Overview

This session is dedicated to association rule mining.
Have fun!

1 Apriori

For the transactional database below, iterate through the Apriori algorithm using a min. support threshold $s = 2$.

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

2 PCY

Here is a collection of twelve baskets. Each contains three of the six items 1 through 6.

$$\{1,2,3\}\{2,3,4\}\{3,4,5\}\{4,5,6\}$$

$$\{1,3,5\}\{2,4,6\}\{1,3,4\}\{2,4,5\}$$

$$\{3,5,6\}\{1,2,4\}\{2,3,5\}\{3,4,6\}$$

On the first pass of the PCY algorithm, we use a hash table with 11 buckets, and the set $\{i, j\}$ is hashed to bucket $i \times j \ mod \ 11$, where mod is the *modulo* operation, i.e. finding the remainder after division by 11. For example, $hash\left(\{5,6\}\right) = (5 \times 6) \ mod \ 11 = 30 \ mod \ 11 = 8$, because $30 = 11 \times 2 + \mathbf{8}$.

The support threshold is 4.

1. Perform the first pass of PCY.
   (a) Compute the support for each individual item.
   (b) Which pairs hash to which buckets? Compute the frequencies of the buckets.
   (c) Which buckets are frequent?
2. Generate all candidate pairs (2-itemsets): which ones are counted on the second pass of PCY?

3 FP Growth

Using the transactional database below, construct its frequent pattern tree (FP tree) using a minimum support threshold, $s = 3$.

TID	Items
1	A, B, C, E, F
2	A, C, D, E, F
3	A, B, C, G, I
4	A, B, C, G
5	B, E, F, H, I, J

4 Thought Question

Pattern explosion is a well-known problem in frequent itemset mining. High support thresholds typically result only in few well-known patterns; but for low support thresholds, the number of frequent itemsets can easily be orders of magnitude larger than the number of transactions. Knowledge discovery in such humongous itemset collections is virtually impossible.
What are the causes of pattern explosion? Can you think of a way to solve or at least alleviate this issue?