

# Itemsets

Mining Massive Datasets

Carlos Castillo

Topic 05

# Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapters 4, 5) – [slides by Lijun Zhang](#)
- Mining of Massive Datasets 2<sup>nd</sup> edition (2014) by Leskovec et al. ([Chapter 6](#)) - [slides](#)
- Data Mining Concepts and Techniques, 3<sup>rd</sup> edition (2011) by Han et al. (Chapter 6)
- Introduction to Data Mining 2<sup>nd</sup> edition (2019) by Tan et al. (Chapters 5, 6) – [slides ch5](#), [slides ch6](#)

# Market Basket Analysis

- Understand customers
  - Purchasing habits, sensitivity to price, promotions
- Understand products
  - Co-purchases, fast/slow movers
- Take action: promotions, store layout, ...

# Transactions contain items, which can be grouped into itemsets

- Transactions
  - Sets of items bought by customers
- The Goal
  - Determine associations between groups of items bought by customers
- Quantification of the Level of Association
  - Frequencies of sets of items
- The Discovered Sets of Items
  - Large itemsets, frequent itemsets, or frequent patterns

# “Transaction” is a general concept

Items	Transactions
Groceries	Grocery cart
University courses	Transcript of courses taken
Guests	Party
Movies	Actor
Symptoms	Patient
Streamed songs	Streaming subscriber
Words	Document
Liked photos	Instagram account

# Applications

- Supermarket Data
  - Target marketing, shelf placement
- Text Mining
  - Identifying co-occurring terms
- Generalization to Dependency-oriented Data Types
  - Web log analysis, software bug detection
- Other Major Data Mining Problems
  - Clustering, classification, and outlier analysis

# Association rules

- Generated from **frequent itemsets**
- Formulation  $X \Rightarrow Y$ 
  - {Soy latte}  $\Rightarrow$  {Brown Sugar}
  - {Kale, Quinoa}  $\Rightarrow$  {Almond milk}
- Applications
  - Promotion
  - Shelf placement
- Conditional Probability  $P(Y|X) = \frac{P(X \cap Y)}{P(X)}$

# Association rule mining

- $U$  is a set of  $d$  items
- $T$  is a set of  $n$  transactions  $T_1, T_2, \dots, T_n$  with  $T_i \subseteq U$
- **Itemset**: a set of items
- **k-itemset**: a set of  $k$  items

How many different k-itemsets exist?  $2^k$



# Binary representation of a transaction

tid	Set of items	Binary representation
1	Bread, Jam, Juice	110010
2	Tofu, Juice, Tomatoes	000111
3	Bread, Strawberries, Tofu, Juice	101110
4	Tofu, Juice, Tomatoes	000111
5	Strawberries, Juice, Tomatoes	001011

# Definitions

- **Support of itemset  $I$** , written  $sup(I)$ :  
the fraction of transactions in the database  
 $T = \{T_1 \dots T_n\}$  that contain  $I$  as a subset.
- Frequent itemset mining with support minsup:  
Given a set of transactions  $T = \{T_1, \dots, T_n\}$ ,  
where  $T_i \subseteq U$ ,  
find all itemsets  $I_j$  such that  $sup(I_j) \geq minsup$

# Example

tid	Set of items	Binary representation
1	Bread, Jam, Juice	110010
2	Tofu, Juice, Tomatoes	000111
3	Bread, Strawberries, Tofu, Juice	101110
4	Tofu, Juice, Tomatoes	000111
5	Strawberries, Juice, Tomatoes	001011

- $\text{sup}(\{\text{Bread, Juice}\}) = 2/5 = 0.4$
- $\text{sup}(\{\text{Strawberries, Tomatoes}\}) = 1/5 = 0.2$
- If  $\text{minsup}=0.3$ ,  $\{\text{Bread, Juice}\}$  is a frequent itemset

# Try it!

## Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

- Write the support of every 2-itemset or 3-itemset occurring in this database
- Indicate which are frequent itemsets if  $\text{minsup} = 1/2$

# Properties

- The smaller minsup is, the larger the number of frequent itemsets

Support monotonicity property:

*if  $J \subseteq I$ ,  $\text{sup}(J) \geq \text{sup}(I)$  WHY?*

# Properties

- The smaller minsup is, the larger the number of frequent itemsets
- **Support monotonicity property:**  
*if  $J \subseteq I$ ,  $\text{sup}(J) \geq \text{sup}(I)$*
- Confusingly, some authors refer to this as the support anti-monotonicity property
- **Downward closure property:** every subset of a *frequent* itemset is also *frequent*

# Closed itemset

An itemset is **closed** if all itemsets containing it are (strictly) less frequent

tid	Set of items
1	Bread, Jam, Juice
2	Tofu, Juice, Tomatoes
3	Bread, Strawberries, Tofu, Juice
4	Tofu, Juice, Tomatoes
5	Strawberries, Juice, Tomatoes

Find a closed itemset in this set of transactions

# Closed itemset

An itemset is **closed** if all itemsets containing it are less frequent

tid	Set of items
1	Bread, Jam, Juice
2	Tofu, Juice, Tomatoes
3	Bread, Strawberries, Tofu, Juice
4	Tofu, Juice, Tomatoes
5	Strawberries, Juice, Tomatoes

- Example closed itemset: {Bread, Juice}
- $\text{sup}(\{\text{Bread, Juice}\}) = 2$ 
  - $\text{sup}(\{\text{Bread, Juice, Jam}\}) = 1$
  - $\text{sup}(\{\text{Bread, Juice, Strawberries}\}) = 1$
  - $\text{sup}(\{\text{Bread, Juice, Tofu}\}) = 1$



# Maximal itemset

An itemset is **maximal** if

- it is closed and
- it has support  $\geq$  minsup

tid	Set of items
1	Bread, Jam, Juice
2	Tofu, Juice, Tomatoes
3	Bread, Strawberries, Tofu, Juice
4	Tofu, Juice, Tomatoes
5	Strawberries, Juice, Tomatoes

## Try it!

- Find three **maximal** frequent patterns at minsup=0.4
- Tip: first find all frequent patterns at minsup=0.4
  - The total number is 11 (?)

# Maximal itemset

An itemset is **maximal** if

- it is closed and
- it has support  $\geq$  minsup

tid	Set of items
1	Bread, Jam, Juice
2	Tofu, Juice, Tomatoes
3	Bread, Strawberries, Tofu, Juice
4	Tofu, Juice, Tomatoes
5	Strawberries, Juice, Tomatoes

- **Maximal** frequent patterns at minsup=0.4  
{Bread, Juice}, {Strawberries, Juice}, {Tofu, Juice, Tomatoes}
- Frequent patterns at minsup=0.4
  - The total number is 11 (?)
  - All of them are subsets of maximal frequent patterns

# Maximal itemset

An itemset is **maximal** if

- it is closed and
- it has support  $\geq$  minsup

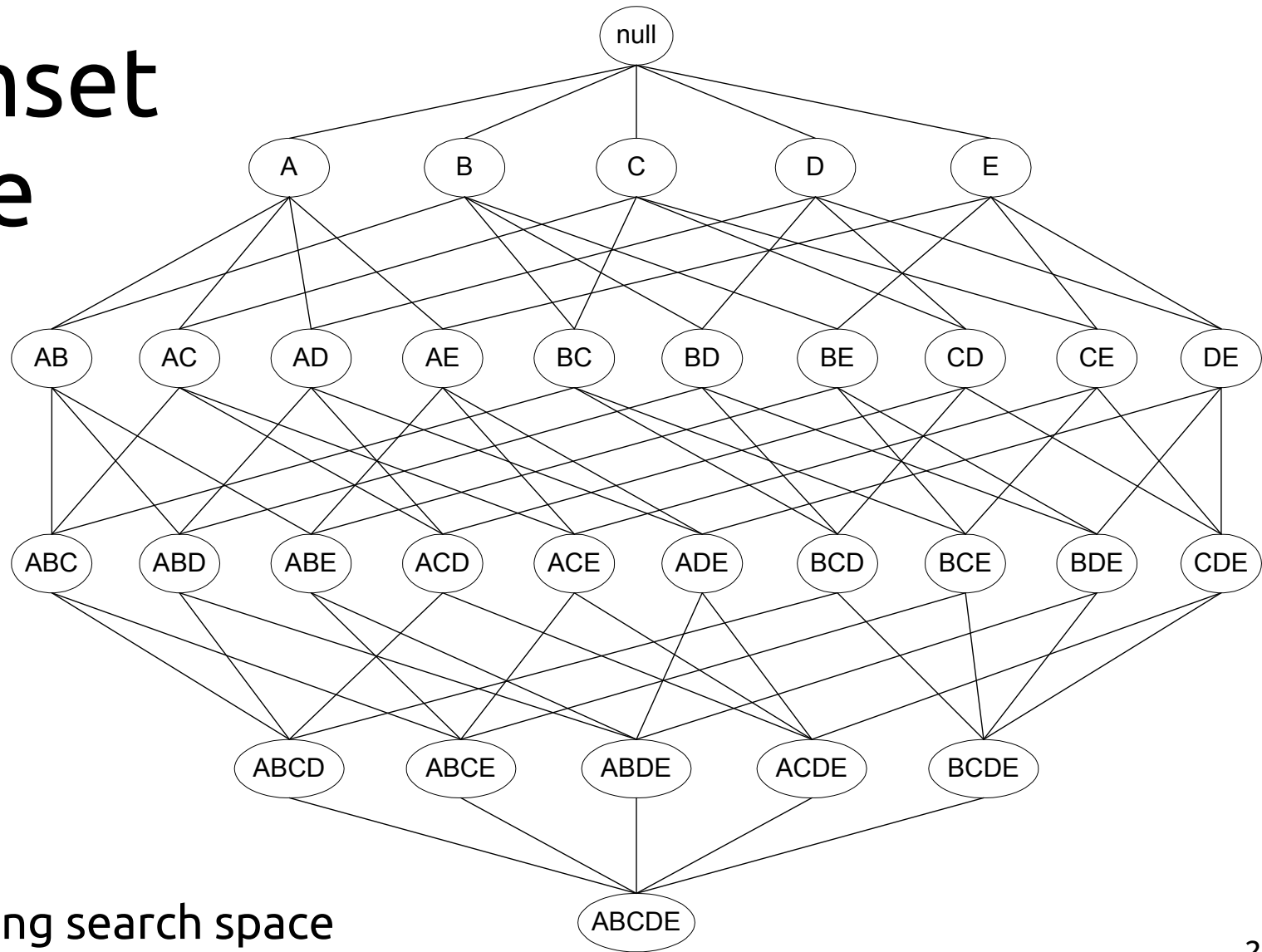
tid	Set of items
1	Bread, Jam, Juice
2	Tofu, Juice, Tomatoes
3	Bread, Strawberries, Tofu, Juice
4	Tofu, Juice, Tomatoes
5	Strawberries, Juice, Tomatoes

## Maximal patterns

{Bread, Juice}, {Strawberries, Juice}, {Tofu, Juice, Tomatoes}

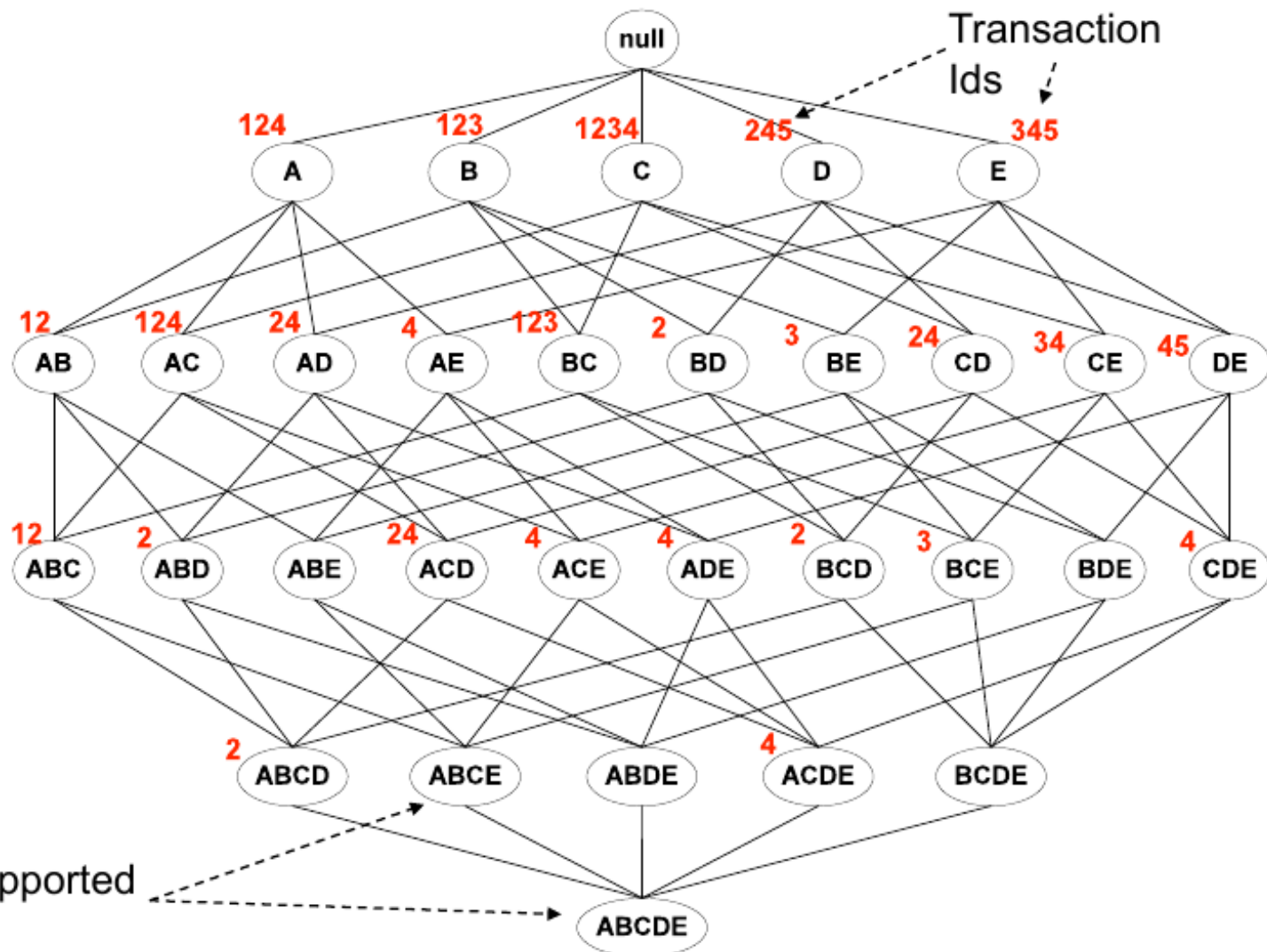
... are **condensed representations of frequent patterns**, but do not retain information about the support of their subsets.

# The itemset lattice



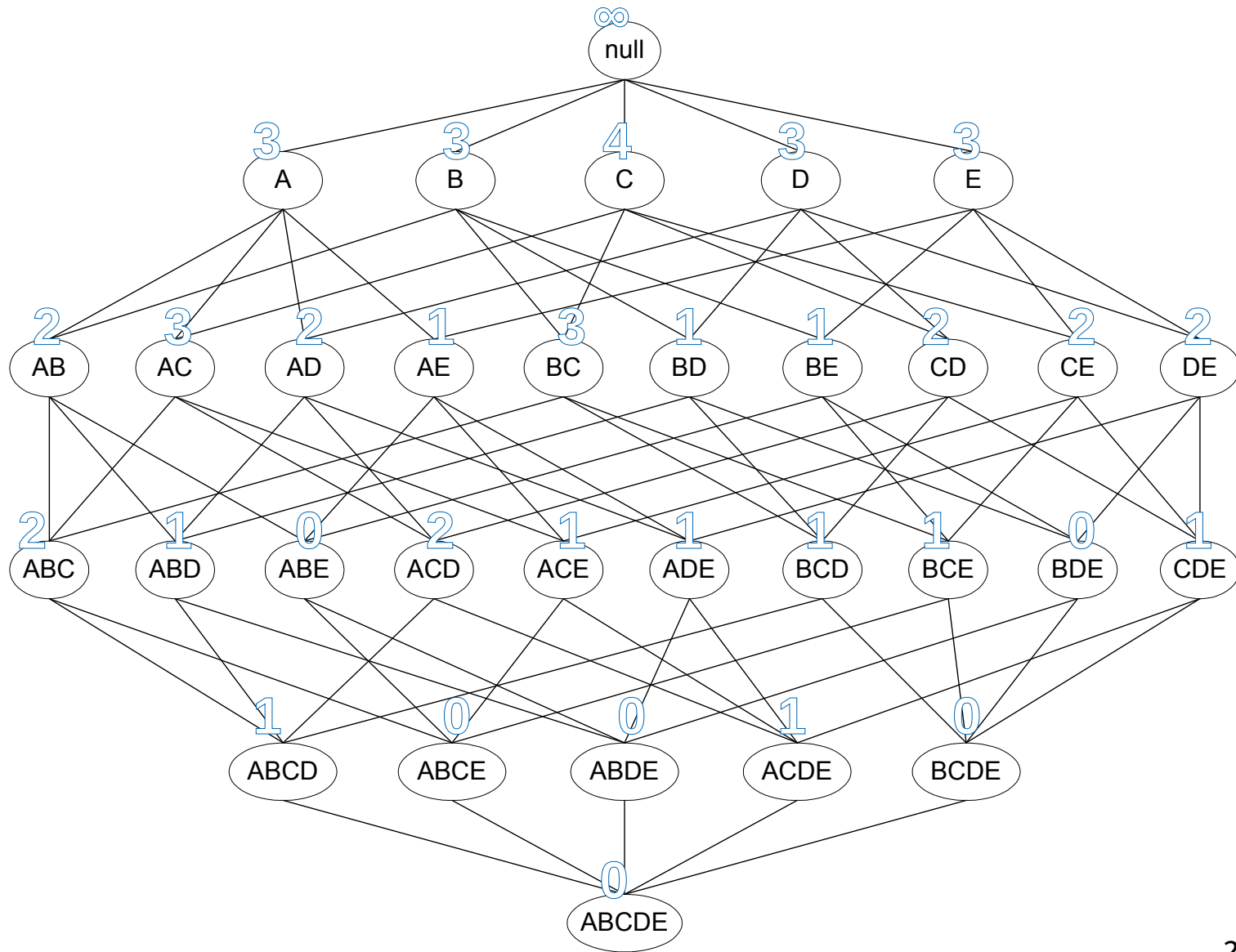
$2^{|U|}$  nodes representing search space

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



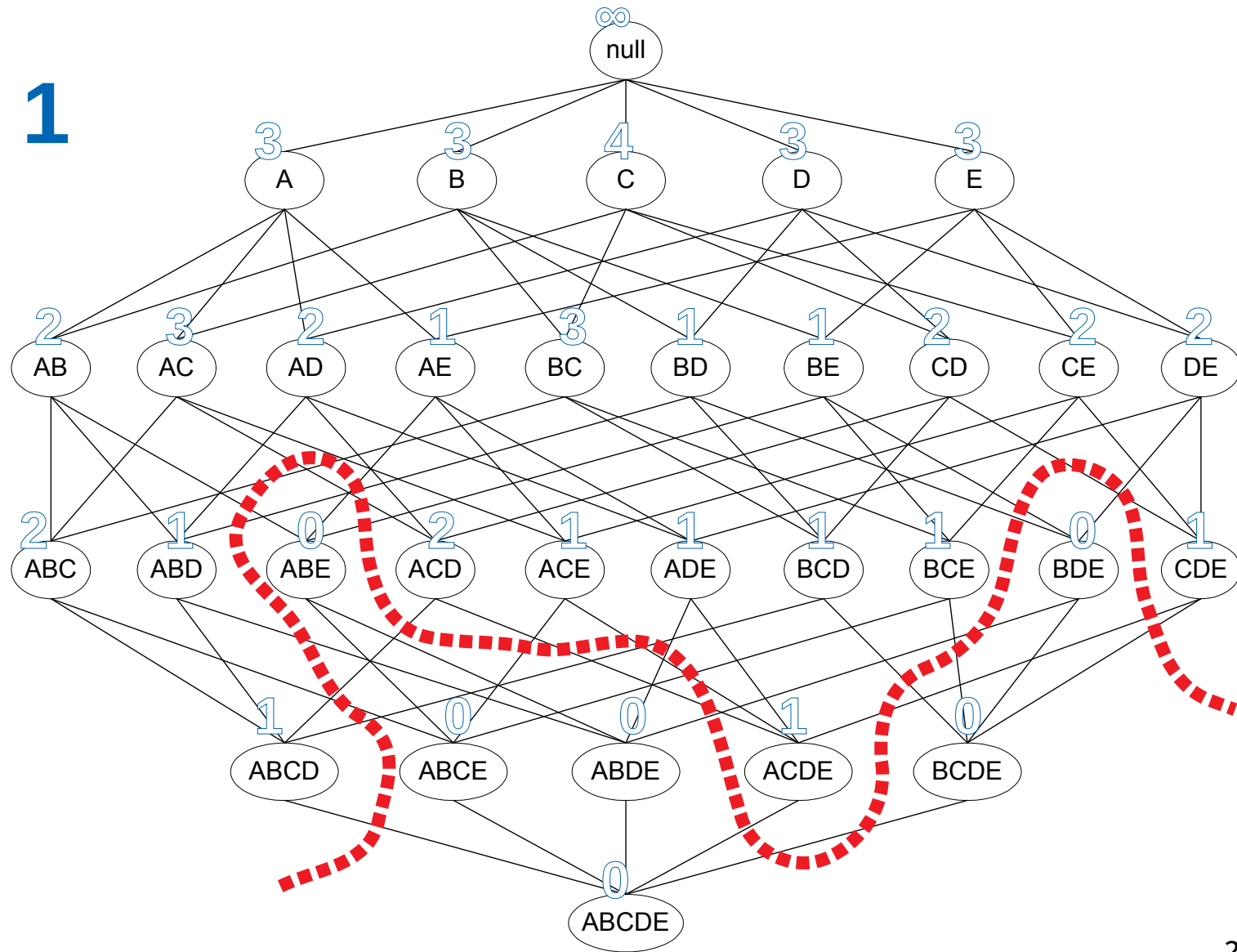
# Support of each itemset

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



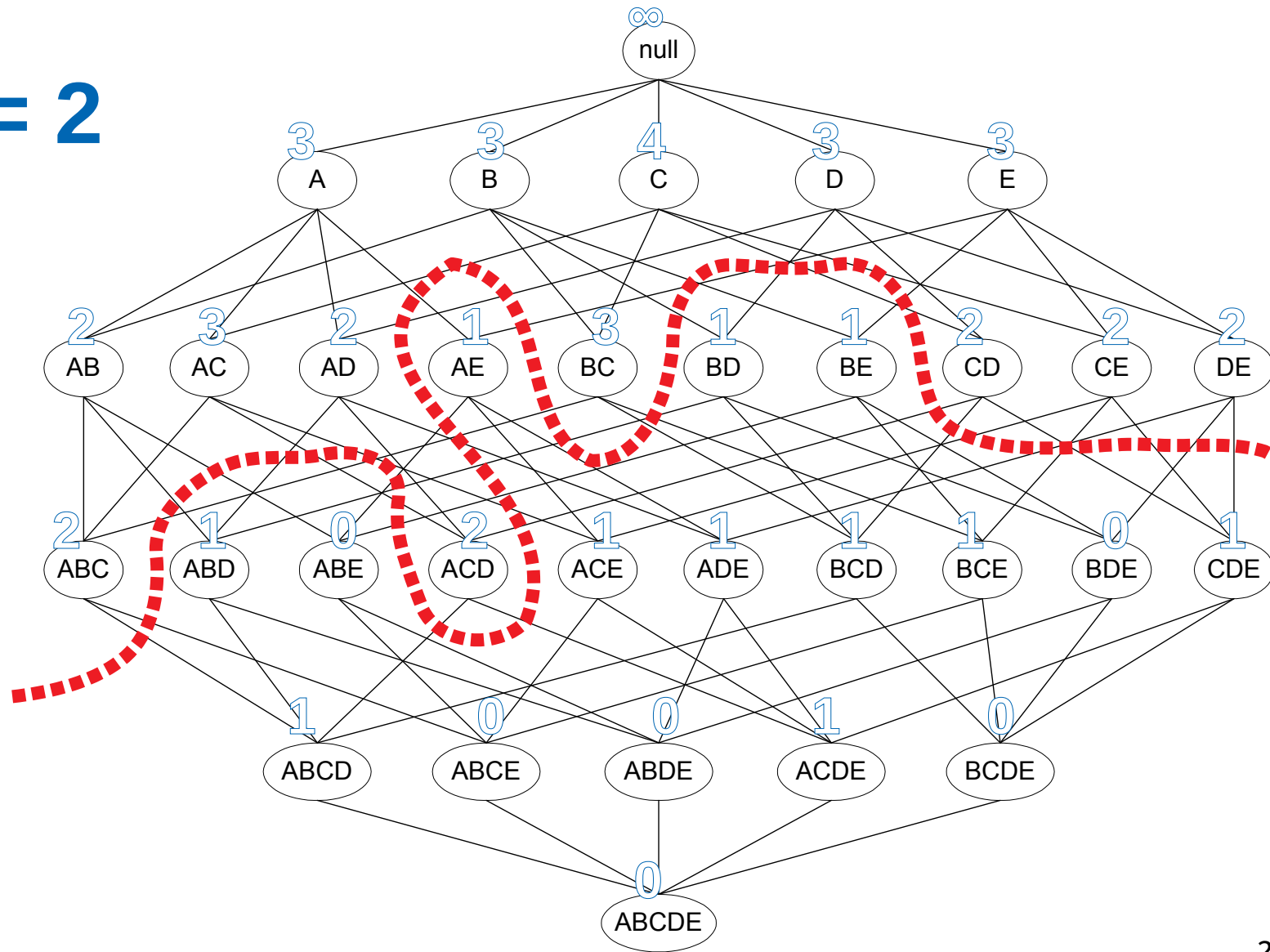
# minsup = 1

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



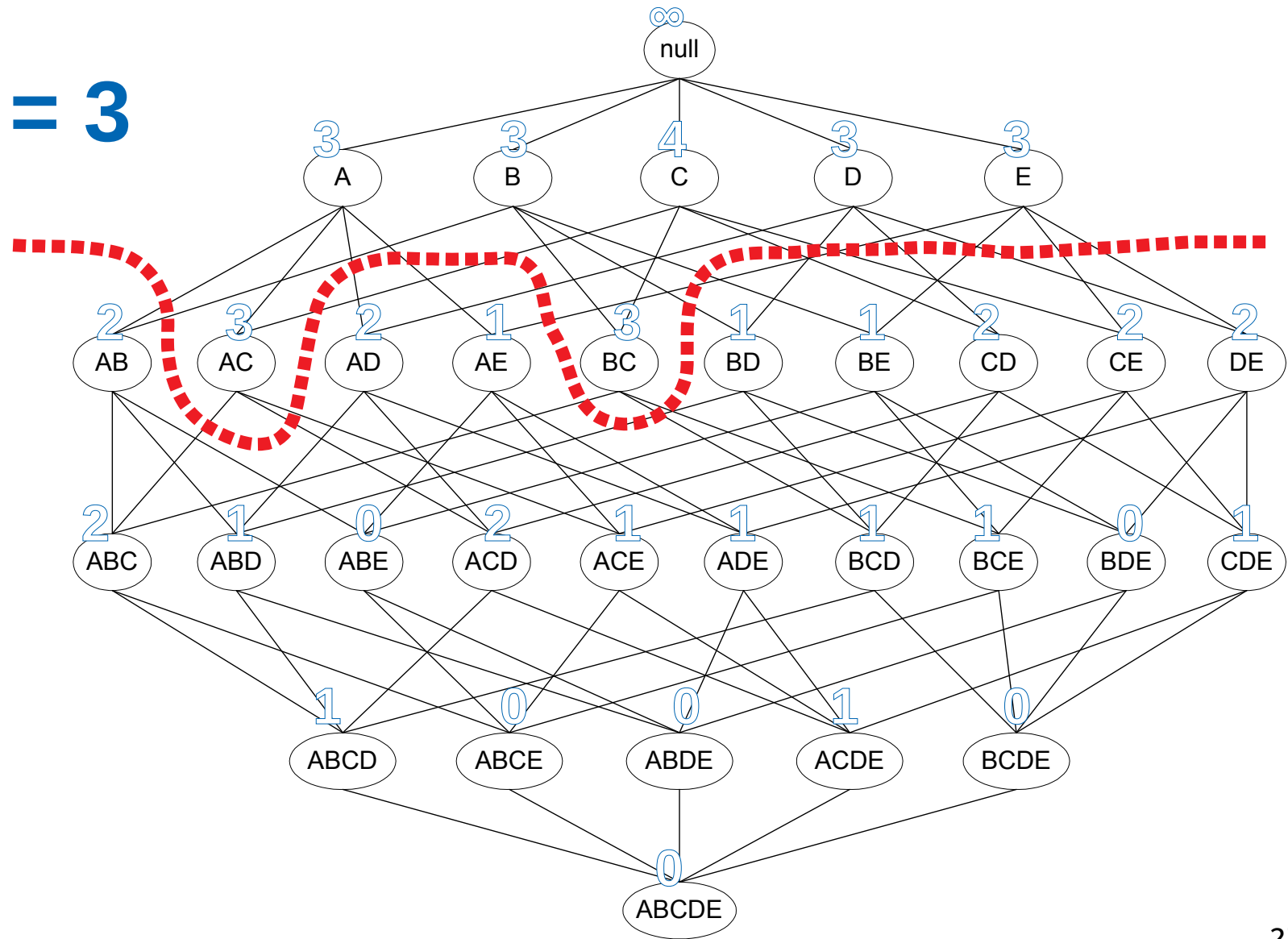
# minsup = 2

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE





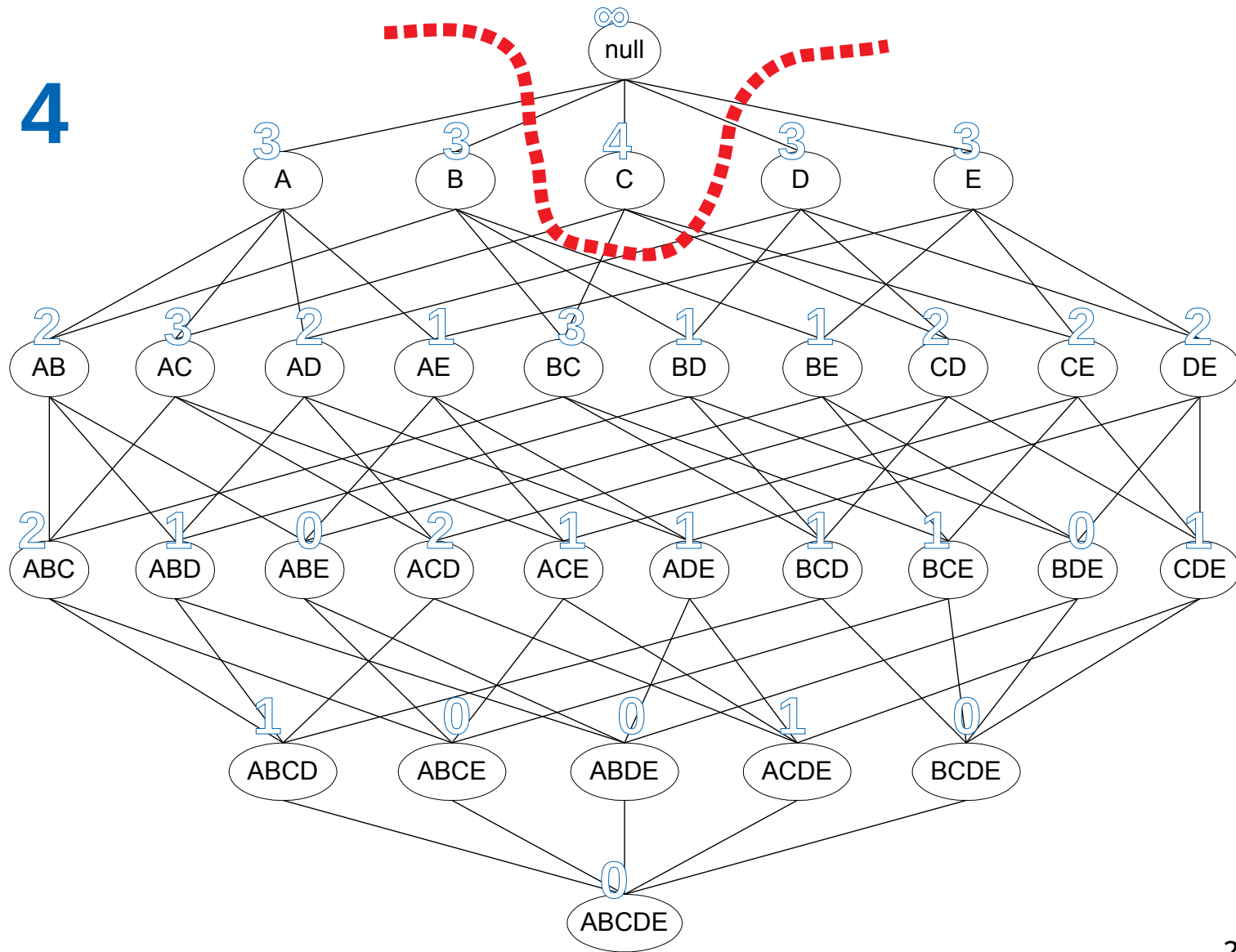
# minsup = 3



TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

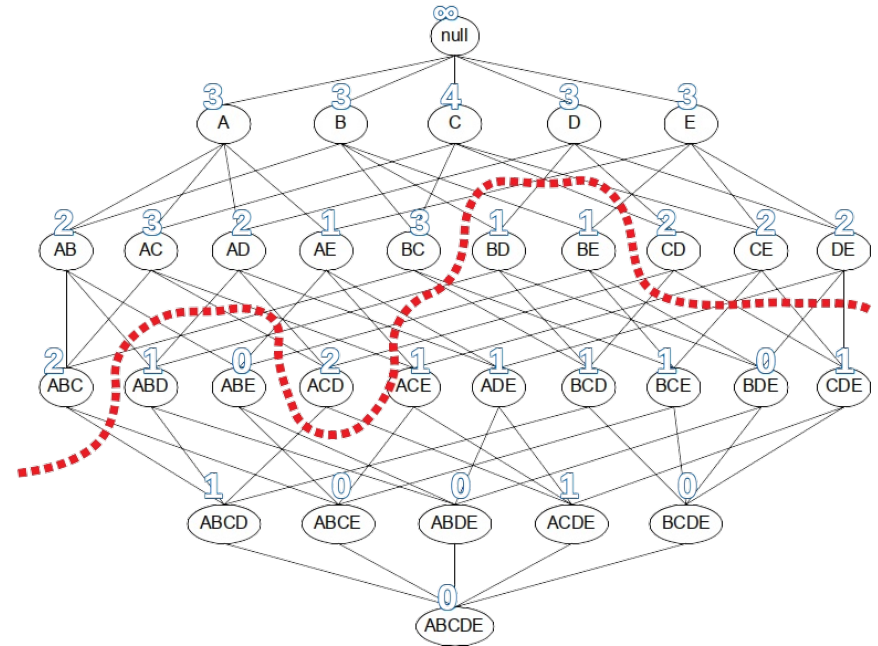
minsup = 4

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



# The border is a graph cut and ...

- All itemsets **above** the border are **frequent**
- All itemsets **below** the border are **not frequent**
- All **maximal** frequent itemsets are adjacent to the border
- Any border respects the **downward closure** property



# Association rules

# What is a rule

- A rule is of the form  $X \Rightarrow Y$   
X and Y are itemsets
- X is the antecedent, Y is the consequent
- The **confidence** of the rule is:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$

# Confidence of a rule

- The **confidence** of the rule  $X \Rightarrow Y$  is:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$

- This is the conditional probability of  $X \cup Y$  occurring in a transaction, given that  $X$  occurs in the transaction

# Confidence of a rule (cont.)

tid	Set of items
1	Bread, Jam, Juice
2	Tofu, Juice, Tomatoes
3	Bread, Strawberries, Tofu, Juice
4	Tofu, Juice, Tomatoes
5	Strawberries, Juice, Tomatoes

$\text{conf}(\{\text{tofu}, \text{juice}\} \Rightarrow \{\text{tomatoes}\}) = ?$

# Lift of a rule

- The **lift** of the rule  $X \Rightarrow Y$  is:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X) \text{sup}(Y)}$$

- This is the ratio between the observed support and the expected support if  $X$  and  $Y$  were independent



# Try it!



$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$

$$\text{lift}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X) \text{sup}(Y)}$$

Rule	Support $\text{sup}(X \cup Y)$	Confidence	Lift
$A \Rightarrow D$			
$C \Rightarrow A$			
$A \Rightarrow C$			
$B \& C \Rightarrow D$			

# Association rule (minsup, minconf)

- Let  $X, Y$  be two itemsets; the rule  $X \Rightarrow Y$  is an **association rule** of minimum support **minsup** and minimum confidence **minconf** if:

$$\text{sup}(X \Rightarrow Y) \geq \text{minsup}$$

and

$$\text{conf}(X \Rightarrow Y) \geq \text{minconf}$$

# Summary

# Things to remember

- Itemset, k-itemset, transaction
- Support, confidence, lift
- Maximal and closed itemsets
- Association rule of minsup and minconf

# Exercises for this topic

- Data Mining, The Textbook (2015) by Charu Aggarwal
  - Exercises 4.9 → 1-3, 5, 7-8
  - Exercises 5.7 → 1-5
- Mining of Massive Datasets 2<sup>nd</sup> edition (2014) by Leskovec et al.
  - Exercises 6.1.5 → 6.1.1-6.1.7
- Introduction to Data Mining 2<sup>nd</sup> edition (2019) by Tan et al.
  - Exercises 5.10 → 2-7