfloor k_*/2\rfloor, 2k]} Z_\ell$ and $\max_{\ell \in [k_*, 2k]} Z_\ell$. Let us have a closer look at the latter case, the former being analogous. To begin with, instead of starting the sum in (36) from $\ell = 1$, we start it from $\ell = k_*$. Now, as in (37), summing over the first ℓ_* terms starting from k_* form a decreasing sequence. Thus, on the one hand,

$$
\begin{aligned}
&\sum_{\ell=k_*+1}^{2k} \exp\left(-(1+o(1))\left((\ell-1)\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}(\ell-1)^2 np^3\right)\right) \\
&\geq \sum_{\ell=k_*+1}^{k_*+\ell_*} \exp\left(-(1+o(1))\left((\ell-1)\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}(\ell-1)^2 np^3\right)\right) \\
&\geq \ell_* \exp\left(-(1+o(1))(k_*+\ell_*)\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}(k_*+\ell_*)^2 np^3\right)\right) \\
&= \left(\frac{1}{np^3}\right)^{(1-\sqrt{\zeta/2}-1/8+o(1))/2}
\end{aligned}
$$

and, on the other hand,

$$
\begin{aligned}
&\sum_{\ell=k_*+1}^{2k} \exp\left(-(1+o(1))\left((\ell-1)\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}(\ell-1)^2 np^3\right)\right) \\
&\leq \sum_{\ell=k_*+1}^{2k} \exp\left(-(1+o(1))(\ell-1)\left(\sqrt{\zeta np^3 \log(1/(np^3))} + \frac{1}{2}k_* np^3\right)\right) \\
&\leq \frac{\exp\left(-(1+o(1))k_*\left(\sqrt{\zeta np^3 \log(1/(np^3))} + k_* np^3/2\right)\right)}{1 - \exp\left(-(1+o(1))\left(\sqrt{\zeta np^3 \log(1/(np^3))} + k_* np^3/2\right)\right)}
\end{aligned}
$$

27

$$= \left(\frac{1}{np^3}\right)^{(1-\sqrt{\zeta/2}-1/8+o(1))/2}.$$

As a consequence, $\prod_{\ell=k_*}^{2k} \mathbb{P}(Z_\ell \leq T_n) = \exp(-(np^3)^{(\zeta-1+\sqrt{\zeta/2}+1/8)/2+o(1)})$. As $n \to \infty$, solving the equation $\zeta + \sqrt{\zeta/2} - 7/8 = 0$ leads to

$$\frac{\max_{\ell \in [k_*, 2k]} Z_\ell - np}{\sqrt{np \log(1/(np^3))}} \xrightarrow{\mathbb{P}} \frac{9}{8} - \frac{1}{\sqrt{2}} \approx 0.418,$$

and similarly, replacing k_* by $k_*/2$,

$$\frac{\max_{\ell \in [\lfloor k_*/2 \rfloor, 2k]} Z_\ell - np}{\sqrt{np \log(1/(np^3))}} \xrightarrow{\mathbb{P}} \frac{33}{32} - \frac{1}{\sqrt{8}} \approx 0.678.$$

Corollary 3.26. *A.a.s. $Z^{(1)} = Z^{(1)}_*$ and $Z^{(2)} = Z^{(2)}_*$.*

Proof. We show that a.a.s. $Z^{(2)}_* > \max_{\ell > k_*} Z_j$, which implies the statement of the corollary. Firstly, Lemma 3.24 together with the second conclusion Remark 3.25 implies that a.a.s. $Z^{(1)} > \max_{\ell \in [k_*/2, 2k]} Z_j$. Similarly, Lemma 3.24 and the first conclusion of Remark 3.25 imply that a.a.s. $\max_{\ell \in [\lfloor k_*/2 \rfloor, 2k]} Z_\ell > \max_{\ell \in [k_*, 2k]} Z_\ell$, and thus finishes the proof of the corollary. $\square$

Next, we estimate $Z^{(1)}_* - Z^{(2)}_*$. The following lemma is a general result for binomial random variables.

Lemma 3.27. *Fix $n \in \mathbb{N}$ and $t \in [n]$. Then, the function $s \in \mathbb{N} \cap [t, n-1] \mapsto \mathbb{P}(Z^{(2)}_* \leq Z^{(1)}_* - t \mid Z^{(1)}_* = s)$ is increasing in s.*

Proof. Firstly, define $\ell_1 = \min\{\ell \in [k_*], Z^{(1)}_* = Z_\ell\}$. Note that by Corollary 3.26, ℓ_1 a.a.s. coincides with $\min\{\ell \in [2k], Z^{(1)} = Z_\ell\}$. We show that for every $\ell \in [k_*]$, the function $s \in \mathbb{N} \mapsto \mathbb{P}(Z^{(2)}_* \leq Z^{(1)}_* - t \mid Z^{(1)}_* = s, \ell_1 = \ell)$ is increasing, which implies the lemma. Let us condition on the event $\ell_1 = \ell$. We have

$$\mathbb{P}(Z^{(2)}_* \leq Z^{(1)}_* - t \mid Z^{(1)}_* = s, \ell_1 = \ell) = \prod_{j \in [\ell-1]} \mathbb{P}(Z_j \leq s - t \mid Z_j \leq s - 1)$$

$$\cdot \prod_{j=\ell+1}^{k_*} \mathbb{P}(Z_j \leq s - t \mid Z_j \leq s),$$

where we used the independence of the random variables $(Z_j)_{j=1}^{k_*}$. Now, let us fix $j \in [\ell+1, k_*]$ (the case when $j \in [\ell - 1]$ is treated analogously). On the one hand, given positive integers t and $s \geq t$,

$$\mathbb{P}(Z_j \leq s - t \mid Z_j \leq s) = \frac{\mathbb{P}(Z_j \leq s - t)}{\mathbb{P}(Z_j \leq s)}$$

and

$$\mathbb{P}(Z_j \leq s + 1 - t \mid Z_j \leq s + 1) = \frac{\mathbb{P}(Z_j \leq s + 1 - t)}{\mathbb{P}(Z_j \leq s + 1)}.$$

On the other hand, it is well-known that binomial random variables have log-concave probability mass functions. Hence, the cumulative distribution function of Z_j, say F, is also log-concave (see for instance Proposition 1-1 (ii) in [27]). Then, observing that for $\lambda = (t+1)^{-1}$ we have that $s = \lambda(s-t) + (1-\lambda)(s+1)$ and $s + 1 - t = (1-\lambda)(s-t) + \lambda(s+1)$, it follows that

$$\log F(s-t) + \log F(s+1) \leq \log F(s+1-t) + \log F(s),$$

which is equivalent to

$$\frac{\mathbb{P}(Z_j \leq s - t)}{\mathbb{P}(Z_j \leq s)} = \frac{F(s-t)}{F(s)} \leq \frac{F(s+1-t)}{F(s+1)} = \frac{\mathbb{P}(Z_j \leq s + 1 - t)}{\mathbb{P}(Z_j \leq s + 1)},$$

thereby concluding the proof of the lemma. $\square$

By Lemma 3.24 one can define a sequence of positive real numbers $(\varepsilon_n)_{n\geq 1}$ converging to 0 and such that, on the one hand, $\varepsilon_n \geq (\log(np))^{-1/2}$ for all sufficiently large n, and moreover

$$M_n = np + \sqrt{(1 - \tfrac{\varepsilon_n}{2})np\log(1/(np^3))}$$

is a.a.s. smaller than $Z_*^{(1)}$. Also, define $\gamma_n = (np)^{1/2-\varepsilon_n}$.

Lemma 3.28.
$$\mathbb{P}\big(Z_*^{(2)} \leq Z_*^{(1)} - 2\gamma_n\big) = 1 - o(1).$$

Proof. As in the proof of Lemma 3.27, let $\ell_1 = \min\{\ell \in [k_*], Z_*^{(1)} = Z_\ell\}$ and note that, by Corollary 3.26, a.a.s. $\ell_1 = \min\{\ell \in [2k], Z^{(1)} = Z_\ell\}$. Using Lemma 3.27 and the definition of the sequence $(\varepsilon_n)_{n\geq 1}$ (for the second inequality) and Corollary 3.26 (for the equality) below, we get that

$$
\begin{aligned}
\mathbb{P}\big(Z_*^{(2)} \leq Z_*^{(1)} - 2\gamma_n\big) &\geq \mathbb{P}(Z_*^{(1)} \geq M_n)\mathbb{P}(Z_*^{(2)} \leq Z_*^{(1)} - 2\gamma_n \mid Z_*^{(1)} \geq M_n) \\
&\geq (1 - \mathbb{P}(Z_*^{(1)} \leq M_n - 1))\mathbb{P}(Z_*^{(2)} \leq Z_*^{(1)} - 2\gamma_n \mid Z_*^{(1)} = M_n) \\
&= (1 - o(1))\sum_{\ell=1}^{k_*} \mathbb{P}(\ell_1 = \ell \mid Z_*^{(1)} = M_n) \prod_{j\in[k_*]\setminus\{\ell\}} \mathbb{P}(Z_j \leq M_n - 2\gamma_n \mid \ell_1 = \ell, Z_\ell = M_n) \\
&\geq (1 - o(1))\prod_{\ell=1}^{k_*} \mathbb{P}(Z_\ell \leq M_n - 2\gamma_n \mid Z_\ell \leq M_n), \quad\quad (39)
\end{aligned}
$$

where for the last inequality we used that by independence of $(Z_\ell)_{\ell=1}^{k_*}$, for every $\ell \in [k_*]$, the product in the third line rewrites as

$$\prod_{j=1}^{\ell-1} \mathbb{P}(Z_j \leq M_n - 2\gamma_n \mid Z_j \leq M_n - 1) \prod_{j=\ell+1}^{k_*} \mathbb{P}(Z_j \leq M_n - 2\gamma_n \mid Z_j \leq M_n).$$

In particular, it is at least

$$\prod_{j\in[k_*]\setminus\{\ell\}} \mathbb{P}(Z_j \leq M_n - 2\gamma_n \mid Z_j \leq M_n),$$

which is uniformly bounded from below by (39). Moreover, using that $\mathbb{P}(Z_\ell \leq M_n) = 1 - o(1)$ for every $\ell \in [k_*]$, the product in (39) rewrites as

$$
\begin{aligned}
\prod_{\ell=1}^{k_*} \big(1 - \mathbb{P}\big(Z_\ell \in [M_n - 2\gamma_n + 1, M_n] \mid Z_\ell \leq M_n\big)\big) \\
= \prod_{\ell=1}^{k_*} (1 - (1 + o(1))\mathbb{P}(Z_\ell \in [M_n - 2\gamma_n + 1, M_n])). \quad\quad (40)
\end{aligned}
$$

Let us show that for every j, the terms $(\mathbb{P}(Z_j = M_n - \ell))_{\ell=0}^{2\gamma_n}$ are all of the same order. In fact, we only show that the terms $\mathbb{P}(Z_j = M_n)$ and $\mathbb{P}(Z_j = M_n - 2\gamma_n)$ are of the same order, the computation for the remaining ones being analogous. Indeed, recalling that $q = 1 - p$, note that

$$\frac{\mathbb{P}(Z_\ell = M_n - 2\gamma_n)}{\mathbb{P}(Z_\ell = M_n)} = \frac{M_n(M_n - 1)\cdots(M_n - 2\gamma_n + 1)q^{2\gamma_n}}{(z_\ell - M_n + 1)\cdots(z_\ell - M_n + 2\gamma_n)p^{2\gamma_n}}, \quad\quad (41)$$

and also

$$M_n^{2\gamma_n}\Big(1 - \frac{1}{M_n}\Big)\cdots\Big(1 - \frac{2\gamma_n - 1}{M_n}\Big) = M_n^{2\gamma_n}\exp\Big(-(1 + o(1))\frac{4\gamma_n^2}{2M_n}\Big) = (1 - o(1))M_n^{2\gamma_n}.$$

Furthermore, since $\gamma_n p = o(1)$, $q^{2\gamma_n} = \exp(-(1+o(1))2\gamma_n p) = 1 - o(1)$. Therefore, (41) is equal to

$$(1-o(1))\frac{M_n^{2\gamma_n}}{p^{2\gamma_n}}\prod_{i=1}^{2\gamma_n}\frac{1}{z_\ell - M_n + i} = (1-o(1))\prod_{i=1}^{2\gamma_n}\frac{np + \sqrt{(1-\frac{\varepsilon_n}{2})np\log(1/(np^3))}}{np - (1+o(1))\ell n p^2}$$

$$= \left(1 + (1+o(1))\left(\sqrt{\tfrac{1}{np}(1-\tfrac{\varepsilon_n}{2})\log(1/(np^3))} + \ell p\right)\right)^{2\gamma_n}$$

$$= \exp\left((2\gamma_n + o(\gamma_n))\left(\sqrt{\tfrac{1}{np}(1-\tfrac{\varepsilon_n}{2})\log(1/(np^3))} + \ell p\right)\right)$$

$$= \exp(O((np)^{-\varepsilon_n/2}))$$

$$= 1 + o(1),$$

where in the second-to-last equality we used that $\gamma_n \ell p \le \gamma_n k_* p = o(1)$.

Using this observation in (40) implies that

$$\prod_{\ell=1}^{k_*}(1 - (1+o(1))\mathbb{P}\left(Z_\ell \in [M_n - 2\gamma_n + 1, M_n]\right)) = \prod_{\ell=1}^{k_*}(1 - (1+o(1))2\gamma_n\mathbb{P}\left(Z_\ell = M_n\right)). \qquad (42)$$

Finally, let us fix $\ell \in [k_*]$ and find the order of $2\gamma_n\mathbb{P}(Z_\ell = M_n)$. Recall that for $m = m(s)$ such that $1 \ll m \ll s$ as $s \to \infty$, it holds that

$$\binom{s}{m} \sim \left(\frac{se}{m}\right)^m \frac{1}{\sqrt{2\pi m}}\exp\left(-\frac{m^2}{2s} + O\left(\frac{m^3}{s^2}\right)\right).$$

Thus, since $M_n = (1+o(1))np$, we have

$$2\gamma_n\mathbb{P}(Z_\ell = M_n) = 2\gamma_n\binom{z_\ell}{M_n}p^{M_n}q^{z_\ell - M_n}$$

$$= 2(np)^{1/2-\varepsilon_n}\left(\frac{z_\ell pe}{M_n}\right)^{M_n}\frac{1+o(1)}{\sqrt{2\pi np}}q^{z_\ell - M_n}\exp\left(-\frac{M_n^2}{2z_\ell} + O\left(\frac{M_n^3}{z_\ell^2}\right)\right)$$

$$= (np)^{-\varepsilon_n}\left(\frac{z_\ell p}{M_n}\right)^{M_n}(1+o(1))q^{z_\ell - M_n}\sqrt{\frac{2}{\pi}}\exp\left(M_n - \frac{M_n^2}{2z_\ell} + O\left(\frac{M_n^3}{z_\ell^2}\right)\right). \qquad (43)$$

Using that $1 + x = \exp(x - \frac{x^2}{2} + O(x^3))$ as $x \to 0$ in order to bound from above $\frac{z_\ell p}{M_n} = 1 + \frac{z_\ell p - M_n}{M_n}$, we have

$$\left(\frac{z_\ell p}{M_n}\right)^{M_n} = \exp\left(z_\ell p - M_n - \frac{(z_\ell p - M_n)^2}{2M_n} + O\left(\frac{|z_\ell p - M_n|^3}{M_n^2}\right)\right), \qquad (44)$$

$$q^{z_\ell - M_n} = \exp\left(-\left(p + \frac{p^2}{2} + O(p^3)\right)|z_\ell - M_n|\right). \qquad (45)$$

Using that $p^3 = o(1/n)$, $z_\ell = (1+o(1))n$, $M_n = (1+o(1))np$, and observing that both $p^3 z_\ell$ and $p^2 M_n$ are of order $O(np^3) = o(1)$, the exponent in the right-hand side of (45) is

$$-\left(p + \frac{p^2}{2}\right)(z_\ell - M_n) + o(1) = -z_\ell p - \frac{(z_\ell p - M_n)^2}{2z_\ell} + \frac{M_n^2}{2z_\ell} + o(1).$$

Hence, combining (43), (44) and (45), we obtain that $2\gamma_n\mathbb{P}(Z_\ell = M_n)$ is equal to

$$(np)^{-\varepsilon_n}\exp\left(-\frac{1}{2}(z_\ell p - M_n)^2\left(\frac{1}{z_\ell} + \frac{1}{M_n}\right) + O\left(\frac{M_n^3}{z_\ell^2} + \frac{|z_\ell p - M_n|^3}{M_n^2}\right) + O(1)\right).$$

Next, observe that $\frac{M_n^3}{z_\ell^2} = (1+o(1))np^3 = o(1)$ and

$$z_\ell p - M_n = -(\ell - 1)np^2 + \tfrac{1}{2}(\ell - 1)(\ell - 2)np^3 - \sqrt{(1 - \tfrac{\varepsilon_n}{2})np\log(\tfrac{1}{np^3})}$$

$$= -np^2\left(\ell - 1 + 2\sqrt{2(1 - \tfrac{\varepsilon_n}{2})}k_* + O(k_*^2 p)\right).$$

Thus, $\frac{|z_\ell p - M_n|^3}{M_n^2} = O(k_*^3 np^4) = o(1)$ and $\frac{(z_\ell p - M_n)^2}{z_\ell} = O(k_*^2 np^4) = o(1)$. Moreover, using that $M_n = (1 - O(k_*p))np$ and $k_*^3 p = O((\log(1/(np^3)))^{3/2}/\sqrt{n^3 p^7}) = o(1)$, we get that

$$\frac{(z_\ell p - M_n)^2}{2M_n} \geq (1 + O(k_*p))\left(4\left(1 - \frac{\varepsilon_n}{2}\right)k_*^2 + O(k_*^2 p)\right)np^3 = \frac{1}{2}\left(1 - \frac{\varepsilon_n}{2}\right)\log\left(\frac{1}{np^3}\right) + o(1).$$

Hence,

$$2\gamma_n \mathbb{P}(Z_j = M_n) \leq (np)^{-\varepsilon_n} \exp\left(-\tfrac{1}{2}(1 - \tfrac{\varepsilon_n}{2})\log(\tfrac{1}{np^3}) + o(1)\right)$$

$$= O((np)^{-\varepsilon_n}(np^3)^{1/2 - \varepsilon_n/4}) = O\left(\frac{(np)^{-\varepsilon_n/2}}{k_*}\right), \tag{46}$$

where for the last equality we used that $(n^2 p^4)^{-\varepsilon_n/4} \leq 1$ and $(np)^{-\varepsilon_n/4} \ll \frac{1}{\sqrt{\log(1/(np^3))}}$.

Using that $1 - x = \exp(-(1 + o(1))x)$ as $x \to 0$, we conclude that (42) rewrites as

$$\exp\left(-(1 + o(1))\sum_{\ell=1}^{k_*} 2\gamma_n \mathbb{P}(Z_\ell = M_n)\right) = \exp(-O((np)^{-\varepsilon_n/2})) = 1 - o(1),$$

which finishes the proof of the lemma. $\qquad\square$

Now, recall that by definition $\mathfrak{B}^{(1)} = \max_{\ell \in [2k]} \mathfrak{B}_1(\ell)$, and define also

$$\mathfrak{B}^{(2)} = \max_{\ell \in [2k]\setminus\{\ell_1\}} \mathfrak{B}_1(\ell),$$

where $\ell_1 = \min\{\ell \in [k_*] : Z_*^{(1)} = Z_\ell\}$.

Corollary 3.29. *Under the coupling from Lemma 3.22, a.a.s. $\mathfrak{B}_1(\ell_1) = \mathfrak{B}^{(1)}$ and $\mathfrak{B}^{(1)} - \mathfrak{B}^{(2)} \geq \gamma_n$.*

Proof. Recall that, under the coupling from Lemma 3.22, we have that a.a.s.

$$|\mathfrak{B}_1(\ell_1) - Z_{\ell_1}| \leq (np)^{2/5} \tag{47}$$

and, for $\widehat{\ell} \in [2k] \setminus \{\ell_1\}$ such that $|\mathfrak{B}^{(2)} - Z_{\widehat{\ell}}| = \min_{\ell \in [2k]\setminus\{\ell_1\}} |\mathfrak{B}^{(2)} - Z_\ell|$,

$$|\mathfrak{B}^{(2)} - Z_{\widehat{\ell}}| \leq (np)^{2/5}. \tag{48}$$

By Corollary 3.26 and Lemma 3.28 we know that a.a.s. $|Z^{(1)} - Z^{(2)}| \geq 2\gamma_n \gg (np)^{2/5}$, which together with (47) and (48) directly implies the first statement of the corollary. For the second statement, since by definition of $Z^{(1)}$ and $Z^{(2)}$ we have $Z^{(1)} - Z_{\widehat{\ell}} \geq Z^{(1)} - Z^{(2)} = |Z^{(1)} - Z^{(2)}|$, the triangle inequality implies that a.a.s.

$$|\mathfrak{B}^{(1)} - \mathfrak{B}^{(2)}| \geq |Z^{(1)} - Z_{\widehat{\ell}}| - |\mathfrak{B}_1(\ell_1) - Z_{\ell_1}| - |Z_{\widehat{\ell}} - \mathfrak{B}^{(2)}| \geq 2\gamma_n - 2(np)^{2/5} \geq \gamma_n,$$

which finishes the proof of the corollary. $\qquad\square$

At this stage, we have the tools necessary to analyze the difference between the number of vertices in Level 3 that, at the second round of ALAP, receive the first- and second-most-represented label among the Level 3 vertices. Our next result bounds the difference in terms of the number of vertices in Level 3 possessing label $\ell_1 = \min\{\ell \in [2k] : Z_*^{(1)} = Z_\ell\}$ (recall that, by Corollary 3.26, a.a.s. $\ell_1 = \min\{\ell \in [k_*] : Z^{(1)} = Z_\ell\}$) and establishes that ℓ_1 is in fact the most represented label among Level 3 vertices after the second round of ALAP.

Lemma 3.30. *There is a constant $c_2 > 0$ such that the event "in Level 3 there are $\mathfrak{C}_2(\ell_1) \geq \frac{n}{2k}$ vertices with label ℓ_1 after the second round of ALAP, and moreover, the number of vertices with any label in $[k] \setminus \{\ell_1\}$ in Level 3 after the second round is at least by $c_2 p^{1/2}(np)^{-\varepsilon_n}\mathfrak{C}_2(\ell_1)$ less than the number of vertices with label ℓ_1", that is,*

$$\{\mathfrak{C}_2(\ell_1) \geq \tfrac{n}{2k} \text{ and } \forall \ell \in [k] \setminus \{\ell_1\}, \mathfrak{C}_2(\ell) \leq (1 - c_2 p^{1/2}(np)^{-\varepsilon_n})\mathfrak{C}_2(\ell_1)\}$$

holds a.a.s.

Proof. Let us condition on the a.a.s. event that $\mathfrak{B} \leq 3knp$ (see Lemma 3.10). Since in ALAP, conditionally on the edges exposed at the first round, every vertex in Level 3 receives label ℓ_1 independently and with probability at least $1/k$, it follows that $\mathfrak{C}_2(\ell_1)$ dominates a binomial random variable with parameters $n - 2k - 3knp = (1 - o(1))n$ and $1/k$, so by Chernoff's bound

$$\mathbb{P}(\mathfrak{C}_2(\ell_1) \leq n/2k) \leq \exp\left(-\frac{(n/2k - \mathbb{E}\mathfrak{C}_2(\ell_1))^2}{2\mathbb{E}\mathfrak{C}_2(\ell_1)}\right) = \exp(-\Omega(n/k)) = o(1).$$

Now, fix $\ell \in [k] \setminus \{\ell_1\}$. Then, by the above inequality, a.a.s. the number of vertices in Level 3 receiving a label among $\{\ell_1, \ell\}$ after the second round is at least $\mathfrak{C}_2(\ell_1) \geq n/2k$. Let us condition on this event and on the set $C_2(\{\ell_1, \ell\})$ of these vertices.

Now, by Corollary 3.29 we know that a.a.s. $\mathfrak{B}^{(1)} - \mathfrak{B}^{(2)} \geq \gamma_n$, and moreover, combining Lemma 3.24 and Corollary 3.29 implies that $\mathfrak{B}^{(1)} \leq M_n^+ = np + \sqrt{\frac{3}{2}np\log(1/(np^3))}$. Let us condition on these events as well. Then, for a vertex u in the third level, Lemma 2.6 implies that

$$\frac{\mathbb{P}(u \in C_2(\ell_1) \mid u \in C_2(\{\ell_1, \ell\}))}{\mathbb{P}(u \in C_2(\ell) \mid u \in C_2(\{\ell_1, \ell\}))} \geq$$
$$\frac{\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell_1), p) > \mathrm{Bin}(\mathfrak{B}_1(\ell), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell_1), p) = \mathrm{Bin}(\mathfrak{B}_1(\ell), p))}{\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) > \mathrm{Bin}(\mathfrak{B}_1(\ell_1), p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) = \mathrm{Bin}(\mathfrak{B}_1(\ell_1), p))}.$$

By replacing the denominators in the latter fraction with the sum of the corresponding numerator and denominator and using the inequality $\mathfrak{B}^{(1)} - \mathfrak{B}^{(2)} \geq \gamma_n$ we obtain that, in our procedure, the probability that a vertex in $C_2(\{\ell_1, \ell\})$ gets label ℓ_1, conditionally on obtaining either ℓ or ℓ_1, is bounded from below by

$$\alpha_\ell := \mathbb{P}\big(\mathrm{Bin}(\mathfrak{B}^{(1)}, p) > \mathrm{Bin}(\mathfrak{B}^{(1)} - \gamma_n, p)\big) + \tfrac{1}{2}\mathbb{P}\big(\mathrm{Bin}(\mathfrak{B}^{(1)}, p) = \mathrm{Bin}(\mathfrak{B}^{(1)} - \gamma_n, p)\big).$$

Then, Remark 2.4 implies that α_ℓ is bounded from below by $\frac{1}{2} + \Omega\left(\frac{\gamma_n p}{\sqrt{\mathfrak{B}^{(1)}p}}\right) = \frac{1}{2} + \Omega\left(\frac{p^{1/2}}{(np)^{\varepsilon_n}}\right)$. Hence, using that conditionally on $\mathfrak{C}_2(\{\ell_1, \ell\})$ we have $\mathbb{E}\mathfrak{C}_2(\ell_1) \geq \alpha_\ell \mathfrak{C}_2(\{\ell_1, \ell\})$, by Chernoff's bound the number of vertices in $C_2(\{\ell_1, \ell\})$ getting label ℓ_1 in our procedure satisfies

$$\mathbb{P}\left(\mathfrak{C}_2(\ell_1) \leq \frac{1}{2}\left(\frac{1}{2} + \alpha_\ell\right)\mathfrak{C}_2(\{\ell_1, \ell\})\right) = \exp\left(-\Omega\left(\left(\alpha_\ell - \frac{1}{2}\right)^2\mathfrak{C}_2(\{\ell_1, \ell\})\right)\right) = o(1/n).$$

Hence, by a union bound we conclude that a.a.s. for every $\ell \in [k] \setminus \{\ell_1\}$, the difference between the number of vertices with label ℓ_1 and ℓ in Level 3 after the second round is at least $(\alpha_\ell - 1/2)\mathfrak{C}_2(\{\ell_1, \ell\}) = \Omega(p^{1/2}(np)^{-\varepsilon_n}\mathfrak{C}_2(\ell_1))$. $\qquad\square$

32

It remains to analyze the effect of the second round of ALAP over the vertices in Level 2. For every $\ell \in [k] \setminus \ell_1$, denote by $\bar{B}_{\ell_1,\ell}$ the vertices in B outside the basins $B_1(\ell_1), B_1(\ell)$ as well as the basins $B_1(i)$ where at least one of $v_i v_{\ell_1}$ and $v_i v_\ell$ is an edge in G_n.

Lemma 3.31. *There exists a constant $c_3 > 0$ such that a.a.s. the following holds: for every $\ell \in [k] \setminus \{\ell_1\}$, the number of vertices with label ℓ_1 after the second round in $\bar{B}_{\ell_1,\ell}$ is at least by $c_3 \gamma_n n^{-1/2} |B_2(\ell_1) \cap \bar{B}_{\ell_1,\ell}| \geq c_3 \gamma_n p \sqrt{n}/2$ larger than the number of vertices with label ℓ after the second round in $\bar{B}_{\ell_1,\ell}$ of ALAP.*

Proof. We reveal all edges with two endvertices in A and condition on the following a.a.s. events. The first of them is the event that, for every label in $\ell \in [2k]$, $\mathfrak{A}_1(\ell) \leq 10$ (this holds a.a.s. by Observation 3.14). Moreover, we condition on the event that, for every $\ell \in [2k]$, $2np \geq \mathfrak{B}_1(\ell)$, and that $\mathfrak{B}_1([k+1, 2k]) \geq 2knp/3$. Both of these events are a.a.s. by Chernoff's bound (applied as in Lemma 3.10), and a union bound over all $2k$ vertices in A in the first case. Finally, we also condition on the event of Corollary 3.29. Note that all three events are measurable in terms of the edges between two vertices in A and the ones between v_ℓ and $B_1(\ell)$ for all $\ell \in [2k]$.

For every vertex $v \in B_1([k+1, 2k])$, note that the indicator variable of the event $v \in B_2(\ell_1)$ stochastically dominates a Bernoulli random variable with success probability $1/k$ since v is not connected by an edge to any of $(v_j)_{j=1}^k$ and $B_1(\ell_1)$ is larger than all other basins by definition. Note also that all vertices in $B_1([k+1, 2k])$ are attributed label ℓ_1 independently of each other. Thus, the number W of such vertices stochastically dominates a binomial random variable $\text{Bin}(2knp/3, 1/k)$, and by Chernoff's bound $\mathbb{P}(W \leq 3np/5) \leq e^{-\Omega(np)}$. We conclude that a.a.s., our procedure attributes label ℓ_1 to at least $3np/5$ vertices in Level 2, that is, $\mathfrak{B}_2(\ell_1) \geq 3np/5$.

Now, we show that a.a.s. for every $\ell \in [k] \setminus \{\ell_1\}$, among the vertices in $B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}$ there are more vertices with label ℓ_1 than with label ℓ. Fix $\ell \in [k] \setminus \{\ell_1\}$. First, by the preceding paragraph and the fact that no vertex in A has more than 10 neighbors in A, a.a.s.,

$$|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}| \geq (1 - o(1))|B_1([k+1, 2k]) \cap B_2(\ell_1)| \geq np/2.$$

We condition on the set $B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}$ and on the event $|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}| \geq np/2$. Given a vertex $v \in B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}$, recall that each of its edges towards the vertices with label ℓ_1 and ℓ on the first two levels remains unexposed, and the numbers a_1 and a_2 of these vertices are among $\mathfrak{B}_1(\ell_1) \pm 11$ and $\mathfrak{B}_1(\ell) \pm 11$, respectively, where, thanks to the conditioning in the beginning of the proof, 11 is an upper bound on the number of vertices on Level 1 with any given label. Hence, applying Remark 2.4 for a_1, a_2, $X_1 \in \text{Bin}(a_1, p)$ and $X_2 \in \text{Bin}(a_2, p)$, we get that

$$\mathbb{P}(v \in B_2(\ell_1) \mid v \in B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}) \geq \mathbb{P}(X_1 > X_2) + \frac{1}{2}\mathbb{P}(X_1 = X_2)$$

$$= \frac{1}{2} + \Omega\left(\frac{\gamma_n p}{\sqrt{a_1 p}}\right) \geq \frac{1}{2} + 2c_3 \gamma_n n^{-1/2},$$

where $c_3 > 0$ is a sufficiently small absolute constant.

We conclude that the number of vertices in $B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}$ receiving label ℓ_1 by our procedure dominates the sum of $|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}| \geq np/2$ Bernoulli random variables with success probability $\frac{1}{2} + 2c_3 \gamma_n n^{-1/2}$. Recalling that we are working conditionally on the set $B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}$ and the event $|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}| \geq np/2$, and that $\gamma_n = (np)^{1/2-o(1)}$, Chernoff's bound implies that

$$\mathbb{P}\big(|B_2(\ell_1) \cap \bar{B}_{\ell_1,\ell}| \leq |B_2(\ell) \cap \bar{B}_{\ell_1,\ell}| + 2c_3 \gamma_n n^{-1/2}|B_2(\{\ell_1, \ell\} \cap \bar{B}_{\ell_1,\ell})|\big)$$

$$= \mathbb{P}\big(|B_2(\ell_1) \cap \bar{B}_{\ell_1,\ell}| \leq (\tfrac{1}{2} + c_3 \gamma_n n^{-1/2})|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}|\big)$$

$$\leq \mathbb{P}\big(\text{Bin}(|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}|, \tfrac{1}{2} + 2c_3 \gamma_n n^{-1/2}) \leq (\tfrac{1}{2} + c_3 \gamma_n n^{-1/2})|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}|\big)$$

$$\leq \exp\left(-\frac{(c_3 \gamma_n n^{-1/2}|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}|)^2}{3|B_2(\{\ell_1, \ell\}) \cap \bar{B}_{\ell_1,\ell}|}\right)$$

$$\leq \exp(-(np)^{1-o(1)}p) = o(1/n).$$

The statement follows by taking a union bound over all $\ell \in [k] \setminus \{\ell_1\}$. $\qquad\square$

It remains to analyze the number of vertices in $B \setminus \bar{B}_{\ell_1,\ell}$ which obtain a label in $\{\ell_1, \ell\}$ at the second round of ALAP. In fact, we will concentrate our effort on showing that the vertices in $B \setminus \bar{B}_{\ell_1,\ell}$ getting label ℓ at the second round is a.a.s. smaller than the difference ensured by Lemma 3.30.

Lemma 3.32. *A.a.s. for every $\ell \in [2k] \setminus \{\ell_1\}$, there are at most $200n^{3/2}p^{5/2+\varepsilon}/100$ vertices with label ℓ in $B \setminus \bar{B}_{\ell_1,\ell}$ after the second round of ALAP.*

Proof. Fix $m := \lceil n^{-1/2}p^{-3/2-\varepsilon}/100 \rceil$ and observe that $m \ll k$. By Chernoff's bound together with a union bound, a.a.s. $\mathfrak{B}_1(\ell) = np - O(\sqrt{np}\log m)$ for all $\ell \in [m]$ (with uniform constant in the $O(\cdot)$) and, moreover, a.a.s. every vertex in B has up to 10 neighbors in A by Observation 3.14. Moreover, conditionally on the latter a.a.s. event, another Chernoff bound implies that the (up to 22) basins of vertices v_i which are incident to $\{v_\ell, v_{\ell_1}\}$ in A have total size up to $22 \cdot 3np$; we condition on the above three events.

Now, we fix a vertex $v \in B \setminus \bar{B}_{\ell_1,\ell}$ for some $\ell \leq m$, and bound from above the probability of the event $\{v \in B_2(\ell)\}$. To do so, we expose the edges from v to $B_1(\ell)$; note that v has up to 10 neighbors in A but we ignore these for the moment. Observe that there is a family of independent random variables $(X_i)_{i=1}^m$ with distributions respectively $(\mathrm{Bin}(\mathfrak{B}_1(i) - \mathbb{1}_{v \in B_1(i)}, p))_{i=1}^m$ which are stochastically dominated by the number of edges from v to vertices with label i not yet exposed after the first round. Our first step is to find the expected value of the maximum M of $(X_i)_{i=1}^m$ as well as its fluctuations; while this index i of the maximal X_i does not necessarily correspond to the label obtained by v after the second round, it will be sufficient to compare the probability that a label $\ell \leq m$ is attributed to v for different values of ℓ. For $c \in \mathbb{R}$, fix $x_c = x_c(n) := np^2 + \sqrt{2np^2 \log(m/\sqrt{\log m})} + c\sqrt{2np^2/\log(m/\sqrt{\log m})}$. Then,

$$\mathbb{P}(X_i \geq x_c) = \sum_{j=x_c}^{\mathfrak{B}_1(i)} \binom{\mathfrak{B}_1(i)}{j} p^j (1-p)^{\mathfrak{B}_1(i)-j}. \tag{49}$$

Instead of directly computing the above sum, we will compare consecutive terms first. Given j, we have

$$\binom{\mathfrak{B}_1(i)}{j+1} p^{j+1}(1-p)^{\mathfrak{B}_1(i)-j-1} \Big/ \binom{\mathfrak{B}_1(i)}{j} p^j (1-p)^{\mathfrak{B}_1(i)-j} = \left(\frac{\mathfrak{B}_1(i)+1}{j+1} - 1\right)\frac{p}{1-p}. \tag{50}$$

Moreover, one can easily check that

$$\mathfrak{B}_1(i)p = np^2 + O(\sqrt{np^3}\log n) = np^2 + o(\sqrt{np^2/\log m})$$

and $p = o(1/\sqrt{np^2 \log m})$, which implies that (50) is equal to

$$\frac{np^2}{j+1} + o\left(\frac{1}{\sqrt{np^2 \log m}}\right) = 1 + \frac{np^2 - j - 1}{j+1} + o\left(\frac{1}{\sqrt{np^2 \log m}}\right).$$

In particular, for any $d > c$, we have that

$$\prod_{j=x_c}^{x_d} \left(1 + \frac{np^2 - j - 1}{j+1} + o\left(\frac{1}{\sqrt{np^2 \log m}}\right)\right) = \exp\left((1+o(1))\sum_{j=x_c}^{x_d} \frac{np^2 - j - 1}{j+1}\right)$$

$$= \exp\left((1+o(1))\sum_{j=x_c}^{x_d} \frac{np^2 - j - 1}{np^2}\right)$$

$$= \exp\left((1+o(1))\frac{(np^2 - x_c)^2 - (np^2 - x_d)^2}{2np^2}\right)$$

34

$$= e^{2(c-d)+o(1)}.$$

As a result, using a classical approximation of the sum (49) by an integral and the fact that $px_c = o(1)$, we obtain that

$$\mathbb{P}(X_i \geq x_c) = (1 + o(1)) \binom{\mathfrak{B}_1(i)}{x_c} p^{x_c} (1 - p)^{\mathfrak{B}_1(i) - x_c} \cdot \sqrt{\frac{2np^2}{\log(m/\sqrt{\log m})}} \int_0^\infty e^{-2u} du$$

$$= (1 + o(1)) \left(\frac{e\mathfrak{B}_1(i)p}{x_c} \right)^{x_c} \exp\left(-\frac{x_c^2}{2\mathfrak{B}_1(i)} \right) \frac{1}{\sqrt{2\pi x_c}} e^{-p\mathfrak{B}_1(i)} \cdot \sqrt{\frac{np^2}{2\log m}}$$

$$= (1 + o(1)) \left(\frac{enp^2}{x_c} \right)^{x_c} e^{-np^2} \cdot \sqrt{\frac{1}{4\pi \log m}}$$

$$= (1 + o(1)) \left(1 + \frac{np^2 - x_c}{x_c} \right)^{x_c} e^{x_c - np^2} \cdot \sqrt{\frac{1}{4\pi \log m}}. \tag{51}$$

Finally, Taylor expansion of the first term in (51) shows that

$$\left(1 + \frac{np^2 - x_c}{x_c} \right)^{x_c} = e^{np^2 - x_c - (x_c - np^2)^2/2x_c + O((x_c - np^2)^3/x_c^2)}$$

$$= (1 + o(1)) e^{np^2 - x_c} \cdot e^{-2c} \sqrt{\log m}/m,$$

which means that (49) equals $(1 + o(1))e^{-2c}/(\sqrt{4\pi}m)$, and the leading order of this expression does not depend on $i \in [m]$.

Moreover, for every $i \in [m]$ and every integer value, X_i takes this value with probability $O(1/\sqrt{np^2}) = o(1/m)$. As a result, a.a.s. a single variable among $X_1, \ldots, X_m$ attains the maximum and, for any constant $C \in \mathbb{N}$,

$$\mathbb{P}\left(\exists j \in [m] \setminus \{i\}, X_j \geq M - C \, \Big| \, \max_{j \in [m]} X_j = X_i = M \right) = O\left(\frac{m}{\sqrt{np^2}} \right) = o(1). \tag{52}$$

We conclude that, first, the edges between v and A a.a.s. do not influence the label that v obtains at the second round of ALAP, and second, each of the first m labels is attributed to v with the same probability, up to a factor of $1 + o(1)$: indeed, on the one hand, for every $\varepsilon > 0$, there is a $c = c(\varepsilon) > 0$ such that $M = \max_{i \in [m]} X_i \in [x_{-c}, x_c]$ with probability at least $1 - \varepsilon$ and, on the other hand, conditionally on this event, for every $i, j \in [m]$,

$$\mathbb{P}(X_i = M) = (1 + o(1))\mathbb{P}(X_j = M) = (1 + o(1))/m.$$

Denote by Y_ℓ the number of vertices in $B \setminus \bar{B}_{\ell_1,\ell}$ which obtain label ℓ at the second round. In particular, thanks to our conditioning that $B \setminus \bar{B}_{\ell_1,\ell}$ contains at most 22 basins of total size at most $22 \cdot 3np$ and the fact that every label in $[m]$ is obtained with approximately the same probability by the vertices in B, the mean of Y_ℓ is of order $\Theta(|B \setminus \bar{B}_{\ell_1,\ell}|/m) = \Theta(np/m)$, but also at most $22 \cdot 3np/m \leq 100np/m$, say.

Next, we show that a.a.s., for all $\ell \in [m]$, Y_ℓ is "close" to its expectation. The argument is similar to Step 2 in the proof of Lemma 3.15. Note that $Y_\ell = \sum_{v \in B_1(\ell)} \mathbb{1}_{v \in B_2(\ell)}$. Then, we have that

$$\mathbb{V}(Y_\ell) = \sum_{u,v \in B_1(\ell)} \left(\mathbb{E}[\mathbb{1}_{u \in B_2(\ell)} \mathbb{1}_{v \in B_2(\ell)}] - \mathbb{E}[\mathbb{1}_{v \in B_2(\ell)}] \mathbb{E}[\mathbb{1}_{v \in B_2(\ell)}] \right)$$

$$= (1 + o(1))\mathbb{E}[X_\ell] + \sum_{u,v \in B_1(\ell): u \neq v} \left(\mathbb{E}[\mathbb{1}_{u \in B_2(\ell)} \mathbb{1}_{v \in B_2(\ell)}] - \mathbb{E}[\mathbb{1}_{u \in B_2(\ell)}] \mathbb{E}[\mathbb{1}_{v \in B_2(\ell)}] \right). \tag{53}$$

Using a transformation similar to the one from equations (26)–(30), we deduce that for all pairs of different vertices u, v in $B_1(\ell)$, $\mathbb{E}[\mathbb{1}_{u \in B_2(\ell)} \mathbb{1}_{v \in B_2(\ell)}]$ rewrites as

$$q\mathbb{E}[\mathbb{1}_{u \in B_2(\ell)} \mathbb{1}_{v \in B_2(\ell)} \mid uv \notin G_n] + p\mathbb{E}[\mathbb{1}_{u \in B_2(\ell)} \mathbb{1}_{v \in B_2(\ell)} \mid uv \in G_n]$$

$$= q\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \notin G_n]\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n]$$
$$+ p\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \in G_n]\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n]$$
$$= q\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \notin G_n]^2 + p\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \in G_n]^2,$$

while $\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)}]$ rewrites as

$$(q\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \notin G_n] + p\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \in G_n])$$
$$\cdot (q\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n] + p\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n]).$$

This implies that the general term in the sum in (53) rewrites as

$$pq(\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{u\in B_2(\ell)} \mid uv \in G_n])$$
$$\cdot (\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n])$$
$$= pq(\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n])^2.$$

Finally, to deduce the analogue of (31), we show that

$$\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n] = (1 + o(1))\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n] = (1 + o(1))\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)}]. \tag{54}$$

Fix $j \in [m] \setminus \{\ell\}$. Since by (52) a constant number of edges do not modify the probability of obtaining a label in $[m]$ apart from lower order terms, the probabilities

$$\frac{\mathbb{P}(v \in B_2(\ell) \mid uv \in G_n)}{\mathbb{P}(v \in B_2(\ell) \mid uv \in G_n) + \mathbb{P}(v \in B_2(j) \mid uv \in G_n)} \quad \text{and} \quad \frac{\mathbb{P}(v \in B_2(\ell))}{\mathbb{P}(v \in B_2(\ell)) + \mathbb{P}(v \in B_2(j))} \tag{55}$$

are the same up to a factor of $1 + o(1)$, and the comparison conditionally on the event $uv \notin G_n$ instead of $uv \in G_n$ is also of the same order, which proves (54).

We conclude that $\mathbb{V}(Y_\ell) = (1 + o(1))(\mathbb{E}Y_\ell + o(p(\mathbb{E}Y_\ell)^2)) = o(p(\mathbb{E}Y_\ell)^2))$, where we used that $\mathbb{E}Y_\ell = \Theta((\mathbb{E}Y_\ell)^2/(np/m)) = o(p(\mathbb{E}Y_\ell)^2)$. Finally, recalling that $\mathbb{E}Y_\ell \leq 100np/m$, for every $\ell \in [m]$,

$$\mathbb{P}\Big(Y_\ell \geq \frac{200np}{m}\Big) \leq \frac{\mathbb{V}(Y_\ell)}{(\mathbb{E}Y_\ell)^2} = o(p).$$

Since ties are broken towards smaller labels in the first round, the above display actually holds for all $\ell \in [2k]$. Since $kp = o(1)$, a union bound finishes the proof. $\square$

Remark 3.33. The same argument (up to minor modifications in the definitions of a and b in the proof of Lemma 3.32 due to v possibly being in a different basin) shows that a.a.s. for all $\ell \in [k] \setminus \{\ell_1\}$, $\mathfrak{B}_2(\ell) \leq 200np \log n$. Indeed, for $\ell \leq \frac{k}{\log n}$, the argument of the proof of Lemma 3.32 can be applied directly to bound from above the probability of the complementary event, and for larger values of ℓ, this probability can only decrease. The result then follows by a union bound over all $\ell \in [k] \setminus \{\ell_1\}$.

Lemma 3.34. *Let $(\Omega_\ell)_{\ell=1}^k$ be subsets of B such that for every $\ell \in [k]\setminus\{\ell_1\}$, $|\Omega_\ell| - |\Omega_{\ell_1}| \leq 200n^{3/2}p^{5/2+\varepsilon/100}$ and $|\Omega_\ell| \leq 200np \log n$. Then, a.a.s. for every vertex v in C and every $\ell \in [k]$, it holds that $|N(v) \cap \Omega_\ell| - |N(v) \cap \Omega_{\ell_1}| \leq 400n^{3/2}p^{7/2+\varepsilon/100}$.*

Proof. Fix a vertex $v \in C$. By Chernoff's bound, with probability $1 - o(n^{-2})$, we have that both

$$|N(v) \cap \Omega_\ell| \leq p|\Omega_\ell| + \sqrt{p \cdot 200np(\log n)^3}$$
$$|N(v) \cap \Omega_{\ell_1}| \geq p|\Omega_{\ell_1}| - \sqrt{p \cdot 200np(\log n)^3},$$

and therefore also

$$|N(v) \cap \Omega_\ell| - |N(v) \cap \Omega_{\ell_1}| \leq p(|\Omega_\ell| - |\Omega_{\ell_1}|) + 2\sqrt{200np^2(\log n)^3} \leq 400n^{3/2}p^{7/2+\varepsilon/100}$$

where we used that $n^{3/2}p^{7/2+\varepsilon/100} \gg \sqrt{np^2(\log n)^3}$ implied by $np \geq n^{5/8+\varepsilon}$. The lemma follows by a union bound over the complementary events for all $v \in C$ and $\ell \in [k] \setminus \{\ell_1\}$. $\square$

Proof of the second point in Theorem 1.1. We first show the desired conclusion for the third round of ALAP and then make the connection with LPA. Note that, by Lemma 3.31, Lemma 3.32 and Remark 3.33, the assumptions of Lemma 3.34 with $\Omega_i = B_2(i)$ are satisfied.

Recall that attributing the labels of the vertices in C based only on their edges towards B leaves all edges in C unexposed. Using this, we prove that the surplus coming from the neighbors with label ℓ_1 in Level 3 is far larger than $400n^{3/2}p^{7/2+\varepsilon}/100$ for any vertex (note that based on the conclusion of Lemma 3.34 with the above choice of $(\Omega_i)_{i=1}^k$, this is sufficient to conclude the proof). Indeed, fix a vertex $v \in C$ and, for every $\ell \in [k]$, let Y_ℓ be the number of neighbors of v in $C_2(\ell)$, that is, $Y_\ell = |N(v) \cap C_2(\ell)|$. Then, by Lemma 3.30, a.a.s.,

$$\mathfrak{C}_2(\ell_1) \geq \frac{n}{2k}.$$

We condition on this event. Then, an application of Chernoff's bound shows (for positive real numbers $(\varepsilon_n)_{n \geq 1}$ chosen as before the statement of Lemma 3.28) that

$$\mathbb{P}(Y_{\ell_1} - \mathbb{E}Y_{\ell_1} \leq -p^{1/2}(np)^{-2\varepsilon_n}\mathbb{E}Y_{\ell_1}) \leq \exp\left(-\frac{p(np)^{-4\varepsilon_n}\mathbb{E}Y_{\ell_1}}{3}\right)$$
$$\leq \exp\left(-\frac{p^2(np)^{-4\varepsilon_n}n}{6k}\right) = o\left(\frac{1}{kn}\right),$$

while, for every other label $\ell \in [k] \setminus \{\ell_1\}$, we have that

$$\mathbb{P}\left(Y_\ell - \mathbb{E}Y_\ell \geq p^{1/2}(np)^{-2\varepsilon_n}\mathbb{E}Y_{\ell_1}\right) \leq \exp\left(-\frac{p(np)^{-4\varepsilon_n}\mathbb{E}Y_{\ell_1}}{3}\right) = o\left(\frac{1}{kn}\right).$$

Using Lemma 3.30, we conclude that, with probability $1 - o(\frac{1}{kn})$, for every vertex $v \in C$, by taking into consideration its neighbors in $B \cup C$, the number of neighbors of v with label ℓ_1 after round 2 that are in Level 3 is at least by

$$Y_{\ell_1} - Y_\ell - 400n^{3/2}p^{7/2+\varepsilon}/100 \geq \mathbb{E}Y_{\ell_1} - \mathbb{E}Y_\ell - 2p^{1/2}(np)^{-2\varepsilon_n}\mathbb{E}Y_{\ell_1} - 400n^{3/2}p^{7/2+\varepsilon}/100$$
$$= \Omega(p^{1/2}(np)^{-\varepsilon_n}\mathbb{E}Y_{\ell_1}) - 2p^{1/2}(np)^{-2\varepsilon_n}\mathbb{E}Y_{\ell_1} - 400n^{3/2}p^{7/2+\varepsilon}/100 = \Omega(p^{1/2}(np)^{-\varepsilon_n}\mathbb{E}Y_{\ell_1})$$

larger than the number of neighbors with label ℓ, where the last equality uses that $\mathbb{E}Y_{\ell_1} \geq \frac{np}{2k}$ and so

$$p^{1/2}(np)^{-\varepsilon_n}\mathbb{E}Y_{\ell_1} \geq n^{3/2+o(1)}p^{7/2} \gg n^{3/2}p^{7/2+\varepsilon}/100.$$

In particular, a union bound over the complementary events for all vertices in Level 3 and all labels $\ell \in [k] \setminus \{\ell_1\}$ implies that a.a.s. after the third round all vertices in C have a surplus of $\Omega(p^{1/2}(np)^{-\varepsilon_n}\frac{np}{2k}) \gg Kp$ neighbors with label ℓ_1 compared to neighbors with any other label (with K defined as in Lemma 3.1). Since by Lemma 3.1 a.a.s. the labels of the vertices $V \setminus V([K])$ after the second round coincide in ALAP and LPA, and Chernoff's bound implies that a.a.s. every vertex has at most $O(Kp)$ neighbors among $V([K])$, a.a.s. all vertices in Level 3 receive label ℓ_1 after round 3 both in ALAP and in LPA. As there are more than $0.9n$ vertices in Level 3, the proof follows by Lemma 3.2. □

4 Proof of Lemma 2.6

In this proof, we work with an integer $M \in [0, n]$ and under the convention that, when a fraction has a positive numerator and denominator 0, it is equal to infinity. In this section, we present the delayed proof of Lemma 2.6. First of all, by rewriting the claimed inequality in the equivalent form

$$\frac{p_X(M, \rho)}{p_X(-1, \rho)} \geq \frac{p_Y(M, \rho)}{p_Y(-1, \rho)},$$

we observe that it is sufficient to show the lemma for $n' = n - 1$ (which we do assume): indeed, to deduce the inequality for smaller values of n', it is enough to apply induction.

We are going to show that, for every $M \geq 0$,

$$\frac{p_X(M, \rho)}{p_Y(M, \rho)} \geq \frac{p_X(M - 1, \rho)}{p_Y(M - 1, \rho)}. \tag{56}$$

Using that $p_X(M - 1, \rho) > p_X(M, \rho)$ and $p_Y(M - 1, \rho) > p_Y(M, \rho)$ for all integer $M \in [0, n - 1]$ and real $\rho \geq 0$, (56) is equivalent to

$$\frac{p_X(M, \rho)}{p_Y(M, \rho)} \geq \frac{p_X(M - 1, \rho) - p_X(M, \rho)}{p_Y(M - 1, \rho) - p_Y(M, \rho)}. \tag{57}$$

Note that the conclusion of the lemma is obtained by consecutive applications of (56); nevertheless, for technical reasons, we concentrate on showing (57) instead. To this end, we show the inequalities

$$\frac{p_X(n, \rho)}{p_Y(n, \rho)} \geq \frac{p_X(M - 1, \rho) - p_X(M, \rho)}{p_Y(M - 1, \rho) - p_Y(M, \rho)} \tag{58}$$

and

$$\forall i \in [n - M], \quad \frac{p_X(M + i - 1, \rho) - p_X(M + i, \rho)}{p_Y(M + i - 1, \rho) - p_Y(M + i, \rho)} \geq \frac{p_X(M + i - 2, \rho) - p_X(M + i - 1, \rho)}{p_Y(M + i - 2, \rho) - p_Y(M + i - 1, \rho)}. \tag{59}$$

Indeed, (58) and (59) imply that simultaneously

$$\frac{p_X(n, \rho)}{p_Y(n, \rho)} \geq \frac{p_X(M - 1, \rho) - p_X(M, \rho)}{p_Y(M - 1, \rho) - p_Y(M, \rho)},$$

$$\frac{p_X(n - 1, \rho) - p_X(n, \rho)}{p_Y(n - 1, \rho) - p_Y(n, \rho)} \geq \frac{p_X(M - 1, \rho) - p_X(M, \rho)}{p_Y(M - 1, \rho) - p_Y(M, \rho)},$$

$$\dots$$

$$\frac{p_X(M, \rho) - p_X(M + 1, \rho)}{p_Y(M, \rho) - p_Y(M + 1, \rho)} \geq \frac{p_X(M - 1, \rho) - p_X(M, \rho)}{p_Y(M - 1, \rho) - p_Y(M, \rho)},$$

and (57) follows by summing the numerators and the denominators of the fractions on the left-hand side.

Again, (58) is clear since $p_Y(n, \rho) = 0$ while all other terms are positive real numbers. Set $N = M + i - 1$. Then, the left-hand side of (59) rewrites as

$$\frac{\frac{1}{\rho+1}\mathbb{P}(X = N > Y) + \frac{1}{\rho+2}\mathbb{P}(X = N = Y) + (1 - \frac{1}{\rho+1})\mathbb{P}(X = N + 1 > Y) + (\frac{1}{2} - \frac{1}{\rho+2})\mathbb{P}(X = N + 1 = Y)}{\frac{1}{\rho+1}\mathbb{P}(Y = N > X) + \frac{1}{\rho+2}\mathbb{P}(Y = N = X) + (1 - \frac{1}{\rho+1})\mathbb{P}(Y = N + 1 > X) + (\frac{1}{2} - \frac{1}{\rho+2})\mathbb{P}(Y = N + 1 = X)},$$

while the right-hand side is given by

$$\frac{\frac{1}{\rho+1}\mathbb{P}(X = N - 1 > Y) + \frac{1}{\rho+2}\mathbb{P}(X = N - 1 = Y) + (1 - \frac{1}{\rho+1})\mathbb{P}(X = N > Y) + (\frac{1}{2} - \frac{1}{\rho+2})\mathbb{P}(X = N = Y)}{\frac{1}{\rho+1}\mathbb{P}(Y = N - 1 > X) + \frac{1}{\rho+2}\mathbb{P}(Y = N - 1 = X) + (1 - \frac{1}{\rho+1})\mathbb{P}(Y = N > X) + (\frac{1}{2} - \frac{1}{\rho+2})\mathbb{P}(Y = N = X)}.$$

We divide each of the numerator and the denominator of the above two displays by $(1 - p)^{2n-1}$. As a result, each of them can be seen as a polynomial of $x = \frac{p}{1-p}$. For concreteness, we assume that the first display rewrites as $P(x)/Q(x)$ while the second display rewrites as $R(x)/S(x)$ where

$$A(x) = A_1(x) + \rho A_2(x) \qquad \text{and} \qquad B(x) = B_1(x) + \rho B_2 r(x)$$

with

$$A_1(x) = 2\binom{n}{N}\binom{n-1}{N}x^{2N}, \qquad A_2(x) = \binom{n}{N+1}\binom{n-1}{N+1}x^{2N+2},$$

$$B_1(x) = 2\binom{n}{N-1}\binom{n-1}{N-1}x^{2N-2}, \quad B_2(x) = \binom{n}{N}\binom{n-1}{N}x^{2N},$$

38

and

$$P_1(x) = \binom{n}{N} \sum_{i=0}^{N-1} \binom{n-1}{i} x^{N+i}, \qquad P_2(x) = \binom{n}{N+1} \sum_{i=0}^{N} \binom{n-1}{i} x^{N+i+1},$$

$$Q_1(x) = \binom{n-1}{N} \sum_{i=0}^{N-1} \binom{n}{i} x^{N+i}, \qquad Q_2(x) = \binom{n-1}{N+1} \sum_{i=0}^{N} \binom{n}{i} x^{N+i+1},$$

$$R_1(x) = \binom{n}{N-1} \sum_{i=0}^{N-2} \binom{n-1}{i} x^{N+i-1}, \quad R_2(x) = \binom{n}{N} \sum_{i=0}^{N-1} \binom{n-1}{i} x^{N+i},$$

$$S_1(x) = \binom{n-1}{N-1} \sum_{i=0}^{N-2} \binom{n}{i} x^{N+i-1}, \qquad S_2(x) = \binom{n-1}{N} \sum_{i=0}^{N-1} \binom{n}{i} x^{N+i},$$

and finally,

$$P(x) = \frac{P_1(x)}{\rho+1} + \frac{\rho P_2(x)}{\rho+1} + \frac{A(x)}{2\rho+4}, \quad Q(x) = \frac{Q_1(x)}{\rho+1} + \frac{\rho Q_2(x)}{\rho+1} + \frac{A(x)}{2\rho+4},$$

$$R(x) = \frac{R_1(x)}{\rho+1} + \frac{\rho R_2(x)}{\rho+1} + \frac{B(x)}{2\rho+4}, \quad S(x) = \frac{S_1(x)}{\rho+1} + \frac{\rho S_2(x)}{\rho+1} + \frac{B(x)}{2\rho+4}.$$

Note that actually $P_1(x) = R_2(x)$ and $Q_1(x) = S_2(x)$, but we nevertheless keep both notations for simplicity of the exposition. In the sequel, we abbreviate notation and write $P_1, Q_1, \ldots$ instead of $P_1(x), Q_1(x), \ldots$, leaving the dependence on x implicit for better readability.

To derive (59), we show that, for all $x, \rho \geq 0$, $PS \geq QR$ or, equivalently, that

$$(\rho+1)^2(2\rho+4)PS \geq (\rho+1)^2(2\rho+4)QR. \tag{60}$$

Observe that

$$
\begin{aligned}
&(\rho+1)^2(2\rho+4)PS \\
&= P_1\big[(2\rho+4)S_1 + \rho(2\rho+4)S_2 + (\rho+1)B_1 + \rho(\rho+1)B_2\big] \\
&\quad + P_2\big[\rho(2\rho+4)S_1 + \rho^2(2\rho+4)S_2 + \rho(\rho+1)B_1 + \rho^2(\rho+1)B_2\big] \\
&\quad + (A_1 + \rho A_2)\big[(\rho+1)S_1 + \rho(\rho+1)S_2\big] + (\rho+1)^2 AB/(2\rho+4) \\
&= \rho^3 \cdot \big[2P_2S_2 + P_2B_2 + S_2A_2\big] \\
&\quad + \rho^2 \cdot \big[2P_1S_2 + P_1B_2 + 2P_2S_1 + 4P_2S_2 + P_2B_1 + P_2B_2 + S_1A_2 + S_2A_1 + S_2A_2\big] \\
&\quad + \rho \cdot \big[2P_1S_1 + 4P_1S_2 + P_1B_1 + P_1B_2 + 4P_2S_1 + P_2B_1 + S_1A_1 + S_1A_2 + S_2A_1\big] \\
&\quad + 1 \cdot \big[4P_1S_1 + P_1B_1 + S_1A_1\big] + (\rho+1)^2 AB/(2\rho+4),
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&(\rho+1)^2(2\rho+4)QR \\
&= \rho^3 \cdot \big[2Q_2R_2 + Q_2B_2 + R_2A_2\big] \\
&\quad + \rho^2 \cdot \big[2Q_1R_2 + Q_1B_2 + 2Q_2R_1 + 4Q_2R_2 + Q_2B_1 + Q_2B_2 + R_1A_2 + R_2A_1 + R_2A_2\big] \\
&\quad + \rho \cdot \big[2Q_1R_1 + 4Q_1R_2 + Q_1B_1 + Q_1B_2 + 4Q_2R_1 + Q_2B_1 + R_1A_1 + R_1A_2 + R_2A_1\big] \\
&\quad + 1 \cdot \big[4Q_1R_1 + Q_1B_1 + R_1A_1\big] + (\rho+1)^2 AB/(2\rho+4),
\end{aligned}
$$

The remainder of this proof is dedicated to showing the following simpler auxiliary inequalities for every $x \geq 0$:

I1 $P_1 S_1 \geq Q_1 R_1$,

I2 $P_2 S_2 \geq Q_2 R_2$,

I3 $P_1 S_2 + P_2 S_1 \geq Q_1 R_2 + Q_2 R_1$,

and use them to derive that, for every $x \geq 0$, each of the following inequalities holds as well:

J1 $2P_2 S_2 + P_2 B_2 + S_2 A_2 \geq 2Q_2 R_2 + Q_2 B_2 + R_2 A_2$,

J2 $2P_1 S_2 + P_1 B_2 + 2P_2 S_1 + 4P_2 S_2 + P_2 B_1 + P_2 B_2 + S_1 A_2 + S_2 A_1 + S_2 A_2$
$$\geq 2Q_1 R_2 + Q_1 B_2 + 2Q_2 R_1 + 4Q_2 R_2 + Q_2 B_1 + Q_2 B_2 + R_1 A_2 + R_2 A_1 + R_2 A_2,$$

J3 $2P_1 S_1 + 4P_1 S_2 + P_1 B_1 + P_1 B_2 + 4P_2 S_1 + P_2 B_1 + S_1 A_1 + S_1 A_2 + S_2 A_1$
$$\geq 2Q_1 R_1 + 4Q_1 R_2 + Q_1 B_1 + Q_1 B_2 + 4Q_2 R_1 + Q_2 B_1 + R_1 A_1 + R_1 A_2 + R_2 A_1,$$

J4 $4P_1 S_1 + P_1 B_1 + S_1 A_1 \geq 4Q_1 R_1 + Q_1 B_1 + R_1 A_1$.

Showing I1. We will show that all coefficients of the polynomial $P_1 S_1 - Q_1 R_1$ are positive. First, we consider the terms x^{2N-1+k} with $k \in [0, N-2]$. For any such k, the coefficient in front of x^{2N-1+k} rewrites

$$\sum_{i,j \geq 0,\, i+j=k} \left(\binom{n}{N} \binom{n-1}{i} \binom{n-1}{N-1} \binom{n}{j} - \binom{n-1}{N} \binom{n}{i} \binom{n}{N-1} \binom{n-1}{j} \right). \tag{61}$$

Then, by the inequality $\binom{n}{N}\binom{n-1}{N-1} \geq \binom{n-1}{N}\binom{n}{N-1}$ and the symmetry of the sum with respect to i, j, the positivity of the last expression follows. In order to prepare the terrain for the case $k \in [N-1, 2N-3]$, we further transform (61). Using that, for every $\ell \geq 0$, $\binom{n-1}{\ell} = \frac{n-\ell}{n}\binom{n}{\ell}$, we obtain that the above coefficient rewrites

$$\sum_{i,j \geq 0,\, i+j=k} \binom{n}{N}\binom{n}{i}\binom{n}{N-1}\binom{n}{j} \frac{(n-i)(n-N+1) - (n-j)(n-N)}{n^2},$$

which, thanks to the inherent symmetry between i and j above, rewrites

$$\frac{1}{2}\sum_{i,j \geq 0,\, i+j=k} \binom{n}{N}\binom{n}{i}\binom{n}{N-1}\binom{n}{j} \frac{(n-i+n-j)((n-N+1)-(n-N))}{n^2},$$

or equivalently

$$\frac{1}{2}\sum_{i,j \geq 0,\, i+j=k} \binom{n}{N}\binom{n}{i}\binom{n}{N-1}\binom{n}{j} \frac{2n-k}{n^2} > 0. \tag{62}$$

Next, we turn to the coefficients in front of the terms x^{2N-1+k} for $k \in [N-1, 2N-3]$ where i could be equal to $N-1$ but j cannot, hence the loss of the symmetry used above. Nevertheless, using that the term corresponding to $i = N-1, j = k-(N-1)$ is the only term without symmetric counterpart, we obtain that the coefficient in front of x^{2N-1+k} rewrites as a sum of two terms. Following (62), the first term contains the 'symmetric part' of the sum, which is equal to

$$\frac{1}{2}\sum_{i,j \leq N-2,\, i+j=k} \binom{n}{N}\binom{n}{i}\binom{n}{N-1}\binom{n}{j} \frac{2n-k}{n^2},$$

and further rewrites as

$$\binom{n}{N}\binom{n}{N-1}\binom{n}{N-1}\binom{n}{k-N+1} \sum_{i,j \leq N-2,\, i+j=k} \frac{\binom{n}{i}\binom{n}{j}}{\binom{n}{N-1}\binom{n}{k-N+1}} \frac{n-k/2}{n^2}. \tag{63}$$

40

The second term is

$$\binom{n}{N}\binom{n-1}{N-1}^2\binom{n}{k-N+1} - \binom{n-1}{N}\binom{n}{N-1}^2\binom{n-1}{k-N+1},$$

which further rewrites as

$$\binom{n}{N}\binom{n}{N-1}^2\binom{n}{k-N+1}\frac{(n-N+1)^2 - (n-N)(n+N-k-1)}{n^2}$$
$$= \binom{n}{N}\binom{n}{N-1}^2\binom{n}{k-N+1}\frac{1-(n-N)(2N-k-3)}{n^2}. \tag{64}$$

Finally, using that minimizing the expression $\binom{n}{i}\binom{n}{j}$ under the constraints $i, j \leq N-2$ and $i+j = k$ requires $|i-j|$ to be largest possible (that is, $i = N-2$ and $j = k-N+2$ or vice versa), we derive that $\binom{n}{i}\binom{n}{j} \geq \binom{n}{N-2}\binom{n}{k-N+2} \geq \binom{n}{N-1}\binom{n}{k-N+1}$. This yields the inequality

$$\sum_{i,j \leq N-2,\, i+j=k} \frac{\binom{n}{i}\binom{n}{j}}{\binom{n}{N-1}\binom{n}{k-N+1}}\frac{n-k/2}{n^2} \geq \frac{(2N-k-3)(n-k/2)}{n^2} \geq \frac{(2N-k-3)(n-N)}{n^2},$$

which shows that the sum of (63) and (64) is positive and finishes the proof of **I1**.

Showing I2. The proof of this second step follows from the proof of the first step by replacing N with $N+1$ in all places.

Showing I3. First, since $P_1 = R_2$ and $Q_1 = S_2$, we have that $P_1 S_2 = Q_1 R_2$. Thus, it remains to study the difference $P_2 S_1 - Q_2 R_1$. As for **I1**, we begin by studying the terms x^{2N+k} with $k \in [0, N-2]$. For any such k, the coefficient in front of x^{2N+k} rewrites as

$$\sum_{i,j \geq 0,\, i+j=k} \left(\binom{n}{N+1}\binom{n-1}{i}\binom{n-1}{N-1}\binom{n}{j} - \binom{n-1}{N+1}\binom{n}{i}\binom{n}{N-1}\binom{n-1}{j} \right).$$

Again, the positivity of the last sum follows from the symmetry of the sum with respect to i, j and the inequality $\binom{n}{N+1}\binom{n-1}{N-1} \geq \binom{n-1}{N+1}\binom{n}{N-1}$. Moreover, using that, for every $\ell \geq 0$, $\binom{n-1}{\ell} = \frac{n-\ell}{n}\binom{n}{\ell}$, we obtain that the above coefficient rewrites

$$\sum_{i,j \geq 0,\, i+j=k} \binom{n}{N+1}\binom{n}{i}\binom{n}{N-1}\binom{n}{j}\frac{(n-i)(n-N+1) - (n-j)(n-N-1)}{n^2},$$

which, thanks to the inherent symmetry between i and j above, rewrites

$$\frac{1}{2}\sum_{i,j \geq 0,\, i+j=k} \binom{n}{N+1}\binom{n}{i}\binom{n}{N-1}\binom{n}{j}\frac{(n-i+n-j)((n-N+1)-(n-N-1))}{n^2},$$

or equivalently

$$\sum_{i,j \geq 0,\, i+j=k} \binom{n}{N+1}\binom{n}{i}\binom{n}{N-1}\binom{n}{j}\frac{2n-k}{n^2} > 0.$$

Next, we turn to the coefficients in front of the terms x^{2N+k} for $k \in [N-1, 2N-2]$ where i could be equal to $N-1$ or N but j cannot, hence the loss of the symmetry used above. We consider three subcases.

Case 3.1: $k = N - 1$. Then, the term corresponding to $i = N - 1$, $j = 0$ is the only term without symmetric counterpart. Hence, the coefficient in front of x^{3N-1} rewrites as the sum of two terms. By the latter computations, the (symmetric) first term is given by

$$\sum_{i,j \geq 1,\, i+j=N-1} \binom{n}{N+1}\binom{n}{i}\binom{n}{N-1}\binom{n}{j}\frac{2n-k}{n^2}, \tag{65}$$

while the second term is

$$\binom{n}{N+1}\binom{n-1}{N-1}^2 - \binom{n-1}{N+1}\binom{n}{N-1}^2,$$

which further rewrites as

$$\binom{n}{N+1}\binom{n}{N-1}^2\left(\frac{(n-N+1)^2}{n^2} - \frac{(n-N-1)n}{n^2}\right)$$
$$= \binom{n}{N+1}\binom{n}{N-1}^2\frac{4-(n-N-1)(N-3)}{n^2}. \tag{66}$$

Finally, using that $\binom{n}{i}\binom{n}{j} \geq \binom{n}{N-1}$ for all $i,j \geq 1$ with $i+j = N-1$ together with the fact that

$$\sum_{i,j \geq 1,\, i+j=N-1} \frac{2n-k}{n^2} = \frac{(N-2)(2n-N+1)}{n^2} \geq \frac{(N-3)(n-N-1)}{n^2},$$

showing that the sum of (65) and (66) is positive and finishing the proof in this case.

Case 3.2: $k \in [N, 2N-3]$. In this case, there are two terms without symmetric counterparts, namely, the ones corresponding to $i = N-1, j = k-N+1$ and $i = N, j = k-N$. The coefficient in front of x^{2N+k} coming from the choice of $i = N-1, j = k-N+1$ is

$$\binom{n}{N+1}\binom{n-1}{N-1}^2\binom{n}{k-N+1} - \binom{n-1}{N+1}\binom{n}{N-1}^2\binom{n-1}{k-N+1},$$

which rewrites as

$$\binom{n}{N+1}\binom{n}{N-1}^2\binom{n}{k-N+1}\left(\frac{(n-N+1)^2}{n^2} - \frac{(n-N-1)(n-k+N-1)}{n^2}\right)$$
$$= \binom{n}{N+1}\binom{n}{N-1}^2\binom{n}{k-N+1}\frac{4-(n-N-1)(2N-4-k)}{n^2}. \tag{67}$$

The coefficient in front of x^{2N+k} coming from the choice of $i = N, j = k-N$ is

$$\binom{n}{N+1}\binom{n-1}{N}\binom{n-1}{N-1}\binom{n}{k-N} - \binom{n-1}{N+1}\binom{n}{N}\binom{n}{N-1}\binom{n-1}{k-N},$$

which rewrites as

$$\binom{n}{N+1}\binom{n}{N}\binom{n}{N-1}\binom{n}{k-N}\left(\frac{(n-N)(n-N+1)}{n^2} - \frac{(n-N-1)(n-k+N)}{n^2}\right)$$
$$= \binom{n}{N+1}\binom{n}{N}\binom{n}{N-1}\binom{n}{k-N}\frac{2-(n-N-1)(2N-2-k)}{n^2}. \tag{68}$$

We now consider two subcases. If $k = 2N-3$, (67) is positive and the inequality $\binom{n}{N-1}\binom{n}{N-2} \geq \binom{n}{N}\binom{n}{N-3}$ ensures that the sum of (67) and (68) remains positive.

42

Otherwise, suppose that $k \in [N, 2N-4]$. Again, minimizing the expression $\binom{n}{i}\binom{n}{j}$ under the constraints $i, j \leq N-2$ and $i+j = k$ requires $|i-j|$ to be largest possible (that is, $i = N-2$ and $j = k-N+2$ or vice versa). As a result, $\binom{n}{i}\binom{n}{j} \geq \binom{n}{N-2}\binom{n}{k-N+2} \geq \binom{n}{N-1}\binom{n}{k-N+1} \geq \binom{n}{N}\binom{n}{k-N}$ for all $i, j \leq N-2$ with $i+j = k$. Combining this observation with the inequality

$$\sum_{i,j \leq N-2, i+j=k} \frac{2n-k}{n^2} = \frac{(2N-k-3)(2n-k)}{n^2}$$

$$\geq \frac{(2N-k-3)(2n-(2N-2))}{n^2}$$

$$\geq \frac{((2N-2-k)+(2N-4-k))(n-N-1)}{n^2}, \tag{69}$$

ensures that the sum of the symmetric term (for $i, j \leq N-2$ with $i+j=k$), (67) and (68) is positive, and finishes the proof in this case.

Case 3.3: $k = 2N - 2$. Then, the coefficient in front of x^{2N+k} in $P_2 S_1 - Q_2 R_1$ is given by (68) and is positive.

Before we continue with the proof of **J1**–**J4**, we simplify the expressions $P_1 - Q_1 = R_2 - S_2$, $P_2 - Q_2$ and $R_1 - S_1$. First, using that, for every $\ell \geq 0$, $\binom{n-1}{\ell} = \frac{n-\ell}{n}\binom{n}{\ell}$, we obtain that

$$P_1 - Q_1 = R_2 - S_2 = \sum_{i=0}^{N-1}\left(\binom{n}{N}\binom{n-1}{i} - \binom{n-1}{N}\binom{n}{i}\right)x^{N+i} = \sum_{i=0}^{N-1}\binom{n}{N}\binom{n}{i}\frac{N-i}{n}x^{N+i}. \tag{70}$$

Furthermore, since P_2 (respectively, Q_2) is obtained from P_1 (respectively, Q_1) by replacing N by $N+1$, we have

$$P_2 - Q_2 = \sum_{i=0}^{N}\binom{n}{N+1}\binom{n}{i}\frac{N+1-i}{n}x^{N+i+1}. \tag{71}$$

Simplifying $R_1 - S_1$ follows a similar computation:

$$R_1 - S_1 = \sum_{i=0}^{N-2}\left(\binom{n}{N-1}\binom{n-1}{i} - \binom{n-1}{N-1}\binom{n}{i}\right)x^{N+i-1} = \sum_{i=0}^{N-2}\binom{n}{N-1}\binom{n}{i}\frac{N-1-i}{n}x^{N+i-1}. \tag{72}$$

Showing J1. By **I2**, we already know that $P_2 S_2 \geq Q_2 R_2$, so it remains to show that

$$(P_2 - Q_2)B_2 - (R_2 - S_2)A_2 \geq 0. \tag{73}$$

By (70) and (71), the desired inequality is equivalent to

$$\sum_{i=0}^{N}\binom{n}{N+1}\binom{n}{i}\binom{n}{N}\binom{n-1}{N}\frac{N+1-i}{n}x^{3N+i+1} - \sum_{i=0}^{N-1}\binom{n}{N}\binom{n}{i}\binom{n}{N+1}\binom{n-1}{N+1}\frac{N-i}{n}x^{3N+i+2} \geq 0.$$

By ignoring the summand for $i = 0$ in the first sum above and changing the variable from $i \in [0, N-1]$ to $i - 1 \in [1, N]$ in the second sum, the last display is implied by the inequality

$$\sum_{i=1}^{N}\left(\binom{n}{N+1}\binom{n}{i}\binom{n}{N}\binom{n-1}{N} - \binom{n}{N}\binom{n}{i-1}\binom{n}{N+1}\binom{n-1}{N+1}\right)\frac{N+1-i}{n}x^{3N+i+1} \geq 0,$$

which is implied by the fact that, for every $i \in [N]$,

$$\frac{\binom{n}{i}}{\binom{n}{i-1}} \cdot \frac{\binom{n-1}{N}}{\binom{n-1}{N+1}} = \frac{n-i+1}{i} \cdot \frac{N+1}{n-1-N} \geq \frac{n-N+1}{N} \cdot \frac{N+1}{n-1-N} > 1.$$

43

Showing J2. By **I2** and **I3**, we know respectively that $P_2 S_2 \geq Q_2 R_2$ and $P_1 S_2 + P_2 S_1 \geq Q_1 R_2 + Q_2 R_1$. Thus, in order to prove **J3**, it suffices to show that

$$(P_1 - Q_1) B_2 + (P_2 - Q_2)(B_1 + B_2) - (R_1 - S_1) A_2 - (R_2 - S_2)(A_1 + A_2) \geq 0.$$

Using (73) and recalling that $P_1 - Q_1 = R_2 - S_2$, it is enough to show that

$$(P_1 - Q_1)(B_2 - A_1) + (P_2 - Q_2) B_1 - (R_1 - S_1) A_2 \geq 0. \tag{74}$$

To prove (74), observe that combining the fact that $A_1 = 2B_2$, (70) and a change of variable ensures that

$$(P_1 - Q_1)(B_2 - A_1) = -(P_1 - Q_1) B_2 = - \sum_{i=0}^{N-1} \binom{n}{N} \binom{n}{i} \frac{N-i}{n} \binom{n}{N} \binom{n-1}{N} x^{3N+i}$$

$$= - \sum_{j=1}^{N} \binom{n}{N} \binom{n}{j-1} \frac{N+1-j}{n} \binom{n}{N} \binom{n-1}{N} x^{3N+j-1}. \tag{75}$$

Furthermore, (71) and (72) ensure respectively that

$$(P_2 - Q_2) B_1 = 2 \sum_{i=0}^{N} \binom{n}{N+1} \binom{n}{i} \frac{N+1-i}{n} \binom{n}{N-1} \binom{n-1}{N-1} x^{3N+i-1}, \tag{76}$$

$$(R_1 - S_1) A_2 = \sum_{i=0}^{N-2} \binom{n}{N-1} \binom{n}{i} \frac{N-1-i}{n} \binom{n}{N+1} \binom{n-1}{N+1} x^{3N+i+1}$$

$$= \sum_{j=2}^{N} \binom{n}{N-1} \binom{n}{j-2} \frac{N+1-j}{n} \binom{n}{N+1} \binom{n-1}{N+1} x^{3N+j-1}. \tag{77}$$

Henceforth, we adopt the convention that $\binom{b}{a} = 0$ if $a < 0$. Then, by combining (75), (76) and (77), the coefficient in front of x^{3N+i-1} in (74) is non-negative if and only if

$$2 \binom{n}{N+1} \binom{n}{i} \binom{n}{N-1} \binom{n-1}{N-1} - \binom{n}{N-1} \binom{n}{i-2} \binom{n}{N+1} \binom{n-1}{N+1} - \binom{n}{N} \binom{n}{i-1} \binom{n}{N} \binom{n-1}{N} \geq 0. \tag{78}$$

For $i = 0$, the inequality is obvious. For $i \in [N]$, observe that

$$\frac{\binom{n}{i}}{\binom{n}{i-1}} \cdot \frac{\binom{n}{N-1}}{\binom{n}{N}} \cdot \frac{\binom{n}{N+1} \binom{n-1}{N-1}}{\binom{n}{N} \binom{n-1}{N}} = \frac{n-i+1}{i} \cdot \frac{N}{n-N+1} \cdot \frac{N}{N+1} \geq \frac{N}{N+1},$$

and

$$\frac{\binom{n}{N-1}}{\binom{n}{N}} \cdot \frac{\binom{n}{i-2}}{\binom{n}{i-1}} \cdot \frac{\binom{n}{N+1}}{\binom{n}{N}} \cdot \frac{\binom{n-1}{N+1}}{\binom{n-1}{N}} = \frac{N}{n-N+1} \cdot \frac{i-1}{n-i+2} \cdot \frac{n-N}{N+1} \cdot \frac{n-N-1}{N+1} \leq \frac{i-1}{N+1} \leq \frac{N-1}{N+1}.$$

By combining the last two inequalities, it follows that the right hand side of (78) dominates

$$\left(\frac{2N}{N+1} - \frac{N-1}{N+1} - 1 \right) \binom{n}{N} \binom{n}{i-1} \binom{n}{N} \binom{n-1}{N} = 0,$$

which finishes the proof of **J2**.

Showing J3. By **I1** and **I3**, we know respectively that $P_1 S_1 \geq Q_1 R_1$ and $P_1 S_2 + P_2 S_1 \geq Q_1 R_2 + Q_2 R_1$. Thus, in order to prove **J3**, it suffices to show that

$$(P_1 - Q_1)(B_1 + B_2) + (P_2 - Q_2)B_1 - (R_1 - S_1)(A_1 + A_2) - (R_2 - S_2)A_1 \geq 0.$$

By recalling that $P_1 - Q_1 = R_2 - S_2$ and using (74), it is enough to show that

$$(P_1 - Q_1)B_1 - (R_1 - S_1)A_1 \geq 0. \tag{79}$$

However, this inequality follows similarly to (73): indeed, note that $P_2, Q_2, R_2, S_2, A_2, B_2$ are obtained respectively from $P_1, Q_1, R_1, S_1, \frac{1}{2}A_1, \frac{1}{2}B_1$ by replacing N with $N + 1$.

Showing J4. This inequality follows by combining **I1** and (79). This concludes the proof of Lemma 2.6.

5 Concluding remarks

The focus of the current paper was the rigorous analysis of a variant of LPA on the binomial random graph $\mathcal{G}(n, p)$. We showed that as long as $np \geq n^{5/8+\varepsilon}$, a.a.s. a unique label survives after 5 iterations of the algorithm. The proof distinguished two regimes that required the use of different techniques. In the regime $np = \Omega(n^{2/3})$, the fact that the sizes of the basins were typically at distance at least $\mathbb{V}(\mathfrak{B}_1(\ell)) = \Omega(n^{1/3})$ from each other was crucial. In this case, the surviving label was among the $O(1)$ initial ones (and if $np \gg n^{2/3}$, it is the first one). For smaller values of p, a finer understanding of the gap between the largest and the second largest basin was needed. In this case, a closer look at the proof shows that the surviving label is distributed over a range of $\Theta((np^3 \log(1/(np^3)))^{-1/2})$ initial labels.

We finish with several further comments:

1. The last part of the proof of the second point of Theorem 1.1 is the bottleneck of our argument when $n^{5/8} \ll np \ll n^{2/3}$. In particular, this is the place where the exponent $5/8$ appears. Nevertheless, one may improve this constant by reusing the idea from Lemmas 3.15 and 3.31 ensuring the lower bound on $\mathfrak{B}_2(1)$ and $\mathfrak{B}_2(\ell_1)$, respectively. More precisely, one may similarly define a set of labels $[k_1]$ such that no vertex in C carries label in $[k] \setminus [k_1]$ after the third round. Then, partition Level 3 into two sets: $C_2([k_1])$ and $S = C \setminus C_2([k_1])$. By designing a suitable alternative procedure exposing only edges in C incident to $C_2([k_1])$, we find a label $\ell_2 \in [k_1]$ (most likely different from ℓ_1) that appears most often. As will turn out, $\mathfrak{C}_2([k_1]) \gg \mathfrak{B}$ by the choice of k_1, so the difference between $\mathfrak{C}_3(\ell_2)$ and $\mathfrak{C}_3(\ell)$ (for $\ell \in [k_1] \setminus \ell_2$) will grow larger compared to $\mathfrak{C}_2(\ell_1) - \mathfrak{C}_2(\ell)$. Thus, for suitably large p, we may similarly show that after 4 rounds, label ℓ_2 is carried by $n - o(n)$ vertices in S. In fact, this argument can also be bootstrapped: if the differences in size between $S(\ell_2)$ and $S(\ell)$ are still small, one may look for an integer k_2 such that the largest set after the fourth round has label in $[k_2]$. In that case, partition S into $S([k_2])$ and its complement, and explore the edges incident to $S([k_2])$ before exploring the rest. As the formal proof of this additional step would increase the technicality of the paper without contributing new ideas, we omit the details. It is not clear (to us) how much the lower bound on np could be improved this way; at some point, we expect other bottlenecks to appear as well.

2. As mentioned in the introduction, empirical evidence reported in [19, 25] suggests that the behavior of the label propagation algorithm on $\mathcal{G}(n, p)$ exhibits a threshold behavior around $np = n^{1/5}$. The same article [19] shows that there is an $\epsilon > 0$ such that when $\log(n) \ll np \leq n^{\epsilon}$ the algorithm terminates with $\Omega((np)^3)$ label classes, each of size $O(n/(np)^3)$. We hope that the insights on which our contribution relies might also help estimating the range of values of ϵ for which the claim still holds.

3. We showed that when $np = cn^{2/3}$, the a.a.s. unique label that survives after 5 rounds is a tight random variable. In fact, with a little bit of extra work, one could show that this label is distributed as the index of the maximum of $(N_i - c(i-1))_{i \geq 1}$ where $(N_i)_{i \geq 1}$ is a sequence of i.i.d. normal variables of expectation 0 and variance 1.

References

[1] Barber, M. J. and Clark, J. W.: Detecting network communities by propagating labels under constraints, *Physical Review E*, **80 (2)**, (2009), 026129.

[2] Bedi, P. and Sharma, C.: Community detection in social networks, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **6 (3)**, (2016), 115–135.

[3] Benjamini, I. and Chan, S. and O'Donnell, R. and Tamuz, O. and Tan L-Y.: Convergence, unanimity and disagreement in majority dynamics on unimodular graphs and random graphs, *Stochastic Processes and their Applications*, **126 (9)**, (2016), 2719–2733.

[4] Bollobás, B.: Random graphs, *Cambridge University Press*, 2001.

[5] Clifford, P. and Sudbury, A.: A model for spatial conflict, *Biometrika*, **60 (3)**, (1973), 581–588.

[6] Chellig, J. and Fountoulakis, N.: Majority dynamics on random graphs: the multiple states case, *arxiv.org/abs/2311.09078*, (2023).

[7] Cruciani, E. and Natale, E. and Scornavacca, G.: On the metastability of quadratic majority dynamics on clustered graphs and its biological implications, *Bulletin of EATCS*, **2 (125)**, 2018

[8] Cordasco, G. and Gargano, L.: Label propagation algorithm: a semi-synchronous approach, *International Journal of Social Network Mining*, **1 (1)**, (2012), 3–26.

[9] DeGroot, M. H.: Reaching a consensus, *Journal of the American Statistical Association*, **69 (345)**, (1974), 118–121.

[10] Esseen, C.-G.: On the Liapunoff limit of error in the theory of probability, *Arkiv för Matematik, Astronomi och Fysik*, **A28**, (1942), 1–19.

[11] Fountoulakis, N. and Kang, M. and Makai, T.: Resolution of a conjecture on majority dynamics: Rapid stabilization in dense random graphs, *Random Structures & Algorithms*, **57 (4)**, (2020), 1134–1156.

[12] Goles, E. and Olivos, J.: Periodic behaviour of generalized threshold functions, *Discrete Mathematics*, **30 (2)**, (1980), 187–189.

[13] Gregory, S.: Finding overlapping communities in networks by label propagation, *New journal of Physics*, **12 (10)**, (2010), 103018.

[14] Harenberg, S. and Bello, G. and Gjeltema, L. and Ranshous, S. and Harlalka, J. and Seay, R. and Padmanabhan, K. and Samatova, N.: Community detection in large-scale networks: a survey and empirical evaluation, *Wiley Interdisciplinary Reviews: Computational Statistics*, **6(6)**, (2014), 426–439.

[15] Holley, R. A. and Liggett, T. M.: Ergodic theorems for weakly interacting infinite systems and the voter model, *Annals of Probability*, (1975), 643–663.

[16] Janson, S. and Łuczak, T. and Ruciński, A.: Random graphs, *John Wiley & Sons*, 2011.

[17] Kamiński, B. and Prałat, P. and Théberge, F.: Mining Complex Networks, *Chapman and Hall/CRC*, 2021.

[18] Karoński, M. and Frieze, A.: Introduction to Random Graphs, *Cambridge University Press*, 2016.

[19] Knierim, C. and Lengler, J. and Pfister, P. and Schaller, U. and Steger, A.: The Maximum Label Propagation Algorithm on Sparse Random Graphs, *APPROX-RANDOM, Leibniz International Proceedings in Informatics*, **58**, (2019), 1–15.

[20] Kothapalli, K. and Pemmaraju, S. V. and Sardeshmukh, V.: On the analysis of a label propagation algorithm for community detection, *International Conference on Distributed Computing and Networking, Springer*, (2013), 255–269.

[21] Leung, I. X. Y. and Hui, P. and Lio, P. and Crowcroft, J.: Towards real-time community detection in large networks, *Physical Review E*, **79 (6)**, 066107, 2009.

[22] Mossel, E. and Neeman, J. and Tamuz, O.: Majority dynamics and aggregation of information in social networks, *Autonomous Agents and Multi-Agent Systems*, **28 (3)**, (2014), 408–429.

[23] Mossel, E. and Tamuz, O.: Opinion exchange dynamics, *Probability Surveys*, **14**, (2017), 155–204.

[24] Newman, M.: Networks, Oxford University Press, 2018.

[25] Pfister, P.: Processes on Random Graphs and other Random Processes, *PhD Thesis, ETH Zürich*, 2020.

[26] Raghavan, U. N. and Albert, R. and Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks, *Physical review E*, **76 (3)**, (2007), 036106.

[27] Rosling, K.: Inventory cost rate functions with nonlinear shortage costs, *Operations Research*, **50 (6)**, (2002), 1007–1017.

[28] Sah, A. and Sawhney, M.: Majority Dynamics: The Power of One, *Israel Journal of Mathematics*, (2024), 1–49.

[29] Slud, E. V.: Distribution inequalities for the binomial law, *The Annals of Probability*, **5 (3)**, (1977), 404–412.

[30] Tamir, R.: Fast convergence to unanimity in dense Erdős-Rényi graphs, *arXiv.org/abs/2210.05992*, 2022.

[31] Tran, L. and Vu, V.: Reaching a Consensus on Random Networks: The Power of Few, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM*, **176**, (2020), 20:1–20:15.

[32] Wormald, N. C.: Models of random regular graphs, *London Mathematical Society Lecture Note Series, Cambridge University Press*, (1999), 239–298.

[33] Yang, Z. and Algesheimer, R. and Tessone, C. J.: A comparative analysis of community detection algorithms on artificial networks, *Scientific reports*, **6 (1)**, (2016), 1–18.

[34] Zehmakan, A. N.: Opinion forming in Erdős–Rényi random graph and expanders, *Discrete Applied Mathematics*, **277**, (2020), 280–290.