Bipartite independence number and balanced coloring

Debsoumya Chakraborti*

Discrete Mathematics Group, Institute for Basic Science (IBS), Daejeon, South Korea

Email: debsoumya@ibs.re.kr

July 7, 2021

Abstract

In this paper, we establish a couple of results on extremal problems in bipartite graphs. Firstly, we show that every sufficiently large bipartite graph with average degree $\Delta$ and with $n$ vertices on each side has a balanced independent set containing $(1 - \epsilon) \frac{\log \Delta}{\Delta} n$ vertices from each side for small $\epsilon > 0$. Secondly, we prove that the vertex set of every sufficiently large balanced bipartite graph with maximum degree at most $\Delta$ can be partitioned into $(1 + \epsilon) \frac{\Delta}{\log \Delta}$ balanced independent sets. Both of these results are algorithmic and best possible up to a factor of 2, which might be hard to improve as evidenced by the phenomenon known as ‘algorithmic barrier’ in the literature. The first result improves a recent theorem of Axenovich, Sereni, Snyder, and Weber in a slightly more general setting. The second result improves a theorem of Feige and Kogan about coloring balanced bipartite graphs.

1 Introduction

This paper first deals with a bipartite analogue of the Turán’s theorem [39] for complete graphs, which is regarded as a cornerstone of extremal graph theory (see, e.g., [22] for a survey). Next, we discuss a bipartite analogue of the celebrated Johansson-Molloy Theorem on the chromatic number of a triangle free graph with given maximum degree (see, e.g., [9], [33], and [34]). Some seemingly simple problems in the bipartite setting (such as finding the smallest possible ‘bipartite independence number’ of a bipartite graph with maximum degree three) are not yet resolved despite some effort (see, e.g., [3] and [14]). In this paper, we address a few of such problems.

Suppose that we are given a bipartite graph $G = (U \cup V, E)$ with a prescribed vertex bipartition $(U, V)$ and edge set $E$. A balanced bipartite independent set (or bi-hole) of size $t$ in $G$ is a pair $(X,Y)$ where $X \subseteq U$ and $Y \subseteq V$ such that $|X| = |Y| = t$ and there are no edges in $E$ with one endpoint in $X$ and the other in $Y$. The size of the largest bi-hole, referred to as bipartite independence number, can be viewed as a natural bipartite analogue of the standard independence number. Our first main result is the following.

*This work was supported by the Institute for Basic Science (IBS-R029-C1)
Theorem 1.1. For each $\epsilon > 0$, there exists $\Delta_0 = \Delta_0(\epsilon)$ such that the following holds. For each $\Delta \geq \Delta_0$, there is $N_0 = N_0(\Delta)$ such that if $G$ is a balanced bipartite graph with average degree $\Delta \geq \Delta_0$ and with $n \geq N_0$ vertices on each side, then $G$ contains a bi-hole of size $(1 - \epsilon)\frac{\log \Delta}{\Delta} n$.

We next turn our attention to the bipartite analogue of the standard notion ‘chromatic number’. A coloring of the vertices of a bipartite graph $G$ is called balanced if each of the color classes induces a bi-hole. The coloring number, $\chi_B(G)$, is defined to be the minimum number of colors needed for a balanced coloring of a given bipartite graph $G$. We now state our second main result.

Theorem 1.2. For each $\epsilon > 0$, there exists $\Delta_0 = \Delta_0(\epsilon)$ such that the following holds. For each $\Delta \geq \Delta_0$, there is $N_0 = N_0(\Delta)$ such that if $G$ is a balanced bipartite graph with maximum degree $\Delta \geq \Delta_0$ and with $n \geq N_0$ vertices on each side, then $\chi_B(G) \leq (1 + \epsilon)\frac{\log \Delta}{\Delta} n$.

Theorem 1.3 (see [3]). For each $0 < \epsilon < 1$, there exists $\Delta_0 = \Delta_0(\epsilon)$ such that the following holds. If $G$ is a balanced bipartite graph with maximum degree $\Delta \geq \Delta_0$ and with $n \geq (1 + \epsilon)2\Delta$ vertices on each side, then $\chi_B(G) \leq 20\Delta \cdot \frac{\log \Delta}{\Delta} n$.

Theorem 1.4 (see [19]). For each $0 < \epsilon < 1$, there exists $\Delta_0$ such that the following holds. If $G$ is a balanced bipartite graph with maximum degree $\Delta \geq \Delta_0$ and with $n \geq (1 + \epsilon)2\Delta$ vertices on each side, then $\chi_B(G) \leq \frac{20\Delta}{\epsilon^2 \log \Delta}$. 

Their upper bound comes from considering the random bipartite graph $G_{n,n,\Delta/n}$ (the random bipartite graph $G_{n,n,p}$ is a bipartite graph with $n$ vertices on each side where each of the possible $n^2$ edges are present independently with probability $p$). Our proof of Theorem 1.3 is algorithmic and matches the best bound that can be achieved by an efficient algorithm to find a large bi-hole of $G_{n,n,\Delta/n}$. We elaborate in the concluding remarks why further improving this seems hard.

Returning to the balanced coloring of bipartite graphs, Feige and Kogan [19] observed that the coloring number of bipartite graphs behaves quite differently from the usual chromatic number of graphs. For example, removing an independent set from a graph never increases its chromatic number. However, removing a bi-hole from a bipartite graph may increase its coloring number. In fact, the remaining graph may not have a balanced coloring at all. This behavior poses some challenges in estimating coloring number in general. Our Theorem 1.2 improves the following result of Feige and Kogan [19].
We essentially removed the factor of $\frac{20}{\Delta}$ from the above result. Our proof of Theorem 1.2 is algorithmic and gives a bound that is best possible up to a factor of 2 (one can easily get a lower bound of $\frac{\Delta}{(2+\epsilon)\log \Delta}$ by using Theorem 1.3). Again, for this coloring problem, our bound matches the best known bound that can be achieved by an efficient algorithm in the random bipartite graph $G_{n,n,\Delta/n}$.

We observe that one cannot strengthen the bounded maximum degree to a bounded average degree condition in Theorem 1.2. This can be easily seen from the following fact: If a balanced bipartite graph $G$ with $2n$ vertices contains a vertex $v$ with degree $n$ (i.e., $v$ is connected by edge with all the vertices from the opposite partition), then $G$ does not have a balanced coloring.

This paper is organized in the following way. We start with a list of preliminary tools in the next section that will be helpful throughout the paper. We give a proof of Theorem 1.1 in Section 3 by analyzing a natural randomized algorithm to find a large bi-hole in a given bipartite graph. We next give a more sophisticated randomized algorithm in Section 4 to bound the coloring number of a balanced bipartite graph with bounded degree to prove Theorem 1.2. This proof uses several technical claims which will be proved in the subsequent section. Finally, we end with a few concluding remarks in Section 6, where we elaborate some of the points from the introduction.

2 Preliminaries

We start with a remark that Ehard, Mohr, and Rautenbach [14] gave an easy proof of Theorem 1.1 with a worse bound of $\frac{\log \Delta}{8\Delta}n$. We next state a couple of simple results regarding balanced coloring from the existing literature, which will be useful to us later.

Observation 2.1. [19] A bipartite graph $G$ has a balanced coloring if and only if the bipartite complement of $G$ contains a perfect matching.

Lemma 2.2. [8] If $G$ is a balanced bipartite graph with maximum degree $\Delta$ and $n \geq 2\Delta$ vertices on each side, then $\chi_B(G) \leq 2\Delta + 1$.

This lemma gives a weaker upper bound on $\chi_B$ for Theorem 1.2. Although it appeared in [8], we still prove it to keep our paper self-contained.

Proof of Lemma 2.2. Let $G$ be a bipartite graph $G$ with maximum degree $\Delta$ and $n \geq 2\Delta$ vertices on each side. Consider the bipartite complement $G'$ of $G$. Now using the fact that $G'$ has minimum degree at least $n - \Delta$ and $n \geq 2\Delta$, we deduce that the Hall’s conditions hold for $G'$. Thus, by Observation 2.1, $G$ has a balanced coloring. Now, let $M = \{e_1, e_2, \ldots, e_n\}$ be a perfect matching of $G'$. We now show that we can greedily color the vertices of $G$ using $2\Delta + 1$ colors so that both the vertices corresponding to each edge of $M$ gets the same color. Indeed, let we have already colored the vertices corresponding to $e_1, e_2, \ldots, e_t$ for some $t < n$. Now, the total number of neighbors of the vertices $u, v$ in $e_{t+1}$ is at most $2\Delta$, thus there must be at least one color left which is used in none of the neighbors of $u$ and $v$. We can use that color for both $u$ and $v$. Thus, each color appears the same number of times in both parts, proving Lemma 2.2. □
We need some probabilistic tools to prove Theorems 1.1 and 1.2. We start with a few of the most frequently used probabilistic bounds.

**Lemma 2.3 (Markov’s Inequality).** If $X$ is a nonnegative random variable and $t > 0$, then,

$$P[X \geq t] \leq \frac{E(X)}{t}.$$  

**Lemma 2.4 (Chebyshev Inequality).** If $X$ is a random variable with a finite mean and variance, then, for $t > 0$,

$$P[|X - E(X)| \geq t] \leq \frac{\text{Var}(X)}{t^2}.$$  

We next state the Chernoff bound, due to Chernoff [12] and Okamoto [36]. We use the version stated by Janson [28, Theorem 1].

**Lemma 2.5 (The Chernoff bound).** Let $X = \sum_{i=1}^{n} X_i$, where $X_i$ are independent Bernoulli variables with $P[X_i = 1] = p_i$. Let $\mu = E(X) = \sum_{i=1}^{n} p_i$. Then for $t \geq 0$,

1. $P[X \geq \mu + t] \leq e^{-\frac{t^2}{3\mu + \frac{3}{2}t}}$ and
2. $P[X \leq \mu - t] \leq e^{-\frac{t^2}{2\mu}}$.

We also need a recent extension [23] of Chernoff bounds to the case when some dependencies between the random variables are allowed. We use the version due to Jukna [29]. To state it, we need the following definition.

**Definition.** A family $Y_1, \ldots, Y_r$ of random variables is read-$k$ if there exists a sequence $X_1, \ldots, X_m$ of independent random variables, and a sequence $S_1, \ldots, S_r$ of subsets of $[m] = \{1, \ldots, m\}$ such that

- each $Y_i$ is some function of $(X_i : j \in S_i)$, and
- no element of $[m]$ appears in more than $k$ of the $S_i$’s.

**Theorem 2.6 (Chernoff bound for dependent random variables, [29]).** Let $Y_1, \ldots, Y_r$ be a family of read-$k$ indicator variables with $P[Y_i = 1] = p_i$, and let $p$ be the average of $p_1, \ldots, p_r$. Then for any $\epsilon > 0$,

$$P[|(Y_1 + \cdots + Y_r) - pr| \geq \epsilon r] \leq 2e^{-2\epsilon^2 r/k}.$$  

We use the asymmetric version of the local lemma [16]. We state the version from [2].

**Lemma 2.7 (The local lemma, [2]).** Let $A_1, \ldots, A_n$ be events in an arbitrary probability space. A directed graph $D = (V, E)$ on the set of vertices $V = [n]$ is called a dependency digraph for the events $A_1, \ldots, A_n$ if for each $i, 1 \leq i \leq n$, the event $A_i$ is mutually independent of all the events $\{A_j : (i, j) \not\in E\}$. Suppose that $D = (V, E)$ is a dependency digraph for the above events and suppose there are real numbers $x_1, \ldots, x_n$ such that $0 \leq x_i < 1$ and $P[A_i] \leq x_i \prod_{(i,j) \in E} (1 - x_j)$ for all $1 \leq i \leq n$. Then, with positive probability no event $A_i$ holds.
We want to mention that there are algorithmic versions of local lemma (see, e.g., [35] and [37]). Thus, when we use the local lemma, we can have an efficient randomized algorithm to get the desirable choice of events. This point will not be further discussed inside the proofs.

Throughout the paper, we omit the use of floor and ceiling signs for the sake of clarity of presentation. For an event $A_n$ that depends on $n$, we say that $A_n$ occurs ‘w.h.p.’, if the probability of $A_n$ tends to one as $n$ tends to infinity.

3 Finding large bipartite independent sets

Let $G = (U \cup V, E)$ be an $n$ by $n$ bipartite graph with $|E| = \Delta n$. First, remove exactly $\epsilon^2 n$ vertices from both side to make sure that the maximum degree of the induced graph on the remaining vertices is at most $\Delta^{\epsilon/2}$. Thus, it is enough to prove Theorem 1.1 with the extra assumption that the maximum degree of the underlying graph is at most $\Delta^{\epsilon/2}$. This will be crucial in applying certain concentration bounds while analyzing our randomized algorithm.

We can assume that $0 < \epsilon < 1/10$. Throughout the proof, wherever needed, we will use that $\Delta$ is sufficiently large with respect to $\epsilon$ and $n$ is sufficiently large with respect to $\Delta$.

The algorithm is very simple and natural. First, we pick the vertices in $U$ independently with probability $(1 - \epsilon/2) \log \Delta/\Delta$. Let $U'$ denote the set of all the vertices picked from $U$. Let $V'$ denote the set of vertices in $V$ which do not have any neighbor in $U'$. To prove Theorem 1.1 it is enough to show that the sizes of $|U'|$ and $|V'|$ are both at least $(1 - \epsilon) \log \Delta/\Delta n$ with positive probability. These are shown in the following couple of claims.

Claim 3.1. W.h.p., we have that $|U'| \geq (1 - \epsilon) \log \Delta/\Delta n$.

Proof. Let $X_u$ denote the indicator random variable for the event that the vertex $u \in U$ is picked. It is clear that $|U'| = \sum_{u \in U} X_u$. A straightforward application of the Chernoff bound (i.e., Lemma 2.5) yields our claim.

Claim 3.2. W.h.p., we have that $|V'| \geq (1 - \epsilon) \log \Delta/\Delta n$.

Proof. For each vertex $v \in V$, let $Y_v$ denote the indicator random variable for the event that the vertex $v \in V$ is picked. It is clear that $|V'| = \sum_{v \in V} Y_v$. We first compute the expected size of $|V'|$. For each $v \in V$, the probability that none of its neighbors are picked is exactly $(1 - \epsilon/2) \log \Delta/\Delta)^{d(v)}$, where $d(v)$ is the degree of $v$. Now, using Jensen’s inequality, we have the following.

$$
\mathbb{E}(|V'|) = \sum_{v \in V} \left(1 - \epsilon/2\right)^{\log \Delta/\Delta}^{d(v)} \geq n \left(1 - \epsilon/2\right)^{\log \Delta/\Delta} \Delta \\
\geq ne^{-(1-\epsilon/4) \log \Delta} \\
= \frac{n}{\Delta^{1-\epsilon/4}}.
$$

We next use Theorem 2.4 to show concentration of the random variable $|V'|$. We claim that the family of random variables $\{Y_v : v \in V\}$ is read-$\Delta^{\epsilon/2}$. It is clear by observing the following facts.
• $X_u, u \in U$ are independent random variable,
• for each $v \in V$, $Y_v$ is a function of $(X_u : u \in N(v))$, and
• no vertex $u \in U$ is adjacent to more than $\frac{\Delta}{\epsilon}$ vertices in $V$.

Thus, a straightforward application of Theorem 2.6 on the random variables $Y_v, v \in V$ shows us that $P[|V'| \leq (1 - \epsilon)\frac{\log \Delta}{\Delta}n] \leq e^{-\Omega(\Delta)}$. This finishes the proof of Theorem 1.1.

4 Balanced colorings of bipartite graphs

In this section, we prove Theorem 1.2 through a series of claims. We later prove these claims in the next section.

**Proof of Theorem 1.2.** We can assume that $0 < \epsilon < \frac{1}{10}$. Let $G = (U \cup V, E)$ be an $n$ by $n$ bipartite graph with maximum degree $\Delta$. Similar to the previous section, wherever needed, we use that $\Delta$ is sufficiently larger with respect to $\epsilon$ and $n$ is sufficiently large with respect to $\Delta$. Fix a set $Q$ of $q = (1 + \epsilon/2)\frac{\Delta}{\log \Delta}$ colors. We first color the vertices in $U$ independently and uniformly at random with the colors in $Q$. We obtain the following fact by a simple application of the Chernoff bound similar to the proof of Claim 3.1 (we omit the details).

**Claim 4.1.** W.h.p., for every color $c \in Q$, the number of vertices in $U$ with color $c$, denoted by $|U_c|$, satisfies that $nq - o(n) \leq |U_c| \leq nq + o(n)$.

Next, we assign a set $Q_v \subseteq Q$ of available colors to each $v \in V$. Let $C_v$ denote the set of all colors that are already used by some neighbor of $v$. We set $Q_v = Q \setminus C_v$. We now color each $v \in V$ independently and uniformly at random with the colors in $Q_v$. Note that some of the vertices $v \in V$ may remain uncolored, if the corresponding set of available colors $Q_v$ is empty. However, we will show that this does not happen for too many vertices in $V$.

**Claim 4.2.**

1. W.h.p., for every pair of colors $c_1, c_2 \in Q$, the numbers of vertices in $V$ with color $c_1$ and the numbers of vertices in $V$ with color $c_2$ differ by $o(n)$.

2. W.h.p., for every color $c \in Q$, the number of vertices in $V$ with color $c$ is at least $\left(1 - \frac{\log \Delta}{\epsilon^2 \log \Delta} \right) \frac{n}{q}$.

Our strategy is to finish by coloring all the uncolored vertices in $V$ and recoloring some of the vertices in $U$ and $V$ (to make the coloring balanced) by the remaining $\frac{\epsilon^2 \log \Delta}{2}$ colors which are not in $Q$. To this end, we denote by $S = S_V$ the set of all the uncolored vertices in $V$. If the size of $S$ is small, then we can greedily finish the coloring as demonstrated next.

Suppose that $|S| \leq \frac{n}{32}$. Denote by $U_c$ (analogously $V_c$) the set of all vertices in $U$ ($V$) that are colored with $c$. By Claims 4.1 and 4.2 we have the following.

$$|U_c| - |V_c| \geq -o(n). \quad (4.1)$$
For every color $c \in Q$, if $|U_c| < |V_c|$, then arbitrarily uncolor some vertices of $V_c$ to make sure that the number of vertices colored with $c$ in both parts is exactly $|U_c|$ (this step is necessary to make sure every color class contains the same number of vertices from $U$ and $V$). After this step, update the sets $V_c$, i.e., throw away the vertices from $V_c$ that got uncolored (note that $|U_c| \geq |V_c|$ for all $c \in Q$ at this stage). Due to (4.1), we have uncolored at most $o(n)$ vertices of $V$, denote by $S_0$ the set of all vertices that got uncolored. Let $S' = S \cup S_0$, clearly $|S'| \leq \frac{2n}{\Delta^2}$. We now wish to color all the vertices in $S'$ and recolor some vertices of $U$ with a new color $c^*$. More precisely, for every color $c \in Q$, we recolor exactly $|U_c| - |V_c|$ vertices of $U_c$ by using $c^*$. To do this, the only thing we need to verify is that there are at least $|U_c| - |V_c|$ vertices in $U_c$ that do not have any neighbor in $S'$. Indeed, the number of vertices in $U$ with at least one neighbor in $S'$ is at most $\frac{2n}{\Delta}$, and we have that $\frac{2n}{\Delta} < |V_c|$ by Claim 4.2 and the assumption that $|S| \leq \frac{n}{\Delta^2}$. Thus, we have successfully colored $G$ with $q + 1$ colors such that every color class induces a bi-hole.

Thus, from now on, we assume that $|S| \geq \frac{n}{\Delta^2}$. In this case, we desire to get a set $S_U \subset U$ with the same size as $S$ (remember that we want a balanced coloring) such that the maximum degree of the graph induced by $(S_U, S)$ is small enough to apply Lemma 2.2 and finish the coloring using the remaining $\frac{\Delta}{2 log \Delta}$ colors not in $Q$. To achieve this, We start by showing that very few vertices of $U$ have many neighbors in $S$.

Claim 4.3. W.h.p., for every color $c \in Q$, at most $\frac{100n \sqrt{\log \Delta}}{\epsilon^2 \Delta}$ of the vertices $u$ in $U$ satisfies the following two properties.

- $u$ is colored with $c$ and
- $u$ has more than $\frac{\Delta}{\log^{3/2} \Delta}$ neighbors in $S$.

Suppose now, we fix an instance satisfying all the high probability events. Denote by $U^*_c$ the set of all vertices in $U_c$ with at most $\frac{\Delta}{\log^{3/2} \Delta}$ neighbors in $S$. By Claims 4.1 and 4.3, $|U^*_c| \geq \frac{n \log \Delta}{2 \Delta}$. Let $a_c = |U_c| - |V_c|$. By Claims 4.1 4.2 and the assumption that $|S| \geq \frac{n}{\Delta^2}$, we have that

$$0 \leq a_c \leq \frac{100}{\epsilon^2 \log^2 \Delta} \cdot \frac{n}{q} + o(n). \tag{4.2}$$

We next show that we can choose exactly $a_c$ vertices from $U^*_c$ for all $c \in Q$ to form $S_U$ so that no vertices from $S$ has more than $\frac{\Delta}{\log^{3/2} \Delta}$ neighbors in $S_U$.

Claim 4.4. There exists $S_U$ consisting of exactly $a_c$ vertices of $U^*_c$ for all $c$, such that the balanced graph induced by $(S_U, S)$ has maximum degree at most $\frac{\Delta}{\log^{3/2} \Delta}$.

Finally, it follows from Lemma 2.2 and the fact that $|S_U| = |S_V| \geq \frac{n}{\Delta^2} > \frac{2\Delta}{\log^{3/2} \Delta}$ that there is a balanced coloring of the induced graph on $(S_U, S_V)$ by the remaining $\frac{\Delta}{2 \log \Delta}$ colors that are not used yet. This finishes the proof of Theorem 1.2 modulo the claims.

□
5 Proofs of intermediate claims

In this section, we complete the proof of Theorem 1.2 by showing the validity of the claims of the last section.

Proof of Claim 4.2. For every color $c \in Q$, let $Z_c$ be the random variable denoting the number of vertices in $V$ with color $c$. Define $Z = \sum_{c \in Q} Z_c$. Observe that $Z = \sum_{v \in V} I_v$, where $I_v$ is the indicator random variable for the set $Q_v$ being non-empty. Hence,

$$E(Z) = \sum_{v \in V} E(I_v) = \sum_{v \in V} P[Q_v \neq \emptyset].$$

(5.1)

For each vertex $v \in V$, the probability that $Q_v$ is empty is the same as the probability that all the colors of $Q$ appear in the neighborhood of $v$. To estimate this probability, consider the following process: Start with an empty set $S_0 = \emptyset$, then at each time step $t > 0$, we generate a uniformly random color $c_t$ from $Q$ independently of previous choices and define $S_t = S_{t-1} \cup \{c_t\}$ (note that this is a set, hence even if a color comes more than once, it appears only once in $S_t$). Define $T$ to be the random variable that counts the minimum number of time step $t$ such that $|S_t| = q$. Now, observe that

$$P[Q_v = \emptyset] = P[T \leq d(v)] \leq P[T \leq \Delta].$$

(5.2)

The random variable $T$ is well-studied and estimating it is known as ‘coupon collector’s problem’ in the literature (see, e.g., [31]). To keep our paper self-contained, we estimate the lower tail of $T$ by a simple application of Chebyshev inequality.

Lemma 5.1. $P[T \leq \Delta] < \frac{50}{\epsilon^2 \log^2 \Delta}$.

Proof. For each $1 \leq j \leq q$, we define the random variables $T_j$ denoting the minimum number of time step $t$ such that $|S_t| = j$ (define $T_0 = 0$). Clearly, $T_q = T$. Note that the random variable $T_j - T_{j-1}$ denotes the time needed for a new color to be added in our collection as $j$-th color. Thus, $T_j - T_{j-1}$ has a geometric distribution with probability $\frac{q-j+1}{q}$. Remember that a random variable with geometric distribution with probability $p$ has expectation $\frac{1}{p}$ and variance $\frac{1-p}{p^2}$. It follows that

$$E(T) = \sum_{j=1}^{q} E(T_j - T_{j-1}) = \sum_{j=1}^{q} \frac{q}{q-j+1} \geq q \int_{1}^{q+1} \frac{1}{x} dx \geq q \log q.$$

(5.3)

Furthermore, observe that the random variables $T_j - T_{j-1}, j \in [q]$ are independent and thus, we have the following.

$$\text{Var}(T) = \sum_{j=1}^{q} \text{Var}(T_j - T_{j-1}) \leq \sum_{j=1}^{q} \frac{q^2}{(q-j+1)^2} \leq q^2 \left(1 + \int_{1}^{q} \frac{1}{x^2} dx\right) < 2q^2.$$

(5.4)

Using (5.3), (5.4), and Lemma 2.4, we have the following.
For each application of Theorem 2.6, we can finish the proof of Claim 4.2. 
\[ \sum_v \] is colored with \( c \) denoting the color chosen for \( u \).

The number of vertices \( V \) \( \subseteq \mathbb{Z} \) of Claim 4.2, we use Theorem 2.6 to show the concentration of each \( Z_c \) around its mean.

Thus, by (5.1), (5.2), and Lemma 5.1, we have the following.

For every color \( c \in Q \), let \( Y_v \) be the indicator random variable for the event that \( v \) is colored with \( c \). Clearly, \( Z_c = \sum_v Y_v \). To apply Theorem 2.6, we wish to show that the family of random variables \( \{ Y_v : v \in V \} \) is read-\( \Delta \). For \( u \in U \), let \( X_u \) be the random variable denoting the color chosen for \( u \). In order to model the random variables \( Y_v \) conveniently, for \( v \in V \), let \( X'_v \) be independent random variables with continuous uniform distribution on the interval \([0, 1]\). For the convenience of our analysis, we now specify how we assign colors to \( U \).

Thus, the family of random variables \( \{ Y_v : v \in V \} \) is read-\( \Delta \). Finally, by a simple application of Theorem 2.6, we can finish the proof of Claim 4.2.

**Proof of Claim 4.3.** For every color \( c \in Q \), let \( Z_c \) be the random variable denoting the number of vertices \( u \in U \) with color \( c \) and more than \( \frac{\Delta}{\log^{3/2} \Delta} \) neighbors in \( S \). Define \( Z = \sum_{c \in Q} Z_c \). Observe that \( Z = \sum_{u \in U} A_u \), where \( A_u \) is the indicator random variable for the event that \( u \) has more than \( \frac{\Delta}{\log^{3/2} \Delta} \) neighbors in \( S \). For \( u \in U \), define the random variable \( B_u = \sum_{v \in N(u)} I^c_v \), where \( I^c_v \) is the indicator random variable for the set \( Q_v \) being empty. Thus, for each \( u \in U \), we have that \( A_u = 1 \) if and only if \( B_u > \frac{\Delta}{\log^{3/2} \Delta} \). Now, using (5.2) and Lemma 5.1, we have the following.

\[
\mathbb{E} (B_u) = \sum_{v \in N(u)} \mathbb{E} (I^c_v) = \sum_{v \in N(u)} \mathbb{P} [Q_v = \emptyset] < \frac{50 \Delta}{\epsilon^2 \log^2 \Delta}. \tag{5.5}
\]

Thus, by (5.5) and a simple application of Markov’s inequality (Lemma 2.3), we have the following.

\[
\mathbb{E} (A_u) = \mathbb{P} [A_u = 1] = \mathbb{P} \left[ B_u > \frac{\Delta}{\log^{3/2} \Delta} \right] < \frac{50}{\epsilon^2 \log^{1/2} \Delta}.
\]
Thus, $\mathbb{E}(Z) = \sum_{u \in U} \mathbb{E}(A_u) < \frac{50n}{c^2 \log^{1/2} \Delta}$. By symmetry, every $Z_c$ has the same distribution. Hence, by the linearity of expectation, we have that $\mathbb{E}(Z_c) = \frac{\mathbb{E}(Z)}{q} < \frac{50n \log^{1/2} \Delta}{q^2 \Delta^2}$. We next complete the proof of our claim by using Theorem 2.6 to show the concentration of each $Z_c$ around its mean.

Fix a color $c \in Q$. For $u \in U$, let $\mathcal{Y}_u$ be the indicator random variable for the event that $u$ has color $c$ and $u$ has more than $\frac{\Delta}{\log^{1/2} \Delta}$ neighbors in $S$. Clearly, $Z_c = \sum_{u \in U} \mathcal{Y}_u$. We now wish to show that the family of random variables $\{\mathcal{Y}_u : u \in U\}$ is read-$(\Delta^2 + 1)$. Remember that $X_u$ is the random variable denoting the color of $u \in U$. For convenience, for $u \in U$, define $\Gamma(u)$ to be the set of all vertices in $U$ at distance exactly two from $u$. Now, observe the following:

- $\{X_u : u \in U\}$ are independent random variables,
- for each $u \in U$, the random variable $\mathcal{Y}_u$ is a function of $X_u$ and $(X_{u'} : u' \in \Gamma(u))$, and
- for each $u \in U$, the random variable $X_u$ affects at most $|\Gamma(u)| + 1 \leq \Delta^2 + 1$ many random variables in $\{\mathcal{Y}_u : u \in U\}$.

Thus, the family of random variables $\{\mathcal{Y}_u : u \in U\}$ is read-$(\Delta^2 + 1)$ and a simple application of Theorem 2.6 like before yields Claim 4.3.

**Proof of Claim 4.4.** We make use of the local lemma to prove this claim. Include every $u \in U$ independently in a set $S'_U$ with probability $p := \frac{1}{\log^{1/4} \Delta}$. For every $v \in S_V$, assign a bad event $B_v$ which denotes that $v$ has more than $\frac{\Delta}{\log^{1/2} \Delta}$ neighbors in $S'_U$. For every color $c \in Q$, assign a bad event $A_c$ which denotes that $|S'_U \cap U_c^*| \leq \frac{n}{\Delta \log^{7/8} \Delta}$. Let us first calculate the probabilities of these bad events. For convenience, denote by $\mathcal{B}(n, p)$ the binomial distribution with the parameters $n$ and $p$. By the Chernoff bound (Lemma 2.5), we obtain the following.

\[
\mathbb{P}[B_v] \leq \mathbb{P}\left[ \mathcal{B}(d(v), p) \geq \frac{\Delta}{\log^{3/2} \Delta} \right] \leq \mathbb{P}\left[ \mathcal{B}(\Delta, p) \geq \frac{\Delta}{\log^{3/2} \Delta} \right] \leq e^{-\Delta^{3/4}}. \tag{5.6}
\]

\[
\mathbb{P}[A_c] \leq \mathbb{P}\left[ \mathcal{B}(|U_c^*|, p) \leq \frac{n}{\Delta \log^{7/8} \Delta} \right] \leq \mathbb{P}\left[ \mathcal{B}\left( \frac{n \log \Delta}{2\Delta}, p \right) \leq \frac{n}{\Delta \log^{7/8} \Delta} \right] \leq e^{-\frac{n}{\Delta \log^{7/8} \Delta}}. \tag{5.7}
\]

For $v \in S_V$, let $\Gamma(v)$ denote the set of all vertices in $S_V$ which are in distance exactly 2 from $v$. Clearly, $|\Gamma(v)| \leq \Delta^2$ for all $v \in S_V$. Note that $B_v$ is mutually independent of all the events $\{B_{v'} : v' \notin \Gamma(v)\}$. To verify the hypothesis of Lemma 2.7 set $x_v := e^{-\sqrt{\Delta}}$ for each $v \in S_V$ and $x_c := e^{-n/\Delta^2}$ for each $c \in Q$. We now have the following for each $v \in S_V$.

\[
x_v \prod_{v' \in \Gamma(v)} (1 - x_{v'}) \prod_{c \in Q} (1 - x_c) \geq e^{-\sqrt{\Delta}} \left( 1 - e^{-\sqrt{\Delta}} \right)^{\Delta^2} \left( 1 - e^{-n/\Delta^2} \right)^q \\
\geq \frac{1}{2} e^{-\sqrt{\Delta}} \geq \mathbb{P}[B_v], \tag{5.8}
\]
where in the last step we have used (5.6). Similarly, we have the following for each \( c \in Q \).

\[
x_c \prod_{v \in S_V} (1 - x_v) \prod_{c \in Q} (1 - x_c) \geq e^{-n/\Delta^2} \left( 1 - e^{-\sqrt{\Delta}} \right)^n \left( 1 - e^{-n/\Delta^2} \right)^q
\]

\[\geq e^{-n/\Delta^2} \cdot e^{-n/\Delta^2} \cdot \frac{1}{2} \geq \mathbb{P}[A_c], \tag{5.9}\]

where in the last step we have used (5.7). Thus, by (5.8), (5.9), and using Lemma 2.7, we have a choice of \( S'_U \) such that none of \( B_v \) and \( A_c \) holds. Now, for each \( c \in Q \), choose \( a_c \) vertices from \( S'_U \cap U^*_c \) and include them in our desirable set \( S_U \) (this can be done because of (11.2)). It is clear that we still have the property that no vertices in \( S_V \) has more than \( \frac{\Delta}{\log^{3/2} \Delta} \) neighbors in \( S_U \). Remember that for each \( c \in Q \), all vertices in \( U^*_c \) have at most \( \frac{\Delta}{\log^{3/2} \Delta} \) neighbors in \( S_V \). Thus, we have proved Claim 4.4.

This finishes the proof of Theorem 1.2.

6 Concluding remarks

How good is the estimate of Lemma 5.1? If one can put a significantly better bound in this lemma, then it might be possible to prove Theorem 1.2 avoiding Claims 4.3 and 4.4 (thus, the local lemma would not be needed). There are some ‘central limit theorem’ type results on coupon collector’s problem (see, e.g., [17] and [31]). However, these results do not seem to help us in improving Lemma 5.1.

We remark that finding the largest bi-hole of a bipartite graph is a NP-hard problem. To see this and some inapproximability results on the bipartite independence number, the interested readers can have a look at [18]. Naturally, one can expect the problem of finding coloring number of a bipartite graph to be even harder.

We next discuss why the current known upper bound of Theorem 1.1 and lower bound of Theorem 1.2 can be hard to improve by considering the appropriate random bipartite graphs. To show the upper bound of Theorem 1.3, the authors [3] essentially proved that the random bipartite graph \( G_{n,n,\Delta/n} \) cannot have a bi-hole of size \((2 + \epsilon) \frac{\log \Delta}{\Delta} \) w.h.p. It can be shown (using essentially same arguments as in [20] or [21]) that this upper bound is asymptotically tight for the bipartite independence number of \( G_{n,n,\Delta/n} \) w.h.p. Thus, it is not possible to improve the lower bound for Theorem 1.1 by considering random graphs. Similarly, it is shown in [11] that the coloring number of the random bipartite graph \( G_{n,n,\Delta/n} \) is concentrated around \( \frac{\Delta}{2 \log \Delta} \) w.h.p. Thus, perhaps the lower bound on \( \chi_B(G) \) for Theorem 1.2 cannot be improved by considering random bipartite graphs.

We next reason why we believe that improving the current gap of lower and upper bounds in Theorems 1.1 and 1.2 can be challenging. Before discussing it, we mention the situation for a similar problem in graphs (not restricted to bipartite graphs). The best known lower and upper bounds for the largest possible chromatic number of a triangle-free graph with bounded maximum degree have a multiplicative gap of two. However, it is believed to be hard to improve this gap (see, e.g., [1], [33], and [40]). We experience similar situation in the bipartite setting as demonstrated next.
A simple greedy algorithm obtains a bi-hole of size \((1 - \epsilon)\frac{\log \Delta}{\Delta} n\) in the random bipartite graph \(G_{n,n,\Delta/n}\) w.h.p. (e.g., the same method as in Exercise 6.7.20 of [21] works here). However, no efficient (polynomial time) algorithm (deterministic or randomized) is known to find a significantly larger bi-hole (see, e.g., [1] and [10]). This shows some difficulty of improving Theorem 1.1, it seems especially challenging to find an efficient algorithm to find a significantly larger bi-hole in Theorem 1.1 (because, an algorithm for Theorem 1.1 will likely find a similar sized bi-hole in \(G_{n,n,\Delta/n}\)). On the other hand, since there is no efficient algorithm known to find a bi-hole in \(G_{n,n,\Delta/n}\) of size significantly larger than \(\frac{\log \Delta}{\Delta} n\), we do not have any efficient algorithm to color \(G_{n,n,\Delta/n}\) using significantly less than \(\frac{n}{\log \Delta}\) colors. Our bound of Theorem 1.2 matches this and extends this to efficiently color any bipartite graph with maximum degree \(\Delta\) with about \(\frac{\Delta}{\log \Delta}\) colors.

We next briefly discuss about some related problems to Theorem 1.1 in the literature. We would suggest the readers to have a look at Section 2 of [3] to see a more detailed description of various connections with Theorem 1.1 or 1.3. As mentioned in [3], they are related to the bipartite version of the Erdős-Hajnal conjecture (see, e.g., [4] and [15]), the bipartite Ramsey numbers (see, e.g., [10] and [13]), and the Zarankiewicz function (see, e.g., [5], [6], [22], [24], and [25]). To see the connection with the bipartite Ramsey number, for bipartite graphs \(H_1\) and \(H_2\), let the bipartite Ramsey number \(br(H_1, H_2)\) be the smallest \(N\) such that any red-blue edge-coloring of the complete bipartite graph \(K_{N,N}\) contains either a red copy of \(H_1\) or a blue copy of \(H_2\). For results on this topic, see, e.g., Beineke and Schwenk [7], Caro and Rousseau [10], Conlon [13], Hattingh and Henning [26], Irving [27], Lin and Li [32], and Thomason [38]. As an application of Theorem 1.1 we obtain that \(br(K_{1,\Delta}, K_{n,n}) \lesssim \frac{\Delta}{\log \Delta} n\) for sufficiently large but fixed \(\Delta\) and growing \(n\).

We end with suggesting two directions for future research. Firstly, it will be interesting to study multi-partite analogues of Theorems 1.1 and 1.2. For example, one can define 'tri-hole' in a tripartite graph to be an independent set with the same number of vertices in all the three parts. It might be worth estimating the size of the largest tri-hole in a tripartite graph with bounded average degree or bounded local degree. The straightforward extensions of the methods used in this paper do not seem to work for \(k\)-partite graphs when \(k \geq 3\).

There is a recent result by Kogan [30] on a generalization of the notion of bipartite independence number. They bounded the largest \(k\) for which a given \(n\) by \(n\) bipartite graph has a \(k\) by \(k\) induced \(d\)-degenerate subgraph. This can be studied in the context of Theorem 1.1. For example, it is worth investigating if one can improve the trivial bound obtained by Theorem 1.1 to get a significantly larger balanced \(d\)-degenerate subgraph.

**Acknowledgements**

We are thankful to Rutger Campbell and Sang-il Oum for helping us to improve the writing of this paper.
References

[1] D. Achlioptas and A. Coja-Oghlan, Algorithmic barriers from phase transitions, *Proceedings of FOCS*, 2008, 793–802.

[2] N. Alon and J. H. Spencer, The Probabilistic Method (4th edition), Wiley 2016.

[3] M. Axenovich, J.-S. Sereni, R. Snyder, and L. Weber, Bipartite independence number in graphs with bounded maximum degree, *Siam J. Discrete Math.*, **35**(2) (2021), 1136–1148.

[4] M. Axenovich, C. Tompkins, and L. Weber, Large homogeneous subgraphs in bipartite graphs with forbidden induced subgraphs, *J. Graph Theory*, **97** (2021), 34–46.

[5] C. Balbuena, P. García-Vázquez, X. Marcote, and J. C. Valenzuela, New results on the Zarankiewicz problem, *Discrete Math.*, **307** (2007), 2322–2327.

[6] C. Balbuena, P. García-Vázquez, X. Marcote, and J. C. Valenzuela, Extremal $K_{(s,t)}$-free bipartite graphs, *Discrete Math. Theor. Comput. Sci.*, **10** (2008), 35–48.

[7] L. W. Beineke and A. J. Schwenk, On a bipartite form of the Ramsey problem, in Proceedings of the Fifth British Combinatorial Conference, University of Aberdeen, Aberdeen, 1975, 17–22.

[8] A. Ben-Dor, T. Hartman, R. M. Karp, B. Schwikowski, R. Sharan, and Z. Yakhini, Towards optimally multiplexed applications of universal arrays, *J. Comput. Biol.*, **11**(2–3) (2004), 477–493.

[9] A. Bernshteyn, The Johansson-Molloy Theorem for DP-coloring, *Random Struct. Algorithms*, **54**(4) (2019), 653–664.

[10] Y. Caro and C. Rousseau, Asymptotic bounds for bipartite Ramsey numbers, *Electron. J. Combin.*, **8** (2001), R17.

[11] D. Chakraborti, Coloring number of random bipartite graphs, in preparation.

[12] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Statatistics*, **23**(4) (1952), 493–507.

[13] D. Conlon, A new upper bound for the bipartite Ramsey problem. *J. Graph Theory*, **58**(4) (2008), 351–356.

[14] S. Ehard, E. Mohr, and D. Rautenbach, Biholes in balanced bipartite graphs, 2020, arXiv preprint arXiv:2004.03245.

[15] P. Erdős, A. Hajnal, and J. Pach, A Ramsey-type theorem for bipartite graphs, *Combinatorics, Geom-binatorics*, **10** (2000), 64–68.
[16] P. Erdős and L. Lovász, Problems and results on 3-chromatic hypergraphs and some related questions, in Infinite and Finite Sets II, A. Hajnal, R. Rado, and V. T. Sós, eds., North–Holland, Amsterdam, 1975, 609–627.

[17] P. Erdős and A. Rényi, On a classical problem of probability theory, A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei, 6 (1961), 215–220.

[18] U. Feige and S. Kogan, Hardness of approximation of the balanced complete bipartite subgraph problem, Technical Report MCS04-04, The Weizmann Institute, 2004.

[19] U. Feige and S. Kogan, Balanced coloring of bipartite graphs, J. Graph Theory, 64 (2010), 277–291.

[20] A. M. Frieze, On the independence number of random graphs, Discrete Math., 81 (1990), 171–175.

[21] A. Frieze and M. Karoński, Introduction to Random Graphs, Cambridge University Press, 2015.

[22] Z. Füredi and M. Simonovits, The history of degenerate (bipartite) extremal graph problems, in Erdős Centennial, Bolyai Soc. Math. Stud. 25, Springer-Verlag, Berlin, 2013, 169–264.

[23] D. Gavinsky, S. Lovett, M. E. Saks, and S. Srinivasan, A tail bound for read-k families of functions, Random Struct. Algorithms, 47(1) (2015), 99–108.

[24] J. R. Griggs and J. Ouyang, (0,1)-matrices with no half-half submatrix of ones, European J. Combin., 18 (1997), 751–761.

[25] J. R. Griggs, M. Simonovits, and G. R. Thomas, Extremal graphs with bounded densities of small subgraphs, J. Graph Theory, 29 (1998), 185–207.

[26] J. H. Hattingh and M. A. Henning, Bipartite Ramsey theory, Util. Math., 53 (1998), 217–230.

[27] R. W. Irving, A bipartite Ramsey problem and the Zarankiewicz numbers, Glasg. Math. J., 19 (1978), 13–26.

[28] S. Janson, On concentration of probability, Contemporary combinatorics, 10(3) (2002), 1–9.

[29] S. Jukna, Extremal Combinatorics, Springer, Berlin (2001), available at http://www.thi.informatik.uni-frankfurt.de/~jukna/EC_Book_2nd/comment6.html

[30] S. Kogan, A note on a Caro–Wei bound for the bipartite independence number in graphs, Discrete Math., 344(4) (2021), 112285.

[31] D. Levin, Y. Peres, and E. Wilmer, Markov chains and mixing times, AMS, Providence, 2017.
[32] Q. Lin and Y. Li, Bipartite Ramsey numbers involving large $K_{n,n}$, *European J. Combin.*, 30 (2009), 923–928.

[33] M. Molloy, The list chromatic number of graphs with small clique number, *J. Combin. Theory Ser. B*, 134 (2019), 264–284.

[34] M. Molloy, B. Reed, Graph Colouring and the Probabilistic Method, Springer, 2002.

[35] R. Moser and G. Tardos, A constructive proof of the general Lovász local lemma, *J. ACM*, 57(2) (2010), 1–15.

[36] M. Okamoto, Some inequalities relating to the partial sum of binomial probabilities, *Annals of the Institute of Statistical Mathematics*, 10(1) (1959), 29–35.

[37] W. Pegden, An extension of the Moser-Tardos algorithmic local lemma, *Siam J. Discrete Math.*, 28 (2014), 911–917.

[38] A. Thomason, On finite Ramsey numbers, *European J. Combin.*, 3 (1982), 263–273.

[39] P. Turán, On an extremal problem in graph theory (in Hungarian), *Mat. Fiz. Lapok*, 48 (1941), 436–452.

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