CS 225: Pseudorandomness	**Prof. Salil Vadhan**
Problem Set 2	
Assigned: Fri. Feb. 11, 2011	Due: Fri. Feb. 25, 2011 (1 PM sharp)

- You must *type* your solutions. LaTeX, Microsoft Word, and plain ascii are all acceptable. Submit your solutions *via email* to `cs225-hw@seas.harvard.edu`. If you use LaTeX, please submit both the compiled file (`.pdf`) and the source (`.tex`). Please name your files `PS2-yourlastname.*`.
- Strive for clarity and conciseness in your solutions, emphasizing the main ideas over low-level details. Do not despair if you cannot solve all the problems! Difficult problems are included to stimulate your thinking and for your enjoyment, not to overwork you. *'ed problems are extra credit.

Problem 2.9.(Spectral Graph Theory) Let M be the random-walk matrix for a d-regular *undirected* graph $G = (V, E)$ on n vertices. We allow G to have self-loops and multiple edges. Recall that the uniform distribution (or all-ones vector) is an eigenvector of M of eigenvalue $\lambda_1 = 1$. Prove the following statements. (Hint: for intuition, it may help to think about what the statements mean for the behavior of the random walk on G.)

1. All eigenvalues of M have absolute value at most 1.
2. G is disconnected $\iff$ 1 is an eigenvalue of multiplicity at least 2.
3. Suppose G is connected. Then G is bipartite $\iff$ -1 is an eigenvalue of M.
4. G connected $\Rightarrow$ all eigenvalues of M other than λ_1 are at most $1 - 1/\text{poly}(n, d)$. To do this, it may help to first show that the second largest eigenvalue of M (not necessarily in absolute value) equals
$$\max_x \langle xM, x\rangle = 1 - \frac{1}{d} \cdot \min_x \sum_{(i,j) \in E} (x_i - x_j)^2,$$
where the maximum/minimum is taken over all vectors x of length 1 such that $\sum_i x_i = 0$, and $\langle x, y\rangle = \sum_i x_i y_i$ is the standard inner product. For intuition, consider restricting the above maximum/minimum to $x \in \{+\alpha, -\beta\}^n$ for $\alpha, \beta > 0$.
5. G connected and nonbipartite $\Rightarrow$ all eigenvalues of M (other than 1) have absolute value at most $1 - 1/\text{poly}(n, d)$ and thus $\lambda(G) \le 1 - 1/poly(n, d)$.

6* Establish the (tight) bound $1 - \Omega(1/d \cdot D \cdot n)$ in Part 4, where D is the diameter of the graph. Conclude that $\gamma(G) = \Omega(1/d^2 n^2)$ if G is connected and nonbipartite.

Problem 3.1.(Derandomizing RP versus BPP) Show that **prRP** = **prP** implies that **prBPP** = **prP**, and thus also that **BPP** = **P**. (Hint: Look at the proof that **NP** = **P** ⇒ **BPP** = **P**.)

Problem 3.2.(Designs) Designs (also known as packings) are collections of sets that are nearly disjoint. In Chapter 7, we will see how they are useful in the construction of pseudorandom generators. Formally, a collection of sets $S_1, S_2, \ldots, S_m \subset [d]$ is called an (ℓ, a)-*design* (for integers $a \leq \ell \leq d$) if

- For all i, $|S_i| = \ell$.
- For all $i \neq j$, $|S_i \cap S_j| < a$.

For given ℓ, we'd like m to be large, a to be small, and d to be small. That is, we'd like to pack many sets into a small universe with small intersections.

1. Prove that if $m \leq \binom{d}{a} / \binom{\ell}{a}^2$, then there exists an (ℓ, a)-design $S_1, \ldots, S_m \subset [d]$.
   Hint: Use the Probabilistic Method. Specifically, show that if the sets are chosen randomly, then for every $S_1, \ldots, S_{i-1}$,
   $$\mathop{\mathrm{E}}_{S_i}\left[\#\{j < i : |S_i \cap S_j| \geq a\}\right] < 1.$$
2. Conclude that for every constant $\gamma > 0$ and every $\ell, m \in \mathbb{N}$, there exists an (ℓ, a)-design $S_1, \cdots, S_m \subseteq [d]$ with $d = O\left(\frac{\ell^2}{a}\right)$ and $a = \gamma \cdot \log m$. In particular, setting $m = 2^\ell$, we fit exponentially many sets of size ℓ in a universe of size $d = O(\ell)$ while keeping the intersections an arbitrarily small fraction of the set size.
3. Using the Method of Conditional Expectations, show how to construct designs as in Parts 1 and 2 *deterministically* in time poly(m, d).

Problem 3.6.(Frequency Moments of Data Streams) Given one pass through a huge "stream" of data items $(a_1, a_2, \ldots, a_k)$, where each $a_i \in \{0, 1\}^n$, we want to compute statistics on the distribution of items occurring in the stream while using small space (not enough to store all the items or maintain a histogram). In this problem, you will see how to compute the *2nd frequency moment* $f_2 = \sum_a m_a^2$, where $m_a = \#\{i : a_i = a\}$.

The algorithm works as follows: Before receiving any items, it chooses t random *4-wise independent* hash functions $H_1, \ldots, H_t : \{0, 1\}^n \to \{+1, -1\}$, and sets counters $X_1 = X_2 = \cdots = X_t = 0$. Upon receiving the i'th item a_i, it adds $H_j(a_i)$ to counter X_j. At the end of the stream, it outputs $Y = (X_1^2 + \cdots + X_t^2)/t$.

Notice that the algorithm only needs space $O(t \cdot n)$ to store the hash functions H_j and space $O(t \cdot \log k)$ to maintain the counters X_j (compared to space $k \cdot n$ to store the entire stream, and space $2^n \cdot \log k$ to maintain a histogram).

1. Show that for every data stream $(a_1, \ldots, a_k)$ and each j, we have $\mathrm{E}[X_j^2] = f_2$, where the expectation is over the choice of the hash function H_j.

2. Show that $\mathrm{V}ar[X_j^2] \leq 2f_2^2$.

3. Conclude that for a sufficiently large constant t (independent of n and k), the output Y is within 1% of f_2 with probability at least .99.