Discrepancy of random graphs and hypergraphs

Jie Ma* Humberto Naves† Benny Sudakov‡

Abstract

Answering in a strong form a question posed by Bollobás and Scott, in this paper we determine the discrepancy between two random k-uniform hypergraphs, up to a constant factor depending solely on k.

1 Introduction

A *hypergraph* H is an ordered pair $H = (V, E)$, where V is a finite set (the *vertex set*), and E is a family of distinct subsets of V (*the edge set*). The hypergraph H is k-uniform if all its edges are of size k. In this paper we consider only k-uniform hypergraphs. The *edge density* of a k-uniform hypergraph H with n vertices is $\rho_H = e(H)/\binom{n}{k}$. We define the *discrepancy* of H to be

$$\mathrm{disc}(H) = \max_{S \subseteq V(H)} \left| e(S) - \rho_H \binom{|S|}{k} \right|, \tag{1}$$

where $e(S) = e(H[S])$ is the number of edges in the sub-hypergraph induced by S. The discrepancy can be viewed as a measure of how uniformly the edges of H are distributed among the vertices. This important concept appears naturally in various branches of Combinatorics and has been studied by many researchers in recent years. The discrepancy is closely related to the theory of quasi-random graphs (see [7]), as the property $\mathrm{disc}(G) = o(|V(G)|^2)$ implies the quasi-randomness of the graph G.

Erdős and Spencer [9] proved that for $k \geq 2$, any k-uniform hypergraph H with n vertices has a subset S satisfying $\left|e(S) - \frac{1}{2}\binom{|S|}{k}\right| \geq cn^{\frac{k+1}{2}}$, which implies the bound $\mathrm{disc}(H) \geq cn^{\frac{k+1}{2}}$ for k-uniform hypergraphs H of edge density $\frac{1}{2}$. Erdős, Goldberg, Pach and Spencer [8] obtained a similar lower bound for graphs of edge density smaller than $\frac{1}{2}$. These results were later generalized by Bollobás and Scott in [3], who proved the inequality $\mathrm{disc}(H) \geq c_k\sqrt{r}n^{\frac{k+1}{2}}$ for k-uniform hypergraphs H, whenever $r = \rho_H(1-\rho_H) \geq 1/n$. The random hypergraphs show that all the aforementioned lower bounds are optimal up to constant factors. For more discussion and general accounts of discrepancy, we refer the interested reader to Beck and Sós [2], Bollobás and Scott [3], Chazelle [6], Matoušek [11] and Sós [12].

A similar notion is the relative discrepancy of two hypergraphs. Let G and H be two k-uniform hypergraphs over the same vertex set V, with $|V| = n$. For a bijection $\pi : V \to V$, let G_π be obtained from G by permuting all edges according to π, i.e., $E(G_\pi) = \pi(E(G))$. The *overlap* of G and H

*Department of Mathematics, UCLA, Los Angeles, CA 90095. Email: jiema@math.ucla.edu. Research supported in part by AMS-Simons travel grant.

†Department of Mathematics, UCLA, Los Angeles, CA 90095. Email: hnaves@math.ucla.edu.

‡Department of Mathematics, ETH, 8092 Zurich, Switzerland and Department of Mathematics, UCLA, Los Angeles, CA 90095. Email: bsudakov@math.ucla.edu. Research supported in part by NSF grant DMS-1101185 and by a USA-Israel BSF grant.

with respect to π, denoted by $G_\pi \cap H$, is a hypergraph with the same vertex set V and with edge set $E(G_\pi) \cap E(H)$. The *discrepancy of G with respect to H* is

$$\mathrm{disc}(G,H) = \max_\pi \left| e(G_\pi \cap H) - \rho_G \rho_H \binom{n}{k} \right|, \tag{2}$$

where the maximum is taken over all bijections $\pi : V \to V$. For random bijections π, the expected size of $E(G_\pi) \cap E(H)$ is $\rho_G \rho_H \binom{n}{k}$; thus $\mathrm{disc}(G,H)$ measures how much the overlap can deviate from its average. In a certain sense, the definition (2) is more general than (1), because one can write $\mathrm{disc}(H) = \max_{1 \le i \le n} \mathrm{disc}(G_i, H)$, where G_i is obtained from the complete i-vertex k-uniform hypergraph by adding $n-i$ isolated vertices.

Bollobás and Scott introduced the notion of relative discrepancy in [4] and showed that for any two n-vertex graphs G and H, if $\frac{16}{n} \le \rho_G, \rho_H \le 1 - \frac{16}{n}$, then $\mathrm{disc}(G,H) \ge c \cdot f(\rho_G, \rho_H) \cdot n^{\frac{3}{2}}$, where c is an absolute constant and $f(x,y) = x^2(1-x)^2y^2(1-y)^2$. As a corollary, they proved a conjecture in [8] regarding the *bipartite discrepancy* $\mathrm{disc}(G, K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil})$. Moreover, they also conjectured that a similar bound holds for k-uniform hypergraphs, namely, there exists $c = c(k, \rho_G, \rho_H)$ for which $\mathrm{disc}(G,H) \ge c n^{\frac{k+1}{2}}$ holds for any k-uniform hypergraphs G and H satisfying $\frac{1}{n} \le \rho_G, \rho_H \le 1 - \frac{1}{n}$.

In their paper, Bollobás and Scott also asked the following question (see Problem 12 in [4]). Given two random n-vertex graphs G and H with constant edge probability p, what is the expected value of $\mathrm{disc}(G,H)$? In this paper, we solve this question completely for general k-uniform hypergraphs. Let $\mathcal{H}_k(n,p)$ denote the random k-uniform hypergraph on n vertices, in which every edge is included independently with probability p. We say that an event happens *with high probability*, or w.h.p. for brevity, if it happens with probability at least $1 - n^{-\omega(1)}$, where here and later $\omega(1)$ denotes an arbitrary function tending to infinity together with n.

Theorem 1.1. *For positive integers n and k, let $N = \binom{n-\frac{n}{k}}{k-1}$. Let G and H be two random hypergraphs distributed according to $\mathcal{H}_k(n,p)$ and $\mathcal{H}_k(n,q)$ respectively, where $\frac{\omega(1)}{N} \le p \le q \le \frac{1}{2}$.*

(1) DENSE CASE – *If $pqN > \frac{1}{30} \log n$, then w.h.p.* $\mathrm{disc}(G,H) = \Theta_k\left(\sqrt{pq\binom{n}{k} n \log n}\right)$;

(2) SPARSE CASE – *If $pqN \le \frac{1}{30} \log n$, let $\gamma = \frac{\log n}{pqN}$; then*

(2.1) if $pN \ge \frac{\log n}{5 \log \gamma}$, then w.h.p. $\mathrm{disc}(G,H) = \Theta_k\left(\frac{n \log n}{\log \gamma}\right)$.

(2.2) if $pN < \frac{\log n}{5 \log \gamma}$, then w.h.p. $\mathrm{disc}(G,H) = \Theta_k\left(p\binom{n}{k}\right)$.

The previous theorem also provides tight bounds when p and/or $q \ge \frac{1}{2}$, as we shall see in the concluding remarks. The result of Theorem 1.1 in the sparse range is closely related to the recent work of the third author with Lee and Loh [10]. Among other results, the authors of [10] show that two independent copies G, H of the random graph $G(n,p)$ with $p \ll \sqrt{\log n / n}$ w.h.p. have overlap of order $\Theta\left(n \frac{\log n}{\log \gamma}\right)$, where $\gamma = \frac{\log n}{p^2 n}$. Hence $\mathrm{disc}(G,H) = \Theta\left(n \frac{\log n}{\log \gamma}\right)$ holds, since in this range of edge probability, $n \frac{\log n}{\log \gamma}$ is larger than the average overlap $p^2\binom{n}{2}$. Our proof in the sparse case borrows some ideas from [10]. On the other hand, one can not use their approach for all cases; hence some new ideas were needed to prove Theorem 1.1.

It will become evident from our proof that the problem of determining the discrepancy can be essentially reduced to the following question. Let $K > 0$, and let X be a binomial random variable with

parameters m and ρ. What is the maximum value of $\Lambda = \Lambda(m, \rho, K)$ satisfying $\mathbb{P}\big[X - m\rho > \Lambda\big] \geq e^{-K}$? This question is related to the rate function of binomial distribution. In all cases, the discrepancy in the statement of Theorem 1.1 is w.h.p.

$$\operatorname{disc}(G, H) = \Theta_k\left(n \cdot \Lambda\left(p\binom{n-1}{k-1}, q, \log n\right)\right). \tag{3}$$

Note that $p\binom{n-1}{k-1}$ is roughly the size of the neighborhood of a vertex in the hypergraph G.

The rest of this paper is organized as follows. Section 2 contains a high level outline of our proof. It also includes the definition of the *probabilistic discrepancy* $\operatorname{disc}_P(G, H)$. Section 3 contains a list of inequalities and technical lemmas used throughout the paper. In particular, we demonstrate that $\operatorname{disc}_P(G, H)$ w.h.p. does not deviate too much from $\operatorname{disc}(G, H)$. In Section 4, we establish the upper bound for $\operatorname{disc}(G, H)$ based on a similar bound for $\operatorname{disc}_P(G, H)$. In Section 5, we give a detailed proof of the lower bound for $\operatorname{disc}(G, H)$. The final section contains some concluding remarks. In this paper, the function log refers to the natural logarithm and all asymptotic notation symbols (Ω, O, o and Θ) are with respect to the variable n. Furthermore, the k-subscripts in these symbols indicate the dependence on k in the relevant constants.

2 Outline of the proof

In this section, we describe the main ideas in the proof of Theorem 1.1. In order to determine $\operatorname{disc}(G, H)$, we introduce a related quantity, the *probabilistic discrepancy* $\operatorname{disc}_P(G, H)$. Let G and H be two random hypergraphs over the same vertex set V, distributed according to $\mathcal{H}_k(n, p)$ and $\mathcal{H}_k(n, q)$, respectively. The *probabilistic discrepancy* of G with respect to H is defined by

$$\operatorname{disc}_P(G, H) = \max_{\pi} \left| e(G_\pi \cap H) - pq\binom{n}{k} \right|,$$

where the maximum is taken over all bijections $\pi : V \to V$. In Section 4, we show that $\operatorname{disc}_P(G, H)$ is w.h.p. very close to $\operatorname{disc}(G, H)$, hence, to bound $\operatorname{disc}(G, H)$, it suffices to show corresponding bounds for $\operatorname{disc}_P(G, H)$.

The proof of the upper bound for $\operatorname{disc}_P(G, H)$ is fairly standard. In case (2.2) of the main theorem, the proof is trivial, as w.h.p. $e(G) < 2p\binom{n}{k}$. For the remaining cases, we remark that for any fixed permutation $\pi : V \to V$, the overlap $G_\pi \cap H$ is a random hypergraph distributed according to $\mathcal{H}_k(n, pq)$. The upper bound then follows from a straightforward union bound argument over all possible permutations π, together with the application of concentration inequalities for the binomial distribution. The remaining details of this particular argument are presented in Section 4.

The main contribution of the paper is the proof of the lower bound. In Section 5, we show that w.h.p. there exists a permutation π such that the corresponding overlap $e(G_\pi \cap H)$ is much bigger than $pq\binom{n}{k}$. Note that $e(G_\pi \cap H) > pq\binom{n}{k}$, so the discrepancy is "positive" here. In the proof, we fix an arbitrary set $L \subseteq V$ of size $|L| = \frac{n}{k}$, and restrict the set of possible permutations to bijections permuting only the elements of L. Then, we gradually expose the edges (belonging to both G and H) in two rounds. In the first round, we expose the edges having exactly one vertex in L, while keeping unexposed the edges having zero or at least two vertices in L. This way, the overall contribution to the discrepancy from the edges exposed in the first round is exactly the sum of the contributions from each individual choice of $\pi(x)$. To be more precise, let R be the set of all $(k-1)$-subsets of $V \setminus L$;

for each $u \in L$, let $N_G(u)$ be the collection of all $(k-1)$-sets $T \in R$ such that $\{u\} \cup T$ is an edge of G, and let $N_H(u)$ be defined similarly; finally, for each pair $u, v \in L$, let $\mathrm{codeg}(u, v)$ denote the size of $N_G(u) \cap N_H(v)$. The total number of edges in the overlap $G_\pi \cap H$ having exactly one vertex in L is precisely the sum

$$\sum_{x \in L} \mathrm{codeg}(x, \pi(x)). \tag{4}$$

See Figure 1 for more details. The size $|L| = \frac{n}{k}$ was appropriately chosen to maximize the number of edges having precisely one vertex in L. Additionally, we remark that $|R| = \binom{n-\frac{n}{k}}{k-1}$, which is exactly the value of N in the statement of Theorem 1.1. The following inequality will be used extensively later in the paper. It relates N and the binomial coefficient $\binom{n}{k}$ for large enough n, as

$$\frac{1}{3}\binom{n}{k} \leq N\frac{n}{k} \leq \frac{1}{2}\binom{n}{k}.$$

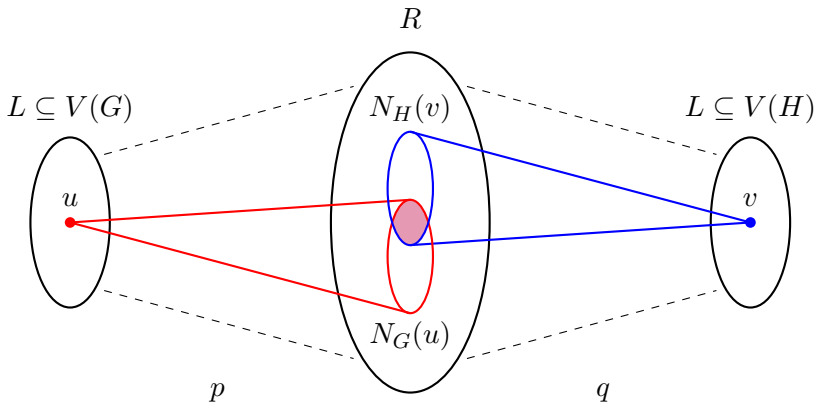

Figure 1: Edges of G and H having one vertex in L.

Having found the bijection π with big overlap in the exposed edges (we have not yet explained how to obtain such bijection), the final step would be to expose the remaining edges of both hypergraphs (second round exposure) and compute the overall discrepancy. The potential "loss" in this final step will be w.h.p. much smaller than the "gain" we already obtained in the previous steps.

It remains to explain how to obtain the bijection π. We define the *connection* graph $\Gamma = \Gamma(G, H)$ as follows. The set of vertices of Γ is the union of two disjoint copies of L, which we will refer to as L_G and L_H, respectively. We will add an edge between $u \in L_G$ and $v \in L_H$ in Γ when $\mathrm{codeg}(u, v)$ is sufficiently large. The notion of large here will vary, depending on which case (dense or sparse) we are trying to prove. Because of (4), in order to maximize the overlap, it will suffice to show the existence of a large matching in auxiliary graph Γ.

In the dense case, we prove that we can find a nearly regular subgraph of Γ (i.e., all the degrees are roughly the same) and thus the existence of the desired bijection π easily follows from well-known theorem of Vizing. For more details, see Section 5.1. In the sparse case, the proof is slightly different. To find the matching in Γ, we divide L_G into chunks, each having size $n^{2/5}$. Then, for each chunk in L_G, we expose the neighborhoods of its vertices to R and w.h.p. we show that these neighborhoods can be made disjoint by removing very few edges. Finally, we start matching the vertices in L_H with the vertices in L_G. This is done by exposing the neighborhood of a vertex in L_H (one by one, according

to an arbitrary predetermined order), and matching it with a high codegree vertex in L_G. The details of this construction are contained in Section 5.2.

3 Auxiliary results

In this section we list and prove some useful concentration inequalities about the binomial and hypergeometric distributions. In addition, we prove that $\mathrm{disc}_P(G,H)$ (defined in the previous section) is w.h.p. very close to $\mathrm{disc}(G,H)$. Lastly, we prove a corollary from the well-known Vizing's Theorem which asserts the existence of a linear-size matching in *nearly regular* graphs (i.e., the maximum degree is close to the average degree). We will not attempt to optimize our constants, preferring rather to choose values which provide a simpler presentation. Let us start with classical Chernoff-type estimates for the tail of the binomial distribution (see, e.g., [1]).

Lemma 3.1. *Let $X = \sum_{i=1}^{l} X_i$ be the sum of independent zero-one random variables with average $\mu = \mathbb{E}[X]$. Then for all non-negative $\lambda \le \mu$, we have $\mathbb{P}[|X-\mu| > \lambda] \le 2e^{-\frac{\lambda^2}{4\mu}}$.*

The following lower tail inequality (see [1]) is due to Janson.

Lemma 3.2. *Let $A_1, A_2, \ldots, A_l$ be subsets of a finite set Ω, and let R be a random subset of Ω for which the events $r \in R$ are mutually independent over $r \in \Omega$. Define X_j to be the indicator random variable of $A_j \subset R$. Let $X = \sum_{j=1}^{l} X_j$, $\mu = \mathbb{E}[X]$, and $\Delta = \sum_{i \sim j} \mathbb{E}[X_i \cdot X_j]$, where $i \sim j$ means that X_i and X_j are dependent (i.e., A_i intersects A_j). Then for any $\lambda > 0$,*

$$\mathbb{P}[X \le \mu - \lambda] < e^{-\frac{\lambda^2}{2\mu+\Delta}}.$$

Next, we establish that the difference between $\mathrm{disc}(G,H)$ and $\mathrm{disc}_P(G,H)$ is w.h.p. very small. This difference is, in fact, much smaller than any bound stated in Theorem 1.1. Thus, to prove bounds for $\mathrm{disc}(G,H)$, it suffices to show corresponding bounds for $\mathrm{disc}_P(G,H)$.

Lemma 3.3. *Let G and H be two random hypergraphs over the same vertex set V, distributed according to $\mathcal{H}_k(n,p)$ and $\mathcal{H}_k(n,q)$, respectively. With probability at least $1-4e^{-\sqrt{n}}$, the inequality $|\mathrm{disc}(G,H) - \mathrm{disc}_P(G,H)| \le 2\varepsilon$ holds, where $\varepsilon = 4n^{\frac{1}{4}}\sqrt{pq\binom{n}{k}}$.*

Proof. Since $p\binom{n}{k} = \Omega(n)$, applying Lemma 3.1 to the random variable $e(G)$ for $\lambda = 2n^{\frac{1}{4}}\sqrt{p\binom{n}{k}} \le p\binom{n}{k}$ yields

$$\mathbb{P}\left[\left|e(G) - p\binom{n}{k}\right| \le 2n^{\frac{1}{4}}\sqrt{p\binom{n}{k}}\right] \ge 1 - 2e^{-\sqrt{n}}.$$

Similarly, we have $\mathbb{P}\left[|e(H) - q\binom{n}{k}| \le 2n^{\frac{1}{4}}\sqrt{q\binom{n}{k}}\right] \ge 1 - 2e^{-\sqrt{n}}$. Therefore, with probability at least $1-4e^{-\sqrt{n}}$, $|\rho_G - p| \le 2n^{\frac{1}{4}}\left(p/\binom{n}{k}\right)^{1/2}$ and $|\rho_H - q| \le 2n^{\frac{1}{4}}\left(q/\binom{n}{k}\right)^{1/2}$. But if $|AB - A_0B_0| \ge \epsilon_1\epsilon_2 + |A_0|\epsilon_2 + |B_0|\epsilon_1$, then either $|A - A_0| \ge \epsilon_1$ or $|B - B_0| \ge \epsilon_2$. Together, these inequalities imply

$$\left|\rho_G\rho_H\binom{n}{k} - pq\binom{n}{k}\right| \le 4\sqrt{pqn} + 2pn^{\frac{1}{4}}\sqrt{q\binom{n}{k}} + 2qn^{\frac{1}{4}}\sqrt{p\binom{n}{k}} \le 2\varepsilon,$$

completing the proof of the lemma. □

In the proof of the dense case of the main theorem we will need a lower bound for the tail of the hypergeometric distribution. To prove it we use the following well-known estimates for the binomial coefficient.

Proposition 3.4. *Let $H(p) = -p\log p - (1-p)\log(1-p)$ (the binary entropy), then for any integer $m > 0$ and real $p \in (0,1)$ satisfying $pm \in \mathbb{Z}$ we have*

$$\frac{\sqrt{2\pi}}{e^2} \leq \binom{m}{pm}\sqrt{mp(1-p)}e^{-mH(p)} \leq \frac{e}{2\pi}.$$

Proof. This can be derived from Stirling's formula $\sqrt{2\pi m}\left(\frac{m}{e}\right)^m \leq m! \leq e\sqrt{m}\left(\frac{m}{e}\right)^m$. □

Lemma 3.5. *Let d_1, d_2, Δ and N be integers and K be a real parameter such that $1 \leq d_1, d_2 \leq \frac{2N}{3}$, $1 \leq K \leq \frac{d_1 d_2}{100N}$ and $\Delta = \sqrt{\frac{d_1 d_2 K}{N}}$. Then*

$$\sum_{t \geq \frac{d_1 d_2}{N} + \Delta} \frac{\binom{d_1}{t}\binom{N-d_1}{d_2-t}}{\binom{N}{d_2}} \geq e^{-40K}.$$

Proof. For convenience, we write $f(t) = \binom{d_1}{t}\binom{N-d_1}{d_2-t}/\binom{N}{d_2}$. In order to show the desired lower bound of the hypergeometric sum, it suffices to prove that

$$f(t) \geq \frac{4e^{-40K}}{\sqrt{\frac{d_1 d_2}{N} + \Delta}},$$

for every integer $t = \frac{d_1 d_2}{N} + \theta\Delta$ with $1 \leq \theta \leq 2$. Indeed, to see this, note that there are at least $\lfloor \Delta \rfloor \geq \frac{\Delta}{2}$ integers between $\frac{d_1 d_2}{N} + \Delta$ and $\frac{d_1 d_2}{N} + 2\Delta$ and

$$\Delta > \frac{1}{2}\sqrt{\Delta^2 + \Delta} \geq \frac{1}{2}\sqrt{\frac{d_1 d_2}{N} + \Delta}.$$

Next we prove the bound for $f(t)$. For our choice of Δ, the inequality $\Delta \leq \frac{d_1}{15}$ is true since

$$\Delta = \sqrt{\frac{d_1 d_2 K}{N}} = d_1\sqrt{\frac{d_2}{N}\cdot\frac{K}{d_1}} \leq d_1\sqrt{\frac{d_2}{N}\cdot\frac{d_2}{100N}} = \frac{d_1}{10}\cdot\frac{d_2}{N} \leq \frac{d_1}{15}.$$

Similarly $\Delta \leq \frac{d_2}{15}$. Let $x = \frac{d_2}{N}$, $y = \frac{\theta\Delta}{d_1}$ and $z = \frac{\theta\Delta}{N-d_1}$. Then $t = (x+y)d_1$ and $d_2 - t = (x-z)(N-d_1)$. But $0 < x + y < 1$, because $0 < x \leq \frac{2}{3}$ and $0 < y \leq \frac{2\Delta}{d_1} < \frac{1}{3}$. Furthermore, $0 < x - z < 1$, because $\frac{z}{x} = \frac{\theta\Delta N}{d_2(N-d_1)} \leq \frac{3\theta\Delta}{d_2} \leq \frac{2}{5}$ and $x \leq \frac{2}{3}$. By Proposition 3.4, we have

$$f(t) = \frac{\binom{d_1}{(x+y)d_1}\binom{N-d_1}{(x-z)(N-d_1)}}{\binom{N}{xN}} \geq \frac{4\pi^2}{e^5}\sqrt{R}e^{-L},$$

where $L = -d_1 \cdot H(x+y) - (N-d_1)\cdot H(x-z) + N \cdot H(x)$ and

$$R = \frac{x(1-x)N}{(x-z)(1-x+z)(x+y)(1-x-y)d_1(N-d_1)} \geq \frac{1}{(x+y)d_1} \geq \frac{1}{2}\cdot\frac{1}{\frac{d_1 d_2}{N} + \Delta}.$$

Here we used $z \leq x$ for the first the inequality; and we used $\theta \leq 2$ and the identity $(x+y)d_1 = t =$

$\frac{d_1 d_2}{N} + \theta\Delta$ for the second inequality. Because $d_1 y = (N - d_1)z = \theta\Delta$ and $\log(1+s) \le s$, we obtain

$$\begin{aligned} L &= d_1 \left[(x+y)\log\left(1+\frac{y}{x}\right) + (1-x-y)\log\left(1-\frac{y}{1-x}\right)\right] \\ &+ (N-d_1)\left[(x-z)\log\left(1-\frac{z}{x}\right) + (1-x+z)\log\left(1+\frac{z}{1-x}\right)\right] \\ &\le d_1\left[\frac{(x+y)y}{x} - \frac{(1-x-y)y}{1-x}\right] + (N-d_1)\left[-\frac{(x-z)z}{x} + \frac{(1-x+z)z}{1-x}\right] \\ &= \theta\Delta\cdot(y+z)\cdot\left(\frac{1}{x}+\frac{1}{1-x}\right) = \frac{\theta^2\Delta^2 N^3}{d_1(N-d_1)d_2(N-d_2)} \le 36K. \end{aligned}$$

Thus we always have $f(t) \ge \frac{4\pi^2}{\sqrt{2}e^5}\cdot\frac{e^{-36K}}{\sqrt{\frac{d_1 d_2}{N}+\Delta}} \ge \frac{4e^{-40K}}{\sqrt{\frac{d_1 d_2}{N}+\Delta}}$, completing the proof. □

The next lemma will be used to prove the lower bound in the sparse case of Theorem 1.1 and was inspired by an analogous result in [10].

Lemma 3.6. *For positive integers n and k, let $N = \binom{n-\frac{n}{k}}{k-1}$, $\frac{\omega(1)}{N} \le p \le q \le \frac{1}{2}$ and suppose that $pqN \le \frac{1}{30}\log n$. Define $\gamma = \frac{\log n}{pqN}$. Let $N_1, \ldots, N_s \subseteq B$ be $s \ge n^{1/3}$ disjoint sets of size $(1+o(1))Np$, and consider the random set B_q, obtained by taking each element of B independently with probability q. Then w.h.p., there is an index i for which*

(1) $|B_q \cap N_i| \ge \frac{\log n}{6\log\gamma}$ if $pN \ge \frac{\log n}{5\log\gamma}$;

(2) $N_i \subseteq B_q$ if $pN < \frac{\log n}{5\log\gamma}$.

Proof. If $pN \ge \frac{\log n}{5\log\gamma}$, let $t = \frac{\log n}{6\log\gamma}$. Clearly $1-q \ge e^{-3q/2}$ when $q \le 1/2$. For a fixed index i, the probability that $|B_q \cap N_i| \ge t$ is at least $\binom{|N_i|}{t} q^t (1-q)^{|N_i|-t}$. Using the bounds $\binom{a}{n} \ge (\frac{a}{b})^b$ for $a \ge b$, and $\frac{1}{30}\log n \ge Npq = \frac{\log n}{\gamma}$, we obtain

$$\begin{aligned} \binom{|N_i|}{t} q^t (1-q)^{|N_i|-t} &\ge \left(\frac{(1+o(1))Npq}{t}\right)^t e^{-2pqN} \ge \left(\frac{5\log\gamma}{\gamma}\right)^{\frac{\log n}{6\log\gamma}} n^{-1/15} \\ &\ge n^{-1/6}\cdot n^{-1/15} \ge n^{-0.3}. \end{aligned}$$

Hence the expected number of indices i such that $|B_q \cap N_i| \ge t$ is at least $sn^{-0.3} \ge n^{1/30}$. Since the sets N_i are disjoint, these events are independent for different choices of i. Therefore by Lemma 3.1 w.h.p. we can find such an index (actually many).

If $pN < \frac{\log n}{5\log\gamma}$, then $q = \frac{\log n}{\gamma pN} > \frac{5\log\gamma}{\gamma} \ge \gamma^{-1}$. Therefore the probability that some $N_i \subseteq B_q$ is

$$q^{|N_i|} \ge \gamma^{-(1+o(1))Np} \ge \gamma^{-\frac{\log n}{4\log\gamma}} = n^{-1/4},$$

and we can complete the proof as in the first case. □

The last lemma in this section, which can be easily derived from Vizing's Theorem, will be used to find a linear-size matching in nearly regular graphs.

Lemma 3.7. *Every graph G with maximum degree $\Delta(G)$, contains a matching of size at least $\frac{e(G)}{\Delta(G)+1}$.*

Proof. By Vizing's Theorem, the graph G has a proper edge coloring $f : E(G) \to \{1, 2, \ldots, \Delta(G)+1\}$. For each color $1 \le c \le \Delta(G)+1$, the edges $f^{-1}(c)$ form a matching in G. By the pigeonhole principle, there is a color c such that $f^{-1}(c)$ has at least $\frac{e(G)}{\Delta(G)+1}$ edges. □

4 Upper bounds

In this section we prove the upper bound for the discrepancy in Theorem 1.1. By Lemma 3.3, it suffices to prove the corresponding bounds for $\mathrm{disc}_P(G, H)$ instead.

Lemma 4.1. *Let G and H be as in Theorem 1.1. Then w.h.p.* $\mathrm{disc}_P(G, H)$ *satisfies the stated upper bounds of Theorem 1.1.*

Proof. Since the number of edges of G is distributed binomially and $p\binom{n}{k} = \Omega(n)$, by Lemma 3.1, we have $e(G) < 2p\binom{n}{k}$ with probability at least $1 - e^{-\Theta(n)}$. Since $\mathrm{disc}_P(G, H)$ is bounded by $\max\left\{e(G), pq\binom{n}{k}\right\}$, this implies the assertion in the case (2.2) of Theorem 1.1.

For any fixed bijection $\pi : V \to V$, the number of edges in $G_\pi \cap H$ is distributed binomially with parameters $\binom{n}{k}$ and pq. If $pq\binom{n}{k} > 4n \log n$ let $\lambda = 2\sqrt{pq\binom{n}{k} n \log n} \le pq\binom{n}{k}$. Then by Lemma 3.1, the probability that $\left|e(G_\pi \cap H) - pq\binom{n}{k}\right| > \lambda$ is at most $2e^{-n \log n}$. On the other hand, if $pq\binom{n}{k} \le 4n \log n$, let $\gamma' = 4e\frac{n \log n}{pq\binom{n}{k}} \ge e > 1$ and $\lambda = \frac{4e^2 n \log n}{\log \gamma'} \ge \frac{4e^2 n \log n}{\gamma'} = epq\binom{n}{k}$. Since $\binom{a}{b} \le \left(\frac{ea}{b}\right)^b$, the probability that $e(G_\pi \cap H) > \lambda$ is at most

$$\binom{\binom{n}{k}}{\lambda} (pq)^\lambda \le \left(\frac{e\binom{n}{k} pq}{\lambda}\right)^\lambda = \left(\frac{4e^2 n \log n}{\gamma' \lambda}\right)^\lambda = \left(\frac{\gamma'}{\log \gamma'}\right)^{-\frac{4e^2 n \log n}{\log \gamma'}} < e^{-n \log n}.$$

In either case, since there are $n!$ possible bijections $\pi : V \to V$, by the union bound

$$\mathbb{P}\left[\mathrm{disc}_P(G, H) > \lambda\right] \le n! \cdot 2e^{-n \log n} \le e^{-n/2},$$

which finishes the proof of the upper bound in case (1). Since γ (defined in Theorem 1.1) satisfies $\gamma = \Theta_k(\gamma')$, this implies upper bound in case (2.1) as well. Finally, observe that we divided the dense and sparse cases in this proof, according to whether $pq\binom{n}{k}$ is bigger (or smaller) than $4n \log n$, a threshold slightly different than the one used in Theorem 1.1. This difference is not essential though, as for p, q satisfying both $pq\binom{n}{k} \le 4n \log n$ and $pqN \ge \frac{1}{30} \log n$, we have $\sqrt{pq\binom{n}{k} n \log n} = \Theta_k\left(\frac{4e^2 n \log n}{\log \gamma'}\right)$. □

5 Lower bounds

In this section we prove the lower bounds in Theorem 1.1. As we previously explained, it is enough to obtain these bounds for $\mathrm{disc}_P(G, H)$. We divide the proof into two cases. The first (*dense case*) will be discussed in the next subsection. The second (*sparse case*) will be discussed in subsection 5.2. Throughout the proofs, we assume that k is fixed and n is tending to infinity.

5.1 Dense Case

Let $N = \binom{n-\frac{n}{k}}{k-1}$ and let p, q be such that $pqN > \frac{1}{30}\log n$. Select an arbitrary set $L \subseteq V$ of size $|L| = \frac{n}{k}$. We prove that w.h.p. there exists an *L-bijection* $\pi : V \to V$ with overlap

$$e(G_\pi \cap H) \geq pq\binom{n}{k} + \Theta_k\left(n \cdot \sqrt{pqN\log n}\right) = pq\binom{n}{k} + \Theta_k\left(\sqrt{pq\binom{n}{k}n\log n}\right), \tag{5}$$

where an L-bijection $\pi : V \to V$ is a bijection from V to V which only permutes the elements of L, i.e., $\pi(x) = x$ for all $x \notin L$.

We start by describing the construction outlined in Section 2 in more details. From the random hypergraph G we construct a random bipartite graph $\widetilde{G}$ with vertex set $L_G \cup R$, where $L_G = L$ and R is the set of all $(k-1)$-tuples in $V \setminus L$. Note that $|R| = N$. The vertices $v_1 \in L_G$ and $\{v_2, v_3, \ldots, v_k\} \in R$ are adjacent if $\{v_1, v_2, \ldots, v_k\}$ forms an edge in the hypergraph G. With slight abuse of notation, we view $\widetilde{G}$ as a sub-hypergraph of G, containing all edges e having exactly one vertex in L, i.e. $|e \cap L| = 1$. Similarly, from the random hypergraph H we construct a random bipartite graph $\widetilde{H}$ with vertex set $L_H \cup R$. Figure 1 shows the resulting bipartite graphs.

Given an L-bijection $\pi : V \to V$, we divide the edge set of $G_\pi \cap H$ into two subsets: the edge set of $\widetilde{G}_\pi \cap \widetilde{H}$ and its complement. To prove our result we first expose the random edges in $\widetilde{G}$ and $\widetilde{H}$, and show how to find an L-bijection π having overlap at least $\Theta_k\left(n \cdot \sqrt{pqN\log n}\right)$ more than the expectation. Then we fix such π and expose all the remaining edges in G and H showing that the contribution of these edges to $G_\pi \cap H$ does not deviate much from the expected contribution. More precisely, let $e_\pi = |E((G - \widetilde{G})_\pi) \cap E(H - \widetilde{H})|$, then $e(G_\pi \cap H) = e(\widetilde{G}_\pi \cap \widetilde{H}) + e_\pi$. Moreover, e_π is distributed according to $\mathrm{Bin}(m, pq)$, where $\frac{1}{2}\binom{n}{k} \leq m = \binom{n}{k} - N\frac{n}{k} \leq \binom{n}{k}$. Thus w.h.p. $|e_\pi - pqm| < \sqrt{pqm} \cdot \log n$, as Lemma 3.1 shows. Also, $\sqrt{pqm} \cdot \log n \ll \sqrt{pq\binom{n}{k}n\log n} = \Theta_k\big(n\sqrt{pqN\log n}\big)$. To obtain (5), it is therefore enough to show that w.h.p. there exists an L-bijection π such that

$$e(\widetilde{G}_\pi \cap \widetilde{H}) \geq \frac{n}{k} \cdot \left(pqN + \Theta_k\left(\sqrt{pqN\log n}\right)\right). \tag{6}$$

since then w.h.p.,

$$\begin{aligned} e(G_\pi \cap H) &= e(\widetilde{G}_\pi \cap \widetilde{H}) + e_\pi \\ &\geq \frac{n}{k}(pqN + \Theta_k(\sqrt{pqN\log n})) + pqm - \sqrt{pqm}\log n \\ &= \frac{n}{k}\Theta_k(\sqrt{pqN\log n}) + pq\binom{n}{k} - \sqrt{pqm}\log n \\ &= pq\binom{n}{k} + \Theta_k\left(\sqrt{pq\binom{n}{k}n\log n}\right). \end{aligned}$$

We define an auxiliary bipartite graph $\Gamma = \Gamma(\widetilde{G}, \widetilde{H})$ as follows. A vertex $u \in L_G$ *survives* if $|\deg_{\widetilde{G}}(u) - pN| \leq 2\sqrt{2pN}$ and similarly, a vertex $v \in L_H$ *survives* if $|\deg_{\widetilde{H}}(v) - qN| \leq 2\sqrt{2qN}$. Let S_G and S_H be the sets of all surviving vertices of $\widetilde{G}$ and $\widetilde{H}$, respectively. Let $s_G = |S_G|$ and $s_H = |S_H|$. The set of vertices of Γ is the union of S_G and S_H. The edges of Γ are defined by the property

$$u \sim_\Gamma v \iff \mathrm{codeg}(u, v) \geq \frac{\deg_{\widetilde{G}}(u)\deg_{\widetilde{H}}(v)}{N} + 10^{-2}\sqrt{pqN\log n},$$

where $\mathrm{codeg}(u,v)$ denotes the *codegree* of $u \in L_G$ and $v \in L_H$, i.e. $\mathrm{codeg}(u,v) = |N_{\widetilde{G}}(u) \cap N_{\widetilde{H}}(v)|$. The graph Γ has many vertices in both parts, as the following simple lemma demonstrates

Lemma 5.1. *W.h.p. each part of Γ has size at least $\frac{n}{4k}$.*

Proof. Let α be the probability that some vertex u survives in L_G. Since $pN \geq 8$, we have that $2\sqrt{2pN} \leq pN$. Thus Lemma 3.1 applied to $\deg_{\widetilde{G}}(u)$ implies $\alpha \geq 1 - 2e^{-2} \geq 1/2$. Since the events that vertices survive are independent, s_G stochastically dominates the binomial distribution with parameters n/k and $1/2$. Thus, again by Lemma 3.1, w.h.p. $s_G \geq n/(4k)$ and a similar estimate holds for s_H. □

To prove (6), we will show that the following two statements hold w.h.p.

(a) Γ has a matching $M = \{(u_1,v_1),\dots,(u_l,v_l)\}$ of size $l = \frac{n}{50k}$;

(b) there exists an L-bijection π such that $\pi(u_i) = v_i$ for all $i = 1,2,\dots,l$, and,

$$\sum_{u \in L_G \setminus \{u_1,u_2,\dots,u_l\}} \mathrm{codeg}(u,\pi(u)) \geq \left(\frac{n}{k} - l\right) pqN - 2\frac{n}{k}\sqrt{pqN}.$$

Indeed, for any two adjacent vertices u, v in Γ, we have

$$\frac{\deg_{\widetilde{G}}(u)\deg_{\widetilde{H}}(v)}{N} \geq \frac{(pN - \sqrt{8pN})(qN - \sqrt{8qN})}{N} \geq pqN - 6\sqrt{pqN}.$$

Thus using (a), (b) and $l = \frac{n}{50k}$ we obtain

$$\begin{aligned}
e(\widetilde{G}_\pi \cap \widetilde{H}) &= \sum_{u \in L_G} \mathrm{codeg}(u,\pi(u)) \geq \sum_{i=1}^{l} \mathrm{codeg}(u_i,v_i) + \left(\frac{n}{k} - l\right) pqN - 2\frac{n}{k}\sqrt{pqN} \\
&\geq \sum_{i=1}^{l} \left[\frac{\deg_{\widetilde{G}}(u_i)\deg_{\widetilde{H}}(v_i)}{N} + 10^{-2}\sqrt{pqN\log n}\right] + \left(\frac{n}{k} - l\right) pqN - 2\frac{n}{k}\sqrt{pqN} \\
&\geq \sum_{i=1}^{l} \left[pqN - 6\sqrt{pqN}\right] + \frac{n}{50k}10^{-2}\sqrt{pqN\log n} + \left(\frac{n}{k} - l\right) pqN - 2\frac{n}{k}\cdot\sqrt{pqN} \\
&\geq \frac{n}{k}\left(pqN + 10^{-4}\sqrt{pqN\log n}\right)
\end{aligned}$$

We need the following lemma in order to prove that (b) holds.

Lemma 5.2. *Let $0 < \alpha < 1$ be any absolute constant. Then with probability at least $1 - e^{-\frac{n}{k}}$, any two subsets $A \subseteq L_G$ and $B \subseteq L_H$ with $|A| = |B| = \frac{\alpha n}{k}$ satisfy*

$$X_{A,B} := \sum_{u \in A, v \in B} \mathrm{codeg}(u,v) \geq \left(\frac{\alpha n}{k}\right)^2 pqN - 2\alpha\left(\frac{n}{k}\right)^2\sqrt{pqN}.$$

Proof. Let $X_{w,u,v}$ be the indicator of $wu \in E(\widetilde{G})$ and $wv \in E(\widetilde{H})$ for $w \in R, u \in A, v \in B$. So $X_{A,B} = \sum_{w \in R, u \in A, v \in B} X_{w,u,v}$ and $\mathbb{E}[X_{w,u,v}] = pq$. Moreover, $X_{w,u,v}$ and $X_{w',u',v'}$ are dependent if and only if $wu = w'u'$ or $wv = w'v'$. Thus, $\mu = \mathbb{E}[X_{A,B}] = \left(\frac{\alpha n}{k}\right)^2 Npq$ and

$$\Delta = \sum_{w \in R, u \in A} \sum_{v,v' \in B} \mathbb{E}[X_{w,u,v} \cdot X_{w,u,v'}] + \sum_{w \in R, v \in B} \sum_{u,u' \in A} \mathbb{E}[X_{w,u,v} \cdot X_{w,u',v}] = \frac{\alpha n}{k}\binom{\frac{\alpha n}{k}}{2} Npq\,(p+q)\,,$$

where μ and Δ are defined as in Lemma 3.2. Let F be the event that there exists at least one pair of subsets $A \subseteq L_G, B \subseteq L_H$ with $|A| = |B| = \frac{\alpha n}{k}$ satisfying $X_{A,B} < (\frac{\alpha n}{k})^2 Npq - 2\alpha(\frac{n}{k})^2\sqrt{Npq}$. By the union bound and by Lemma 3.2, we have

$$\begin{aligned}\mathbb{P}[F] &\leq \sum_{A\in\binom{L_G}{\alpha n}, B\in\binom{L_H}{\alpha n}} \mathbb{P}\left[X_{A,B} < \mu - 2\alpha\left(\frac{n}{k}\right)^2\sqrt{Npq}\right] \leq \binom{\frac{n}{k}}{\frac{\alpha n}{k}}^2 e^{-\frac{\left(2\alpha(\frac{n}{k})^2\sqrt{Npq}\right)^2}{2\mu+\Delta}} \\ &\leq \left(\frac{e}{\alpha}\right)^{\frac{2\alpha n}{k}} e^{-3\frac{n}{k}} \leq e^{-\frac{n}{k}},\end{aligned}$$

since $2\mu + \Delta \leq \frac{4}{3}\left(\frac{\alpha n}{k}\right)^3 Npq$, $\alpha < 1$ and $\alpha\log(e/\alpha) \leq 1$ for all such α. □

Let $M = \{(u_1, v_1), \ldots, (u_l, v_l)\}$ be a matching satisfying (a) and let $A = L_G \setminus \{u_1, u_2, \ldots, u_l\}$ and $B = L_H \setminus \{v_1, v_2, \ldots, v_l\}$. Write $|A| = |B| = \frac{n}{k} - l = \frac{\alpha n}{k}$, where $\alpha = \frac{49}{50}$. Consider $X_{A,B} = \sum_{u\in A, v\in B}\operatorname{codeg}(u, v)$. Then, by Lemma 5.2, with probability at least $1 - e^{-\frac{n}{k}}$, we have

$$\sum_{u\in A, v\in B}\operatorname{codeg}(u, v) \geq \left(\frac{n}{k} - l\right)^2 pqN - 2\frac{n}{k}\left(\frac{n}{k} - l\right)\sqrt{pqN}.$$

Since the complete bipartite graph with parts A, B is a disjoint union of $\frac{n}{k} - l$ perfect matchings, by the pigeonhole principle, there exists a matching M' between A and B such that

$$\sum_{(u,v)\in M'}\operatorname{codeg}(u, v) \geq \frac{\sum_{u\in A, v\in B}\operatorname{codeg}(u, v)}{\frac{n}{k} - l} \geq \left(\frac{n}{k} - l\right)pqN - \frac{2n}{k}\sqrt{pqN}.$$

Then the matching $M \cup M'$ between L_G and L_H gives the desired L-bijection π and proves (b).

To finish the proof we need to establish (a). If Γ is nearly regular, then by Lemma 3.7, Γ would contain a linear-size matching. Unfortunately, it is not clear that this is the case. However, we will show that it is possible to delete some edges of Γ at random and obtain a *pruned graph* Γ', which is nearly regular. Let

$$f(d_1, d_2) := \mathbb{P}\left[u \sim_\Gamma v \middle| \deg_{\widetilde{G}}(u) = d_1, \deg_{\widetilde{H}}(v) = d_2\right],$$

where $|d_1 - pN| \leq 2\sqrt{2pN}$ and $|d_2 - qN| \leq 2\sqrt{2qN}$. Let f_0 be the minimum of $f(d_1, d_2)$ over all pairs (d_1, d_2) in the domain of f. Suppose that $f_0 \geq n^{-\frac{1}{2}}$, which we shall prove later. We keep each edge uv of Γ in Γ' independently with probability $\frac{f_0}{f(d_1,d_2)}$, where $d_1 = \deg_{\widetilde{G}}(u)$ and $d_2 = \deg_{\widetilde{H}}(v)$. Then, we claim that for any vertex $u \in S_G$, $\deg_{\Gamma'}(u)$ is binomially distributed with parameters s_H and f_0. Indeed, by definition, $\mathbb{P}\left[u \sim_{\Gamma'} v \middle| \deg_{\widetilde{G}}(u) = d_1, \deg_{\widetilde{H}}(v) = d_2\right] = f_0$ for all possible d_1, d_2. Moreover, conditioning on the neighbors of u in $\widetilde{G}$ and on the values of the degrees $\deg_{\widetilde{H}}(v_1)$, $\deg_{\widetilde{H}}(v_2)$, …, $\deg_{\widetilde{H}}(v_m)$, the events $u \sim_\Gamma v_1, u \sim_\Gamma v_2$, …, and $u \sim_\Gamma v_m$ are all independent. Therefore, by definition of Γ', it is easy to see that $u \sim_{\Gamma'} v_1$, $u \sim_{\Gamma'} v_2$, …, and $u \sim_{\Gamma'} v_m$ are independent as well. Thus for any $u \in S_G$, $\deg_{\Gamma'}(u) \sim \operatorname{Bin}(s_H, f_0)$ and similarly, $\deg_{\Gamma'}(v) \sim \operatorname{Bin}(s_G, f_0)$ for all $v \in S_H$.

Conditioning on the degrees of all vertices in $\widetilde{G}, \widetilde{H}$, we obtain sets S_G and S_H, which w.h.p. satisfy the assertion of Lemma 5.1, i.e., $|S_G| = s_G \geq \frac{n}{4k}$ and $|S_H| = s_H \geq \frac{n}{4k}$. Thus both $s_G f_0$ and $s_H f_0$ are $\Omega_k(\sqrt{n})$. Since all degrees in Γ' are binomially distributed, Lemma 5.1 together with the union bound imply that w.h.p. all vertices $u \in S_G, v \in S_H$ satisfy

$$\frac{s_H f_0}{2} \leq \deg_{\Gamma'}(u) \leq \frac{3s_H f_0}{2} \quad \text{and} \quad \frac{s_G f_0}{2} \leq \deg_{\Gamma'}(v) \leq \frac{3s_G f_0}{2}.$$

Therefore, the max-degree $\Delta(\Gamma') \le \max\left\{\frac{3s_Hf_0}{2}, \frac{3s_Gf_0}{2}\right\} \le \frac{3nf_0}{2k}$ and $e(\Gamma') \ge \frac{s_Gs_Hf_0}{2} \ge \frac{n^2f_0}{32k^2}$. Thus by Lemma 3.7, Γ' has a matching of size at least $\frac{e(\Gamma')}{\Delta(\Gamma')+1} \ge \frac{n}{50k}$, completing the proof of (a).

It remains to prove the bound $f_0 \ge n^{-\frac{1}{2}}$. Let $K = \frac{\log n}{5000} \ge 1$. Since pN tends to infinity, $p \le q \le 1/2$ and $|d_1 - pN| \le 2\sqrt{2pN}$, we have $1 \le d_1 = (1+o(1))pN \le \frac{2N}{3}$. Similarly $1 \le d_2 = (1+o(1))qN \le \frac{2N}{3}$. Also recall that $pqN \ge \frac{1}{30}\log n$, which implies

$$\frac{d_1d_2}{100N} = (1+o(1))\frac{pqN}{100} \ge (1+o(1))\frac{\log n}{3000} > K.$$

Therefore we can apply Lemma 3.5 with $\Delta = \sqrt{\frac{d_1d_2K}{N}} > \frac{\sqrt{pqN\log n}}{100}$. By the definition of $f(d_1,d_2)$, we have

$$f(d_1,d_2) = \sum_{t \ge \frac{d_1d_2}{N} + \frac{\sqrt{pqN\log n}}{100}} \frac{\binom{d_1}{t}\binom{N-d_1}{d_2-t}}{\binom{N}{d_2}} \ge \sum_{t \ge \frac{d_1d_2}{N} + \Delta} \frac{\binom{d_1}{t}\binom{N-d_1}{d_2-t}}{\binom{N}{d_2}} \ge e^{-40K} > n^{-\frac{1}{2}}.$$

This completes the proof. □

5.2 Sparse case

In this subsection, we prove the lower bound in the sparse case $pqN \le \frac{1}{30}\log n$. Note that, since $p \le q$ and $\binom{n}{k} \le 3N\frac{n}{k}$ in this case, we have $p \le N^{-1/2+o(1)}$ and $pq\binom{n}{k} < n\log n$. The proof runs along the same lines as that of the dense case differing only in the application of Lemma 3.6 to obtain an L-bijection $\pi : V \to V$ whose sum of codegrees $\sum_{u\in L_G}\mathrm{codeg}(u,\pi(u))$ is large. Suppose first that $pN \ge \frac{\log n}{5\log\gamma}$. Recall that $\gamma = \frac{\log n}{pqN} \ge 30$ and thus $\frac{\log n}{6\log\gamma} \ge \frac{\log n}{42\log\gamma} + \frac{\log n}{\gamma} = \frac{\log n}{42\log\gamma} + pqN$. Also, $\sqrt{pqm}\log n \le \sqrt{pq\binom{n}{k}}\log n \ll \frac{\log n}{42\log\gamma}\frac{n}{k}$. Therefore it is enough to find a bijection π between L_G and L_H such that $\sum_{u\in L_G}\mathrm{codeg}(u,\pi(u)) \ge (1+o(1))\frac{n}{k}\cdot\frac{\log n}{6\log\gamma}$. Using such bijection, together with above inequalities and $m + N\frac{n}{k} = \binom{n}{k}$, we obtain that

$$\begin{aligned} e(G_\pi \cap H) &= \sum \mathrm{codeg}(u,\pi(u)) + e_\pi \\ &\ge (1+o(1))\frac{n}{k}\frac{\log n}{6\log\gamma} + pqm - \sqrt{pqm}\log n \\ &\ge (1+o(1))\frac{\log n}{42\log\gamma}\frac{n}{k} + pq\binom{n}{k}. \end{aligned}$$

Analogous to the dense case, we define the connection graph $\Gamma = \Gamma(\widetilde{G},\widetilde{H})$ for the sparse case. But the criterion to add edges to Γ is different – u and v are joined if and only if $\mathrm{codeg}(u,v) \ge \frac{\log n}{6\log\gamma}$. Again, our goal is to find a large matching in Γ, but the strategy will be slightly different this time.

Partition the vertices of L_G into $r = \frac{n}{ks}$ disjoint sets $S_1,\dots,S_r$ each of size $s = n^{2/5}$. We will construct π by applying the following greedy algorithm to each set. Let us start with S_1. The algorithm will reveal the edges emanating from S_1 to R in $\widetilde{G}$ by repeatedly exposing the neighborhood of a vertex in S_1, one at a time. Throughout this process, we construct a subset $S_1' \subseteq S_1$ of size $(1+o(1))|S_1|$ and a family of disjoint sets $N_u \subseteq R$, such that each N_u has size $(1+o(1))Np$ and is contained in the neighborhood of u, for all $u \in S_1'$. At each step, we pick a fresh vertex u in S_1 and expose its neighborhood. If u has a set of $(1+o(1))Np$ neighbors which is disjoint from N_w for all w in the current S_1', denote this particular set by N_u and put u in the set S_1'; otherwise move to

the next fresh vertex in S_1, until there are none left. The union $X = \cup_{w \in S_1'} N_w$ always has size at most $O(pN \cdot s) \leq N^{0.9+o(1)}$. Moreover, every vertex in $R \setminus X$ is adjacent to u independently with probability p. Since $pN \geq \omega(1)$ tends to infinity with n, the set of neighbors of u outside X has size $(1+o(1))|R \setminus X|p = (1+o(1))Np$ with probability $1+o(1)$. Thus, there exists an absolute lower bound $p_0 = 1+o(1)$ such that the event "S_1' contains u" occurs with probability at least p_0, for all u. Furthermore, conditioned on the sizes of $R \setminus X$, these events are independent for different vertices u. A straightforward coupling argument shows that the number of elements in S_1' can be bounded below by a binomial random variable with s trials and probability p_0. Therefore, by Lemma 3.1, w.h.p. $|S_1'| = (1+o(1))|S_1|$. Next, we construct the partial matching for S_1. Consider the disjoint sets N_u, for $u \in S_1'$, each of size $(1+o(1))Np$. Pick an arbitrary vertex v in L_H and expose its neighbors in $\widetilde{H}$. This is a random subset N_v of R, obtained by taking each element independently with probability q. Therefore by case (1) of Lemma 3.6, w.h.p there is a vertex $u \in S_1'$ such that $\operatorname{codeg}(u,v) \geq |N_u \cap N_v| \geq \frac{\log n}{6 \log \gamma}$. Define $\pi(u) = v$, remove u from S_1', remove v from L_H and continue. Note that, as long as there are at least $n^{1/3}$ vertices remaining in S_1', we can match one of them with a newly exposed vertex from L_H such that the codegree of this pair is at least $\frac{\log n}{6 \log \gamma}$. Once the number of vertices in S_1' drops below $n^{1/3}$, leave the remaining vertices unmatched. W.h.p. we can match a $1+o(1)$ fraction of the vertices in S_1.

Continue the above procedure for $S_2, \ldots, S_r$ as well. At the end of the process, we will have matched a $1+o(1)$ fraction of all the vertices in L_G with distinct vertices in L_H such that codegree of every matched pair is at least $\frac{\log n}{6 \log \gamma}$. Therefore the sum of the codegrees of this partial matching is at least $(1+o(1))\frac{n}{k} \cdot \frac{\log n}{6 \log \gamma}$. To obtain the bijection π, one can match the remaining vertices in L_G and L_H arbitrarily.

When $pN < \frac{\log n}{5 \log \gamma}$ the same proof as above together with case (2) of Lemma 3.6 yields a bijection π such that $\sum_{u \in L_G} \operatorname{codeg}(u, \pi(u)) \geq (1+o(1))\frac{n}{k} \cdot pN$. Since $q \leq \frac{1}{2}$, $p \geq \frac{\omega(1)}{N}$ and $m = \binom{n}{k} - N\frac{n}{k}$, this implies

$$\begin{aligned} e(G_\pi \cap H) &\geq (1+o(1))\frac{n}{k}pN + pqm - \sqrt{pqm}\log n \\ &= \Theta_k\left(p\binom{n}{k}\right) + pq\binom{n}{k}. \end{aligned}$$

finishing the analysis of the sparse case. □

6 Concluding remarks

As we stated in the introduction, Theorem 1.1 also yields tight bounds when p and/or $q > \frac{1}{2}$. For any G and H, one can check that $\operatorname{disc}(G, \overline{H}) = \operatorname{disc}(G, H)$, where $\overline{H}$ is the complement of H. Moreover, $\overline{H}$ is distributed according to $\mathcal{H}_k(n, 1-q)$, hence we can reduce the case $q > \frac{1}{2}$ to the case $q' = 1-q \leq \frac{1}{2}$; the same holds when we take the complement of G instead. We remark that one can determine the discrepancy when p is smaller than $\frac{\omega(1)}{N}$, but we chose not to discuss this range here, since the proof is similar to the sparse case and it wouldn't provide any new insight.

The definition of discrepancy can be rephrased as $\operatorname{disc}(G,H) = \max\{\operatorname{disc}^+(G,H), \operatorname{disc}^-(G,H)\}$, where $\operatorname{disc}^+(G,H) = \max_\pi e(G_\pi \cap H) - \rho_G\rho_H\binom{n}{k}$ and $\operatorname{disc}^-(G,H) = \rho_G\rho_H\binom{n}{k} - \min_\pi e(G_\pi \cap H)$ are the *one-sided relative discrepancies*. In fact, all the lower bounds we obtained are for $\operatorname{disc}^+(G,H)$, and some of them are not true for $\operatorname{disc}^-(G,H)$. This is because $\operatorname{disc}^-(G,H) \leq \rho_G\rho_H\binom{n}{k} \simeq pq\binom{n}{k}$ and in

the sparse case, $pq\binom{n}{k}$ could be much smaller than $\mathrm{disc}(G,H)$. Under the same hypothesis and using similar ideas as in Theorem 1.1, one can show that

$$\mathrm{disc}^-(G,H) = \begin{cases} \Theta_k\left(\sqrt{pq\binom{n}{k} n\log n}\right) & \text{if } pqN > \frac{1}{30}\log n; \\ \Theta_k\left(pq\binom{n}{k}\right) & \text{otherwise.} \end{cases}$$

The last equation is related to the lower tail of the binomial distribution.

It would be interesting to determine the exact dependence on k of the relative discrepancy. It also worth mentioning that there are a substantial number of open problems about $\mathrm{disc}(G,H)$ and its related topics in [4].

Acknowledgment. We would like to thank two anonymous referees for the thorough and helpful comments and suggestions on the early version of this paper.

Note added in proof. After this paper was written and submitted for publication, we learned that Bollobás and Scott [5] obtained similar results.

References

[1] N. Alon and J. Spencer, *The Probabilistic Method*, John Wiley Inc., New York (2008).

[2] J. Beck and V.T. Sós, *Discrepancy theory*, in Handbook of Combinatorics, Vol. 2, 1405–1446, Elsevier, Amsterdam, 1995.

[3] B. Bollobás and A. Scott, *Discrepancy in graphs and hypergraphs*, in More sets, graphs and numbers, Ervin Gyori, Gyula O.H. Katona and Laszlo Lovász, eds, pp. 33–56, Bolyai Soc. Math. Stud. **15**, Springer, Berlin, 2006.

[4] B. Bollobás and A. Scott, *Intersection of graphs*, J. Graph Theory **66** (2011), 261–282.

[5] B. Bollobás and A. Scott, *Intersection of hypergraphs*, preprint.

[6] B. Chazelle, *The discrepancy method*, Cambridge University Press, Cambridge, 2000, xviii+463 pp.

[7] F.R.K. Chung, R.L. Graham and R.M. Wilson, *Quasi-random graphs*, Combinatorica **9** (1989), 345–362.

[8] P. Erdős, M. Goldberg, J. Pach and J. Spencer, *Cutting a graph into two dissimilar halves*, J. Graph Theory **12** (1988), 121–131.

[9] P. Erdős and J. Spencer, *Imbalances in k-colorations*, Networks **1** (1971/2), 379–385.

[10] C. Lee, P. Loh and B. Sudakov, *Self-similarity of graphs*, SIAM J. of Discrete Math. **27** (2013), 959–972.

[11] J. Matoušek, *Geometric discrepancy*, Algorithms and Combinatorics **18**, Springer-Verlag, Berlin, 1999, xii+288 pp.

[12] V.T. Sós, *Irregularities of partitions: Ramsey theory, uniform distribution*, in Surveys in Combinatorics (Southampton, 1983), 201–246, London Math. Soc. Lecture Note Ser., 82, Cambridge Univ. Press, Cambridge-New York, 1983.