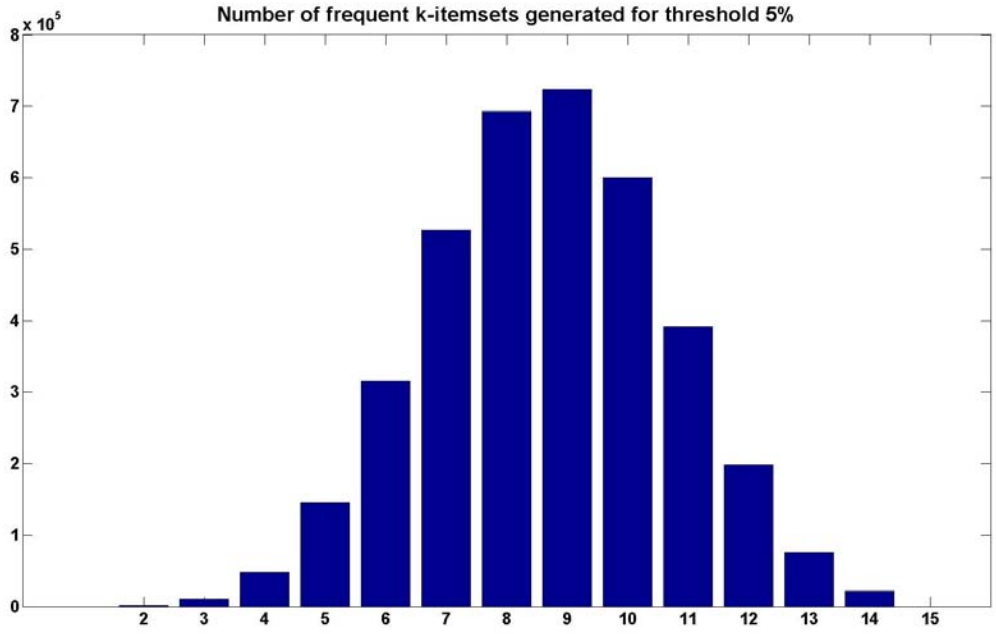
Project 2
CS5523
Robert F.K. Martin
1505151
9 November 2011

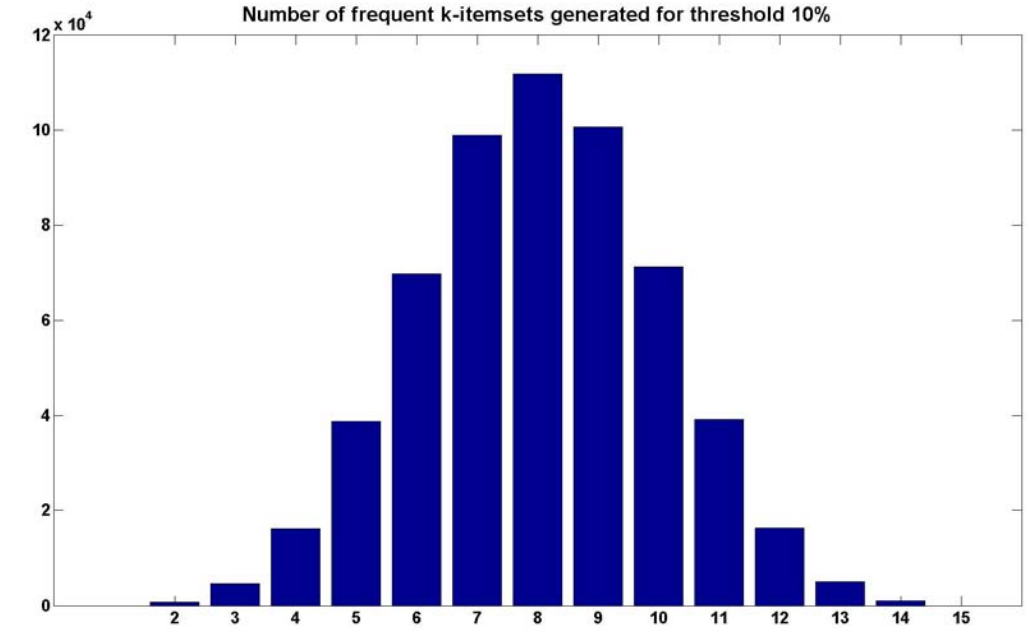
Problem 1a:

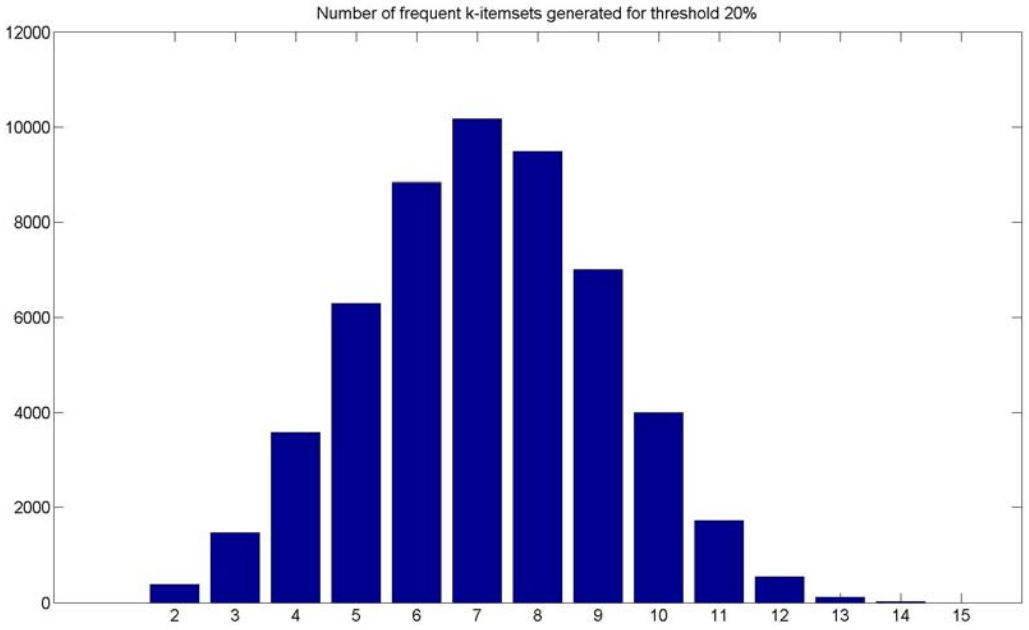
Threshold	reading	filtering	sorting	building	checking	Writing(# of itemsets)	Total
0.01%	0.17	0.0	0.05	0.05	3.33	0.66(410920)	4.26
0.02%	0.86	0.02	0.06	0.05	2.08	0.31(129781)	3.38
0.03%	0.17	0.02	0.03	0.05	1.08	0.05(82108)	1.40

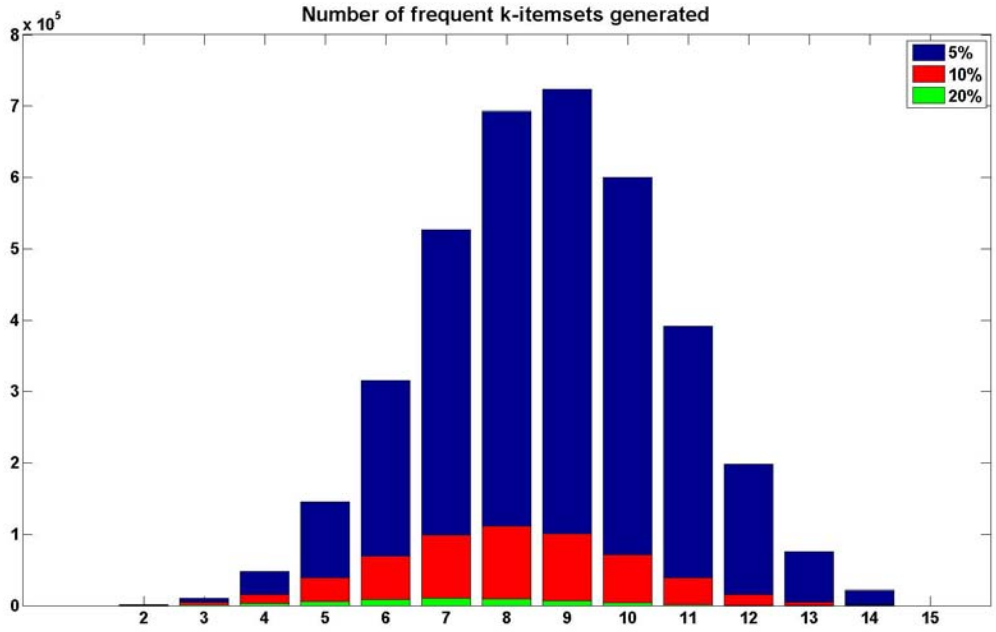
Filtering, sorting, and building seem to be consistent with a small deviation for all support levels. Checking and writing both decrease as the support level increases. The reading time for 0.02% is an anomaly so I will ignore this. As support increases there are fewer frequent itemsets generated. As fewer frequent itemsets are generated there is a decrease in the computation time for generating them. This is reflected in the total times and this correlates nicely with the number of itemsets generated.

Problem 1b:




Number of frequent k-itemsets generated for threshold 10%
x 10^4
12
10
8
6
4
2
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15


Number of frequent k-itemsets generated for threshold 20%
12000
10000
8000
6000
4000
2000
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15



The most obvious thing to note is that the magnitude of the peaks for each successive support threshold decreases. This is intuitive in that increasing the support threshold decreases the overall number of frequent itemsets. Also, the mode of the frequent itemsets decreases slightly as the support threshold increases(from 9 at 5% to 7 at 20%). This also makes sense in that large itemsets are less likely than smaller itemsets and as we increase the threshold fewer of the larger itemsets make the cut.

Problem 1c:

	T1014D100K	Mushroom
Frequent	410920	3755487
Closed	283149	12837
Maximal	127146	1453

We can see that the itemsets decrease from frequent to closed to maximal. And this is exactly as it is shown in the book. Maximal itemsets are a subset of closed itemsets. And closed itemsets are a subset of frequent itemsets.

Problem 2a:

The rule for finding the probability of Team A winning against Team B was 1->5, for which the support is 47.6% and the rule confidence is 40%. Or, of all the games played, Team A played Team B 47.6% of the time and of those games, won 40%.

The rule for finding the probability of Team A winning against Team C was 2->5. The support(% of games played against) was 52.4% and the rule confidence(% of games won) was 45.5%.

Problem 2b:

The probabilities of winning home games against each of the opponents can be found by adding (4) to the antecedent. So, the rule confidence for (1 4)->5 is 36.7% and the confidence for (2 4)->5 is 20%.

Problem 2c:

The probabilities of winning away games against each of the opponents can be found by adding (3) to the antecedent. So, the rule confidence for (1 3)->5 is 70% and the confidence for (2 3)->5 is 66.7%.

Problem 2d:

The results are not necessarily consistent but they are not necessarily inconsistent either. Yes, Team A does much better against its opponents in away games compared to home games and overall, but it also plays a much smaller percentage of its games away from home. The support of (3 5) is 33% which means that its winning percentage on the road is roughly weighted by half when added to the total of its home winning percentage. Since the number of home games is not equal to the number of away games, we really can't compare the results from 2a-2c directly.

Problem 3a:

Item	Support
1	100%
2	35.3%
3	0.5%
4	0.5%
5	0.5%
6	0.7%

Problem 3b:

The only itemset found with a support of 10% was (1 2).

Problem 3c:

The highest threshold I found was 0.466%. This could have been deduced by looking at the minimum threshold in the size-1 itemsets in 3a. If we would have increased the precision in 3a we would have found the minimum support was 7/15 or .46666667%. All itemsets found:

Item	Support
1 4	0.5%
1 3	0.5%
1 5	0.5%
1 5 6	0.5%
5 6	0.5%
1 6	0.7%
1 2	35.3%

Problem 3d:

We still find the most frequent itemset(1 2) but we also found two other low support/high confidence itemsets, one of which is below the threshold found in the previous answer.

Itemset	Support	Confidence

3 4	0.4%	92.9%
5 6	0.5%	70%
1 2	35.3%	35.3%

Problem 3e:

As noted, the itemset (1 2) was found in the original setup of the problem as both items (1) and (2) have high support relative to the other items. Items (3) through (6) have a very low level of support yet the hyperclique program found itemsets that included those items.

If the items correspond to the words {headline,writer,hong,kong,puerto,rico}, then the hyperclique results make sense. “Hong Kong” and “Puerto Rico” are geographical place names that would tend to be associated with each other since there are only a few other times those words would be used independently. For example, RICO is a statute used to indict gang activities but is probably less frequent than overall stories from Puerto Rico. Likewise, “King Kong” is a popular movie but overall less frequently mentioned in the news than “Hong Kong”. So the results show us that although the itemsets have a very low level of support, the items in the itemset have a high correlation with each other.