Itemsets

Mining Massive Datasets

Prof. Carlos Castillo — https://chato.cl/teach

Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Chapters 4, 5) – slides by Lijun Zhang
- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. (Chapter 6) - slides
- Data Mining Concepts and Techniques, 3rd edition (2011) by Han et al. (Chapter 6)
- Introduction to Data Mining 2nd 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
Actors	Movies
Symptoms	Patient
Streamed songs	Streaming subscriber
Words	Document
Liked photos	Instagram account

https://web.stanford.edu/class/cs102/lecturenotes/DataMining.pdf

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⇒Y
  - {Soy latte} ⇒ {Brown Sugar}
  - {Kale, Quinoa} ⇒ {Almond milk}
- Applications
  - Promotion
  - Shelf placement
- Conditional Probability $P(Y|X) = \dfrac{P(X \cap Y)}{P(X)}$

Association rule mining

- U is a set of d items
- T is a set of n transactions $T_1, T_2, \ldots, T_n$ 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

Support of an Itemset

Definitions

- **Support of itemset *I***, written *sup(I)*:

  the fraction of transactions in the database $T = \{T_1 \ldots T_n\}$ that contain *I* as a subset.

- **Frequent itemset mining with support minsup:**

  Given a set of transactions $T = \{T_1, \ldots, T_n\}$,

  where $T_i \subseteq U$, find all itemsets I_j such that $sup(I_j) \geq minsup$

Example

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

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

Exercise: compute support

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 and 3-itemset occurring in this database
- Indicate which are frequent itemsets if $minsup = 1/2$

Spreadsheet links: https://upfbarcelona.padlet.org/chato/hogch321o6pws1fd

Properties

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

Support monotonicity property:

$if\ J \subseteq I,\ sup(J) \geq sup(I)$ WHY?

Properties

- **Support monotonicity property**:

  $$if\ J \subseteq I,\ sup(J) \geq sup(I)$$

- 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 and Maximal Itemsets

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

Maximal itemset

An itemset is **maximal** if:

it is *closed* and

it is *frequent*

(*frequent* means: support ≥ 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

Exercise

Find three **maximal** frequent itemsets at minsup=0.4

Tip: first find all frequent itemsets at minsup=0.4

Maximal itemset

An itemset is **maximal** if:

it is closed and

it has support ≥ 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

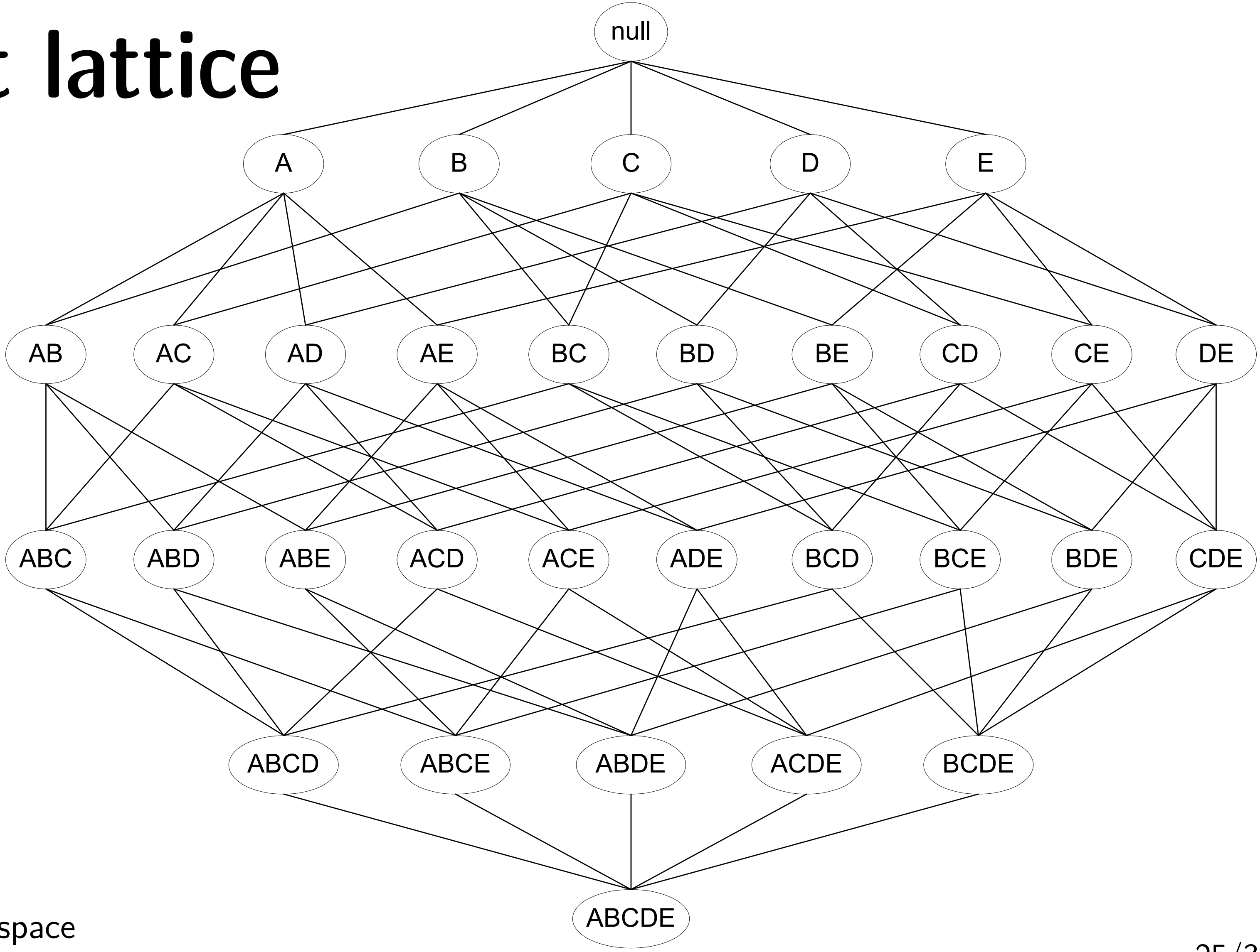
Example **maximal** itemsets:

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

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

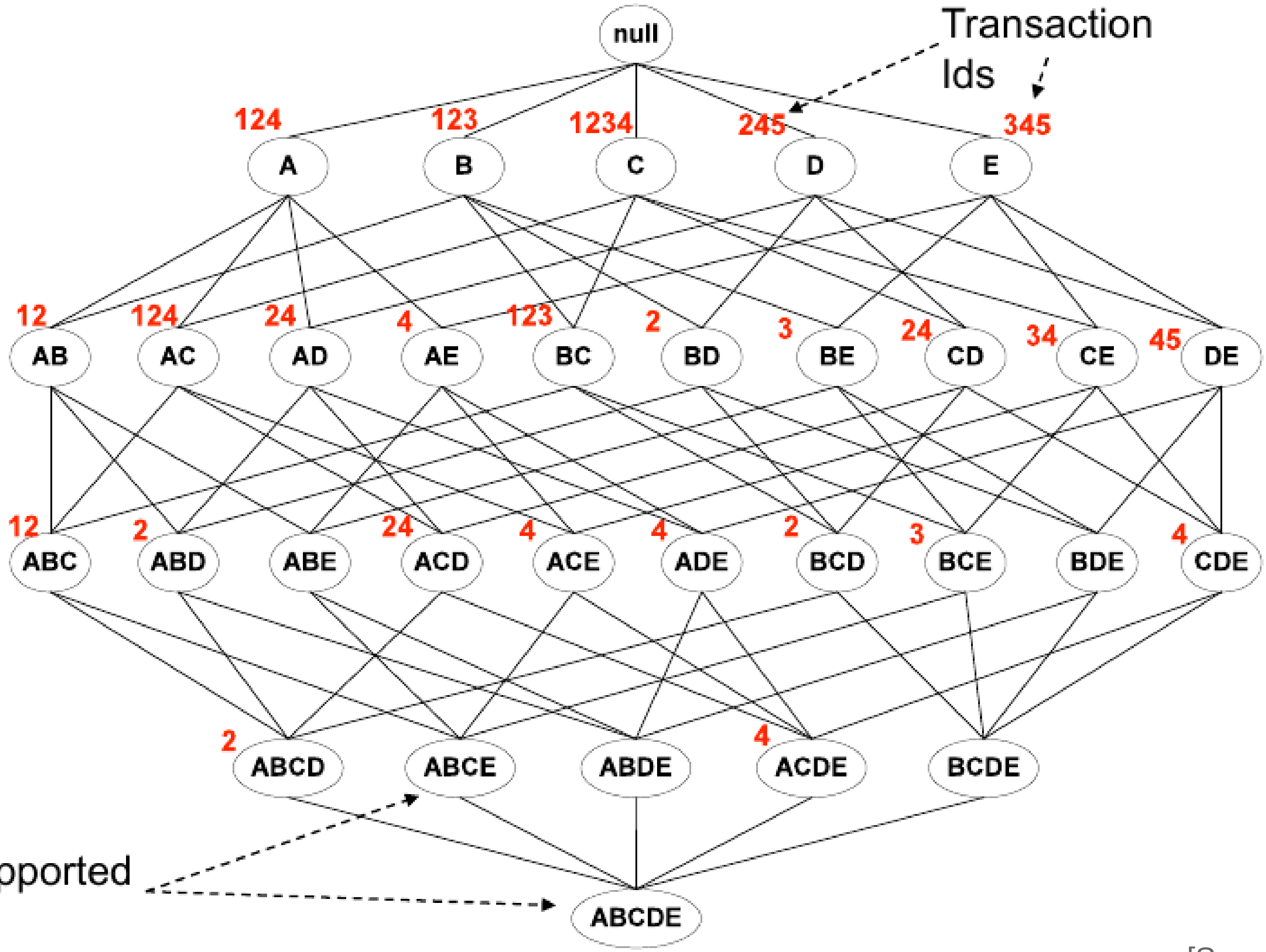
The Itemsets Lattice

The itemset lattice



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

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



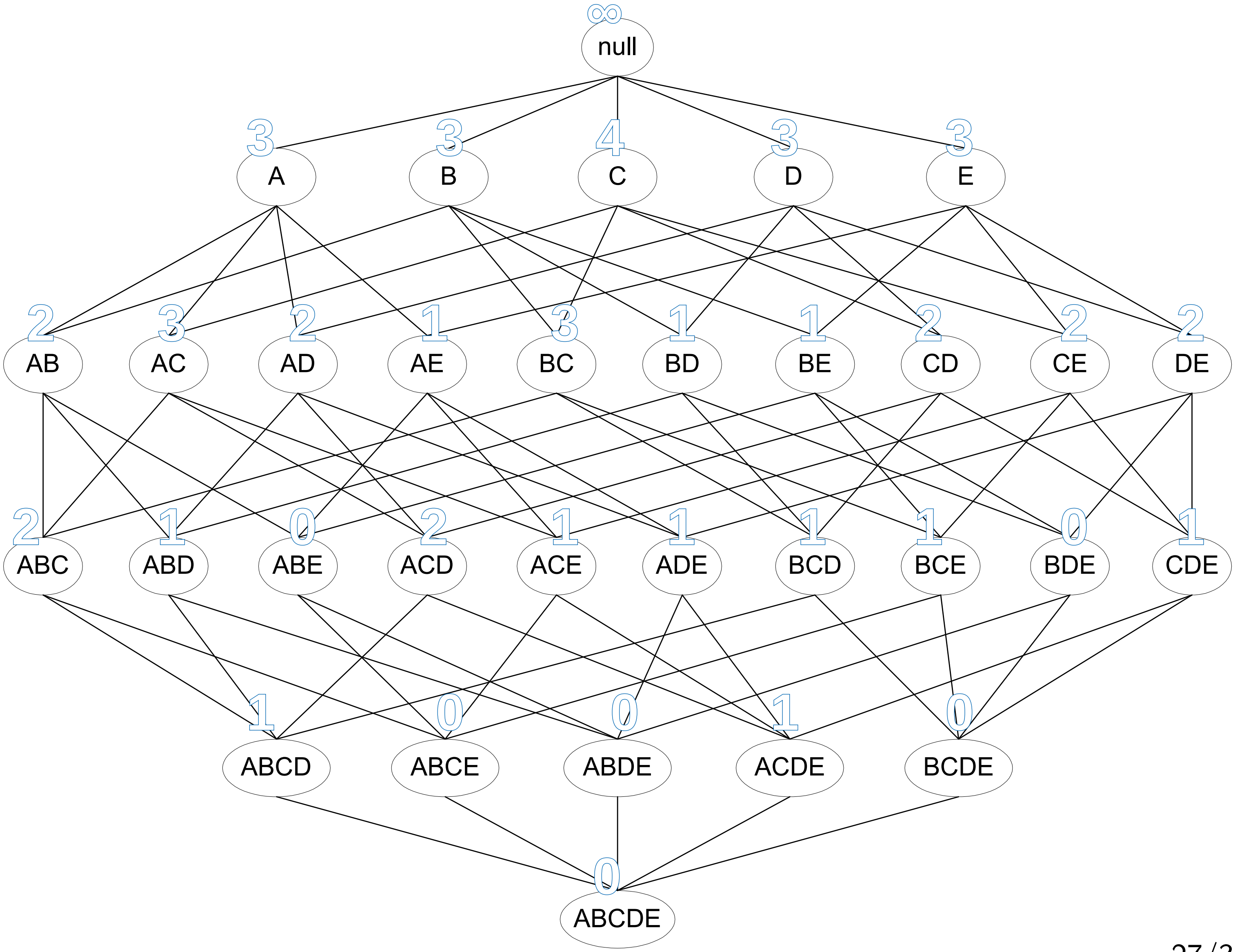
Transaction Ids

Not supported by any transactions

[Source]

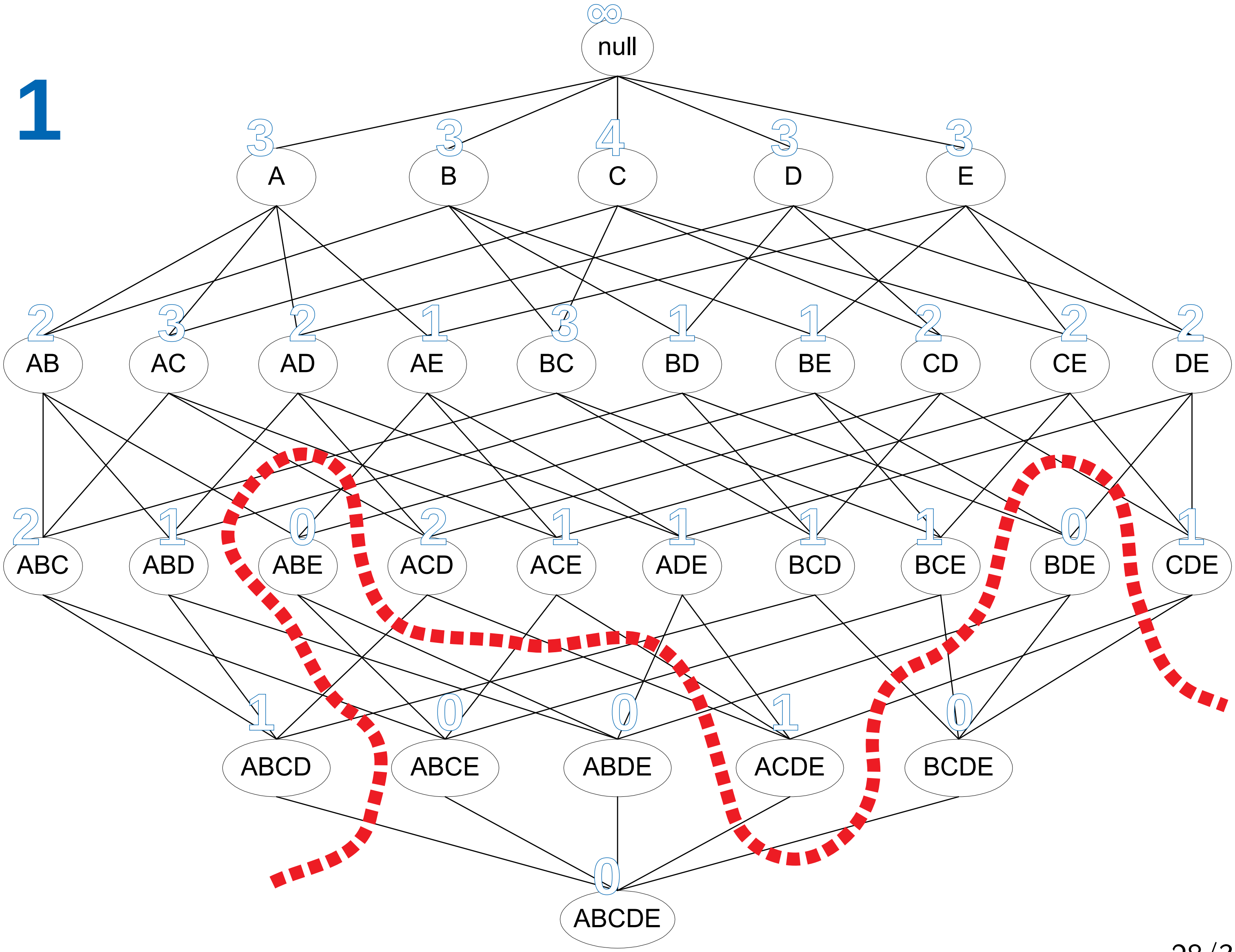
Support of each itemset

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



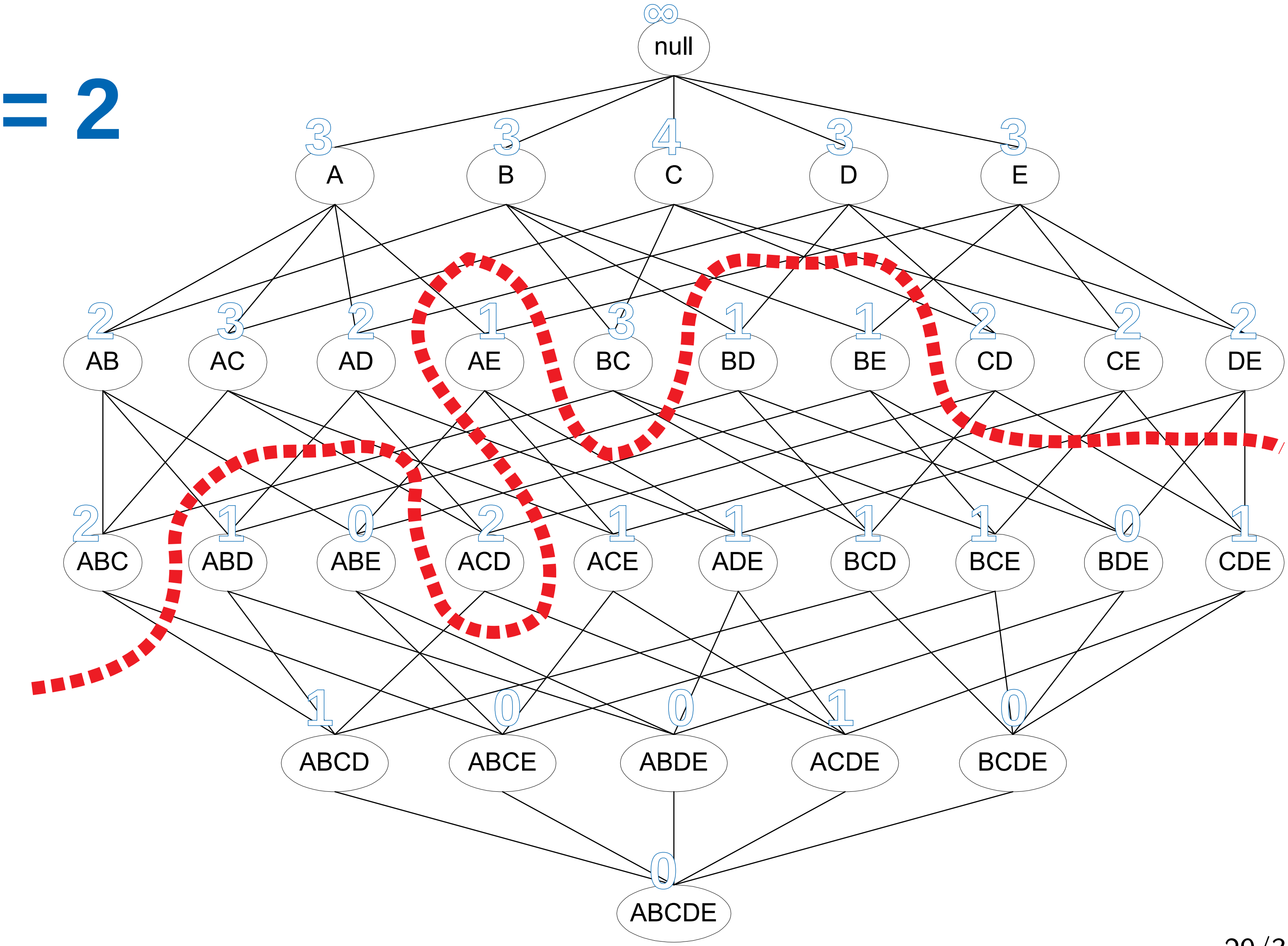
minsup = 1

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

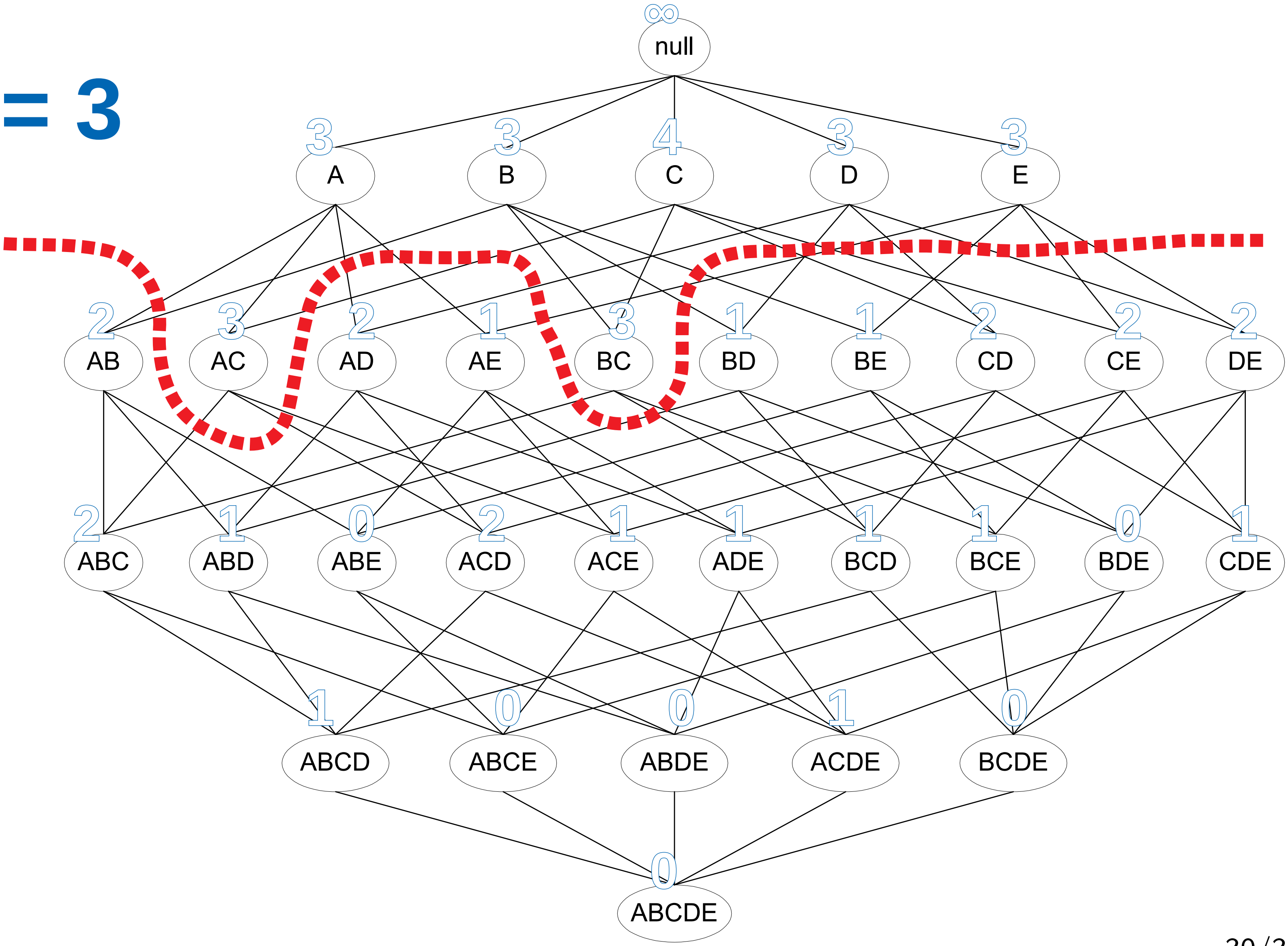


minsup = 2

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



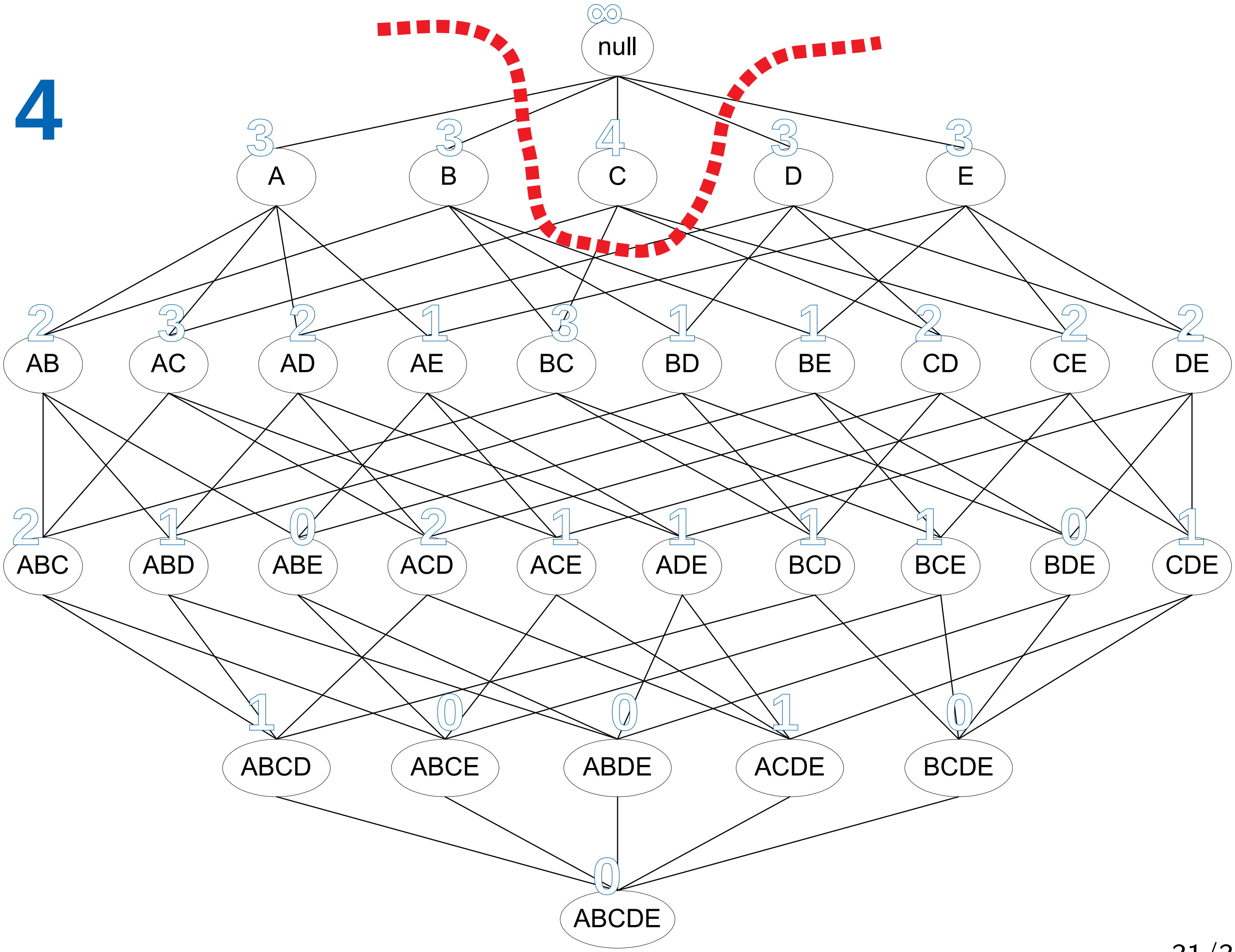
minsup = 3



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

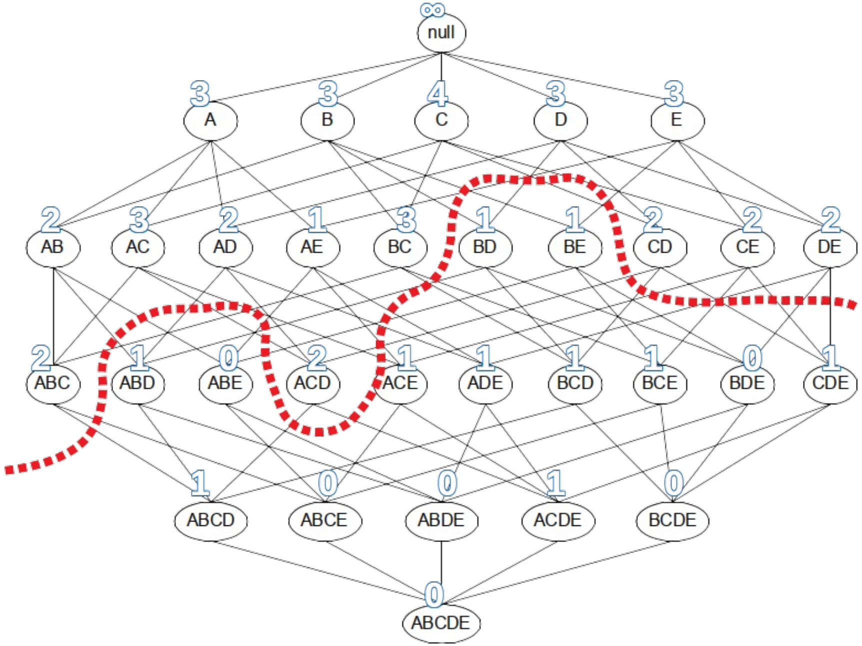
minsup = 4

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



Summary

Things to remember

- Itemset, k-itemset, transaction, support
- Support monotonicity property
- Maximal and closed itemsets
- Itemset lattice

Exercises for TT11-TT12

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