Advanced AI

课程

罗平

Autumn 2016

Pattern Mining over Transaction Database

2-1

FREQUENT PATTERN MINING

1992年，IBM，Rakesh Agrawal

- IBM客户交流会
  - 衔接大客户和IBM研究人员
- 超市：超市扫码机的普及
- 数据
  - 积累大量用户购买记录数据

1992年， IBM， Rakesh Agrawal

- 用户购买数据

购买编号	购买记录
1	B C E
2	A B
3	A B C
4	A B D
5	A B C D E F G

1992年，IBM，Rakesh Agrawal

- 频繁模式：经常被一起购买的商品

购买编号	购买记录
1	B C E
2	A B
3	A B C
4	A B D
5	A B C D E F G

1992年，IBM，Rakesh Agrawal

- 啤酒和尿布的故事

Problem Statement

- Frequent pattern mining
- ***Support***

Transaction No.	Items
1	B C E
2	A B
3	A B C
4	A B D
5	A B C D E F G

The support of {BC} = 3/5

Problem Statement

- ***Frequent patterns***

  support is bigger than a parameter α

Transaction No.	Items
1	B C E
2	A B
3	A B C
4	A B D
5	A B C D E F G

The support of {BC} = 3/5
{BC} is frequent when α = 0.5

Frequent Pattern Mining

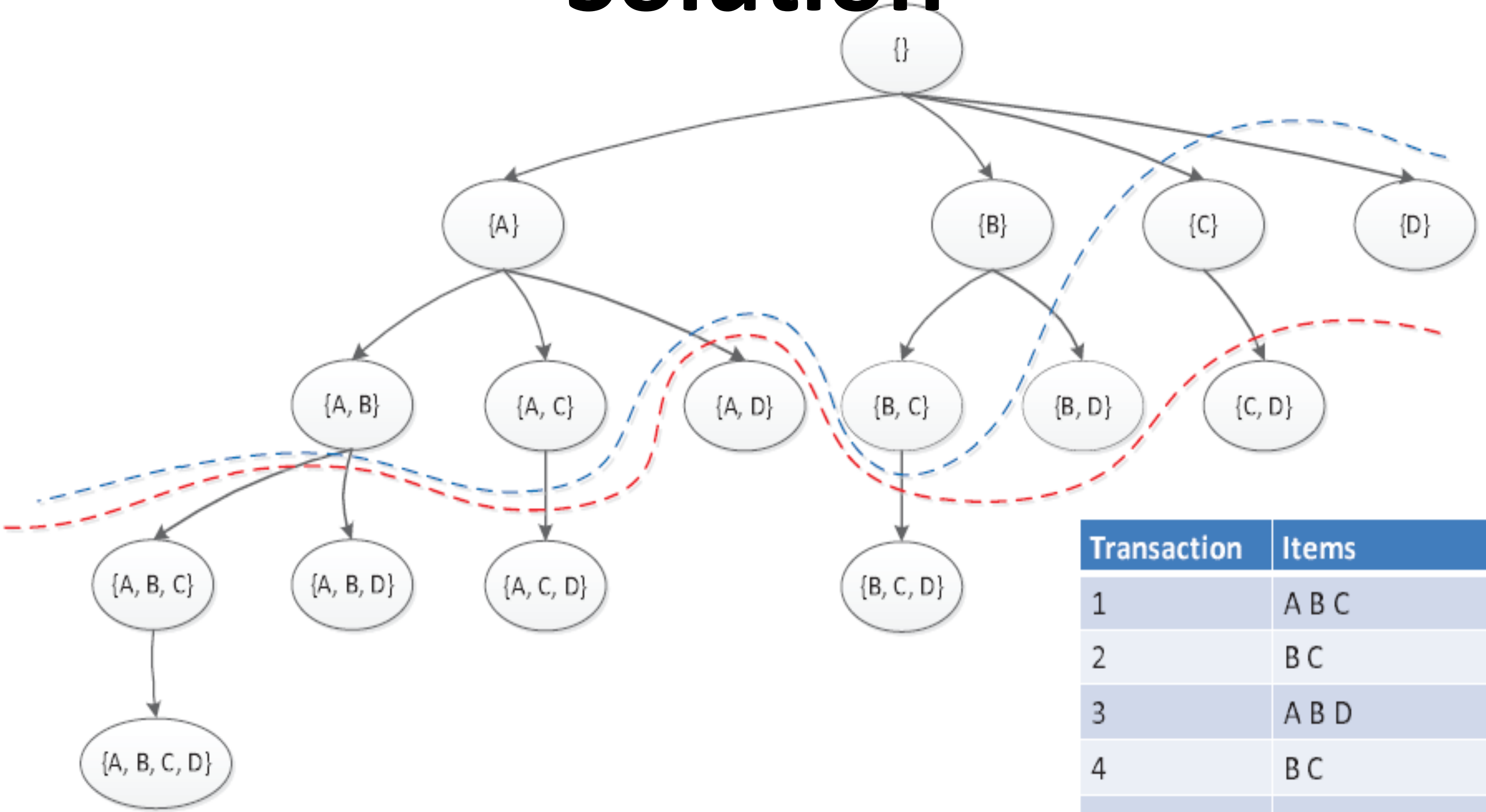
- ***Given: a transaction database, and min_sup*** α
- ***Output: all the frequent patterns***

R. Agrawal, T. Imielinski, A.N. Swami: Mining Association Rules between Sets of Items in Large Databases, SIGMOD 1993. *Won the 2003 SIGMOD Test of Time Award for the most impactful paper over the intervening decade*. Citations.

Frequent Pattern Mining

- *Naïve solution*
- *Check all the patterns (itemset) one by one*

Solution

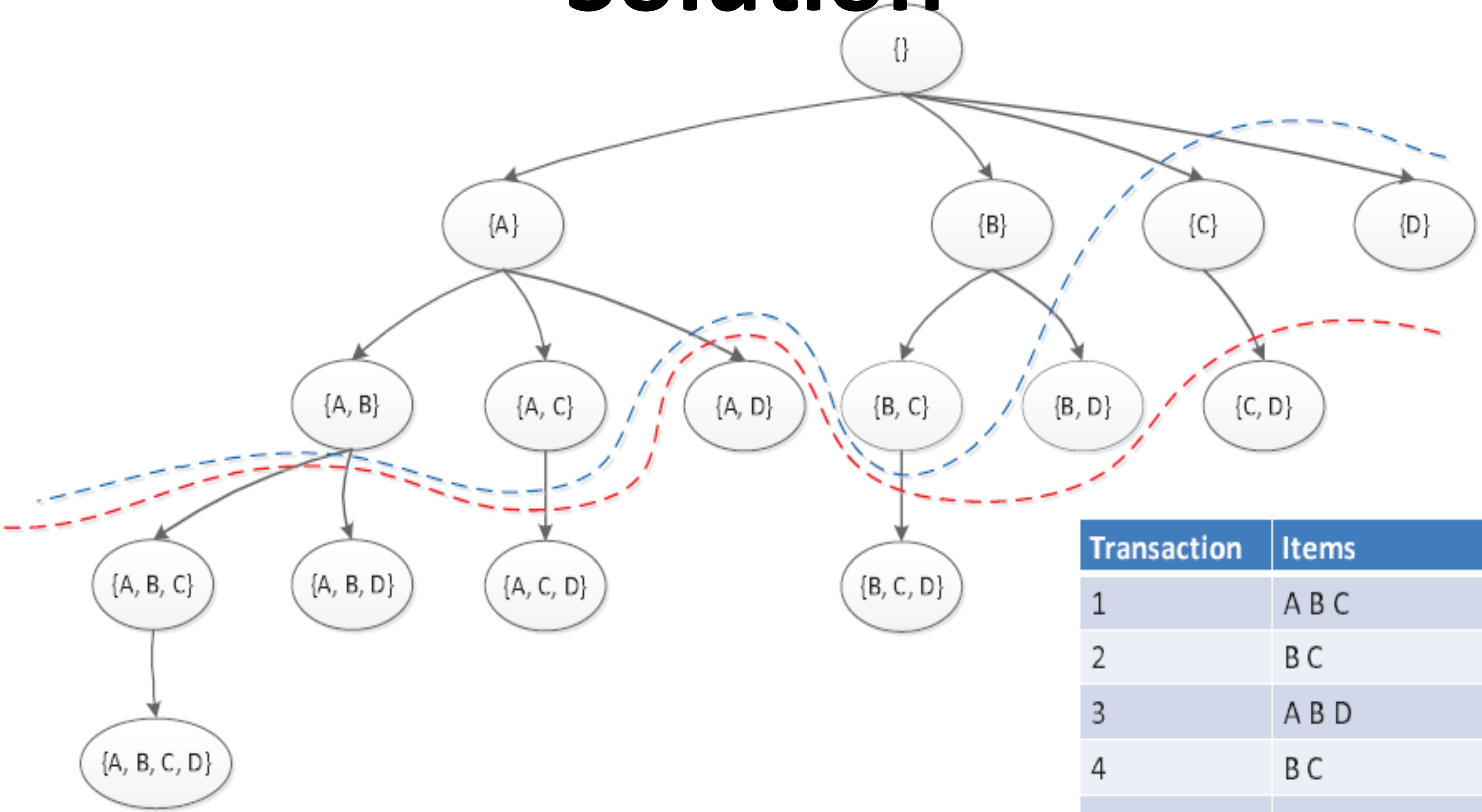


Transaction	Items
1	A B C
2	B C
3	A B D
4	B C
5	A C
6	B C D

lexicographic subset tree: list all the itemsets

(all the items are ranked in a fixed sequence)

Solution



Transaction	Items
1	A B C
2	B C
3	A B D
4	B C
5	A C
6	B C D

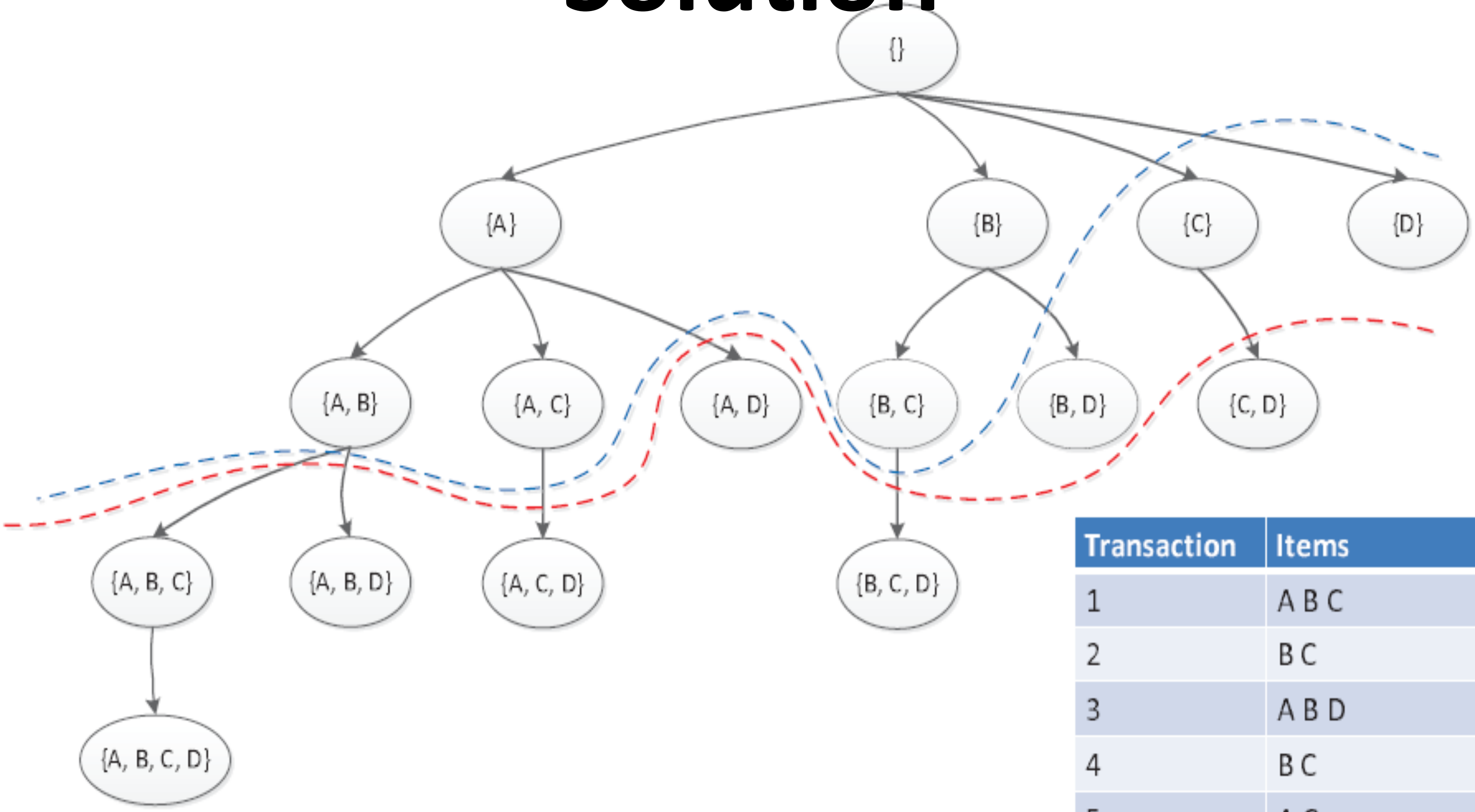
Property on lexicographic subset tree
子树根节点对应的itemset是子树上的任意节点对应的itemset的子集

Frequent Patterns

- ***Anti-Monotone重要性质***
- Frequent itemset的任何子集都是frequent的
- 等价的形式：对于一个itemset，只要它的任意一个子集不frequent，那么它就不frequent
- 推出：如果一个itemset不frequent，那么任何包含它的itemset都不frequent

Anti-monotone: 随着itemset的item增加，它的frequency不增加

Solution

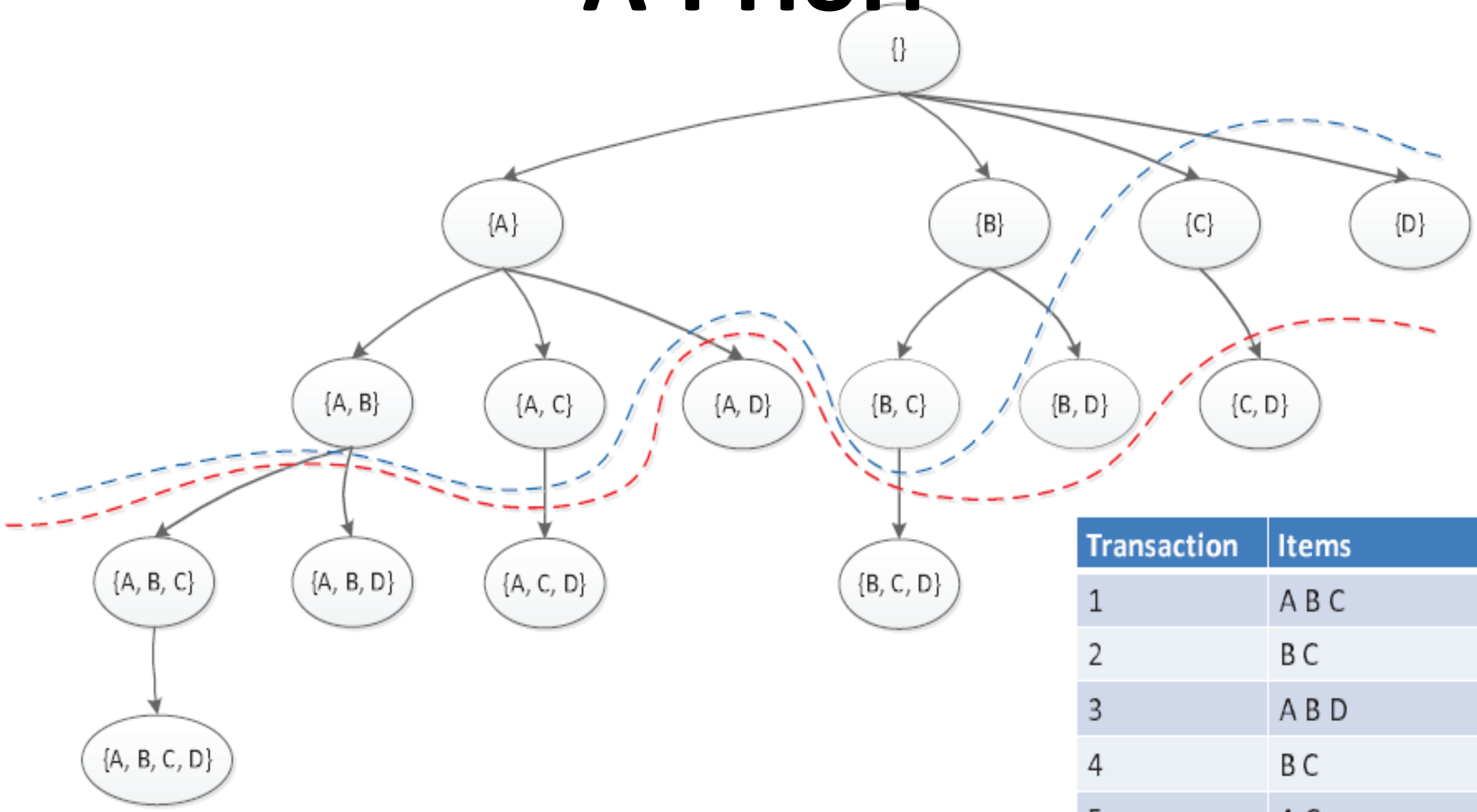


Transaction	Items
1	A B C
2	B C
3	A B D
4	B C
5	A C
6	B C D

Pruning by anti-monotone
The red line in the graph

剪枝的本质：对每棵子树，估算出该棵子树上所有节点的frequency的最大值(upper bound)

A-Priori



Transaction	Items
1	A B C
2	B C
3	A B D
4	B C
5	A C
6	B C D

Breadth-first traverse on the lexicographic subset tree with pruning by anti-monotone

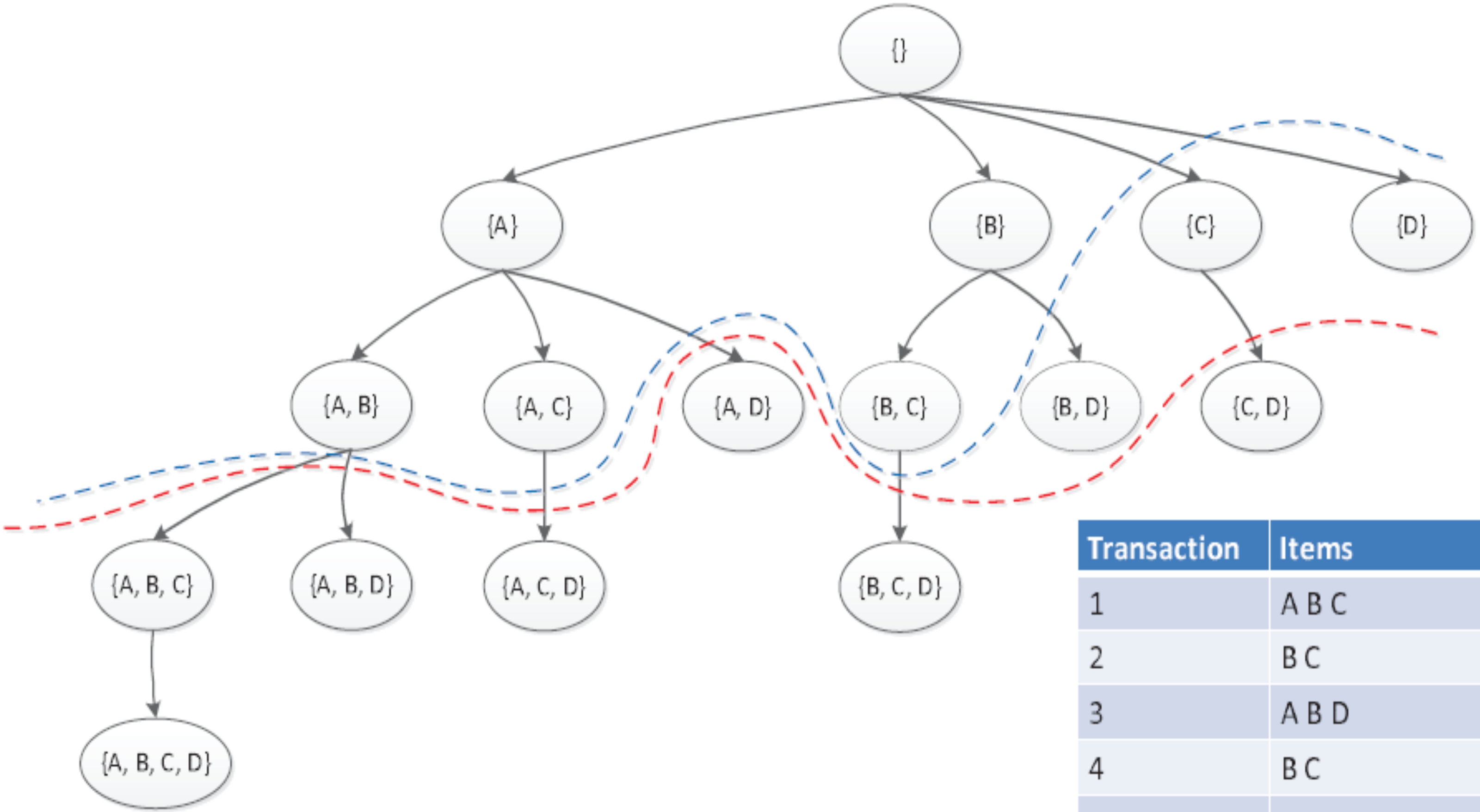
Trick: if BC is infrequent, we don’t need to extend ABC

Maximal Frequent Patterns

- 在某个frequent itemset上，添加任意的item后，都会变为infrequent
- 那么，它是一个maximal frequent itemset
- maximal frequent itemset是最大可能的长

Maximal Frequent Patterns

- 那么，紧贴着红线的节点都是maximal frequent itemsets吗？



Transaction	Items
1	A B C
2	B C
3	A B D
4	B C
5	A C
6	B C D

More Works on Frequent Pattern Mining

- More efficient algorithm for FPM:
  - J. Han, **J. Pei**, Y. Yin, and R. Mao. "Mining Frequent Patterns without Candidate Generation: A Frequent-pattern Tree Approach". *Data Mining and Knowledge Discovery: An International Journal*, Volume 8, Issue 1, pages 53-87, January 2004, Kluwer Academic Publishers.
- – Mining maximal frequent patterns
  - J. Wang, J. Han, and J. Pei, "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets", in Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery andData Mining (KDD'03), Washington, D.C., Aug. 2003.
- – More: parallel, incremental, top-K, … …