Sharp Concentration of Hitting Size for Random Set Systems

Jessie Deering, Anant Godbole, and William Jamieson
Department of Mathematics and Statistics
East Tennessee State University

Lucia Petito
Department of Biostatistics, University of California, Berkeley

November 11, 2018

Abstract

Consider the random set system $\mathcal{A} = R(n, p)$ of $[n] := \{1, 2, \ldots n\}$, where $\mathcal{A} = \{A_j : A_j \in \mathcal{P}([n]), \text{ and } A_j \text{ selected with probability } p = p_n\}$. A set $H \subseteq [n]$ is said to be a hitting set for $\mathcal{A}$ if $\forall A_j \in \mathcal{A} \ |A_j \cap H| \geq 1$. The second moment method is used to exhibit the sharp concentration of the minimal size of H for a variety of values of p.

1   Introduction and Motivation

A set D of vertices in a graph $G = (V, E)$ forms a *dominating set* of G if each $v \in V$ is either in D or adjacent to some $d \in D$. The domination number $\gamma = \gamma(G)$ is the size of the smallest dominating set of G. Given a graph of minimum degree δ, it is proved, e.g., in Alon and Spencer [4] that

$$\gamma(G) \leq \frac{1 + \ln(\delta + 1)}{\delta + 1}. \tag{1}$$

In a result of direct relevence to this paper, Weber [12] proved in 1981 that the domination number of the random graph $G(n, p)$ is sharply concentrated w.h.p. if p is fixed. This result was extended in [13] to the case $p = p_n \to 0$,

1

where a two point concentration was shown to hold for $\gamma(G(n, p_n))$ provided p_n did not decay too rapidly; specifically, $p = 1/\log\log n$ works in the above result; $p = 1/\log n$ does not.

Given a k-uniform hypergraph $H = (V, E_k)$, a *transversal* is a collection T of vertices such that each edge $e \in E_k$ intersects T in at least one vertex. We will denote the transversal number of H by $\tau(H)$. A transversal is also called a *hitting set*, particularly in the Computer Science literature, where it is more typical than not for edges to be of different sizes. Accordingly, we will reserve the terminology "transversal" for k-uniform hypergraphs, and "hitting set" for the general case. In a result that echoes (1), Alon [2] proved that for a k-uniform hypergraph with v vertices and e edges,

$$\tau(H) \leq (1 + o(1))\frac{\log k}{k}(v + e) \ \ (k \to \infty).$$

The Computer Science literature has focused more on complexity issues for hitting sets; see, e.g., [7], [5], [9], and [8]. The connection between *total domination* and transversals has been explored in [11].

If all our edges are of cardinality two, i.e. if we have a graph, let $s \notin T$, T a transversal. Then the only edges containing s, must be between s and t for $t \in T$. Thus T is a minimal hitting set iff T^C is maximal independent. Note that this is also true for arbitrary hypergraphs if independent sets are defined as collections of vertices for which there is no edge that is a subset of these vertices. Also, the sharp two point concentration of the maximal independent set in a random graph has been well understood since the early work of Bollobás and Erdős [6] and Matula [10], and others. In these results on finite point concentration, nothing more than the second moment method was used, though more sophisticated machinery was employed by Alon and Krivelevich [3], and Achlioptas and Naor [1] to show the sharp concentration of the chromatic number of $G(n, p)$. It will turn out that elementary methods will suffice in this paper; we will investigate the sharp concentration of the size of minimal hitting sets (or *hitting number*) for non-uniform hypergraphs.

Our model consists of picking each set $A \subseteq \{1, 2, \ldots, n\}$ with probability $p = p_n$. Let $\mathcal{A}$ be the ensemble of picked sets, which we will call a *random set system* and denote by $R(n, p)$ (to mirror the $G(n, p)$ notation for a random graph). The goal is to discover a class of ps for which the hitting number is close to the intuitive guess of $\lg(p \cdot 2^n)$, where throughout this paper $\lg = \log_2$. In Section 2, we set the stage for when a one or two point concentration holds for the hitting number, and, in Sections 3 and 4, details are provided for two

canonical cases, namely those corresponding to $p = 1/2^{n^\beta}$ and $p = n^\alpha/2^n$.

2 Setting up the Two-Point Concentration

Define the baseline random variable, X_m, to be the number of hitting sets of size m. We start by exploring a lower bound on $|H|$. Clearly

$$\mathbb{E}(X_m) = \binom{n}{m}(1-p)^{2^{n-m}} \leq \binom{n}{m}\exp\{-p2^{n-m}\},$$

since a set of size m is hitting iff we do not pick any of the subsets of its complement to be in the random set system (actually we cannot by definition hit the empty set, so the correct exponent ought to be $2^{n-m} - 1$). Let us set (with hindsight) $m = \lg(p \cdot 2^n) - \varphi(n)$.[1] Thus

$$\mathbb{P}(X_m \geq 1) \leq \mathbb{E}(X_m) \leq \binom{n}{m}\exp\{-2^{\varphi(n)}\} \to 0 \qquad (2)$$

provided that $\binom{n}{m} \ll \exp\{2^{\varphi(n)}\}$, and, using the inequality $1 - p \geq e^{-p/(1-p)}$, with $\epsilon_n = \lg(1-p)^{-1} = O(p)$,

$$\mathbb{E}(X_m) = \binom{n}{m}(1-p)^{2^{n-m}} \geq \binom{n}{m}\exp\{-p2^{n-m+\varepsilon_n}\} \to \infty \qquad (3)$$

if $\binom{n}{m} \gg \exp\{2^{\varphi(n)+\epsilon_n}\}$, where the φ functions in (2) and (3) are different. Since zero-one probability thresholds often occur precisely where the associated expected value transitions from zero to infinity, we anticipate that Equations (2) and (3) occur with near-consecutive values of m.

By Chebychev's inequality, $\mathbb{P}(X_m = 0) \leq \frac{\mathbb{V}(X_m)}{\mathbb{E}^2(X_m)}$, so to establish an upper bound on $|H|$ it would suffice to show that the variance is an order of magnitude smaller than the square of the mean whenever $m \geq m_0$ – for some m_0 to be determined. Since $X_m = \sum_{j=1}^{\binom{n}{m}} I_j$, where the indicator variable I_j equals one iff the jth m-set hits $R(n, p)$, we have that

---

[1]In this paper we will encounter several functions that play a "generic" role. Examples of these functions are $\omega(n)$, $\varphi(n)$, ϵ_n, and μ_n. They are each defined differently in various parts of the paper, but their *role* is always the same, e.g. $\varphi(n)$ will *always* denote how much smaller the hitting set size is than $\lg p \cdot 2^n$ and $\omega(n)$ will always be a function that tends to infinity at an arbitrarily slow rate.

$$\begin{aligned}
\mathbb{V}(X_m) &= \mathbb{E}(X_m^2) - \mathbb{E}^2(X_m) \\
&= \sum_{j=1}^{\binom{n}{m}} \mathbb{E}(I_j^2) - \left(\sum_{j=1}^{\binom{n}{m}} \mathbb{E}(I_j) \right)^2 + \sum_{j \neq k} \mathbb{E}(I_j I_k) \\
&= \mathbb{E}(X_m) - \mathbb{E}^2(X_m) + \sum_{j \neq k} \mathbb{E}(I_j I_k),
\end{aligned}$$

so that

$$\frac{\mathbb{V}(X_m)}{\lambda^2} = \frac{1}{\lambda} - 1 + \frac{\sum_{j \neq k} \mathbb{E}(I_j I_k)}{\lambda^2}, \tag{4}$$

where $\lambda = \lambda_m = \mathbb{E}(X_m)$. Now two sets A, B of size m that intersect in r elements both hit $\mathcal{A}$ iff we do not pick, as part of $\mathcal{A}$, any set that is a subset of A^C or a subset of B^C; there are $2^{n-m} + 2^{n-m} - 2^{n-2m+r}$ of these. Thus, substituting $s = m - r$ and assuming that $\lambda \geq 1$, we have

$$\begin{aligned}
\sum_{j \neq k} \mathbb{E}(I_j I_k) &= \binom{n}{m}^2 \cdot \sum_{r=0}^{m-1} \frac{\binom{m}{r}\binom{n-m}{m-r}}{\binom{n}{m}} (1-p)^{2^{n-m+1} - 2^{n-2m+r}} \\
&= \lambda^2 \sum_{s=1}^{m} \lambda^{-2^{-s}} \binom{m}{s}\binom{n-m}{s}\binom{n}{m}^{2^{-s}-1}, \\
&\leq \lambda^2 \sum_{s=1}^{m} \binom{m}{s}\binom{n-m}{s}\binom{n}{m}^{2^{-s}-1}. \tag{5}
\end{aligned}$$

By (4) and (5) it thus suffices to show that

$$\sum_{s \geq 1} \binom{m}{s}\binom{n-m}{s}\binom{n}{m}^{2^{-s}-1} = 1 + o(1) \tag{6}$$

as $\lambda \to \infty$; this is really a simple statement about the function $m = m(n)$ as $n \to \infty$. Let us set up what it takes to make (6) occur: We first define, with $s_0 = 2(\lg(m \log n))$, the sums

$$\Sigma_1 = \sum_{s \geq s_0} \binom{m}{s}\binom{n-m}{s}\binom{n}{m}^{2^{-s}-1}$$

4

and
$$\Sigma_2 = \sum_{1 \le s \le s_0 - 1} \binom{m}{s}\binom{n-m}{s}\binom{n}{m}^{2^{-s}-1}.$$

In Σ_1, we first bound as follows:
$$\binom{n}{m}^{2^{-s}} \le \left(\frac{ne}{m}\right)^{m/2^s} \le \left(\frac{ne}{m}\right)^{\frac{1}{m(\log n)^2}} = 1 + o(1),$$

so that
$$\Sigma_1 \le (1 + o(1)) \sum_{s \ge s_0 + 1} \frac{\binom{m}{m-s}\binom{n-m}{s}}{\binom{n}{m}} = 1 + o(1),$$

since the sum above represents almost entirely the mass of a hypergeometric variable with mean $\sim m$, provided that $s_0 \ll m$ – which holds if $m \ge \Omega(\log \log n)$. Turning to Σ_2, we have

$$
\begin{aligned}
\Sigma_2 &\le \sum_{1 \le s \le s_0 - 1} \binom{m}{s}\binom{n-m}{s}\binom{n}{m}^{-1/2} \\
&\le \sum_{1 \le s \le s_0 - 1} n^{2s}(n/m)^{-m/2} \\
&\le n^{2s_0 - m/2} m^{m/2} \\
&= e^{(2s_0 - m/2)\log n + (m/2)\log m}.
\end{aligned}
\tag{7}
$$

(where we used the bounds $\max\{\binom{m}{s}, \binom{n-m}{s}\} \le n^s$; $\binom{n}{m} \ge (n/m)^m$ in the second display above.) We wish the estimate in (7) to be of magnitude $o(1)$ and thus need
$$\log m < \left(1 - \frac{8\lg(m\log n)}{m}\right)\log n. \tag{8}$$

It is not too hard to check that (8) holds if m is not too small or too large; specifically one needs
$$\Omega(\log \log n) \le m \le n - \Omega(\log n). \tag{9}$$

So, for ms satisfying (9), we get that the hitting size is at least $m + 1$ if $\lambda = \lambda_m \to 0$, while if $\lambda = \lambda_m \to \infty$, then the hitting size is at most m. We next note that

$$\lambda_{m+1}^2 = \binom{n}{m+1}^2 (1-p)^{2^{n-m}} = \binom{n}{m+1}^2 \binom{n}{m}^{-1} \lambda_m \gg \lambda_m,$$

certainly for all $m \in [1, n-3]$. This leads to the conclusion that either $\lambda_m \to 0$ or $\lambda_{m+1} \to \infty$. If both these hold, then $|H| = m + 1$ w.h.p.; on the other hand if $\lambda_{m-1} \to 0; \lambda_m \to K; \lambda_{m+1} \to \infty$, or $\lambda_m \to 0; \lambda_{m+1} \to K; \lambda_{m+2} \to \infty$ for some $K \in \mathbb{R}^+$, then we have a two point concentration. We summarize the findings of this section in the following result:

Theorem 1. *Consider the random set system $\mathcal{A} = R(n, p)$, where p is unspecified. Let $\mathcal{F}$ denote the interval $[\Omega(\log \log n), n - \Omega(\log n)]$, where the constants in the Ω functions can be readily specified. Let $\ell = \sup\{m = m_n : \lim \mathbb{E}(X_m) = 0\}$ and $h = \inf\{m : \lim \mathbb{E}(X_m) = \infty\}$. Then, for suitable $p = p_n, \ell, h \in \mathcal{F}; h - \ell \in \{1, 2\}$ and $|H| = \ell + 1$ or $|H| = h$ w.h.p.*

It remains to solve for m in terms of p. In the next two sections, we consider the "dense" case, where the hitting size is comparable to n and the "sparse" case, where we will seek to hit a system $\mathcal{A}$ of size satisfying $|\mathcal{A}|^{1/n} \to 1$. For specificity we use the values $p = 1/2^{n\beta}; 0 < \beta < 1$ and $p = n^\alpha/2^n; \alpha > 0$ respectively, even though other choices could have been made, with the analysis being quite similar. In both sections, we seek to find a value of $m = m(p)$ for which $\mathbb{E}(X_m) \to 0$ and $\mathbb{E}(X_{m+1}) \to \infty$ (or $\mathbb{E}(X_{m+2}) \to \infty$).

3   A Dense Case, $p = 1/2^{n\beta}, 0 < \beta < 1$.

With $p = 1/2^{n\beta}$ and $m = (1 - \beta)n - \varphi(n)$, where we restrict $\varphi(n) \leq \lg n$,

$$\mathbb{E}(X_m) \leq \binom{n}{m} \exp\{-2^{\varphi(n)}\} \leq \binom{n}{\beta n}\left(\frac{1-\beta}{\beta}\right)^{\varphi(n)} \exp\{-2^{\varphi(n)}\}.$$

Stirling's formula next yields

$$\begin{aligned}\mathbb{E}(X_m) &\leq \frac{C}{\sqrt{n}}(\beta^{-\beta}(1 - \beta)^{-(1-\beta)})^n \left(\frac{1-\beta}{\beta}\right)^{\varphi(n)} \exp\{-2^{\varphi(n)}\} \\ &\leq \frac{C}{\sqrt{n}}\gamma^{\lg n}\delta^n \exp\{-2^{\varphi(n)}\},\end{aligned}$$

where C is a universal constant, $\gamma = \max\{1, \frac{1-\beta}{\beta}\}$, and $\delta := \beta^{-\beta}(1-\beta)^{\beta-1} \leq 2$. We thus see that

$$\mathbb{P}(X_m \geq 1) \leq \mathbb{E}(X_m) \to 0$$

if
$$2^{\varphi(n)} = n \ln \delta - \frac{1}{2} \ln n + (\ln \gamma)(\lg n) + \ln \omega(n) + \ln C,$$

or if
$$\varphi(n) = \lg \left(n \ln \delta - \frac{1}{2} \ln n + (\ln \gamma)(\lg n) + \ln \omega(n) \right)$$
$$= \lg(n \ln \delta) + o(1).$$

This yields
$$|H| \geq \lfloor (1 - \beta)n - \lg(n \ln \delta) - o(1) \rfloor + 1, \tag{10}$$

where in (10), $\varphi(n) \asymp \lg(n \ln \delta) \leq \lg n$ as stipulated.

For the lower bound, we argue as follows:
$$\mathbb{E}(X_m) = \binom{n}{m}(1-p)^{2^{n-m}} \geq \binom{n}{m} \exp\{-p2^{n-m+\varepsilon_n}\},$$

where $\epsilon_n = \lg(1-p)^{-1} = O(p)$. Setting $m = (1-\beta)n - \varphi(n)$ (we are in search of a different $\varphi(n)$ than in (10)) yields
$$\mathbb{E}(X_m) \geq \binom{n}{\beta n} \exp\{-p2^{n-m+\varepsilon_n}\}$$

if $\beta < 1/2$, and
$$\mathbb{E}(X_m) \geq \binom{n}{\beta n} \left(\frac{1-\beta}{2\beta} \right)^{\lg n} \exp\{-p2^{n-m+\varepsilon_n}\}$$

if $\beta \geq 1/2$. Simplifying as before we get $\mathbb{E}(X_m) \to \infty$ if
$$\varphi(n) = \lg \left(n \ln \delta - \frac{1}{2} \ln n + (\ln \eta)(\lg n) - \ln \omega(n) \right) - \varepsilon_n$$
$$= \lg(n \ln \delta) - o^*(1),$$

where $\eta = \min\{\frac{1-\beta}{2\beta}, 1\}$, and thus
$$|H| \leq \lceil (1 - \beta)n - \lg(n \ln \delta) + o^*(1) \rceil. \tag{11}$$

It is easy to verify that the $\varphi(n)$ functions in (10) and (11) differ by $o(1)$. Thus the worst case scenario is when these quantities straddle an integer, when we have a two point concentration. In the other case, we have that $|H|$ is a constant w.h.p.

We have proved

7

Theorem 2. *Let $H = H(n, \beta)$ be the size of the minimal hitting set of the random set system $R(n, 1/2^{n\beta})$ consisting of the ensemble that is generated when each set in $\mathcal{P}([n])$ is independently picked with probability $p = 2^{-n\beta}$. Then with probability approaching unity, $|H| = h$ or $h + 1$, where*

$$h = \lfloor (1 - \beta)n - \lg(n \ln \delta) - o(1) \rfloor + 1,$$

and where the $o(1)$ is as in the argument leading to (10).

The next result follows immediately:

Corollary 3. *Let $I = I(n, \beta)$ be the size of the maximal independent set of the random set system $R(n, 1/2^{n\beta})$ Then w.h.p., $|I| = i$ or $i + 1$, where*

$$i = n - 2 - \lfloor (1 - \beta)n - \lg(n \ln \delta) - o(1) \rfloor.$$

4 A Sparse Case, $p = n^\alpha/2^n, \alpha > 0$

We now move on to the case $p = n^\alpha/2^n, \alpha > 0$. Notice that as in Theorem 2, the hitting number works out to be a *just a little smaller* than the value $\lg\mathbb{E}|\mathcal{A}| = \lg(p \cdot 2^n)$, which can easily be seen to be the least m such that the set $\{1, 2, \ldots, m\}$ is *expected* to hit all the sets in $\mathcal{A}$.

Theorem 4. *Let $H = H(n, \alpha)$ be the size of the minimal hitting set of the random set system $R(n, n^\alpha/2^n), \alpha > 0$. Then with high probability, $|H| = h$ or $h + 1$, where*

$$h = \lfloor \alpha\lg n - \lg(\alpha\lg n \ln n) - o(1) \rfloor + 1.$$

Proof. Let X_m be as before. We have

$$\mathbb{E}(X_m) \leq \binom{n}{m} \exp\{-p2^{n-m}\} \leq \left(\frac{ne}{m}\right)^m \exp\{-p2^{n-m}\},$$

so, setting $p = n^\alpha/2^n$ and $m = \alpha\lg n - \varphi(n)$ (where we restrict by seeking solutions with $\varphi(n) \leq 3\lg(\lg n)$), we get

$$\mathbb{E}(X_m) \leq \left(\frac{ne}{\alpha\lg n - 3\lg\lg n}\right)^{\alpha\lg n} \exp\{-2^{\varphi(n)}\} \to 0$$

8

if

$$2^{\varphi(n)} = \alpha \lg n \left(\ln n + 1 - \ln(\alpha \lg n - 3 \lg \lg n) \right) + \ln \omega(n)$$

or

$$\begin{aligned} \varphi(n) &= \lg \left(\alpha \lg n \left(\ln n + 1 - \ln(\alpha \lg n - 3 \lg \lg n) \right) + \ln \omega(n) \right) \\ &= \lg(\alpha \lg n \ln n) + o(1). \end{aligned} \tag{12}$$

Note that $\varphi(n) \leq 3 \lg \lg n$ in (12) if n is sufficiently large. Thus

$$|H| \geq \lfloor \alpha \lg n - \lg(\alpha \lg n \ln n) - o(1) \rfloor + 1. \tag{13}$$

Next, setting $\varphi(n) = \lg[\alpha(1 - \mu_n) \lg n \ln n] - \epsilon_n$, where $\mu_n = \frac{5+\alpha}{\alpha} \frac{\lg \lg n}{\lg n}$, and $m = \alpha \lg n - \varphi(n)$, we see that

$$\begin{aligned} \mathbb{E}(X_m) &= \binom{n}{m}(1-p)^{2^{n-m}} \\ &\geq \binom{n}{m} \exp\{-2^{\varphi(n)+\epsilon_n}\} \\ &= \frac{\binom{n}{m}}{n^{\alpha(1-\mu_n)\lg n}} \\ &\geq \frac{(n-m)^m}{C\sqrt{m}} \left(\frac{e}{m}\right)^m \frac{1}{n^{\alpha(1-\mu_n)\lg n}} \\ &\geq \exp\{-m^2/(n-m)\} \frac{1}{C\sqrt{m}} \left(\frac{ne}{m}\right)^m \frac{1}{n^{\alpha(1-\mu_n)\lg n}} \\ &\geq \frac{1}{2C\sqrt{m}} \left(\frac{ne}{\alpha \lg n}\right)^{\alpha \lg n - \varphi(n)} \frac{1}{n^{\alpha(1-\mu_n)\lg n}} \\ &\geq \frac{1}{2C\sqrt{m}} \frac{n^{\alpha\mu_n \lg n - \varphi(n)}}{(\alpha \lg n)^{\alpha \lg n - \varphi(n)}} \\ &\geq \frac{1}{2C\sqrt{m}} \frac{n^{(5+\alpha)\lg \lg n - 3\lg \lg n}}{(\alpha \lg n)^{\alpha \lg n - \varphi(n)}} \\ &\geq n^{\lg \lg n} \\ &\to \infty. \end{aligned}$$

Together with (13), this completes the proof of Theorem 4. $\qquad \square$

9

5 Open Questions

We feel that deriving similar concentrations for hitting set size of random uniform hypergraphs would be of value, as would be results in which subsets of various sizes are picked with (a wide variety of) size-biased probabilities.

6 Acknowledgments

References

[1] D. Achlioptas and A. Naor (2005). "The two possible values of the chromatic number of a random graph," *Ann. Math.* **162,** 1335–1351.

[2] N. Alon (1990). "Transversal numbers of uniform hypergraphs," *Graphs and Combinatorics* **6**, 1–4.

[3] N. Alon and M. Krivelevich (1997). "The concentration of the chromatic number of random graphs," *Combinatorica* **17**, 303–313.

[4] N. Alon and J. Spencer (1992). *The Probabilistic Method.* Wiley, New York.

[5] M. Bläser, M. Hardt, and D. Steurer (2008). "Asymptotically Optimal Hitting Sets Against Polynomials," in ICALP '08: *Proceedings of the 35th International Colloquium on Automata, Languages and Programming, Part I,* pp. 345–356, Springer Verlag, New York.

[6] B. Bollobás and P. Erdős (1976). "Cliques in random graphs," *Math. Proc. Camb. Phil. Soc.* **80,** 419–427.

[7] K. Chandrasekaran, R. Karp, E. Moreno-Centeno, and S. Vempala (2011). "Algorithms for Implicit Hitting Set Problems," Accepted by ACM-SIAM Symposium on Discrete Algorithms (SODA11).

[8] G. Even, D. Rawitz, and S. Shahar (2005). "Hitting sets when the VC-dimension is small," *Information Processing Letters* **95**, 358–362.

[9] L. Li and Y. Jiang (2002). "Computing Minimal Hitting Sets with Genetic Algorithm," in *Proceedings of the 13th International Workshop on Principles of Diagnosis*, 77–80.

[10] D. Matula (1976). "The largest clique size in a random graph," Technical Report, Department of Computer Science, Southern Methodist University, Dallas, Texas.

[11] S. Thomassé and A. Yeo (2007). "Total domination of graphs and small transversals of hypergraphs," *Combinatorica* **27**, 473–487.

[12] K. Weber (1981). "Domination number for almost every graph," *Rostock. Math. Kolloq.* **16**, 31–43.

[13] B. Wieland and A. Godbole (2001). "On the domination number of a random graph," *Electr. J. Combinatorics* **8,** Paper R37, 13 pages.