AN4091019_hw2_p1

1. Program design
1) M bits of bloom filter & U updates are predefined
2) Program will calculate the suppose minimum number of hash function from h = (loge2)m/u
3) Program will test the bloom filter with different number of hash function ranging from h – 4 to h + 4
4) In each test program will calculate the probability of false positive both from random generating data & $P(h) = (1 - (1 - 1/m)^{uh})^{h}$

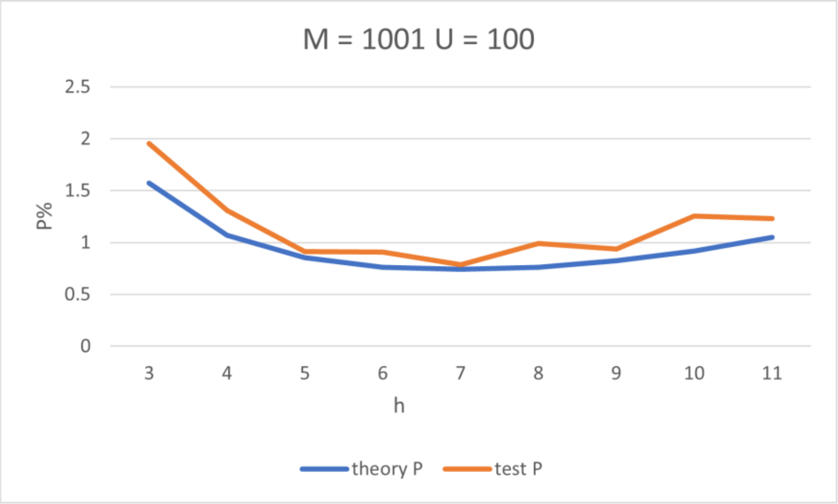
```
// number of time to test false positive
#define TEST 100000
// number of bits in bloom filter
#define M 10001
// number of initial updates
#define U 1000
// number of hash functions
int H;

int main()
{
    // calculate minimum h in from theory
    double tH = log(2) * (double)M / (double)U;
    // test bloom filter with tH - 4 ~ tH + 4 hash functions
    for (H = start; H <= finish; H++)
    {
        // generate random integer insert into bloom filter
        for (int i = 0; i < U; i++)
            generate random data & addData();

        // generate random integer to test collision
        int falsePositive = 0;
        for (int i = 0; i < TEST; i++)
            if (collision)
                falsePositive++;

        // calculate P from theory & test
        p = (double)falsePositive / TEST * 100;
        tP = pow(t, U) * pow(1 - pow(t, U * H), H) * 100;
        print theory p & test p;
    }
    return 0;
}

// generate hash function using hash1 & hash2
unsigned int hash(int i, unsigned int key, int size)
{
    return (hash1(key, size) + hash2(key, size) * i) % size;
}
```

2. Sample output

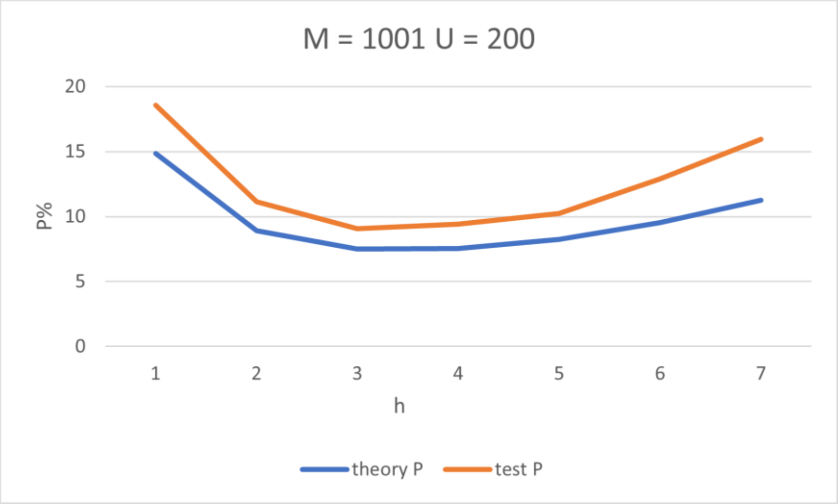
```
m = 10001 u = 2000 h = 1 ~ 7 test 100000 times
theory min h = 3.466
h = 1
theory false positve rate : 14.841%
test false positive rate : 18.080%
h = 2
theory false positve rate : 8.898%
test false positive rate : 11.137%
h = 3
theory false positve rate : 7.519%
test false positive rate : 9.078%
h = 4
theory false positve rate : 7.528%
test false positive rate : 9.634%
h = 5
theory false positve rate : 8.262%
test false positive rate : 10.252%
h = 6
theory false positve rate : 9.533%
test false positive rate : 11.047%
h = 7
theory false positve rate : 11.279%
test false positive rate : 14.232%
```

3. Analysis

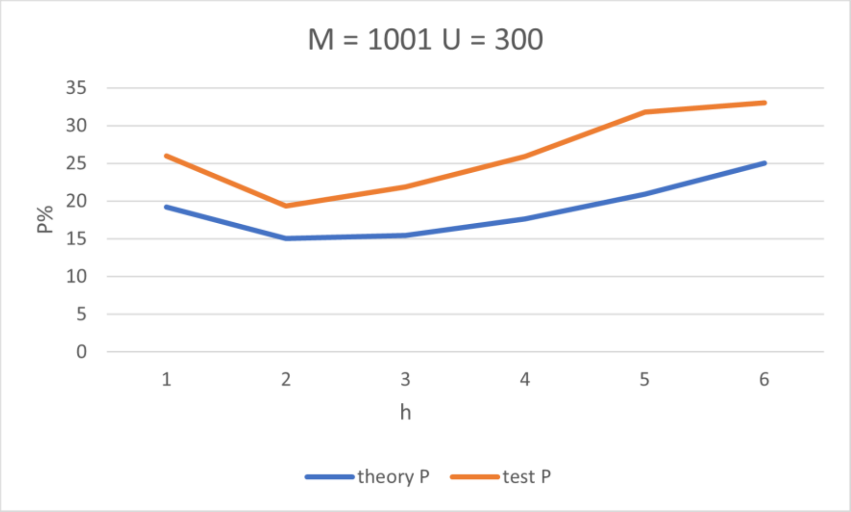
Theory min h = 6.938

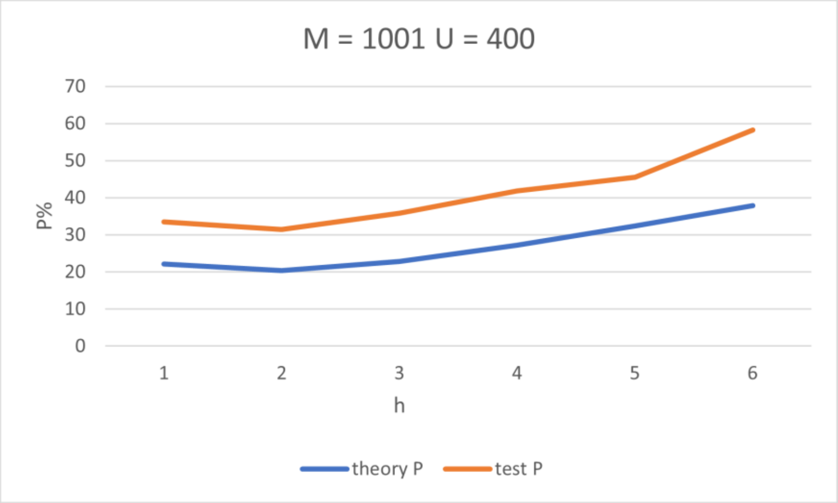


Theory min h = 3.469

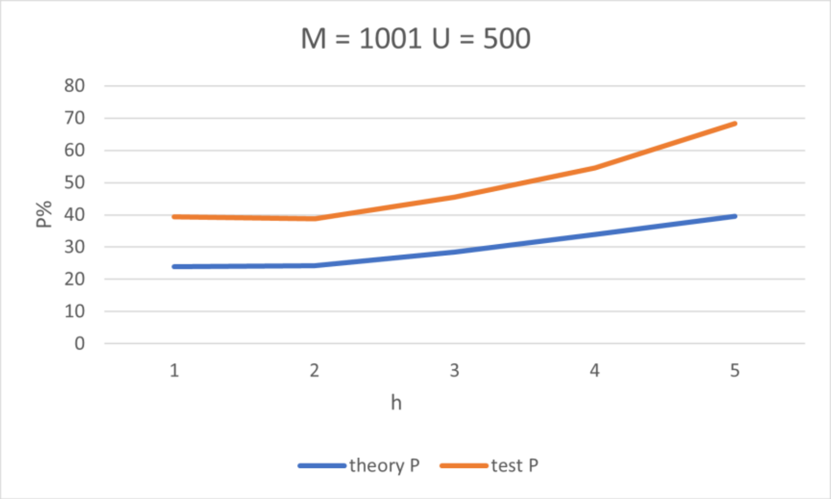


Theory min h = 2.313





Theory min h = 1.735



Theory min h = 1.388

Observe:

1) Test P seems to form local minimum at h = theory min h
2) Test P curve looks like the theory P curve with a slight shift up the y-axis

Conclusion:

1) P(u) exist at h = (loge2)m/u
2) the distance between two curve probably cause by ununiform distribution of the hash function