Label propagation on binomial random graphs

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

[1]Univ. Chile, Santiago, Chile
[2]Univ. Jean Monnet, Saint-Etienne, France
[3]IMC, Pontifícia Univ. Católica, Chile
[4]Toronto Metropolitan University, Toronto, Canada

June 15, 2023

Abstract

We study a variant of the widely popular, fast and often used family of community detection procedures referred to as label propagation algorithms. These mechanisms also exhibit many parallels with models of opinion exchange dynamics and consensus mechanisms in distributed computing.

Initially, given a network, each vertex starts with a random label in the interval $[0,1]$. Then, in each round of the algorithm, 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.

We investigate the performance of this algorithm on the binomial random graph $\mathcal{G}(n,p)$. We show that for $np \geq n^{5/8+\varepsilon}$, the algorithm terminates with a single label a.a.s. (which was previously known only for $np \geq n^{3/4+\varepsilon}$). 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.

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

1 Introduction

In this paper, we consider a class of popular unsupervised learning algorithms for finding communities in complex networks called *label propagation algorithms*. In the specific instance of the algorithm 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 $i \in [n] = \{1,\dots,n\}$, 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 at the beginning, every label is represented only once.) The algorithm ends once the process converges (that is, once no more

*Marcos Kiwi has been partially supported by grant GrHyDy ANR-20-CE40-0002 and BASAL funds for centers of excellence from ANID-Chile (FB210005).

†Dieter Mitsche has been partially supported by grant GrHyDy ANR-20-CE40-0002 and by Fondecyt grant 1220174.

‡Paweł Prałat has been partially supported by NSERC Discovery Grant. Part of this work was done while the author was visiting the Simons Institute for the Theory of Computing.

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 in to detect communities [12, 24]; for more background, see the surveys [2, 13, 30] or any book on mining complex networks such as [16] or [22]. Despite their popularity and the fact that their theoretical analyses were identified as an important research question [1, 6, 20], 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 [21]. Amid the most famous and mathematically interesting models are the deGroot model (see [8] 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, 14]); 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 [11]). Label propagation algorithms are a special case of non-binary majority dynamics.

In [19], 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 [4, 15, 17]. 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 [19] 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.) Our argument below works for a wider range of parameters of np. 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 [18] (see [23] for slightly more details) where the authors showed that there is 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 [18, 23] 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 [10] (see also [26, 29, 31]).

As it is already implicit in [19], 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 notions 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 focus our study in 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.) it is rigorously shown, for a specific variant of label propagation on $\mathcal{G}(n,p)$, that a single label survives. 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 [19].

Interestingly, in the influential paper introducing label propagation as a procedure for community detection, Raghavan, Albert and Kumara [24] 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

[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 [24]. 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.

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 $k = \lceil 15p^{-2}(n^{-1}\log n)^{1/2}\rceil$ 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 $v_1, \dots, v_{2k}$ that initially carry labels $1, \dots, 2k$, 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 $v_1, \dots, v_{\ell-1}$.

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 $\ell \in [2k]\backslash\{\ell_1\}$, we first count the number of vertices in $B\backslash(B_1(\ell)\cup B_1(\ell_1))$ 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 $B_1(\ell)$ 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 [28]). 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.

We propose several couplings that facilitate dealing with intrinsic dependencies inherent to the analysis of label propagation variants. For instance, in Lemma 3.23, the random variables $(\mathfrak{B}_2(\ell))_{\ell=1}^{2k}$ (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 $\text{Bin}(z_\ell, p)$ whose mean is a second order approximation of the expectation of $\mathfrak{B}_2(\ell)$ (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 $\text{Bin}(z_\ell, p)$ again with mean a decreasing function of ℓ (see Lemma 3.25 and Remark 3.26). 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 3.27), and estimate the gap between them (see Lemma 3.29).

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-Essen's and Slud's inequalities.

2 Preliminaries and Notation

2.1 Notation

We use mostly standard asymptotic notations. 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 $[n] = \{1,\dots,n\}$: For given $a, b \in \mathbb{N}$, $a \le b$, we let $[a,b] = \{a,\dots,b\}$. 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 [15]).

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 [27])**.** *Let $X \in \mathrm{Bin}(n,p)$ be a random variable with binomial distribution with parameters n and $p = p(n) \leq 1/4$. 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 $p = o(1)$ 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 $M \in \mathbb{R}$,*

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

In particular,

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

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-Essen's inequality [9]), if $X \in \mathrm{Bin}(a,p)$, then for all $x \in \mathbb{R}$,

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

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,

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

Since $Z_1 \geq \frac{M+X_2-a_1p}{\sqrt{a_1pq}}$ if and only if $Z_1\sqrt{a_1pq} - M + a_1 p \geq X_2$, we get

$$\mathbb{P}(X_1 - X_2 \geq M) \geq -\frac{\zeta}{\sqrt{a_1 p}} + \mathbb{P}\Big(\frac{Z_1\sqrt{a_1pq} - M + (a_1-a_2)p}{\sqrt{a_2pq}} \geq \frac{X_2 - a_2 p}{\sqrt{a_2pq}}\Big)$$

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

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

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

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

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

Observing that

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

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,

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

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

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

On the other hand, since $m_* = \lfloor (a_1+1)p \rfloor$ is a mode of $\mathrm{Bin}(a_1, p)$, 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. □

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 $m \in a_2p \pm \sqrt{a_2pq}$ one has that $\mathbb{P}(X' = m)$ and $\mathbb{P}(X_2 = m + 1)$ are $\Omega(1/\sqrt{a_2p})$, so

$$\mathbb{P}(X' = X_2 - 1) \geq \sum_{m \in a_2p \pm \sqrt{a_2pq}} \mathbb{P}(X' = m)\mathbb{P}(X_2 = m + 1) = \Omega\Big(\frac{1}{\sqrt{a_2p}}\Big).$$

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 $(a_1 - a_2)p \geq N$ (the other assumptions therein remain), for every integer m, (1) and (2) are still satisfied, and moreover

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

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 $A = \{v_1, \dots, v_{2k}\}$, 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. every vertex carries a label in $[k]$. Recall that vertices in Level 1 had initially labels from $[2k]$, that is, $A = V([2k])$. So, in particular, we will show that vertices that initially had a label between $k + 1$ and $2k$ change their label to a smaller one. The reason we choose A of size $2k$ and not of size k will become clear in Lemma 3.32. For now, we just mention that in a certain vertex attribution procedure, the vertices in $B_1([k + 1, 2k])$ will get i.i.d. labels in $[k]$, which will provide us with a needed lower bound on $\mathfrak{B}_2(\ell_1)$ (with ℓ_1 defined as in the outline of the proof).

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,

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

We now formally state the main result of this section.

Lemma 3.1. *Suppose that $n^{5/8+\varepsilon} \le np \le n$. Then, a.a.s., after the second round, each vertex $v_j \in V$ with $j \ge K = \lceil 2(\log n)/p \rceil$ 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 $p \ge \sqrt{15}(\log n/n)^{1/4}$. 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 [19]. 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.

First, we will show that for every pair of vertices $u, v_j \in V \setminus V([k])$ (so in particular $j > k$), with probability $1 - o(n^{-2})$, vertex $u \neq v_j$ has more neighbors in the neighborhood of v_1 than in $N[v_j] \setminus B_1([k])$. 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 $\mathrm{Bin}(n-2, p^2)$. On the other hand, the number of neighbors of u in $N(v_j) \setminus B_1([k])$ is dominated by a binomial random variable Y_j with distribution $\mathrm{Bin}(n-2, q^k p^2)$: indeed, a vertex w is in $N(v_j) \setminus B_1([k])$ if it connects to v_j but does not connect to any vertex in $V([k])$, which happens with probability $q^k p$.

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

$$\mathbb{P}(Y_1 \le Y_j) \le \mathbb{P}(Y_1 - \mathbb{E}Y_1 \le -r/2) + \mathbb{P}(Y_j - \mathbb{E}Y_j \ge r/2). \tag{4}$$

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

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

By using that $np^4k^2 \ge 225\log n$, 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 $j \ge K$. It follows that $|N(v_1) \cap N(v_j)|$ dominates a random variable Z_j with distribution $\mathrm{Bin}(n-2, p^2)$, 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 $\mathrm{Bin}(n-j, pq^{j-1})$. Hence, using that for $j > K$,

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

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

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

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. *Suppose that $n^{5/8} \le np$. 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. Then, after two more rounds, a.a.s. every vertex carries this label.*

Proof. For convenience of notation and with the current setting in mind, we suppose that at least $0.9n$ of the vertices in $\mathcal{G}(n,p)$ have the same label at round R. Moreover, we assume without loss of generality that the label carried by most of the vertices at that point is 1 (the label may vary depending on the range of np), specifically, $\mathfrak{V}_R(1) \ge 0.9n$. Let us first show that a.a.s. $n - \mathfrak{V}_{R+1}(1) \le 300p^{-1}$. Fix $s \le 0.1n$ and a set $S \subseteq V$ of size $|S| = s$. Suppose that for every vertex $u \in S$, $N[u] \cap (V \setminus V_R(1))$ is larger than or equal to $N(u) \cap V_R(1)$. Note that at round $R+1$, every edge uv between a vertex $u \in S$ and a vertex $v \in V \setminus V_R(1)$ influences the labels attributed to u and v only. Hence, the number of edges between S and $V \setminus V_R(1)$ is dominated by $2X$ where X is a binomial random variable with parameters $s(n - \mathfrak{V}_R(1)) \le 0.1sn$ and p, and the number of edges between S and $V_R(1)$ dominates a binomial random variable Y with parameters $s(\mathfrak{V}_R(1) - s) \ge 0.8sn$ and p. Hence, combining Chernoff's bound with a union bound over all sets of size s shows that the probability that a set S as above exists is at most

$$\binom{n}{s}\binom{n}{\mathfrak{V}_R(1)}\mathbb{P}(2X + s \ge Y) \le 4^n(\mathbb{P}(X \ge 0.2snp) + \mathbb{P}(Y \le 0.5snp)) \le 4^n \cdot 2\exp\Big(-\frac{snp}{100}\Big).$$

One may easily check that the right-most expression above is $o(1)$ as long as $s \ge 300p^{-1}$, say. Thus, a.a.s. $n - \mathfrak{V}_{R+1}(1) \le 300p^{-1}$.

Finally, since a.a.s. every vertex has degree $\Omega(np) \gg p^{-1}$, most of the neighbors of every vertex carry label 1 after round $R+1$, which implies that a.a.s. all labels get the same label at round $R+2$, as desired. □

The following corollary is then a direct consequence of Lemma 3.1, Lemma 3.2 and the fact that $n - \frac{2}{p} \ge 0.9n$ for all sufficiently large n.

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

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

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

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 $\mathfrak{B}_1(\ell) \in \mathrm{Bin}(n-2k, q^{\ell-1}p)$, so for $\ell \in [2k]$, since $\ell p \le kp = o(1)$ 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 $n^{2/3} \ll np \le \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, the event*

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

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 $\ell p \le 2kp = o(1)$ 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 $1 - \frac{1}{\sqrt{2}} > \frac{1}{5}$, by Chernoff's bound,

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

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. □

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 $\varepsilon \in (0,1)$ there is a positive integer $L = L(\varepsilon, c)$ such that the event

$$\mathcal{E}_L = \{\text{for every } \ell \in [L+1,2k],\ \mathfrak{B}_1(1) - \mathfrak{B}_1(\ell) \ge \tfrac{1}{\sqrt{2}}(\ell-1)np^2\} \cap \{|\mathfrak{B}_1(1) - \mathbb{E}\mathfrak{B}_1(1)| \le Ln^{1/3}\}$$

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,*

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

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 $\alpha > 0$, $(1+x)^\alpha = 1 + \alpha x + \frac{1}{2}\alpha(\alpha-1)x^2 + O(x^3)$ 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,

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

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 $\mathcal{X}_i = \{j \in [n] : U_j = i\}$ 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 $\frac{1}{L}(1 - (1-\widehat{q})^L) = \widehat{p}$, 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 $i \in [L]$,

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

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. □

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*

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

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,

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

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

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

In particular, there is a $\delta > 0$ such that

$$\min_{i,j\in[L]:i\neq j} |N_i - N_j + (j-i)c^{3/2}| \geq \tfrac{1}{2}\delta c^{3/2}$$

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 1/2 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*

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

holds a.a.s.

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

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

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 get

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

In particular, a union bound yields

$$\mathbb{P}(\exists \ell \in [2k], \mathfrak{B}_1(\ell) \notin \mathbb{E}\mathfrak{B}_1(\ell) \pm \ell\omega) \leq \sum_{\ell=1}^{2k} \exp\left(-\frac{1}{3}\ell^2(np^3)^{1/2}\right) \leq \exp\left(-\frac{1}{4}(np^3)^{1/2}\right) = o(k^{-1}),$$

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 that for all $\ell \in [2k-1]$,

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

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 $\mathfrak{B} \in \mathrm{Bin}(n-2k, 1-q^{2k})$ 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 $U \subseteq [K] \setminus [2k]$ 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 $\frac{1}{2}\mathfrak{C}_2(1)\big(1-\big(\frac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)$ smaller than the number of vertices with label 1", that is,*

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

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, in particular, $(\mathfrak{B}_1(\ell))_{\ell=1}^{2k}$ 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 $np \leq \sqrt{15}n^{3/4}(\log n)^{-1/2}$, 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. Using that on the event $\mathcal{E}$ we have $\mathfrak{B}_1(1)-\mathfrak{B}_1(\ell) \geq (\ell-1)t_n$ gives

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

Also, note that Lemma 2.3 may be applied for $a_1 = \mathfrak{B}_1(1)$, $a_2 = \mathfrak{B}_1(1)-(\ell-1)t_n = (1-o(1))np$, $M=1$ and p; indeed, under the event $\mathcal{E}$ we have that $1 \ll a_2 \leq a_1$ and $\min\{a_2, a_1-a_2\}p \geq t_np = np^3/\sqrt{2} \gg 1$. Since $\frac{(a_1-a_2)p}{\sqrt{(a_1+a_2)pq}} \geq \frac{(\ell-1)t_np}{\sqrt{2pa_1}} = \frac{(\ell-1)np^3}{2\sqrt{pa_1}}$ and $\frac{np}{4} \leq a_2 \leq a_1 \leq (n-2k)p+\omega \leq np(1+p)$, we deduce that

$$\frac{p_1}{p_1+p_\ell} \geq \Phi\Big(\frac{(\ell-1)np^3}{2\sqrt{np^2(1+p)}}\Big) - \frac{4\zeta}{\sqrt{np^2}}, \tag{11}$$

which leads to

$$p_\ell \leq \frac{1-\Phi\Big(\frac{(\ell-1)np^3}{2\sqrt{np^2(1+p)}}\Big)+\frac{4\zeta}{\sqrt{np^2}}}{\Phi\Big(\frac{(\ell-1)np^3}{2\sqrt{np^2(1+p)}}\Big)-\frac{4\zeta}{\sqrt{np^2}}}p_1 = \frac{1-\Phi\Big(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\Big)+\frac{4\zeta}{\sqrt{np^2}}}{\Phi\Big(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\Big)-\frac{4\zeta}{\sqrt{np^2}}}p_1. \tag{12}$$

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

$$\begin{aligned}\Big(\frac{1-\Lambda}{\Lambda}\Big)^{\ell-1} &= \Big(\frac{1-\sqrt{np^4}/5}{1+\sqrt{np^4}/5}\Big)^{\ell-1} = \Big(1-\frac{2}{5}\sqrt{np^4}+O(np^4)\Big)^{\ell-1} \\ &= \exp\Big(-\frac{2}{5}(\ell-1)\sqrt{np^4}+O((\ell-1)np^4)\Big),\end{aligned} \tag{13}$$

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

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

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

$$\frac{1-\Phi\Big(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\Big)+\frac{4\zeta}{\sqrt{np^2}}}{\Phi\Big(\frac{(\ell-1)\sqrt{np^4}}{2\sqrt{1+p}}\Big)-\frac{4\zeta}{\sqrt{np^2}}} = 1-\frac{2(\ell-1)\sqrt{np^4}}{\sqrt{2\pi}}+O\Big((\ell-1)^2np^4+\frac{1}{\sqrt{np^2}}\Big), \tag{14}$$

and using that $\frac{2}{5}\le\frac{2}{\sqrt{2\pi}}$ shows the desired inequality when ε is sufficiently small. On the other hand, when $\varepsilon^{-1}\ge(\ell-1)\sqrt{np^4}\ge\varepsilon$, the inequalities $np^2\gg 1$ and $\ell np^4\le knp^4 = O(\sqrt{np^4\log n}) = o(1)$ 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 $(\ell-1)\sqrt{np^4}\ge\varepsilon^{-1}$ for some sufficiently small ε, using that $\frac{x}{1-x}\le 2x$ when x is small together with (13), the left-hand side of (14) is at most

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

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

Now, suppose that $(\log n)^{-2}\le np^4\le 225\log n$ or equivalently $np\in[n^{3/4}(\log n)^{-1/2},\sqrt{15}n^{3/4}(\log n)^{1/4}]$. Then, using that on the event $\mathcal{E}'$ we have $\mathfrak{B}_1(\ell)-\mathfrak{B}_1(\ell+1)\ge t_n = np^2/\sqrt{2}$ for every $\ell\in[2k-1]$ (by Remark 3.9),

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

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 $\frac{p_\ell}{p_\ell+p_{\ell+1}}\ge\Lambda$, or equivalently $p_{\ell+1}\le\frac{1-\Lambda}{\Lambda}p_\ell$, which by an immediate induction leads to $p_{\ell+1}\le(\frac{1-\Lambda}{\Lambda})^\ell p_1$.

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

$$1=\sum_{\ell=1}^{2k}p_\ell\le p_1\sum_{\ell=0}^{\infty}\Big(\frac{1-\Lambda}{\Lambda}\Big)^\ell = p_1\frac{\Lambda}{2\Lambda-1}. \tag{15}$$

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

$$\begin{aligned}&\mathbb{P}\big(\mathfrak{C}_2(1)\in p_1\mathfrak{C}\pm 2(p_1\mathfrak{C}\log n)^{1/2}\big)=1-o(n^{-1}),\\&\mathbb{P}\big(\mathfrak{C}_2(\ell)\le p_\ell\mathfrak{C}+\max\{2(p_\ell\mathfrak{C}\log n)^{1/2},(\log n)^2\}\big)=1-o(n^{-1}).\end{aligned}\tag{16}$$

In particular, with probability $1-o(n^{-1})$, we get that $\mathfrak{C}_2(1)\ge\frac{3}{4}p_1\mathfrak{C}\ge\frac{2}{3}p_1n\ge\frac{1}{2}(2\Lambda-1)n$, which proves the first part of the lemma. On the other hand, (16) implies that for every $\ell\in[2,2k]$, with probability $1-o(n^{-1})$,

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

[2] https://www.wolframalpha.com/input?key=&i=%281%2F2+%2B+erf%28x%2Fsqrt%282%29%29%2F2%29*%28exp%28-4x%2F5%29%2B1%29-1

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

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,

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

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

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

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

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

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

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

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

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

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

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, (11) and the consequent analysis holds for all $\ell \in [L+1, 2k]$ and p_* instead of p_1. Moreover, (15) must be replaced by

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

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

$$\widehat{\ell}_1 = \min\Big\{\ell \in [L] : \mathfrak{C}_2(\ell) = \max_{i\in[L]}\mathfrak{C}_2(i)\Big\}. \tag{17}$$

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 $\ell \in [L] \setminus \{\widehat{\ell}_1\}$,

$$\mathfrak{C}_2(\widehat{\ell}_1) - \mathfrak{C}_2(\ell) = \Omega((2\Lambda-1)\mathfrak{C}_2(\widehat{\ell}_1)). \tag{18}$$

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 $\Omega(n^{2/3}) = np \le \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, every vertex in Level 2 is connected to at most 7 vertices in Level 1 a.a.s.*

Proof. For any $j \in [n] \setminus [2k]$, the number of neighbors of v_j in Level 1 is $\mathrm{Bin}(2k, p)$. Since $kp = p^{-1}n^{-1/2+o(1)} \le n^{-1/6+o(1)}$, by a union bound over all vertices we conclude that

$$\mathbb{P}(\exists j \in [n] \setminus [2k], |N(v_j) \cap A| \ge 7) \le n\binom{2k}{7}p^7 = O(n(kp)^7) = o(1),$$

as desired. □

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, there is a constant $c_1 > 0$ such that a.a.s. for every $\ell \in [2, 2k]$, the number of vertices $\mathfrak{B}_2(1)$ in Level 2 that carry label 1 after the second round of ALAP is at least $(2\Lambda - 1)\frac{knp}{3}$ and is at least by $c_1\big(1 - \big(\frac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)\mathbb{E}\mathfrak{B}_2(1)$ larger than $\mathfrak{B}_2(\ell)$.*

Proof. Fix $t_n = np^2/\sqrt{2}$. As in the proof of Lemma 3.11, we condition on the events $\mathcal{E}$ (see Lemma 3.4), $\mathcal{F}$ (see 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).

By using the second moment method, we will show that a.a.s. for all $\ell \in [2, 2k]$,

$$\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell) \ge \tfrac{1}{2}\mathbb{E}[\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell)].$$

Step 1. Fix a vertex $u \in B$, its associated (random) set U (as defined in the first round of ALAP), and for every $\ell \in [2k]$, define $\widehat{p}_\ell$ 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 $(\widehat{p}_\ell)_{\ell=1}^{2k}$ depends 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 any such choice and any $\ell \in [2k-1]$, $\widehat{p}_1 - \widehat{p}_{\ell+1}$ is uniformly bounded from below. Fix $\ell \in [2k-1]$ and any vertex u in Level 2. Then, the number of neighbors of u that belong to $B_1(\ell)$ is given by a random variable with distribution $\mathrm{Bin}(\mathfrak{B}_1(\ell), p)$, if $u \notin B_1(\ell)$, and by a random variable with distribution $\mathrm{Bin}(\mathfrak{B}_1(\ell)-1, p)$ otherwise. At the same time, the number of vertices in Level 2 with label $\ell+1$ at distance at most 1 from u is dominated by $|N(u)\cap(\mathfrak{B}_1(\ell+1)\setminus\{u\})|+8$: indeed, $|N(u)\cap(\mathfrak{B}_1(\ell+1)\setminus\{u\})|$ is dominated by a random variable with distribution $\mathrm{Bin}(\mathfrak{B}_1(\ell+1), p)$, and 8 is an upper bound for the number of vertices in $A \cup \{u\}$ at distance at most 1 from u. Hence, using that on the event $\mathcal{E}$ we have $\mathfrak{B}_1(1) - \mathfrak{B}_1(\ell) \ge (\ell-1)t_n$,

$$\frac{\widehat{p}_1}{\widehat{p}_1 + \widehat{p}_\ell} \ge \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(1), p) \ge \mathrm{Bin}(\mathfrak{B}_1(\ell), p) + 8) \ge \mathbb{P}(\mathrm{Bin}(\mathfrak{B}_1(\ell), p) - \mathrm{Bin}(\mathfrak{B}_1(\ell) - (\ell-1)t_n, p) \ge 8). \quad (19)$$

Now, note that Lemma 2.3 may be applied for $a_1 = \mathfrak{B}_1(\ell)$, $a_2 = \mathfrak{B}_1(\ell) - (\ell-1)t_n$, $M = 8$ and p; indeed, under the event $\mathcal{E}$ we have that $1 \ll a_2 \le a_1$ and $\min\{a_2, a_1 - a_2\}p \ge t_n p = np^3/\sqrt{2} \gg 1$. As in the proof of Lemma 3.11, this yields

$$\forall \ell \in [2k-1],\ \widehat{p}_{\ell-1} \le \Big(\frac{1-\Lambda}{\Lambda}\Big)^{\ell}\widehat{p}_1 \text{ and, in particular, } \widehat{p}_1 \ge \frac{2\Lambda-1}{\Lambda}. \quad (20)$$

Step 2. We now 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,

$$\mathbb{E}[(\mathfrak{B}_2(1)-\mathfrak{B}_2(\ell))^2] = \sum_{u\in B}\mathbb{E}[(\mathbb{1}_{u\in B_2(1)}-\mathbb{1}_{u\in B_2(\ell)})^2]+\sum_{u,v\in B:u\neq v}\mathbb{E}[(\mathbb{1}_{u\in B_2(1)}-\mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)}-\mathbb{1}_{v\in B_2(\ell)})] \tag{21}$$

and

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

To bound the first summation in (21), observe that $(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})^2 \leq \mathbb{1}_{u\in B_2(1)} + \mathbb{1}_{u\in B_2(\ell)}$, so

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

Next, recall that $\widehat{p}_{\ell+1} \leq (\frac{1-\Lambda}{\Lambda})^\ell\widehat{p}_1$ for all $\ell \in [2k-1]$. 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 $\mathbb{E}[\mathfrak{B}_2(1)] = \Theta(\widehat{p}_1knp)$ and $\widehat{p}_1 \geq \frac{2\Lambda-1}{\Lambda} \geq 2\Lambda - 1$, we get

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

where the last inequality comes from the fact that when $2\Lambda-1 = \Omega(1)$, then $(2\Lambda-1)^3knp^2 = \Omega(\sqrt{n\log n}) \gg 1$, and when $2\Lambda - 1 = \Omega(\sqrt{np^4})$, then $(2\Lambda - 1)^3knp^2 = \Omega((np^4)^{3/2}\sqrt{n\log n}) \gg (np^3)^2 \gg 1$. By (23), 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 (21) and (22) yields

$$\begin{aligned}&\mathbb{V}[\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell)]\\ &= o\big(p(\mathbb{E}[\mathfrak{B}_2(1) + \mathfrak{B}_2(\ell)])^2\big)\\ &+ \sum_{u,v\in B:u\neq v}\Big(\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)})] - \mathbb{E}[\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} - \mathbb{1}_{v\in B_2(\ell)}]\Big).\end{aligned} \tag{24}$$

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

$$\begin{aligned}&\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)})]\\ &= q\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}) \mid uv \notin G_n]\\ &+ p\mathbb{E}[(\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)})(\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)}) \mid uv \in G_n].\end{aligned} \tag{25}$$

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 first variable is measurable in terms of the edges between u and $B \setminus \{v\}$, and the second variable is measurable in terms of the edges between v and $B \setminus \{u\}$. Hence,

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

$$= \mathbb{E}[\mathbb{1}_{u\in B_2(1)} - \mathbb{1}_{u\in B_2(\ell)} \mid uv \notin G_n]\mathbb{E}[\mathbb{1}_{v\in B_2(1)} - \mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n], \tag{26}$$

and

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

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

$$\begin{aligned}&\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)}]\\ &= p\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \in G_n] + q\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \notin G_n].\end{aligned} \tag{28}$$

Now, consider the general term of summation in (24). Replacing the conditional expectations in (25) by their equivalent in (26)-(27), using (28) twice, and some arithmetic, the general term can be rewriting as

$$pq \cdot \prod_{w\in\{u,v\}} \big(\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \in G_n]\big). \tag{29}$$

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

$$|\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \notin G_n] - \mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)} \mid uv \in G_n]| = o(\mathbb{E}[\mathbb{1}_{w\in B_2(1)} - \mathbb{1}_{w\in B_2(\ell)}]), \tag{30}$$

so the general term in the summation in (24) 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 $\mathbb{V}(\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell)) = o(p(\mathbb{E}[\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell)])^2)$.

To show (30), we need to conduct a thorough case analysis: indeed, u may be in each of $B_1(1)$, $B_1(\ell)$ and $B \setminus (B_1(1) \cup B_1(\ell))\}$, it has between 1 and 7 neighbors with label 1, and between 1 and 7 neighbors with label ℓ in Level 1, and the same set of possibilities holds for v. We choose to analyze the case when $u, v \in B \setminus (B_1(1) \cup B_1(\ell))$ and neither u nor v has any neighbors with label 1 or ℓ after the first round in Level 1 (the minor adjustments for the other cases will be mentioned along the way). Let us concentrate on u. Let $X_i = |N[u] \cap (A_1(i) \cap B_1(i))|$ and observe that there are two independent random variables $\widehat{X}_1 \in \mathrm{Bin}(\mathfrak{B}_1(1), p)$ and $\widehat{X}_\ell \in \mathrm{Bin}(\mathfrak{B}_1(\ell), p)$ and a constant $m \geq 0$ such that $|X_1 - \widehat{X}_1| \leq m$ and $|X_\ell - \widehat{X}_\ell| \leq m$ in the unconditional probability space, in the space conditioned on $uv \in G_n$ and in the space conditioned on $uv \notin G_n$ at the same time. Note that in our case, one may choose $m = 0$, but in general one, due to the (at most 7) edges towards vertices with label 1 (respectively ℓ) in Level 1, the edge uv and the labels of u and v themselves one may need to choose as large as $m = 7 + 1 = 8$.

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

$$\mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell = m) = o(\mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell \geq m) - \mathbb{P}(\widehat{X}_1 - \widehat{X}_\ell < m)) \tag{31}$$

for any fixed constant m (showing that any constant number of edges does not change the probability of receiving a concrete label significantly), (30) is satisfied for $w = u$. The case $w = v$ is analogous.

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

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

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 $\mathbb{E}[\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell)] = \Omega\Big(\big(1 - \big(\tfrac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)\mathbb{E}\mathfrak{B}_2(1)\Big)$. $\square$

Remark 3.16. For $np = cn^{2/3}$, Lemma 3.15 holds by replacing the original statement by "Given $\varepsilon \in (0,1)$ and $L = L(\varepsilon, c)$ (provided by Remark 3.5), conditionally on $\mathcal{E}_L$, there is a constant $c_1 > 0$ such that a.a.s. for every $\ell \in [L+1, 2k]$, the number of vertices $\max_{i\in[L]} \mathfrak{B}_2(i)$ that carry the most represented label in Level 2 after the second round is at least $(2\Lambda - 1)\frac{knp}{3}$ and is at least by $c_1\big(1 - \big(\frac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big) \max_{i\in[L]} \mathbb{E}\mathfrak{B}_2(i)$ larger than the number of vertices carrying the label ℓ.".

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 (31) by

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

for an appropriately small ε, which in turn allows us to replace (30) with

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

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.

Remark 3.17. Fix $np = cn^{2/3}$ for some constant $c > 0$. Then, similarly to Remark 3.13, replacing Lemma 3.4 by Lemma 3.7, and Lemma 2.3 by Remark 2.4 in the proof of Lemma 3.15 implies that conditionally on the event of Lemma 3.7, a.a.s. for every $\ell \in [L] \setminus \{\widehat{\ell}_1\}$ (with $\widehat{\ell}_1$ defined as in Remark 3.13),

$$\mathfrak{B}_2(\widehat{\ell}_1) - \mathfrak{B}_2(\ell) = \Omega((2\Lambda - 1)\mathbb{E}\mathfrak{B}_2(\widehat{\ell}_1)).$$

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

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 3.18. *Suppose that $n^{2/3} \ll np \le \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, after the second round of ALAP, a.a.s. $n - o(n)$ of the vertices in Level 3 have more neighbors in Level 2 with label 1 than with any other label.*

Proof. For every $\ell \in [2k]$, condition on the a.a.s. statement of Lemma 3.15, the event $\mathfrak{B}_2(1) \le (\log n)^{1/3}\mathbb{E}\mathfrak{B}_2(1)$ (which holds a.a.s. as well by Markov's inequality) and the variables $(\mathfrak{B}_2(\ell))_{\ell=1}^{2k}$.

Now, fix a vertex $u \in C$. For every $\ell \in [2k]$, the number of neighbors of u in Level 2 with label ℓ after the second round is a random variable $Y_\ell \in \mathrm{Bin}(\mathfrak{B}_2(\ell), p)$, and moreover $(Y_\ell)_{\ell=1}^{2k}$ are independent variables. For every $\ell \in [2, 2k]$, Chernoff's bound implies

$$\begin{aligned}\mathbb{P}(Y_1 \le Y_\ell) &\le \mathbb{P}\big(Y_1 \le \tfrac{1}{2}\mathbb{E}[Y_1 + Y_\ell]\big) + \mathbb{P}\big(Y_\ell \ge \tfrac{1}{2}\mathbb{E}[Y_1 + Y_\ell]\big)\\ &\le \mathbb{P}\big(Y_1 - \mathbb{E}Y_1 \le -\tfrac{1}{2}\mathbb{E}[Y_1 - Y_\ell]\big) + \mathbb{P}\big(Y_\ell - \mathbb{E}Y_\ell \ge \tfrac{1}{2}\mathbb{E}[Y_1 - Y_\ell]\big)\\ &\le 2\exp\Big(-\tfrac{(\mathbb{E}[Y_1 - Y_\ell])^2}{3\mathbb{E}[Y_1]}\Big) = 2\exp\Big(-\tfrac{(\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell))^2 p}{3\mathfrak{B}_2(1)}\Big).\end{aligned}$$

Finally, having that $\mathfrak{B}_2(1) - \mathfrak{B}_2(\ell) \ge c_1\big(1 - \big(\frac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)\mathbb{E}\mathfrak{B}_2(1)$ and $\mathfrak{B}_2(1) \le (\log n)^{1/3}\mathbb{E}\mathfrak{B}_2(1)$ leads to

$$\begin{aligned}&\mathbb{P}(\exists \ell \in [2, 2k], Y_1 \le Y_\ell)\\ \le\ & 2\sum_{\ell=2}^{2k} \exp\Big(-\frac{c_1^2}{3}\Big(1 - \Big(\tfrac{1-\Lambda}{\Lambda}\Big)^{\ell-1}\Big)^2 \frac{\mathfrak{B}_2(1)p}{(\log n)^{1/3}}\Big)\\ =\ & \sum_{\ell=2}^{\lfloor((2\Lambda-1)\log n)^{-1}\rfloor} \exp\big(-\Omega((2\Lambda-1)^2(\ell-1)^2(\log n)^{-1/3}\mathfrak{B}_2(1)p)\big) + 2k\exp\big(-\Omega((\log n)^{-7/3}\mathfrak{B}_2(1)p)\big)\\ =\ & \exp\big(-\Omega((2\Lambda-1)^2(\log n)^{-1/3}\mathfrak{B}_2(1)p)\big) + 2k\exp\big(-\Omega((\log n)^{-7/3}\mathfrak{B}_2(1)p)\big).\end{aligned}$$

Since $\mathfrak{B}_2(1) \geq \frac{1}{3}(2\Lambda - 1)knp$ by Lemma 3.15, we have that

$$\mathfrak{B}_2(1)p = \Omega((2\Lambda-1)\sqrt{n\log n}) = \Omega(\min\{1,(np^4)^{1/2}\}\sqrt{n\log n}) = \Omega(n^{1/3})$$

and $(2\Lambda - 1)^2(\log n)^{-1/3}\mathfrak{B}_2(1)p = \Omega(\min\{1,(np^4)^{3/2}\}\sqrt{n}(\log n)^{1/6}) = \Omega((\log n)^{1/6})$. It follows that the expected number of vertices that do not satisfy the property of the lemma is bounded from above by $n\exp(-\Omega((\log n)^{1/6})) = o(n)$, which combined with Markov's inequality implies the result. □

Remark 3.19. Suppose that $np = cn^{2/3}$ for some constant $c > 0$. By replacing Lemma 3.15 with Remarks 3.16 and 3.17, 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 have more neighbors with label $\widehat{\ell}_1$ (as defined in Remark 3.13) than neighbors with other label.

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 $\ell \in [2, 2k]$.

Lemma 3.20. *Fix $\mathfrak{N} = \frac{1}{10}(2\Lambda - 1)\mathfrak{C}_2(1)$. Suppose that $n^{2/3} \ll np \leq \sqrt{15}n^{3/4}(\log n)^{1/4}$. Then, a.a.s. $\mathfrak{N} \geq \frac{1}{20}(2\Lambda - 1)^2 n$ 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,

$$\Big\{\mathfrak{C}_2(1) \geq \tfrac{1}{2}(2\Lambda-1)n \text{ and for every } \ell \in [2,2k], \mathfrak{C}_2(1) - \mathfrak{C}_2(\ell) \geq \tfrac{1}{2}\big(1 - \big(\tfrac{1-\Lambda}{\Lambda}\big)^{\ell-1}\big)\mathfrak{C}_2(1)\Big\}, \tag{32}$$

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 $\ell \in [2, 2k]$, (32) 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

$$\mathbb{P}(X_1 - X_2 \leq 3\mathfrak{N}p) \leq \mathbb{P}(X_1 - a_1 p \leq \mathfrak{N}p) + \mathbb{P}(X_2 - a_2 p \geq \mathfrak{N}p) \leq \exp\left(-\frac{\mathfrak{N}^2p^2}{2(a_1p + \mathfrak{N}p)}\right) + \exp\left(-\frac{\mathfrak{N}^2p^2}{2(a_2p + \mathfrak{N}p)}\right).$$

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. □

Remark 3.21. For $np = cn^{2/3}$ for some constant $c > 0$ and $\mathfrak{N}$ as in the statement of Lemma 3.20, consider the event "$\mathfrak{N} \geq \frac{1}{20}(2\Lambda - 1)^2 n$ and every vertex u in Level 3 satisfies $|N(u) \cap C_2(\widehat{\ell}_1)| \geq |N(u) \cap C_2(\ell)| + \mathfrak{p}$ for all $\ell \in [2, 2k]$", with label $\widehat{\ell}_1$ defined as in (17). This event holds a.a.s. conditionally on the label $\widehat{\ell}_1$ being the one with the largest basin (which, for every $\varepsilon > 0$, holds with probability at least $1 - \varepsilon$ as long as L is sufficiently large, see Claim 3.13). The only necessary modification is to replace $\mathfrak{N}$ by arbitrary $\mathfrak{N}' = \mathfrak{N}'(n)$ such that $\mathfrak{N}' = o(\mathfrak{N})$, $\mathfrak{N}' \gg Kp$ and $\mathfrak{N}'^2 p \gg a_1 \log n$ (for example, $\mathfrak{N}' = (\log n)^{-1}\mathfrak{N}$ satisfies these conditions). This is required since we cannot control the constant in front of $(2\Lambda - 1)\mathfrak{C}_2(\widehat{\ell}_1)$ in (18).

Lemma 3.22. *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 3.18, 3.20 and the fact that a.a.s. $\mathfrak{N}p = \Omega((2\Lambda-1)^2 n) \gg Mp$ (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. □

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 3.18 with Remark 3.19 in the proof of Lemma 3.22 shows that conditionally on the event $\mathcal{E}_L \cap \mathcal{G}_{L,\varepsilon}$, a.a.s. only one label survives. However, since $\mathcal{E}_L \cap \mathcal{G}_{L,\varepsilon}$ holds with probability at least $1 - 2\varepsilon$, and ε could be chosen arbitrarily small, the second point of Theorem 1.1 readily follows.

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

Now, we fix $\varepsilon \in (0, 1/24)$ and concentrate on the regime $n^{5/8+\varepsilon} \le np \ll n^{2/3}$ 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 3.23. *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:*

- $|Z_\ell - \mathfrak{B}_1(\ell)| \le (np)^{2/5}$,
- $Z_\ell \in \mathrm{Bin}\,(z_\ell, p)$.

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

$$|\mathbb{E}\mathfrak{B}_1(\ell) - np + (\ell-1)np^2| \le x_\ell. \tag{33}$$

For $\ell = 1$ we know that $\mathfrak{B}_1(1) \in \mathrm{Bin}(n-2k, p)$, so $|\mathbb{E}\mathfrak{B}_1(1) - np| = 2kp \ll x_1$ because $n^{5/8+\varepsilon} \le np$. Suppose that for some $\ell \ge 2$, the statement holds for all $j \in [\ell-1]$. Then, by the induction hypothesis

$$\mathbb{E}\mathfrak{B}_1(\ell) = (n-2k)p - p\sum_{j=1}^{\ell-1} \mathbb{E}\mathfrak{B}_1(j) = (n-2k)p - (\ell-1)np^2 + \frac{1}{2}(\ell-1)(\ell-2)np^3 \pm p\sum_{j=1}^{\ell-1} x_j.$$

Now, on the one hand, $\ell p \le kp \ll 1$ implies that $p\sum_{j=1}^{\ell-1} x_j \le \ell^3 np^4/2 \le x_\ell/2$. Moreover, since $kp \ll np^3$, $2kp + \frac{1}{2}(\ell-1)(\ell-2)np^3 \le x_\ell/2$. We conclude that $2kp + \frac{1}{2}(\ell-1)(\ell-2)np^3 + p\sum_{j=1}^{\ell-1} x_j \le x_\ell$, which proves the statement for ℓ.

Now, define the event

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

Then, (33) and the triangle inequality imply that

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

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

$$\mathbb{P}(\overline{\mathcal{A}_\ell}) \le \mathbb{P}\left(\Big|\sum_{j=1}^{\ell-1}(\mathfrak{B}_1(j) - \mathbb{E}\mathfrak{B}_1(j))\Big| \ge 2\Big(2\log n \sum_{j=1}^{\ell-1} \mathbb{E}\mathfrak{B}_1(j)\Big)^{1/2}\right) \le \exp\left(-(4+o(1))\log n\right) \le \frac{1}{n}.$$

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 $\mathfrak{B}_1([\ell-1]) \ge n - z_\ell - 2k$, 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

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

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

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

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

$$\mathbb{P}\big(Y_\ell \ge (np)^{2/5}\big) \le \frac{1}{n}.$$

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

Remark 3.24. Note that the previous proof can be extended to the range $np \in [n^{1/2+\varepsilon}, n^{2/3}]$ for 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 3.23. Define $Z^{(1)} = \max\{Z_\ell \colon \ell \in [2k]\}$ and $Z^{(2)} = \max\{Z_i \colon \ell \in [2k], Z_\ell < Z^{(1)}\}$. To begin with, we estimate $Z^{(1)}$.

Lemma 3.25.

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

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 $\mathbb{P}\big(\bigcap_{\ell=1}^{2k}\{Z_\ell \le T_n\}\big) = \prod_{\ell=1}^{2k} \mathbb{P}(Z_\ell \le T_n)$. We now provide upper and lower bounds for $\mathbb{P}\big(Z_\ell \le T_n\big)$. On the one hand, the fact that $Z_\ell \in \mathrm{Bin}(z_\ell, p)$ implies that $\mathbb{E}Z_\ell = z_\ell p = (1+o(1))np \gg T_n + 1 - z_\ell p$. Combining this with Chernoff's bound yields

$$\mathbb{P}\big(Z_\ell \ge T_n + 1\big) = \mathbb{P}\big(Z_\ell - \mathbb{E}Z_\ell \ge T_n + 1 - z_\ell p\big) \le \exp\Big(-(1+o(1))\frac{(T_n + 1 - z_\ell p)^2}{2np}\Big).$$

On the other hand, by Slud's inequality (Lemma 2.2)

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

where we recall that Φ is the cumulative density function of a standard normal random variable. Integrating by parts leads to

$$\mathbb{P}\big(Z_\ell \ge T_n + 1\big) \ge \exp\Big(-(1+o(1))\frac{(T_n + 1 - z_\ell p)^2}{2np}\Big).$$

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

$$\frac{1}{2np}(T_n + 1 - z_\ell p)^2 = \frac{1}{2np}\Big(\sqrt{\zeta np \log(1/(np^3))} + (1+o(1))(\ell-1)np^2\Big)^2 \tag{34}$$

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

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

$$\prod_{\ell=1}^{2k} \mathbb{P}(Z_\ell \le T_n) = \prod_{\ell=1}^{2k}(1 - \mathbb{P}(Z_\ell \ge T_n+1)) = \exp\Big(-(1+o(1))\sum_{\ell=1}^{2k}\mathbb{P}(Z_\ell \ge T_n+1)\Big)$$

$$= \exp\Big(-(np^3)^{\zeta/2+o(1)}\sum_{\ell=1}^{2k}\exp\Big(-(1+o(1))\Big((\ell-1)\sqrt{\zeta np^3\log(1/(np^3))} + \frac{1}{2}(\ell-1)^2np^3\Big)\Big)\Big). \tag{35}$$

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

$$\left(\frac{1}{\sqrt{np^3\log(1/(np^3))}}\right)^{1+o(1)} = \frac{1}{(np^3)^{1/2+o(1)}}. \tag{36}$$

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

$$\begin{aligned}
&\sum_{\ell=1}^{2k}\exp\Big(-(1+o(1))\Big((\ell-1)\sqrt{\zeta np^3\log(1/(np^3))} + \frac{1}{2}(\ell-1)^2np^3\Big)\Big)\\
&\le \sum_{\ell=1}^{2k}\exp\Big(-(1+o(1))(\ell-1)\sqrt{\zeta np^3\log(1/(np^3))}\Big)\\
&\le \frac{1}{1-\exp\Big(-(1+o(1))\sqrt{\zeta np^3\log(1/(np^3))}\Big)}\\
&\le \frac{1}{(np^3)^{1/2+o(1)}}.
\end{aligned} \tag{37}$$

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

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

Remark 3.26. Note that (36) and (37) 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 take a closer look at the latter case, the former being analogous. To begin with, instead of starting the sum in (35) from $\ell = 1$, we start it from $\ell = k_*$. Now, as in (36), summing over the first ℓ_* terms (which form a decreasing sequence, but the last one is still a constant factor away from the first one), we obtain that the sum in (35) is bounded from below by

$$\ell_*\exp\Big(-(1+o(1))\Big(k_*\sqrt{\zeta np^3\log(1/(np^3))} + \frac{1}{2}k_*^2np^3\Big)\Big) \ge \frac{1}{(np^3)^{7/16+o(1)}},$$

where for the last equality we used that $\frac{1}{16}\log(1/(np^3)) = \frac{1}{2}k_*^2np^3$.

At the same time, similarly to (37), the same sum is bounded from above by

$$\frac{\exp(-(\frac{1}{2}+o(1))k_*^2np^3)}{1-\exp(-(1+o(1))\sqrt{\zeta np^3\log(1/(np^3))})} \le \frac{1}{(np^3)^{7/16+o(1)}}.$$

As a consequence $\prod_{\ell=k_*}^{2k}\mathbb{P}(Z_\ell \le M) = \exp(-(np^3)^{\zeta/2-7/16+o(1)})$, so

$$\frac{\max_{\ell\in[k_*,2k]} Z_\ell - np}{\sqrt{np\log(1/(np^3))}} \xrightarrow{\mathbb{P}} \frac{7}{8} \text{ as } n\to\infty, \text{ and similarly } \frac{\max_{\ell\in[\lfloor k_*/2\rfloor,2k]} Z_\ell - np}{\sqrt{np\log(1/(np^3))}} \xrightarrow{\mathbb{P}} \frac{31}{32} \text{ as } n\to\infty.$$

Corollary 3.27. *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 3.25 together with the second conclusion Remark 3.26 implies that a.a.s. $Z^{(1)} > \max_{\ell\in[k_*/2,2k]} Z_j$. Similarly, Lemma 3.25 and the first conclusion of Remark 3.26 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. □

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

Lemma 3.28. *Fix* $n \in \mathbb{N}$ *and* $t \in [n]$. *Then, the function* $s \in \mathbb{N}\cap[t, n-1] \mapsto \mathbb{P}(Z_*^{(2)} \le Z_*^{(1)} - t \mid Z_*^{(1)} = s)$ *is increasing in* s.

Proof. Firstly, define $\ell_1 = \min\{\ell \in [k_*],\ Z_*^{(1)} = Z_\ell\}$. Note that by Corollary 3.27, 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 $s \in \mathbb{N} \mapsto \mathbb{P}(Z_*^{(2)} \le Z_*^{(1)} - t \mid Z_*^{(1)} = s, \ell_1 = \ell)$ is increasing, which implies the lemma. Let us condition on the event $\ell_1 = \ell$. We have

$$\mathbb{P}(Z_*^{(2)} \le Z_*^{(1)} - t \mid Z_*^{(1)} = s, \ell_1 = \ell) = \prod_{j\in[\ell-1]} \mathbb{P}(Z_j \le s - t \mid Z_j \le s-1) \prod_{j=\ell+1}^{k_*} \mathbb{P}(Z_j \le s-t \mid Z_j \le s),$$

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 $s \ge t$,

$$\mathbb{P}(Z_j \le s-t \mid Z_j \le s) = \frac{\mathbb{P}(Z_j \le s-t)}{\mathbb{P}(Z_j \le s)} \quad \text{and} \quad \mathbb{P}(Z_j \le s+1-t \mid Z_j \le s+1) = \frac{\mathbb{P}(Z_j \le s+1-t)}{\mathbb{P}(Z_j \le s+1)}.$$

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 [25]). 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

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

which is equivalent to

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

thereby concluding the proof of the lemma. □

By Lemma 3.25 one can define a sequence of positive real numbers $(\varepsilon_n)_{n\ge1}$ converging to 0 and such that, on the one hand, $\varepsilon_n \ge (\log(np))^{-1/2}$ 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 3.29.

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

Proof. Using Lemma 3.28 (for the second inequality) and Corollary 3.27 (for the equality) below, we get that

$$\mathbb{P}\big(Z_*^{(2)} \le Z_*^{(1)} - 2\gamma_n\big) \ge\ \mathbb{P}(Z_*^{(1)} \ge M_n)\mathbb{P}(Z_*^{(2)} \le Z_*^{(1)} - 2\gamma_n \mid Z_*^{(1)} \ge M_n)$$

$$
\begin{aligned}
&\geq\ (1-\mathbb{P}(Z_*^{(1)}\leq M_n-1))\mathbb{P}(Z_*^{(2)}\leq Z_*^{(1)}-2\gamma_n\mid Z_*^{(1)}=M_n)\\
&=\ (1-o(1))\sum_{\ell=1}^{k_*}\mathbb{P}(\ell_1=\ell)\prod_{j\in[k_*]\setminus\{\ell\}}\mathbb{P}(Z_j\leq M_n-2\gamma_n\mid \ell_1=\ell, Z_\ell=M_n)\\
&\geq\ (1-o(1))\prod_{\ell=1}^{k_*}\mathbb{P}(Z_\ell\leq M_n-2\gamma_n\mid Z_\ell\leq M_n),
\end{aligned}
\tag{38}
$$

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 (38). Moreover, using that $\mathbb{P}(Z_\ell\leq M_n)=1-o(1)$ for every $\ell\in[k_*]$, the product in (38) rewrites as

$$
\prod_{\ell=1}^{k_*}\Big(1-\mathbb{P}\big(Z_\ell\in[M_n-2\gamma_n+1,M_n]\mid Z_\ell\leq M_n\big)\Big)=\prod_{\ell=1}^{k_*}\left(1-(1+o(1))\mathbb{P}\left(Z_\ell\in[M_n-2\gamma_n+1,M_n]\right)\right). \tag{39}
$$

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

$$
\frac{\mathbb{P}(Z_\ell=M_n-2\gamma_n)}{\mathbb{P}(Z_\ell=M_n)}=\frac{M_n(M_n-1)\cdots(M_n-2\gamma_n+1)q^{2\gamma_n}}{(z_\ell-M_n+1)\cdots(z_\ell-M_n+2\gamma_n)p^{2\gamma_n}}, \tag{40}
$$

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, (40) is equal to

$$
\begin{aligned}
(1-o(1))\frac{M_n^{2\gamma_n}}{p^{2\gamma_n}}\prod_{i=1}^{2\gamma_n}\frac{1}{z_\ell-M_n+i}&=(1-o(1))\prod_{i=1}^{2\gamma_n}\frac{np+\sqrt{(1-\frac{\varepsilon_n}{2})np\log(1/(np^3))}}{np-(1+o(1))\ell np^2}\\
&=\Big(1+\big(1+o(1)\big)\Big(\sqrt{\tfrac{1}{np}(1-\tfrac{\varepsilon_n}{2})\log(1/(np^3))}+\ell p\Big)\Big)^{2\gamma_n}\\
&=\exp\Big(\big(2\gamma_n+o(\gamma_n)\big)\Big(\sqrt{\tfrac{1}{np}(1-\tfrac{\varepsilon_n}{2})\log(1/(np^3))}+\ell p\Big)\Big)\\
&=\exp(O((np)^{-\varepsilon_n/2}))\\
&=1+o(1),
\end{aligned}
$$

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

Using this observation in (39) implies that

$$
\prod_{\ell=1}^{k_*}\left(1-(1+o(1))\mathbb{P}\left(Z_\ell\in[M_n-2\gamma_n+1,M_n]\right)\right)=\prod_{\ell=1}^{k_*}\big(1-(1+o(1))2\gamma_n\mathbb{P}\big(Z_\ell=M_n\big)\big). \tag{41}
$$

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

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

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

$$\begin{aligned} 2\gamma_n \mathbb{P}(Z_\ell = M_n) = 2\gamma_n \binom{z_\ell}{M_n} p^{M_n} q^{z_\ell - M_n} &= 2(np)^{1/2-\varepsilon_n} \Big(\frac{z_\ell p e}{M_n}\Big)^{M_n} \frac{1+o(1)}{\sqrt{2\pi np}} q^{z_\ell - M_n} \exp\Big(- \frac{M_n^2}{2z_\ell} + O\Big(\frac{M_n^3}{z_\ell^2}\Big)\Big) \\ &= (np)^{-\varepsilon_n} \Big(\frac{z_\ell p}{M_n}\Big)^{M_n} (1+o(1)) q^{z_\ell - M_n} \sqrt{\frac{2}{\pi}} \exp\Big(M_n - \frac{M_n^2}{2z_\ell} + O\Big(\frac{M_n^3}{z_\ell^2}\Big)\Big). \end{aligned} \tag{42}$$

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

$$\Big(\frac{z_\ell p}{M_n}\Big)^{M_n} = \exp\Big(z_\ell p - M_n - \frac{(z_\ell p - M_n)^2}{2M_n} + O\Big(\frac{(z_\ell p - M_n)^3}{M_n^2}\Big)\Big), \tag{43}$$

$$q^{z_\ell - M_n} = \exp\Big(- \Big(p + \frac{p^2}{2} + O(p^3)\Big)(z_\ell - M_n)\Big). \tag{44}$$

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 (44) is

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

Hence, combining (42), (43) and (44), we obtain that

$$2\gamma_n \mathbb{P}(Z_\ell = M_n) = (np)^{-\varepsilon_n} \exp\Big(- \frac{1}{2}(z_\ell p - M_n)^2 \Big(\frac{1}{z_\ell} + \frac{1}{M_n}\Big) + O\Big(\frac{M_n^3}{z_\ell^2} + \frac{(z_\ell p - M_n)^3}{M_n^2}\Big) + O(1)\Big).$$

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

$$\begin{aligned} z_\ell p - M_n &= -(\ell-1)np^2 + \tfrac{1}{2}(\ell-1)(\ell-2)np^3 - \sqrt{(1-\tfrac{\varepsilon_n}{2})np\log(\tfrac{1}{np^3})} \\ &= -np^2\Big(\ell - 1 + 2\sqrt{2(1-\tfrac{\varepsilon_n}{2})}k_* + O(k_*^2 p)\Big). \end{aligned}$$

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

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

Hence,

$$2\gamma_n \mathbb{P}(Z_j = M_n) \leq (np)^{-\varepsilon_n} \exp\big(- \tfrac{1}{2}(1 - \tfrac{\varepsilon_n}{2})\log(\tfrac{1}{np^3}) + o(1)\big) = O((np)^{-\varepsilon_n}(np^3)^{1/2-\varepsilon_n/4}) = O\Big(\frac{(np)^{-\varepsilon_n/2}}{k_*}\Big),$$

where for the last equality we used that $(n^2p^4)^{-\varepsilon_n/4} \leq 1$ 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 (41) rewrites as

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

which finishes the proof of the lemma. □

Now, recall that by definition $\mathfrak{B}^{(1)} = \max_{\ell \in [2k]} \mathfrak{B}_1(\ell)$, and define also $\mathfrak{B}^{(2)} = \max_{\ell \in [2k] \setminus \{\ell_1\}} \mathfrak{B}_1(\ell)$.

Corollary 3.30. *Under the coupling from Lemma 3.23, a.a.s.* $\mathfrak{B}_1(\ell_1) = \mathfrak{B}^{(1)}$ *and* $\mathfrak{B}^{(1)} - \mathfrak{B}^{(2)} \geq \gamma_n$.

Proof. Under the coupling from Lemma 3.23 we have that a.a.s.

$$|\mathfrak{B}_1(\ell_1) - Z_{\ell_1}| \leq (np)^{2/5} \quad \text{and} \quad |\mathfrak{B}^{(2)} - Z^{(2)}| \leq \max_{\ell \in [2k] \setminus \{\ell_1\}} |\mathfrak{B}_1(\ell) - Z_\ell| \leq (np)^{2/5}. \tag{45}$$

By Corollary 3.27 and Lemma 3.29 we know that a.a.s. $|Z^{(1)} - Z^{(2)}| \geq 2\gamma_n \gg (np)^{2/5}$, which together with (45) directly implies the first statement of the corollary. For the second statement, the triangle inequality implies that a.a.s.

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

which finishes the proof of the corollary. □

At this stage, we have the necessary information to analyze the number of vertices in Level 3 that receive label ℓ_1 at the second round of ALAP, and in particular the difference between the first and the second most represented labels in Level 3.

Lemma 3.31. *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, 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

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

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 3.30 we know that a.a.s. $\mathfrak{B}^{(1)} - \mathfrak{B}^{(2)} \geq \gamma_n$, and moreover, combining Lemma 3.25 and Corollary 3.30 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. Hence, in our procedure, the probability that a vertex in $C_2(\{\ell_1, \ell\})$ gets label 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 $\frac{1}{2} + \Omega\Big(\frac{\gamma_n p}{\sqrt{\mathfrak{B}^{(1)} p}}\Big) = \frac{1}{2} + \Omega\Big(\frac{p^{1/2}}{(np)^{\varepsilon_n}}\Big)$. Hence, using that conditionally on $\mathfrak{C}_2(\{\ell_1, \ell\})$ we have $\mathbb{E}\mathfrak{C}_2(\ell_1) = \alpha_\ell \mathfrak{C}_2(\{\ell_1, \ell\})$, by Chernoff's bound the number of vertices in $C_2(\{\ell_1, \ell\})$ getting label ℓ_1 in our procedure satisfies

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

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 $(\alpha_\ell - 1/2)\mathfrak{C}_2(\{\ell_1, \ell\}) = \Omega(p^{1/2} (np)^{-\varepsilon_n} \mathfrak{C}_2(\ell_1))$. □

It remains to analyze the effect of the second round of ALAP over the vertices in Level 2.

Lemma 3.32. *There exists a constant $c_3 > 0$ such that a.a.s. the following holds: for every $\ell \in [k] \setminus \{\ell_1\}$, among $B \setminus B_1(\{\ell_1, \ell\})$ the number of vertices with label ℓ_1 is at least by $c_3\gamma_n n^{-1/2}\mathbb{E}\mathfrak{B}_2(\ell_1) \geq c_3\gamma_n p\sqrt{n}/2$ larger than the number of vertices with label ℓ after the second round 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 $\ell \in [2k]$, $\mathfrak{A}_1(\ell) \leq 2$. Indeed, for a vertex v in A, v has at least two neighbors in A with probability at most $\binom{2k}{2}p^2 = o(1/k)$, so by a union bound a.a.s. there is no such vertex. 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 3.30. 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 $\mathrm{Bin}(2knp/3, 1/k)$, and by Chernoff's bound $\mathbb{P}(W \leq np/2) \leq e^{-\Omega(np)}$. We conclude that a.a.s., our procedure attributes label ℓ_1 to at least $np/2$ vertices in Level 2, that is, $\mathfrak{B}_2(\ell_1) \geq np/2$.

Now, we show that a.a.s. for every $\ell \in [k] \setminus \{\ell_1\}$, among the vertices in $(B_2 \setminus B_1)(\{\ell_1, \ell\}) = B_2(\{\ell_1, \ell\}) \setminus B_1(\{\ell_1, \ell\})$ there are more vertices with label ℓ_1 than with label ℓ. Fix $\ell \in [k] \setminus \{\ell_1\}$. First, by the preceding paragraph a.a.s.

$$|(B_2 \setminus B_1)(\{\ell_1, \ell\})| \geq |B_1([k+1, 2k]) \cap B_2(\ell_1)| \geq np/2$$

for all $\ell \in [k] \setminus \{\ell_1\}$. We condition on the set $(B_2 \setminus B_1)(\{\ell_1, \ell\})$ and on the event $|(B_2 \setminus B_1)(\{\ell_1, \ell\})| \geq np/2$.

Given a vertex $v \in (B_2 \setminus B_1)(\{\ell_1, \ell\})$, recall that $|N(v) \cap B_1(\ell_1)|$ is distributed as $\mathrm{Bin}(\mathfrak{B}_1(\ell_1), p)$. Moreover, since $\mathfrak{A}_1(\ell) \leq 2$, $|N(v) \cap (A_1(\ell) \cup B_1(\ell))|$ is dominated by $\mathrm{Bin}(\mathfrak{B}_1(\ell) + 2, p)$. Hence, applying Remark 2.4 for $a_1 = \mathfrak{B}_1(\ell_1)$, $a_2 = \mathfrak{B}_1(\ell_1) + 2 - \gamma_n \geq \mathfrak{B}_1(\ell) + 2$, $X_1 \in \mathrm{Bin}(a_1, p)$ and $X_2 \in \mathrm{Bin}(a_2, p)$, we get that

$$\mathbb{P}(v \in B_2(\ell_1)) \geq \mathbb{P}(X_1 > X_2) + \frac{1}{2}\mathbb{P}(X_1 = X_2) = \frac{1}{2} + \Omega\left(\frac{\gamma_n p}{\sqrt{a_1 p}}\right) \geq \frac{1}{2} + 2c_3\gamma_n n^{-1/2},$$

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

We conclude that the number of vertices in $(B_2 \setminus B_1)(\{\ell_1, \ell\})$ receiving label ℓ_1 by our procedure dominates the sum of $|(B_2 \setminus B_1)(\{\ell_1, \ell\})| \geq np/2$ Bernoulli random variables with success probability $\frac{1}{2} + 2c_3\gamma_n n^{-1/2}$. Hence, by Chernoff's bound

$$\begin{aligned}
&\mathbb{P}\Big(|B_2(\ell_1) \setminus B_1(\{\ell_1, j\})| \leq \tfrac{1}{2}|(B_2 \setminus B_1)(\{\ell_1, j\})| + c_3\gamma_n n^{-1/2}\mathbb{E}|(B_2 \setminus B_1)(\{\ell_1, j\})|\Big) \\
&\leq \mathbb{P}\Big(|B_2(\ell_1) \setminus B_1(\{\ell_1, j\})| - \mathbb{E}|B_2(\ell_1) \setminus B_1(\{\ell_1, j\})| \leq -c_3\gamma_n n^{-1/2}\mathbb{E}|(B_2 \setminus B_1)(\{\ell_1, j\})|\Big) \\
&\leq \exp\Big(-\tfrac{c_3^2\gamma_n^2}{2n}\mathbb{E}|(B_2 \setminus B_1)(\{\ell_1, j\})|\Big) \leq \exp\Big(-\tfrac{1}{4}c_3^2\gamma_n^2 p\Big) \leq \exp\left(-\frac{c_3^2 np^2}{4(np)^{2\varepsilon_n}}\right) = o(1/n),
\end{aligned}$$

where the last equality follows from our assumption on p. The statement follows by taking a union bound over all $\ell \in [k] \setminus \{\ell_1\}$. □

It remains to analyze the number of vertices in $B_1(\{\ell, \ell_1\})$ that 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 $B_1(\{\ell, \ell_1\})$ getting label ℓ at the second round is a.a.s. smaller than the difference ensured by the previous lemma.

Lemma 3.33. *A.a.s. for every $\ell \in [k] \setminus \{\ell_1\}$, there are at most $\frac{4np}{k}\log n \le np^3(n\log n)^{1/2}$ vertices with label ℓ in $B_1(\{\ell_1, \ell\})$ after the second round of ALAP.*

Proof. By Lemma 3.23 and Chernoff's bound, a.a.s. $\mathfrak{B}_1(\ell) = np - (\ell-1)np^2 + O(\sqrt{np}\log n)$ for all $\ell \in [2k]$. Let us condition on this event. Now, we bound from above the number X_ℓ of vertices in Level 2 with label ℓ after the first round, which remain with label ℓ after the second round, that is, $X_\ell = |B_1(\ell) \cap B_2(\ell)|$. Fix $v \in B_1(\ell)$ and $j_1, j_2 \neq \ell$, and observe that

$$\frac{\mathbb{P}(v \in B_2(j_1))}{\mathbb{P}(v \in B_2(j_1)) + \mathbb{P}(v \in B_2(j_2))} = (1+o(1))\mathbb{P}(|N(v) \cap (A_1(j_1) \cup B_1(j_1))| \ge |N(v) \cap (A_1(j_2) \cup B_1(j_2))|),$$

which is equal to $1/2 + o(1)$ as long as

$$\sqrt{\mathbb{V}(\mathrm{Bin}(\mathfrak{A}_1(j_1) + \mathfrak{B}_1(j_1), p))} = (1+o(1))\sqrt{\mathfrak{B}_1(j_1)p} \gg |\mathfrak{B}_1(j_1) - \mathfrak{B}_1(j_2)|p,$$

that is, the standard deviation of the size of the neighborhoods (restricted to the vertices with labels j_1 and j_2) is of larger order than the expectation of their difference. Using that $|\mathfrak{B}_1(j_1) - \mathfrak{B}_1(j_2)| = |j_1 - j_2|np^2 + O(\sqrt{np}\log n)$, we conclude that for all integers $j_1, j_2 \le \frac{k}{\log n}$ different from ℓ, $\mathbb{P}(v \in B_2(j_1)) = (1+o(1))\mathbb{P}(v \in B_2(j_2))$.

On the other hand, since

$$|N[v] \cap (A_1(\ell) \cup B_1(\ell))| - 1 - \mathbb{1}_{v_\ell \in A_1(\ell)} \in \mathrm{Bin}(\mathfrak{B}_1(\ell) - 1 + \mathfrak{A}_1(\ell) - \mathbb{1}_{v_\ell \in A_1(\ell)}, p)$$

(taking into account that $v \in B_1(\ell)$ and that v_ℓ, which is neighbor of v, can still carry label ℓ after the first round) and the fact that $\sqrt{\mathfrak{B}_1(\ell)p} \gg 1$, we obtain that the probability that v gets label ℓ at the second round is, up to a $1+o(1)$ factor, at most the probability of getting any other label $j \le \frac{k}{\log n}$. Hence, the expectation of X_ℓ is bounded from above by $\frac{2np}{k}\log n \le \frac{1}{2}np^3(n\log n)^{1/2}$. The same bound holds for the expectation of the number of vertices $X_{\ell\to\ell_1}$ in Level 2 which change their label from ℓ to ℓ_1 at the second round, that is, $X_{\ell\to\ell_1} = |B_1(\ell) \cap B_2(\ell_1)|$.

Let us now show that for all ℓ, both X_ℓ and $X_{\ell\to\ell_1}$ are "close" to their expectations a.a.s. The argument is similar to Step 2 in the proof of Lemma 3.15 and will be presented only for X_ℓ, the reasoning for $X_{\ell\to\ell_1}$ being verbatim the same. Note that $X_\ell = \sum_{v\in B_1(\ell)} \mathbb{1}_{v\in B_2(\ell)}$. Then, we have that

$$\begin{aligned}\mathbb{V}(X_\ell) &= \sum_{u,v\in B_1(\ell)} \Big(\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)}\mathbb{1}_{v\in B_2(\ell)}] - \mathbb{E}[\mathbb{1}_{v\in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)}]\Big) \\ &= (1+o(1))\mathbb{E}[X_\ell] + \sum_{u,v\in B_1(\ell): u\neq v} \Big(\mathbb{E}[\mathbb{1}_{u\in B_2(\ell)}\mathbb{1}_{v\in B_2(\ell)}] - \mathbb{E}[\mathbb{1}_{u\in B_2(\ell)}]\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)}]\Big). \qquad (46)\end{aligned}$$

Using a transformation similar to the one from equations (25)- (29), 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

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

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

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

This implies that the general term in the sum in (46) rewrites as

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

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

Finally, to deduce the analogue of (30), we show that

$$\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \notin G_n] = (1+o(1))\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)} \mid uv \in G_n] = (1+o(1))\mathbb{E}[\mathbb{1}_{v\in B_2(\ell)}]. \tag{47}$$

Fix $j \in [k] \setminus \{\ell\}$. We prove that the probabilities

$$\frac{\mathbb{P}(v \in B_2(\ell) \mid uv \in G_n)}{\mathbb{P}(v \in B_2(\ell) \mid uv \in G_n) + \mathbb{P}(v \in B_2(j) \mid uv \in G_n)} \quad \text{and} \quad \frac{\mathbb{P}(v \in B_2(\ell))}{\mathbb{P}(v \in B_2(\ell)) + \mathbb{P}(v \in B_2(j))} \tag{48}$$

are the same up to a factor of $1 + o(1)$. Note that the comparison conditionally on the event $uv \notin G_n$ instead of $uv \in G_n$ is done in the same way, and will be sufficient for us to deduce (47).

Denote for simplicity $a = \mathfrak{A}_1(\ell) - \mathbb{1}_{v_\ell \in A_1(\ell)} + \mathfrak{B}_1(\ell) - 2$ and $b = |N(v) \cap V([\ell+1, 2k]) \cap A_1(j)| + \mathfrak{B}_1(j)$, which count the number of vertices in $A_1(\ell) \cup B_1(\ell)$ (respectively in $A_1(j) \cup B_1(j)$) to which v did not expose its edges conditionally on $uv \in G_n$.

Now, on the one hand, the left-hand side of (48) can be rewritten as

$$\mathbb{P}(\mathrm{Bin}(a,p) + 2 + \mathbb{1}_{v_\ell \in A_1(\ell)} > \mathrm{Bin}(b,p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(a,p) + 2 + \mathbb{1}_{v_\ell \in A_1(\ell)} = \mathrm{Bin}(b,p)), \tag{49}$$

while the right-hand side can be rewritten as

$$\begin{aligned}
&\mathbb{P}(\mathrm{Bin}(a+1,p) + 1 + \mathbb{1}_{v_\ell \in A_1(\ell)} > \mathrm{Bin}(b,p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(a+1,p) + 1 + \mathbb{1}_{v_\ell \in A_1(\ell)} = \mathrm{Bin}(b,p)) \qquad (50)\\
&= \mathbb{P}(\mathrm{Bin}(a,p) + Y + 1 + \mathbb{1}_{v_\ell \in A_1(\ell)} > \mathrm{Bin}(b,p)) + \frac{1}{2}\mathbb{P}(\mathrm{Bin}(a,p) + Y + 1 + \mathbb{1}_{v_\ell \in A_1(\ell)} = \mathrm{Bin}(b,p)),
\end{aligned}$$

where in the second line Y is a Bernoulli random variable with parameter p independent from everything else. Then, if $Y = 1$, (49) and (50) coincide, while if $Y = 0$, it is sufficient to show that both

$$\mathbb{P}(\mathrm{Bin}(a,p) + 2 + \mathbb{1}_{v_\ell \in A_1(\ell)} = \mathrm{Bin}(b,p)) \quad \text{and} \quad \mathbb{P}(\mathrm{Bin}(a,p) + 1 + \mathbb{1}_{v_\ell \in A_1(\ell)} = \mathrm{Bin}(b,p))$$

are of order $o(\mathbb{P}(\mathrm{Bin}(a,p)+2+\mathbb{1}_{v_\ell \in A_1(\ell)} > \mathrm{Bin}(b,p)))$. This is satisfied since, on the one hand, $|b - \mathfrak{B}_1(j)| \le 2$ and therefore a.a.s. $\mathrm{Bin}(b,p) \in np^2 - (j-1)np^3 \pm \sqrt{np^2}\log n$, and on the other hand, $|a - \mathfrak{B}_1(\ell)| \le 2$ and for every $s \in np^2 - (j-1)np^3 \pm \sqrt{np^2}\log n$ and every integer m, we have that

$$\begin{aligned}
\mathbb{P}(\mathrm{Bin}(a,p) = s - m) &= \binom{a}{s-m} p^{s-m} q^{a-(s-m)}\\
&= \binom{a}{s-m-1} p^{s-m-1} q^{a+1-(s-m)} \frac{(a+1-(s-m))p}{(s-m)q}\\
&= (1+o(1))\mathbb{P}(\mathrm{Bin}(a,p) = s-m-1).
\end{aligned}$$

By applying a similar reasoning conditionally on $uv \notin G_n$, we conclude that $\mathbb{V}(X_\ell) = (1+o(1))(\mathbb{E}X_\ell + o(p(\mathbb{E}X_\ell)^2))$, where the latter term dominates the former by our assumption that $np \ge n^{5/8+\varepsilon}$. Finally, recalling that $\mathbb{E}X_\ell \le \frac{2np}{k}\log n$,

$$\mathbb{P}\Big(X_\ell \ge \frac{4np}{k}\log n\Big) \le \frac{\mathbb{V}(X_\ell)}{(\mathbb{E}X_\ell)^2} = o(p).$$

Taking a union bound over all $\ell \in [k] \setminus \ell_1$ finishes the proof of the lemma. $\square$

Remark 3.34. The same argument (up to minor modifications in the definitions of a and b in the proof of Lemma 3.33 due to v possibly being in a different basin) shows that a.a.s. for all $\ell \in [k] \setminus \{\ell_1\}$, $\mathfrak{B}_2(\ell) \le 8np\log n$. Indeed, for $\ell \le \frac{k}{\log n}$, the argument of the proof of Lemma 3.33 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 3.35. *Let $(\Omega_\ell)_{\ell=1}^k$ be subsets of B such that for every $\ell \in [k]\setminus\{\ell_1\}$, $|\Omega_\ell| - |\Omega_{\ell_1}| \le 2np^3\sqrt{n\log n}$ and $|\Omega_\ell| \le 8np\log n$. Then, a.a.s. for every vertex v in C and every $\ell \in [k]$, it holds that $|N_v(\Omega_\ell)| - |N_v(\Omega_{\ell_1})| \le 20\sqrt{np^2\log n}$.*

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

$$\begin{aligned}|N_v(\Omega_\ell)| &\le p|\Omega_\ell| + \sqrt{5p\cdot 8np\log n} = p|\Omega_\ell| + \sqrt{40np^2\log n},\\ |N_v(\Omega_{\ell_1})| &\ge p|\Omega_{\ell_1}| - \sqrt{5p\cdot 8np\log n} = p|\Omega_{\ell_1}| - \sqrt{40np^2\log n},\end{aligned}$$

and therefore also

$$|N_v(\Omega_\ell)| - |N_v(\Omega_{\ell_1})| \le p(|\Omega_\ell| - |\Omega_{\ell_1}|) + 2\sqrt{40np^2\log n} \le 20\sqrt{np^2\log n},$$

where we used that $np^4(n\log n)^{1/2} \ll \sqrt{np^2\log n}$. The lemma follows by a union bound over the complementary events for all $v \in C$ and $\ell \in [k]\setminus\{\ell_1\}$. ☐

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 3.33 and Remark 3.34 the assumptions of Lemma 3.35 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 $20\sqrt{np^2\log n}$ for any vertex (note that based on the conclusion of Lemma 3.35 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, using that

$$\mathfrak{C}_2(\ell_1) = |\Omega_{\ell_1}| \ge \frac{n - \mathfrak{A} - \mathfrak{B}}{k} - 2np^3\sqrt{n\log n} = (1-o(1))\frac{n}{k},$$

an application of Chernoff's bound shows (for positive real numbers $(\varepsilon_n)_{n\ge1}$ chosen as before the statement of Lemma 3.29) that

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

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

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

Using Lemma 3.31, we conclude that with probability $1 - o(\frac{1}{kn})$, for every vertex $v \in C$, the number of neighbors of v with label ℓ_1 after round 2 that are in Level 3 is at least by

$$\begin{aligned}&\mathbb{E}Y_{\ell_1} - \mathbb{E}Y_\ell - 2p^{1/2}(np)^{-2\varepsilon_n}\mathbb{E}Y_{\ell_1} - O(\sqrt{np^2(\log n)^2})\\ =\;& \Omega(p^{1/2}(np)^{-\varepsilon_n}\mathbb{E}Y_{\ell_1}) - 2p^{1/2}(np)^{-2\varepsilon_n}\mathbb{E}Y_{\ell_1} - O(\sqrt{np^2(\log n)^2}) = \Omega(p^{1/2}(np)^{-\varepsilon_n}\mathbb{E}Y_{\ell_1})\end{aligned}$$

larger than the number of neighbors with label ℓ, where the last equality uses that $\mathbb{E}Y_{\ell_1} \ge \frac{np}{2k}$ and that $np \ge n^{5/8+\varepsilon}$. 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 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 3.32 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 [18, 23] 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 [18] 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.

Acknowledgements

The authors thank Ravi Sundaram for calling to their attention the lack of a complete mathematically rigorous understanding of label propagation algorithms, and Yoshiharu Kohayakawa for referring us to [18].

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[15] S. Janson, T. Łuczak, and A. Ruciński. *Random graphs*, volume 45. John Wiley & Sons, 2011.

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

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

[18] C. Knierim, J. Lengler, P. Pfister, U. Schaller, and A. Steger. The maximum label propagation algorithm on sparse random graphs. *APPROX-RANDOM, Leibniz International Proceedings in Informatics*, 58:1–15, 2019.

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

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

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

[22] M. Newman. *Networks*. Oxford university press, 2018.

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

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

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

[26] A. Sah and M. Sawhney. Majority dynamics: The power of one, 2021.

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

[28] R. Tamir. Fast Convergence to Unanimity in Dense Erdős-Rényi graphs, 2022.

[29] L. Tran and V. Vu. Reaching a consensus on random networks: The power of few. In J. Byrka and R. Meka, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM*, volume 176 of *LIPIcs*, pages 20:1–20:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

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

[31] A. N. Zehmakan. Opinion forming in erdős–rényi random graph and expanders. *Discrete Applied Mathematics*, 277:280–290, 2020.