Problem Set 2 Solutions

Point totals are in the margin; the maximum total number of points was 52.

1. Probabilistic method for dominating sets *5pts*

Pick a random subset S of vertices by including each vertex of V in S independently with probability p (a value to be chosen later). Let $T \subseteq U$ be the set of vertices in U that are neither in S nor adjacent to a vertex of S. Then clearly $S \cup T$ is a dominating set for U.

Note that a vertex $u \in U$ belongs to T iff neither it nor any of its (at least d) neighbors belongs to S, which happens with probability at most $(1-p)^{d+1}$. Hence the expected size of the dominating set is

$$\mathrm{E}(|S \cup T|) = \mathrm{E}(|S|) + \mathrm{E}(|T|) \leq np + n(1-p)^{d+1} \leq n(p + e^{-p(d+1)}).$$

Now we choose p so as to minimize this expression. Differentiating the term in the parentheses, we see that we want to take $e^{-p(d+1)} = \frac{1}{d+1}$, or $p = \frac{\ln(d+1)}{d+1}$. Plugging this into the above expression gives $\mathrm{E}(|S \cup T|) \leq n \cdot \frac{\log(d+1)+1}{d+1}$, so we deduce that there must exist a dominating set of at most this size. Indeed, since the size of a dominating set must be an integer, we know there exists a dominating set of size $\lfloor n \cdot \frac{\log(d+1)+1}{d+1} \rfloor$.

2. Locally 2-colorable graphs

(a) Let the r.v. X denote the number of edges in G. Since X is the sum of $\binom{n}{2}$ indicator r.v.'s X_e (one for each possible edge e), we have by linearity of expectation $\mathrm{E}X = \binom{n}{2}p = 8(n-1)$. Moreover, since the X_e are independent, $\mathrm{Var}(X) = \sum_e \mathrm{Var}(X_e) = \binom{n}{2}p(1-p) < \mathrm{E}X$. Now Chebyshev's inequality gives *2pts*

$$\Pr[X > 10(n-1)] \leq \Pr\big[|X - \mathrm{E}X| > \tfrac{1}{4}\mathrm{E}X\big] \leq \frac{\mathrm{Var}(X)}{(\mathrm{E}X/4)^2} = \frac{16}{\mathrm{E}X} = \frac{2}{n-1}.$$

This fraction is less than $\frac{1}{4}$ for all $n \geq 9$.

(b) Fix a coloring; suppose it has r red vertices and $n-r$ green vertices. Let the r.v. Y denote the number of violated edges. Then by linearity $\mathrm{E}Y = (\binom{r}{2} + \binom{n-r}{2})p$. This expression is minimized when $r = n/2$ (assuming n is even), giving $\mathrm{E}Y \geq 2\binom{n/2}{2}p = 4(n-2)$. Now applying the Chernoff bound to Y with $\delta = \frac{3}{4}$ gives *2pts*

$$\Pr[Y \leq n-1] \leq \Pr[Y \leq \tfrac{1}{4}\mathrm{E}Y] \leq \exp\big(-\tfrac{1}{2}(\tfrac{3}{4})^2 4(n-2)\big) = \exp\big(-\tfrac{9}{8}(n-2)\big).$$

Common Issues: Many people did not answer the problem for any arbitrary coloring but rather for a random coloring.

(c) Call a 2-coloring "good" if it has at most $n-2$ violated edges. Part (b) bounds the probability that any given coloring is good. Taking a union bound over all 2^n possible colorings, we get *2pts*

$$\Pr[\exists \text{ a good coloring for } G] \leq 2^n \exp\big(-\tfrac{9}{8}(n-2)\big) \leq \tfrac{1}{4},$$

where the last inequality holds for all $n \geq 9$.

Common Issues: Many people forgot to take the union bound over the 2^n possible colorings. Part (b) only applies to one fixed coloring, so this is necessary.

(d) Let the r.v. Z denote the number of k-cycles in G. Then we can write $Z = \sum_C Z_C$, where the sum is over all possible k-cycles C and Z_C is the indicator r.v. for the event that C is a cycle in G. Since C consists of k edges, we have $\Pr[Z_C = 1] = p^k$. Also, the number of possible C is $\binom{n}{k}\frac{k!}{2k}$; here $\binom{n}{k}$ is the number of ways of choosing the vertices in C, while $k!$ is the number of vertex orderings and $2k$ accounts for the choice of starting point and the sense of traversal of the cycle. By linearity of expectation, we have *2pts*

$$\mathrm{E}Z = \textstyle\sum_C \mathrm{E}Z_C = \binom{n}{k}\frac{k!}{2k}p^k \le (np)^k = 16^k.$$

The expected number of cycles of length at most $\frac{1}{8}\log n$ is thus bounded by

$$\textstyle\sum_{k=3}^{\frac{1}{8}\log n} 16^k \le \frac{16}{15} \times 16^{\frac{1}{8}\log n} = \frac{16}{15}\sqrt{n} < 2\sqrt{n}.$$

(e) By part (d) and Markov's inequality, we have that $\Pr[Z \ge 8\sqrt{n}] \le \frac{1}{4}$. So with probability at least $\frac{3}{4}$, G has at most $8\sqrt{n}$ cycles of length at most $\frac{1}{8}\log n$. Now if we delete an (arbitrary) edge from each of these cycles (a total of at most $8\sqrt{n}$ edges), we are left with a graph whose shortest cycle is longer than $\frac{1}{8}\log n$. But this implies that any induced subgraph on up to $\frac{1}{8}\log n$ vertices is cycle-free, and hence certainly 2-colorable. *2pts*

(f) Taking a union bound over the events in parts (a), (c) and (e), we deduce that, with positive probability (at least $\frac{1}{4}$), the random graph G has all of those properties simultaneously. For such a G, let G' be the graph obtained by deleting at most $8\sqrt{n}$ edges from G, as in part (e). Then G' has all the following properties: *2pts*

(i) G' has at most $10(n-1)$ edges (by part (a));

(ii) G' is not 2-colorable even if we delete any $n-2-8\sqrt{n}$ edges (by part (c); note that when deleting edges to remove cycles we may remove some violated edges);

(iii) the induced subgraph on any $\frac{1}{8}\log n$ vertices of G' is 2-colorable (by part (e)).

Finally, observe that (very crudely) $n-2-8\sqrt{n} > \frac{1}{2}(n-1)$, which by (i) is at least $\frac{1}{20}$ of the number of edges in G'.

Common Issues: Many people failed to take a union bound over parts (a), (c) and (e), instead claiming these events were independent. Many people also forgot to account for the $8\sqrt{n}$ edges that were deleted via part (c).

3. A threshold for isolated vertices

(a) For $i = 1, \ldots, n$, let X_i be the indicator random variable for the event that vertex i is isolated, and let $X = \sum_i X_i$ be the number of isolated vertices. To show that $p = \frac{\ln n}{n}$ is a threshold, we need to show the following two facts: *8pts*

1. if $p \gg \frac{\ln n}{n}$, then $\Pr[X > 0] \to 0$ as $n \to \infty$.
2. if $p \ll \frac{\ln n}{n}$, then $\Pr[X > 0] \to 1$ as $n \to \infty$.

[Here, as in lectures, for functions $f(n), g(n)$, $f \gg g$ means that $f(n)/g(n) \to \infty$ as $n \to \infty$, and $f \ll g$ means that $f(n)/g(n) \to 0$ as $n \to \infty$.]

Clearly $\mathrm{E}X = n(1-p)^{n-1} =: \mu$. Note that

$$\ln \mu = \ln n + (n-1)\ln(1-p) \sim \ln n - (n-1)p. \tag{1}$$

[Here $f \sim g$ means that $f(n)/g(n) \to 1$ as $n \to \infty$.] Thus, as $n \to \infty$, we have $\mu \to 0$ if $p \gg \frac{\ln n}{n}$, and $\mu \to \infty$ if $p \ll \frac{\ln n}{n}$.

For $p \gg \frac{\ln n}{n}$ we immediately have by Markov's inequality that

$$\Pr[X > 0] = \Pr[X \ge 1] \le \mu \to 0 \quad \text{as } n \to \infty,$$

which establishes Fact 1 above.

To establish Fact 2, for $p \ll \frac{\ln n}{n}$ we need to use the second moment method, as follows:

$$\Pr[X = 0] \leq \Pr[|X - \mu| \geq \mu] \leq \frac{\mathrm{Var}X}{\mu^2}.$$

Thus it is sufficient to show that $\frac{\mathrm{Var}X}{\mu^2} \to 0$. To do this, note that

$$\mathrm{E}X^2 = \sum_i \mathrm{E}X_i^2 + \sum_{i \neq j} \mathrm{E}[X_i X_j] = \mu + n(n-1)\mathrm{E}[X_i X_j].$$

Also, for any $i \neq j$ we have $\mathrm{E}[X_i X_j] = (1-p)^{2n-3}$, and thus $n(n-1)\mathrm{E}[X_i X_j] \leq \mu^2/(1-p)$. Hence

$$\frac{\mathrm{Var}X}{\mu^2} \leq \frac{\mu + \mu^2/(1-p) - \mu^2}{\mu^2} = \frac{1}{\mu} + \frac{p}{1-p} \to 0, \tag{2}$$

since $\mu \to \infty$ and $p \to 0$. This establishes Fact 2.

(b) If we set $p = \frac{c \ln n}{n}$ then the above calculations carry over essentially unchanged. In particular, from (1) above we see that $\mu \to 0$ if $c > 1$ and $\mu \to \infty$ if $c < 1$. Thus by Markov's inequality we still get that $\Pr[X > 0] \to 0$ in the former case. And in the latter case the calculation in (2) still holds as well, so we get $\Pr[X = 0] \to 0$ in this case. This shows that the threshold is sharp; the Note in the problem set tells you that the width of the scaling window is actually $\frac{1}{n}$. *2pts*

4. Planted cliques and cryptography

(a) Consider the set of pairs (C, G) where C is a set of k vertices, and G is a graph. Each pair uniquely corresponds to a specific derivation of a graph in $\mathcal{G}'_{n,1/2}$, where G is the graph chosen according to $\mathcal{G}_{n,1/2}$, and C is the set of k vertices chosen to become a clique. Clearly each pair is equally likely, arising with probability $\binom{n}{k}^{-1} 2^{-\binom{n}{2}}$, where the first term is the probability of selecting C, and the second is the probability of choosing G, since all graphs are equally likely in $\mathcal{G}_{n,1/2}$. For a given graph G', the number of pairs (C, G) which would give rise to it is $f(G')2^{\binom{k}{2}}$, since C must be one of the cliques of G', and given the choice of C, G can have any subset of the $\binom{k}{2}$ edges between vertices in C. Thus $\Pr'[G'] = f(G')2^{\binom{k}{2}}\binom{n}{k}^{-1}2^{-\binom{n}{2}} = \frac{f(G')\Pr[G']}{\mu}$. *3pts*

Common Issues: Many people simply claimed that the number of graphs G that could have given rise to a graph G' (in $\mathcal{G}'_{n,1/2}$) is $f(G')2^{\binom{k}{2}}$. This is not true: to see a silly example, let G' be a graph consisting of two disjoint k-cliques; then the number of graphs that could have given rise to G' is $2 \cdot 2^{\binom{k}{2}} - 1$. To see why the -1 is there, note that for either choice of the clique to add, there are $2^{\binom{k}{2}}$ possible graphs that could have yielded G'; however, we are double-counting the graph G', since that graph could have given rise to G' under either choice of which clique to add (because no new edges are needed). When counting, one must be careful to avoid double-counting!

(b) By Chebyshev's inequality, we have *3pts*

$$\Pr[G \text{ is } \alpha\text{-bad}] \leq \Pr[|f(G) - \mu| > (n^\alpha - 1)\mu] \leq \frac{\mathrm{E}(f^2)}{\mu^2(n^\alpha - 1)^2} = \frac{O(n^c \log n)}{(n^\alpha - 1)^2},$$

where in the last step we used the given upper bound on $\frac{\mathrm{E}(f^2)}{\mu^2}$. For large enough n, this last expression is bounded above by $n^{-2\alpha + c + \epsilon}$ for any $\epsilon > 0$.

(c) For $\mathcal{B}_j$ defined as in the hint, we have *3pts*

$$\Pr'[G \text{ is } \alpha\text{-bad}] \quad = \quad \sum_{j=2}^{\infty} \Pr'[G \in \mathcal{B}_j]$$

$$
\begin{aligned}
&\leq \sum_{j=2}^{\infty} n^{(j+1)\alpha/2} \Pr[G \in \mathcal{B}_j] \\
&\leq \sum_{j=2}^{\infty} n^{(j+1)\alpha/2} \Pr[f(G) > n^{j\alpha/2}] \\
&\leq \sum_{j=2}^{\infty} n^{(j+1)\alpha/2} n^{-j\alpha+c+\epsilon} \\
&= n^{\alpha/2+c+\epsilon} \sum_{j=2}^{\infty} n^{-j\alpha/2} \\
&= n^{\alpha/2+c+\epsilon} \times O(n^{-\alpha}) = O(n^{-\alpha/2+c+\epsilon}).
\end{aligned}
$$

The second line here follows from part (a), and the fourth line from part (b).

(d) We assume that $\Pr'[\mathcal{E} \text{ holds for } G] = \Omega(n^{-r})$. Now note that *3pts*

$$
\begin{aligned}
\Pr[\mathcal{E} \text{ holds for } G] &\geq \Pr[\mathcal{E} \text{ holds for } G \text{ and } G \text{ is not } \alpha\text{-bad}] \\
&\geq n^{-\alpha} \Pr'[\mathcal{E} \text{ holds for } G \text{ and } G \text{ is not } \alpha\text{-bad}] \\
&\geq n^{-\alpha} \left(\Pr'[\mathcal{E} \text{ holds for } G] - \Pr'[G \text{ is } \alpha\text{-bad}] \right) \\
&= n^{-\alpha} \left(\Omega(n^{-r}) - O(n^{-\alpha/2+c+\epsilon}) \right).
\end{aligned}
$$

(The second line here uses part (a), and the last line part (c).) Finally, we are free to choose α so that $\alpha/2 - c - \epsilon > r$, which implies that the negative term in the above final expression is $o(n^{-r})$. We conclude that $\Pr[\mathcal{E} \text{ holds for } G] = \Omega(n^{-r-\alpha}) = \Omega(n^{-r'})$, as desired.

Common Issues: Many people failed to take a union bound over the two events ($\mathcal{E}$ holds in G) and (G is not α-bad), instead claiming incorrectly that these events were independent.

(e) Take $\mathcal{E}$ to be the event "a clique of size $(1+\epsilon)\log_2 n$ can be efficiently found in G." The result of part (d) implies that, if this holds for a reasonable fraction of graphs with a planted k-clique, then it also holds for a reasonable fraction of all graphs, which contradicts the conjecture at the beginning of the question. Thus, according to the conjecture, the signature scheme is secure. *1pts*

5. More on unbalancing lights

(a) Since $z \geq 0$ the required inequality, for any fixed $\alpha > 0$, is equivalent to *3pts*

$$
z - \frac{z^3}{\alpha} \leq \frac{2\sqrt{\alpha}}{3\sqrt{3}}.
$$

Differentiating the left-hand side and setting to zero, we get a turning point at $z = \sqrt{\frac{\alpha}{3}}$, and since the second derivative is negative this must be a maximum. There are no other turning points in $z \geq 0$, so the global maximum value is $\frac{2\sqrt{\alpha}}{3\sqrt{3}}$, as desired.

For any random variable Z, applying the above inequality to $|Z|$ and taking expectations we get

$$
\mathrm{E}(|Z|) \geq \frac{3\sqrt{3}}{2\sqrt{\alpha}} \left(\mathrm{E}(Z^2) - \frac{\mathrm{E}(Z^4)}{\alpha} \right).
$$

This holds for any $\alpha > 0$, so we may choose α to maximize the right-hand side. (We may also assume that $\mathrm{E}(Z^2) \neq 0$; otherwise $Z = 0$ with probability 1.) Differentiating, we see that the optimal α is $\alpha = \frac{3\mathrm{E}(Z^4)}{\mathrm{E}(Z^2)}$. Plugging this in and tidying up gives the claimed bound.

(b) Setting $Z = S_n$ in part (a), we need to evaluate $\mathrm{E}(S_n^2)$ and $\mathrm{E}(S_n^4)$. First we have *3pts*

$$\mathrm{E}(S_n^2) = \sum_i \mathrm{E}(X_i^2) + \sum_{i \neq j} \mathrm{E}(X_i X_j) = n.$$

To see this, note that $\mathrm{E}(X_i^2) = 1$ since X_i is ± 1-valued; and $\mathrm{E}(X_i X_j) = \mathrm{E}(X_i)\mathrm{E}(X_j) = 0$ since the X_i are 4-wise independent (and hence certainly pairwise independent). By similar considerations we have

$$\mathrm{E}(S_n^4) = \sum_i \mathrm{E}(X_i^4) + \frac{1}{2}\binom{4}{2} \sum_{i \neq j} \mathrm{E}(X_i^2 X_j^2) = n + 3n(n-1) = 3n^2 - 2n.$$

Note that all other terms have expectation zero because of 4-wise independence and the fact that $\mathrm{E}(X_i) = 0$. (The $\frac{1}{2}\binom{4}{2} = 3$ arises as the number of ways of partitioning the two copies of X_i and X_j among the four factors.) Plugging these two values into the bound from part (a) gives

$$\mathrm{E}(|S_n|) \geq \frac{n^{3/2}}{(3n^2 - 2n)^{1/2}} \geq \sqrt{\frac{n}{3}}.$$

(c) First, suppose we use the randomized algorithm of Lecture 5, except that now the n coin tosses used to determine the column switch settings are only 4-wise independent. The analysis in Lecture 5 relied only on the fact that $\mathrm{E}(|Z_i|)$ is asymptotically at least $\sqrt{\frac{2n}{\pi}}$, where $Z_i = \sum_{j=1}^n X_{ij}$ is the sum of light values for row i. Since the X_{ij} are also 4-wise independent, we may substitute the result of part (b) to deduce that $\mathrm{E}(|Z_i|)$ is asymptotically at least $\sqrt{\frac{n}{3}}$, and hence the asymptotic expected excess is $\frac{1}{\sqrt{3}} n^{3/2}$. (Note that $\frac{1}{\sqrt{3}} \approx 0.58$ while $\sqrt{\frac{2}{\pi}} \approx 0.80$.) *3pts*

Now recall from Lecture 9 that we can construct a family of r d-wise independent fair coin flips using random polynomials of degree $d-1$ over a field of size $q = O(r)$. (Say $q = 2^m$, where $m = \lceil \log_2 r \rceil$.) In our present application, $r = n$ and $d = 4$. The number of points in this sample space (i.e., the number of such polynomials) is only $q^d = O(n^4)$, so instead of running the randomized algorithm we may exhaustively try all sample points in polynomial time. We are guaranteed to find at least one sample point for which the excess is at least the above expected value.

(d) We use the construction in the hint. First, let's check that $\mathrm{E}(|S_n|) = 1$. Note that if $v = (0, \ldots, 0, 0)$ then $Y_i = 0$ for all i, so $S_n = n$. Similarly, if $v = (0, \ldots, 0, 1)$ then $Y_i = 1$ for all i, so $S_n = -n$. Now consider any other value of v, and let v_j be its leftmost non-zero entry (so $j \leq k$). For each $i \in \{0, 1, \ldots, n-1\}$, let i' denote the integer whose binary expansion differs from that of i only in position j. Then it is clear that $Y_i = 1 - Y_{i'}$, and hence $X_i = -X_{i'}$. Thus the contributions to S_n cancel out in pairs, and $S_n = 0$. Putting all this together we get *3pts*

$$\mathrm{E}(|S_n|) = \tfrac{1}{2^{k+1}}(n + n) = 1.$$

To see that the Y_i (and hence the X_i) are 3-wise independent, note first that, by the same argument as above, $\Pr[Y_i = 0] = \Pr[Y_i = 1] = \frac{1}{2}$ for all i. Now fix any three distinct integers $i_1, i_2, i_3 \in \{0, 1, \ldots n-1\}$; we need to show that

$$\Pr[Y_{i_1} = a_1 \wedge Y_{i_2} = a_2 \wedge Y_{i_3} = a_3] = \tfrac{1}{8}$$

for any triple of values $a_j \in \{0, 1\}$. The sample points (vectors v) corresponding to this triple are solutions to the equations $Bv^{\mathrm{T}} = a^{\mathrm{T}}$, where $a = (a_1, a_2, a_3)$, and B is the $3 \times (k+1)$ matrix whose rows are b_{i_1}, b_{i_2} and b_{i_3}. We claim that the rows of B are linearly independent over $GF[2]$; for if not then some combination of them would have to sum to 0, and since all have 1 in the last position this means that some pair must sum to 0, which is not possible since all three rows are distinct. Linear independence means that the system of equations has full rank, so the dimension of the set of solutions is $k + 1 - 3 = k - 2$, i.e., the number of solutions is 2^{k-2}. Since this is independent of the triple a, all eight triples must have the same probability, $\frac{1}{8}$.