Spring 2023
Homework 1
S&DS 684: Statistical Inference on Graphs
Due: Feb 22, 2023
`Prof. Yihong Wu`

Rules:

- It is mandatory to type your solutions in LaTeX; you may use the source file for this PDF posted on the course website. (If you need help with this, let me know.)
- Email your solution in pdf by midnight of the due date to yihong.wu@yale.edu with subject line `Homework XX: your name`.
- Justify your work rigorously. As long as you are able to prove the result or a stronger version, there is no need to follow the hints.

1. (Binomial vs Hypergeometric: stochastic dominance). Binomial and Hypergeometric distributions arise from sampling a finite population with and without replacements, respectively. The next two problems deal with their comparison.

   Consider an urn consisting of N balls in total among which k are red, and $N-k$ are blue. Let X denote the number of red balls obtained by sampling n balls from the urn *without* replacements. Let Y denote the number of red balls obtained by sampling n balls from the urn *with* replacements. Then $X \sim \text{Hypergeometric}(N,k,n)$ and $Y \sim \text{Binom}(n, \frac{k}{N})$. Here N, k, n are integers such that $0 \le k \le N$ and $0 \le n \le N$.

   (a) For any real-valued random variable X and Y, we say that X is *stochastically dominated* by Y, denoted by $X \overset{\text{s.t.}}{\le} Y$, if $F_Y(t) \le F_X(t)$ for every t, where $F_X(t) \triangleq \mathbb{P}[X \le t]$ is the CDF of X. Note that this is a statement about comparing distributions, rather than random variables. Nevertheless, show that $X \overset{\text{s.t.}}{\le} Y$ if and only if if there exists a coupling (joint distribution) between X and Y, that is, a probability space on which X and Y are defined, such that $X \le Y$ almost surely. (Hint: how to generate random variables from uniform distribution?)

   (b) Show that $\text{Bern}(p) \overset{\text{s.t.}}{\le} \text{Bern}(q)$ if $p \le q$. Describe the coupling explicitly.

   (c) Show that both binomial and hypergeometric can be written as a sum of Bernoulli random variables:
   $$X = X_1 + \ldots + X_n, \quad Y = Y_1 + \ldots + Y_n \tag{1}$$
   where X_i's are Y_i's are distributed as $\text{Bern}(\frac{k}{N})$, and Y_i's are independent.

   (d) Show that
   $$\text{Hypergeometric}(N,k,n) \overset{\text{s.t.}}{\le} \text{Binom}\left(n, \frac{k}{N-n}\right).$$
   (Hint: use part (b) and consider the conditional law of X_t given $X_1, \ldots, X_{t-1}$.)

2. (Binomial vs Hypergeometric: convex ordering).

   (a) For any real-valued random variable X and Y, we say that X is *dominated* by Y *in the convex ordering*, denoted by $X \overset{\text{cvx}}{\leq} Y$, if

   $$\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)] \tag{2}$$

   for every convex function f. Again, this is a statement about comparing distributions, rather than random variables. Nevertheless, show that $X \overset{\text{cvx}}{\leq} Y$ if[1] there exists a coupling between X and Y, such that

   $$\mathbb{E}[Y|X] = X, \quad \text{a.s.} \tag{3}$$

   (b) Next we construct such a coupling for binomial and hypergeometric distributions. If you can construct another coupling that works, you can skip these two parts.

   Show that one can simulate sampling with replacements from sampling without replacements as follows: In the context of (1), show that one can generate $(Y_1, \ldots, Y_n)$ from $(X_1, \ldots, X_n)$ by resampling by

   $$Y_i = \begin{cases} X_i, & \text{with probability } 1 - \frac{i-1}{N}, \\ X_m, & \text{with probability } \frac{1}{N}, \quad m = 1, \ldots, i-1. \end{cases} \tag{4}$$

   In other words, show that $(Y_1, \ldots, Y_n)$ defined in (4) are indeed iid $\text{Bern}(\frac{k}{N})$.

   (c) Use (b) to construct an explicit coupling between $X \sim \text{Hypergeometric}(N, k, n)$ and $Y \sim \text{Binom}(n, \frac{k}{N})$, such that (3) holds, thereby proving *Hoeffding's inequality*:

   $$\text{Hypergeometric}(N, k, n) \overset{\text{cvx}}{\leq} \text{Binom}\left(n, \frac{k}{N}\right)$$

   (Hint: To make the coupling symmetric in $Y_1, \ldots, Y_n$, randomize their ordering.)

   (d) Invoke Hoeffding's inequality to compare the variance: $\text{Var}(X) \leq \text{Var}(Y)$.

   (e) Invoke Hoeffding's inequality to show that hypergeometric distribution satisfies the same binomial tail bound:

   $$\mathbb{P}\left[\left|X - \frac{nk}{N}\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n}\right), \quad t > 0.$$

3. (Maximum clique in the planted model) Fix $\epsilon > 0$. Let $G \sim G(n, \frac{1}{2}, k)$, where $k \geq (2+\epsilon)\log_2 n$. Show that with high probability, the unique k-clique in G is the planted one. (Hint: First moment calculation, this time in the planted model not the null model.)

4. (Slightly smarter exhaustive search for planted clique). To find the k-clique K planted in $G \sim G(n, \frac{1}{2}, k)$, exhaustive search over all k-subsets of vertices takes $\binom{n}{k} \sim n^k$ time. Here is an $n^{\Theta(\log n)}$-time algorithm that works for all $k \geq C\log_2 n$ where C is a sufficiently large constant .

   (a) By exhaustive search, we can find a clique T of size $C\log n$. Show that $|T \cap K| \geq (C - 2 - \epsilon)\log_2 n$ with high probability.

   (b) Let S denote the set of all vertices that has at least $3|T|/4$ neighbors in T. Then show that $K \subset S$ with high probability.

   (c) Now S may also contain non-clique vertices, which requires some cleanup. So we report the k highest-degree vertices in the induced subgraph $G[S]$. (Hint: Show $|S \setminus K|$ is relatively small. Be careful with the union bound as T is random.)

[1] Is the "only if" part also correct?