Complexity of Hashing

For simplicity, consider a sequence of $n$ INSERT operations followed by $n$ MEMBER operations, one for each of the items inserted into a hash table of size $m \geq n$.

**Assumptions:**

- The elements to be inserted are chosen at random from the set of all possible elements.
- The hash function is an "ideal" one where for any $i$ in the range of $0 \leq i < m$, if $x$ is randomly chosen, then the probability that $h(x) = i$ is $1/m$.
  (Equivalently, if $x$ and $y$ are randomly chosen, then the probability that $h(x) = h(y)$ is $1/m$.)
(complexity of $n$ INSERTS followed by $n$ MEMBERS continued)

**Space: $O(m+n)$**

The array uses $O(m)$ space and the total space used by linked list vertices is $O(n)$.

**Worst case time: $O(n^2)$**

All elements could hash to the same position.

For each value of $1 \leq i \leq n$, one of the member instructions will examine $i$ linked list buckets, and hence the total number of comparisons is given by:

$$1 + 2 + 3 + \cdots + n = \frac{1}{2} n^2 + \frac{1}{2} n = \Omega\left(n^2\right)$$
Expected time:

**INSERT:**
Since a new item can be inserted at the front of a bucket, each INSERT is $O(1)$ in the worst case.

**MEMBER:**
Since each INSERT is equally likely to hash to any of the $m$ buckets and $n \leq m$ elements are inserted, then on inserting the $i^{th}$ element, the expected length of the list on which it is placed is $(i−1)/m < 1$.

Hence the expected bucket size is $O(1)$, and the expected time for each MEMBER operation is $O(1)$. 
Facts About Hash Tables

If \( n \) items are inserted into an initially empty hash table of size \( n \) (using an ideal hash function):

- The expected number of empty buckets is \( n/e \), where the constant \( e = 2.7182... \) is the natural logarithm base.

- The expected size of the largest bucket is \( O(\log(n)/\log\log(n)) \).

**Note:** If \( n\log(n) \) items are inserted, the expected size of the largest bucket is \( O(\log(n)) \). That is, overcrowding the hash table results in a "flattening" of the distribution of bucket sizes so that the expected size of the largest bucket is on the same order as the average size of a bucket.