Computer Science 2300: Lab 5

Due: April 21, 2010

For Lab 5, you will implement a Bloom Filter and study its performance on the task of storing IP addresses. You will compare its runtime with the provided **chain_hash** program which implements hashing with chaining.

1 Implementing a Bloom Filter

In the provided files, **ip_address_10k.txt** contains 10000 distinct IP addresses. Write a program that implements a Bloom Filter with k hash functions. Use the implemented Bloom Filter to keep N IP addresses, read from **ip_address_10k.txt**. Each hash function should have the following form: hash value $= \sum_{i=1}^{4} a_i * \text{IP}_i$, where IP_i is the i_{th} component of an IP address, and a_i is an integer selected uniformly ar random between 0 and $M - 1$. Then, estimate the false positive rate of the Bloom Filter by checking the 1000 IP addresses in **ip_addr_test_1k.txt** to see which ones lead to positive results from the Bloom Filter (note that **ip_address_10k.txt** and **ip_addr_test_1k.txt** have no IP addresses in common). You must also measure the time required to insert N IP addresses into the Bloom Filter.

2 Analysis and Comparison

Run your program 10 times for each setting of M, N and k, where M = 60013, N = 1000, 2000, 4000, 8000, 10000, and k = 1, 3, 4, 6. Record the time used to insert all elements in the hash table and the average false positive rate of each setting. Then, run **chain_hash** program using the following commands

./chain_hash ip_address_10k.txt ip_addr_test_1k.txt 60013 N 1000 0,

for all $N = 1000, 2000, 4000, 8000, 10000$.

Finally, create a table of all your results and use your favorite plotting software to create two plots: one of the times taken to insert elements into hash table using both algorithms, and one of the false positive rate for each setting. Show your results and your program to your TA to receive credit.