LINGI2364: Mining Patterns in Data
Exercise session 1: Frequent Itemset Mining

Gilles Peiffer

11 February 2020

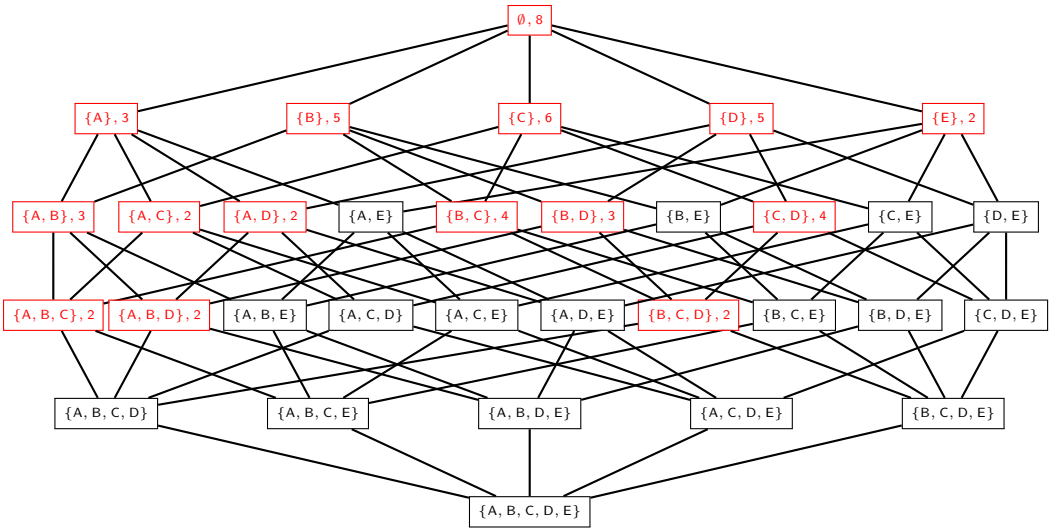
3 Solutions

1. (a)

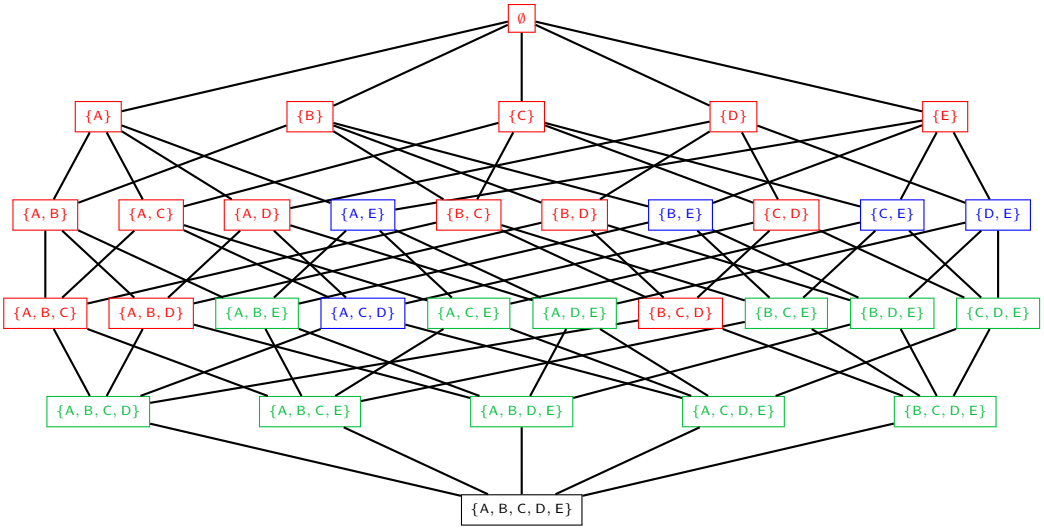
$\mathcal{S}$	$\text{cover}(\mathcal{S})$	$\text{support}(\mathcal{S})$	$\text{frequency}(\mathcal{S})$
$\{C, D\}$	$\{1, 3, 8\}$	3	$3/10$
$\{F\}$	$\{0, 8\}$	2	$1/5$
$\{B, D\}$	$\{2, 3, 5, 6, 9\}$	5	$1/2$

   (b) From the anti-monotonicity property, one can deduce that $\text{cover}(\{B, C, D\}) \subseteq \text{cover}(\{C, D\}) = \{1, 3, 8\}$, but also that $\text{cover}(\{B, C, D\}) \subseteq \text{cover}(\{B, D\}) = \{2, 3, 5, 6, 9\}$. Constraints on the support and frequency can be deduced from this. Combining these results, one can find that $\text{cover}(\{B, C, D\}) \subseteq \text{cover}(\{C, D\}) \cap \text{cover}(\{B, D\}) = \{3\}$, which can again be translated into constraints on the support and frequency.

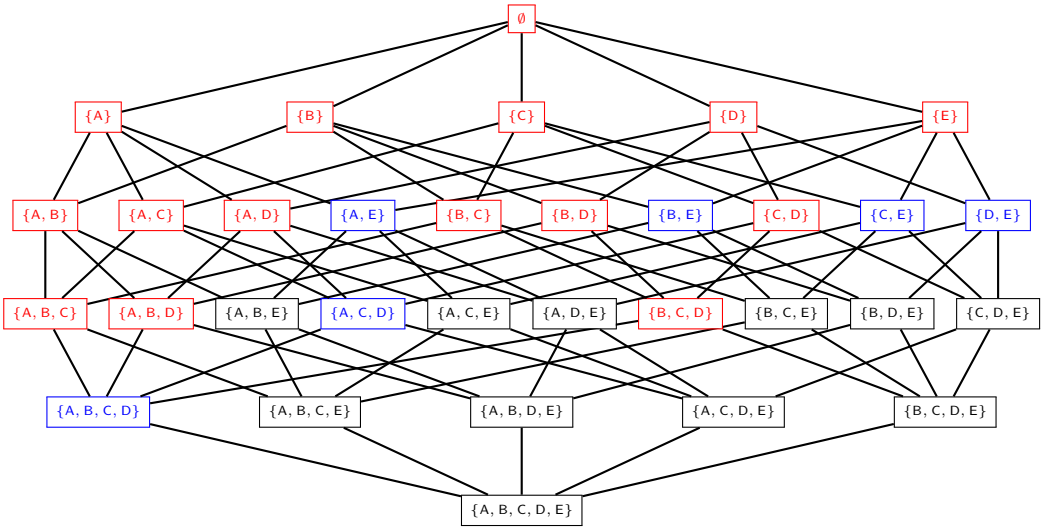
2. (a) The frequent itemsets are as given in red, with their support next to them:



   (b) The Apriori algorithm proceeds from top to bottom. At each level, green means a set was generated but removed after finding an infrequent subset, blue means it was generated, has no infrequent subsets, but was removed after computing its frequency, and finally red means the set is frequent. Black simply denotes the existence of a set which was not generated.



(c) The improved Apriori algorithm also proceeds from top to bottom. The color coding is the same as in the previous exercise.



3. One good strategy would be to apply the Apriori algorithm to the first dataset to find frequent itemsets, and only compute the support on the second dataset for those itemsets which were found to be frequent for the first dataset.

   The maximum frequency constraint can be handled in a similar fashion as the minimum frequency constraint, by using an “inverse Apriori” algorithm which only looks at the subsets of infrequent sets (with θ being the maximum cutoff) in the second dataset, and removes those which are too frequent. One first applies the regular Apriori algorithm using the first dataset, and then applies the inverse Apriori algorithm using the second dataset on the frequent itemsets of the first dataset. This algorithm is a consequence of the anti-monotonicity property.