# MCS 590 – Foundations of Data Science
Spring 2015
Problem Set 3

Lev Reyzin

**Due**: 4/24/15 at the beginning of class

**Instructions:** Atop your problem set, please write your name and list your collaborators.

## Problems

**1.** Show that for a 2-universal hash family,

$$\Pr(h(x) = z) = \frac{1}{M+1}$$

for all $x \in \{1, 2, \ldots, m\}$ and $z \in \{0, 1, 2, ..., M\}$.

**2.** Consider random strings of length n composed of the integers 0 through 9. Represent a string by its set of length k-substrings. What is the resemblance of the sets of length $k$-substrings from two random strings of length $n$ for various values of $k$ as $n$ goes to infinity?

**3.** In 1-dimension, is a center that minimizes the sum of distances to the data points unique? How about in the case of sum of squared distances? (Assume in both cases that the center need not be a data point.)

**4.** For the $k$-median problem, give an upper bound on the ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points to be centers.

**5.** If $A$ is a symmetric matrix with distinct singular values, show that the left and right singular vectors are the same and that $A = VDV^T$.