

FREQUENT PATTERN MINING

FREQUENT PATTERN MINING

- Frequent Pattern mining/ FP discovery
 - Objective: find items that occur frequently together in a database
 - Algorithmically difficult problem
- Association Rule Learning
 - From frequent patterns,
 - Identify statistically relevant associations

MARKET BASKET ANALYSIS

- Typical example: Market Basket Analysis
 - Database: people buying products
 - One reason why supermarkets/shops propose Loyalty programs
- If you buy tomatoes, onions and hamburger patties, you will probably buy hamburger breads
- Famous unexpected association:
 - Beers and Diapers
 - (Probably a legend...)



Association



MARKET BASKET ANALYSIS

- Usage of market basket analysis:
 - Put one object on sale, to favor selling the other ones
 - Sales on burger breads=>consumer buy tomatoes, onion and beef patty
 - Put products close/far away
 - Men buying diapers tempted to buy beers ? Put beers close to diapers
- Relevant in other contexts of course
 - Relation between medical condition and life habits
 - Smoking + cholesterol=>heart disease...
 - High pH + bacterial I=> mosquito development

DATASETS

- Type of data: list of itemsets
 - $1 = \{\text{milk, bread, fruit}\}$
 - $2 = \{\text{butter, eggs, fruit}\}$
 - $3 = \{\text{beer, diapers}\}$
 - $4 = \{\text{milk, bread, butter, eggs, fruit}\}$
 - $5 = \{\text{bread}\}$

| transaction ID | milk | bread | butter | beer | diapers | eggs | fruit |
|----------------|------|-------|--------|------|---------|------|-------|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

DEFINITIONS

- **Items:** $I = \{i_1, i_2, \dots, i_n\}$
 - Unique item (butter, milk, etc)
- **Database** $D = \{t_1, t_2, \dots, t_m\}$
 - Collection of transactions
 - $(t_i \subseteq I)$, arbitrary size
- **Itemset:** set of items of arbitrary size ($X \subseteq I$)

DEFINITIONS

- Absolute Support of itemset X in D :
 - Number of transactions containing X (i.e., $|\{t \in D/X \subseteq t\}|$)
- Relative support (or simply *Support*)
 - Fraction of transactions containing X
 - $\frac{\text{abs_support}(X)}{|D|}$
 - Estimation of $P(X)$
 - Probability for a random transaction to contain X
- **Frequent** itemset:
 - Itemset with support $\geq \text{min_supp}$

SUPPORT

- Support {Milk,bread} = 2/5
- Support {diapers,beer}= 1/5

| transaction ID | milk | bread | butter | beer | diapers | eggs | fruit |
|----------------|------|-------|--------|------|---------|------|-------|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

DEFINITIONS

- Association rule : rule of the form
 - $X \rightarrow Y$
 - $X \subseteq I, Y \subseteq I$
 - $X \cap Y = \emptyset$
 - If X is in a transaction, then Y too
- Support of $X \rightarrow Y$:
 - \Rightarrow Support of itemset $W = X \cup Y$
- For an association to be interesting, we further look at interest scores
 - Else, risk to find spurious associations

SCORES OF INTEREST

CONFIDENCE

- $\text{conf}(X \Rightarrow Y) = P(Y|X) = \frac{\text{supp}(X \cap Y)}{\text{supp}(X)} = \frac{\text{number of transactions containing } X \text{ and } Y}{\text{number of transactions containing } X}$
- Fraction of transactions containing X that also contains Y
 - An itemset/rule can be frequent because its elements are frequent
 - We want to know if Y is frequent when we have X
- Non-symmetric

| transaction ID | milk | bread | butter | beer | diapers | eggs | fruit |
|----------------|------|-------|--------|------|---------|------|-------|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

- Confidence $\text{Milk} \Rightarrow \text{bread} = 2/2 = 1$
- Confidence $\text{bread} \Rightarrow \text{milk} = 2/3$
- Confidence $\text{diapers} \Rightarrow \text{beer} = 1/1$
- Confidence $\text{beer} \Rightarrow \text{diapers} = 1/1$

LIFT

- If Y has high confidence, but is also frequent, confidence is not enough.

- $\text{lift}(X \Rightarrow Y) = \frac{\text{confidence}(X \Rightarrow Y)}{\text{supp}(Y)},$

- Compares Y presence when X with Y in general

- $\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cap Y)}{\text{supp}(X) \times \text{supp}(Y)}$

- Compares observed co-presence with expected co-presence

- $[0, +\infty]$

- X and Y are independent: $\text{lift} = 1$

| transaction ID | milk | bread | butter | beer | diapers | eggs | fruit |
|----------------|------|-------|--------|------|---------|------|-------|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

- Lift Milk=>bread
 - $(2/5)/(6/25)=1.666$
 - $(1)/(3/5)=1.666$
- Lift beer=>diapers
 - $(1/5)/(1/25)=5$
 - $(1)/(1/5)=5$

LEVERAGE

- $\text{leverage}(A \rightarrow C) = \text{support}(A \rightarrow C) - \text{support}(A) \times \text{support}(C)$, range: $[-1, 1]$
 - Difference between the observed frequency of A and C appearing together and the frequency that would be expected if A and C were independent
- 0 indicates independence

| transaction ID | milk | bread | butter | beer | diapers | eggs | fruit |
|----------------|------|-------|--------|------|---------|------|-------|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

- Leverage $\text{Milk} \Rightarrow \text{bread}$
 - $(2/5) - (6/25) = 0.16$
- Leverage $\text{beer} \Rightarrow \text{diapers}$
 - $(1/5) - (1/25) = 0.16$

SCORES

- Other scores exist:
 - Conviction
 - zhangs metric
 - ...
- https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/association_rules/

FREQUENT ITEMSET OBJECTIVE

- Objective: limit the number of rules found
 - Given a minimum support threshold min_sup
 - Given a minimum confidence threshold min_conf
 - Find all association rules with $support \geq min_sup$ and $confidence \geq min_conf$

FREQUENT ITEMSET EXTRACTION

NAIVE APPROACH

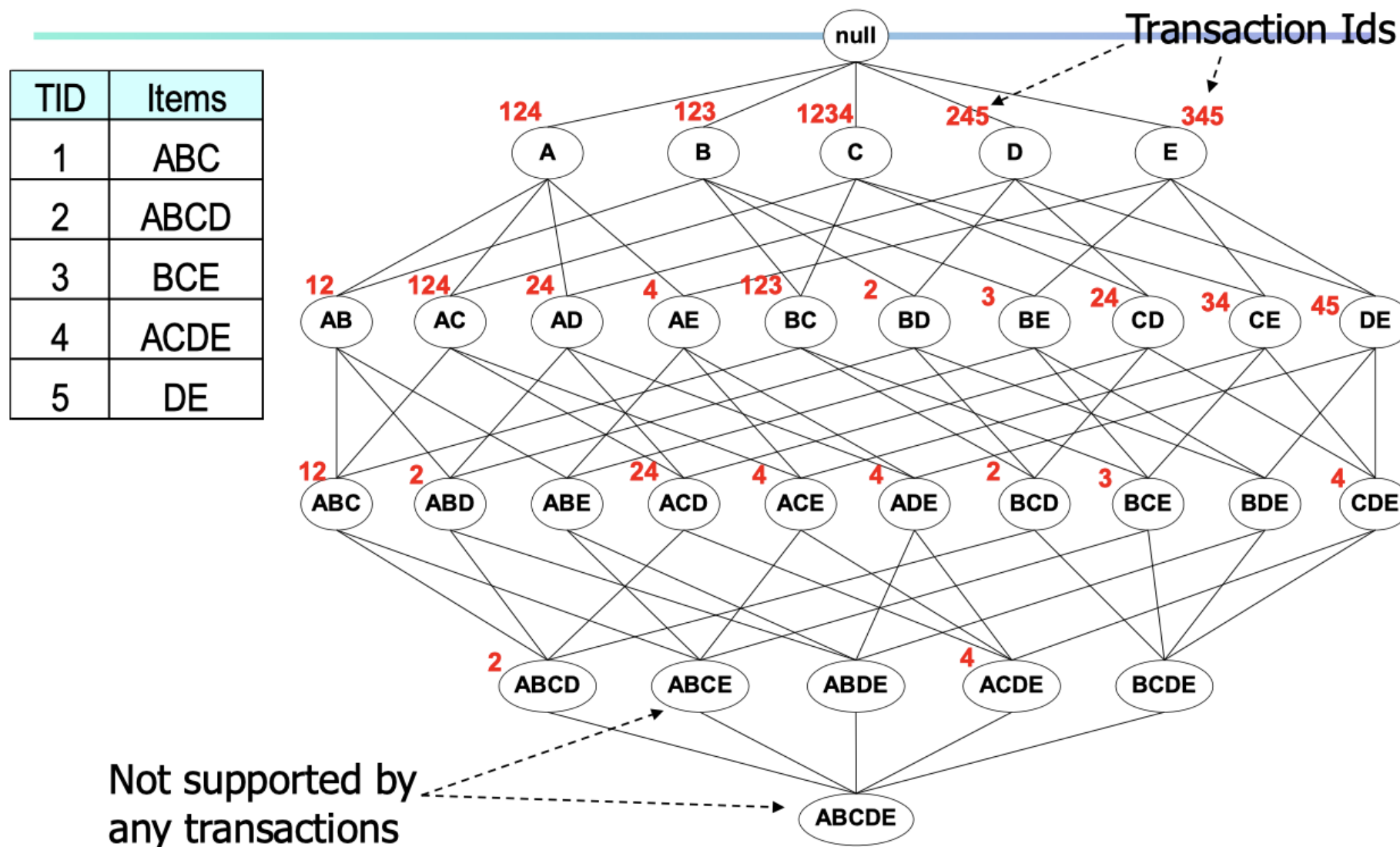
- Naive approach
 - 1) Generate all possible itemsets (size 1, 2, 3, 4 etc.)
 - 2) Compute their support from the database
- Problem: explosion of possible combinations
 - 1000 items
 - 1000 itemsets of size 1
 - $1000 \cdot 999 / 2$ itemsets of size 2
 - ...
 - 2^{100} combinations

SUPPORT PROPERTY

- Anti-monotonic property of support
 - If X_1 is frequent, then $X_2 \subset X_1$ is frequent
 - If X_1 is not frequent, then $X_2, X_1 \subset X_2$ is not frequent
- Computation trick:
 - 1) Find frequent 1-itemsets
 - 2) Find frequent 2-itemsets
 - Among those that contains only frequent 1-itemsets
 - 3) Repeat for all size (or until reaching a threshold)

SUPPORT PROPERTY

Maximal vs Closed Itemsets

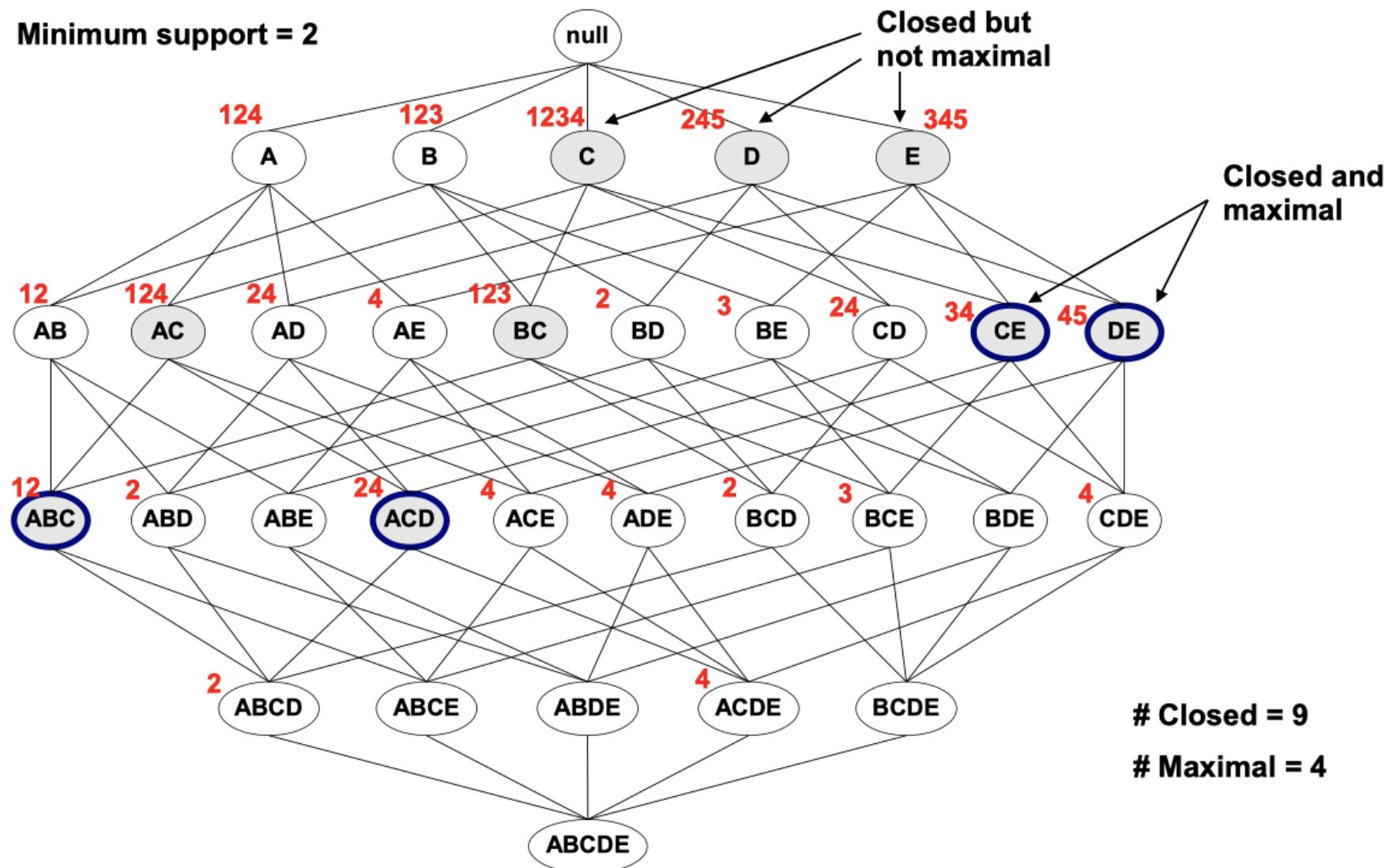


CLOSED AND MAXIMAL

- We define a **closed** pattern as a frequent pattern (support > threshold) with not sub-pattern of equal support
- We defined a **maximal** pattern as a frequent pattern that has no frequent sub-pattern

SUPPORT PROPERTY

Maximal vs Closed Frequent Itemsets



ALGORITHM: APRIORI

APRIORI

Apriori(T, ϵ)

$L_1 \leftarrow \{\text{frequent 1-itemsets}\}$

$k \leftarrow 2$

while L_{k-1} **is not** empty

$C_k \leftarrow \text{Apriori_gen}(L_{k-1}, k)$

for transactions t **in** T

$D_t \leftarrow \{c \text{ in } C_k : c \subseteq t\}$

for candidates c **in** D_t

$\text{count}[c] \leftarrow \text{count}[c] + 1$

$L_k \leftarrow \{c \text{ in } C_k : \text{count}[c] \geq \epsilon\}$

$k \leftarrow k + 1$

return $\text{Union}(L_k)$

Apriori_gen(L, k)

$\text{result} \leftarrow \text{list}()$

for all $p \in L, q \in L$ **where** $p_1 = q_1, p_2 = q_2, \dots, p_{k-2} = q_{k-2}$ **and** $p_{k-1} < q_{k-1}$

$c = p \cup \{q_{k-1}\}$

if $u \in L$ **for all** $u \subseteq c$ **where** $|u| = k-1$

$\text{result.add}(c)$

return result

GOING FURTHER

- Many other works in this domain
 - Sequential Pattern Mining: Take order into account
 - If we first buy a printer, then we will buy ink (and not the opposite)
 - Numeric target value: Find relevant intervals
 - If $\{a,b\} \Rightarrow z \in [12,25]$, if $\{a,c\} \Rightarrow z \in [25,32]$
 - Subgraph frequent itemsets
 - Spatial frequent itemsets
 - ...

FREQUENT PATTERN \Rightarrow GRAPHS

- Frequent patterns can represent another way to do data transformation:
 - Extract rules such as item1 \Rightarrow item2
 - Consider this information as an edge
 - Create a network out of it
 - Can be more relevant than a simple distance