



1. Privacy and Signed Cards (10 points each, 40 points total)

In class we presented a method to collect information about whether students have done some behavior while at the same time maintaining some of their privacy. In this question we'll look at some properties of this protocol in more detail. Recall from class that we looked at a method (which we'll call scheme 1) where students had three cards with a 2/1 split of + and - cards. We will also consider a similar scheme, called scheme 2, in which there is a 11/10 split of cards (21 cards total).

1. What is $\Pr[\text{Yes} \mid +]$ as a function of $\Pr[\text{Yes}]$? Do the calculation for both schemes and submit a plot of these functions (try using Wolfram Alpha for producing nice plots!).
2. Plot the functions $\Pr[\text{Yes} \mid +] / \Pr[\text{Yes}]$ and $\Pr[\text{Yes} \mid +] - \Pr[\text{Yes}]$ for both schemes. State some observations about the maximum and minimum values they take on, and for what values of $\Pr[\text{Yes}]$ they take on those values. How do you interpret these results?
3. Suppose that in a class of 100 people we use scheme 1 and see 75 + cards. What is the most likely probability of Yes? Since you don't know anything about the population, you should assume a uniform prior distribution over possible probabilities of Yes (that is to say, $\Pr[\Pr[\text{Yes}] = p] = 1$ for p between 0 and 1). Hint: $\Pr[75+ \text{ and } 25-]$ is constant. You probably don't need to know its actual value. Also, logarithms are your friend.
4. Do you think that either scheme sufficiently preserves privacy? What about our ability to make good inferences about how frequent behaviors occur? Look at both the case of rare events ($\Pr[\text{Yes}]$ is very close to 0 or 1) and common behaviors ($\Pr[\text{Yes}]$ is near $\frac{1}{2}$).

2. ICM and Submodular Functions (10, 15, 15 points for parts 1-3, respectively. 15 BONUS points each for parts 4-6)

If X is a finite set, we say that a function $F: 2^X \rightarrow \mathbb{R}$ (a function from subsets of X to the reals) is a *set function* over the *ground set* X . Our goal is to find the subset A of X that maximizes F ; in general this is NP-hard. Nemhauser and Wolsey showed that *non-negative, (non-decreasing) monotone submodular set functions* can be approximately maximized using an efficient, greedy algorithm. If $F(A) \geq 0$ for every $A \subseteq X$, we say that F is non-negative. If $F(A) \leq F(B)$ whenever $A \subseteq B \subseteq X$, we say that F is non-decreasing and monotone (from here on we will assume that monotone means non-decreasing and monotone). Finally, if $F(A \cup B) + F(A \cap B) \leq F(A) + F(B)$ for every pair of subsets A and B of X , we say that F is submodular. The amazing thing that Nemhauser and Wolsey proved is that we can efficiently (in time polynomial in $|X|$, assuming that F can be quickly calculated) find a set S such that $F(S) \geq (1 - 1/e) F(B)$ for every $B \subseteq X$. This means that we can quickly find a solution that is at least 63% as good as the optimal solution, regardless of $|X|$! We will develop some theory of set functions and apply it to the problem of optimally starting a rumor under the Independent Cascade Model.

1. Let $X = \{1, 2, \dots, 100\}$. Which of the following functions are non-negative, monotone submodular set functions (NN-M-SM-SF): $F(A) = |A|$ (set cardinality), $F(A) = \max(A)$, $F(A) = \text{avg}(A)$, $F(A) = |A|^2$.

You may assume that $\max(\emptyset) = \text{avg}(\emptyset) = 0$. Make sure to provide proof that F is a NN-M-SM-SF, or demonstrate which properties it fails to satisfy.

2. Prove that a non-negative linear combination of NN-M-SM-SFs is also non-negative, monotone, and submodular. That is to say, if F and G are NN-M-SM-SFs, then for any $\alpha, \beta > 0$, the function $H(A) = \alpha F(A) + \beta G(A)$ is also a NN-M-SM-SF.
3. Prove that the submodular condition $F(A \cup B) + F(A \cap B) \leq F(A) + F(B)$ implies the *economies of scale* principle: If $A \subseteq B \subseteq X$, and $e \in X \setminus B$ (i.e., $e \notin B$), then $F(B \cup \{e\}) - F(B) \leq F(A \cup \{e\}) - F(A)$. Informally, the marginal cost of 'producing' e given that you've already produced $B \supseteq A$ is no more than the marginal cost of producing e if you've only already produced A .

So how does all of this relate to the Independent Cascade Model? It turns out that the function $\sigma(A)$ that maps sets A of initially infected nodes to the expected size of the infection (randomness taken over the 'u infects v' coin flips) is a non-negative, monotone submodular set function. You will prove this fact in three steps:

1. In the description of the ICM, we said that when a node becomes infected it tries to infect all of its neighbors. If you think about how you'd implement the ICM in code (we'll call this program 1), you could generate (pseudo-) random coin flips only when a node becomes infected. Alternatively, you could imagine flipping the coins for *every* edge in the graph first, thereby inducing a subset of edges along which the infection can travel. Any node which is reachable from the set of initially infected nodes via the remaining edges would get infected. We call this approach program 2. Argue why the output (the set of nodes which eventually get infected) of these two programs has the same distribution. Having the same distribution means that on the same set of initially infected nodes A , for every final set of infected nodes B , $\Pr[\text{Out}(\text{Program 1}) = B] = \Pr[\text{Out}(\text{Program 2}) = B]$.
2. Using the idea behind program 2, write down a nice expression for $\sigma(A)$. Hint: Consider conditioning on the outcome of all of the coin flips (the set E of edges remaining) and look at the set of nodes reachable from A using edges in E (you may want to call this set $R(A, E)$ for simplicity).
3. Prove that $\sigma(A)$ is a non-negative, monotone submodular function. Hint: In the last part, you should have come up with a way of writing $\sigma(A)$ as a non-negative sum of functions. What properties do these functions have?

3. Erdos-Reyni Graphs and Gephi (10 points each, 20 pts total)

We can use Gephi to explore some of the properties of Erdos-Reyni random graphs. In class the theorems we presented were of the form 'Property P is true asymptotically almost surely,' but in practical cases it will suffice to look at graphs $G(n, p)$ with $n \approx 1000$.

1. Generate $G(n, p)$ graphs with $n = 1000$ nodes and varying values of p . The values of p you examine are $0.5/n$, $1/n$, $2/n$, $0.5 \ln(n)/n$, $\ln(n)/n$, and $2 \ln(n)/n$. What is the size of the largest component in your random graph instance at each of these p values? You can repeat the experiment for smaller and larger values of n . Compare the connectivity and the size of the big

connected component in your graphs with the results we presented in class. Do they seem to be true for graphs with only 100 nodes?

2. For the $G(n,p)$ instance with $p = 2 \ln(n) / n$, what is the diameter of your graph? What is the degree distribution? Write down the fraction of nodes that have degree d , and compare it to the expected fraction of nodes that would have that degree.