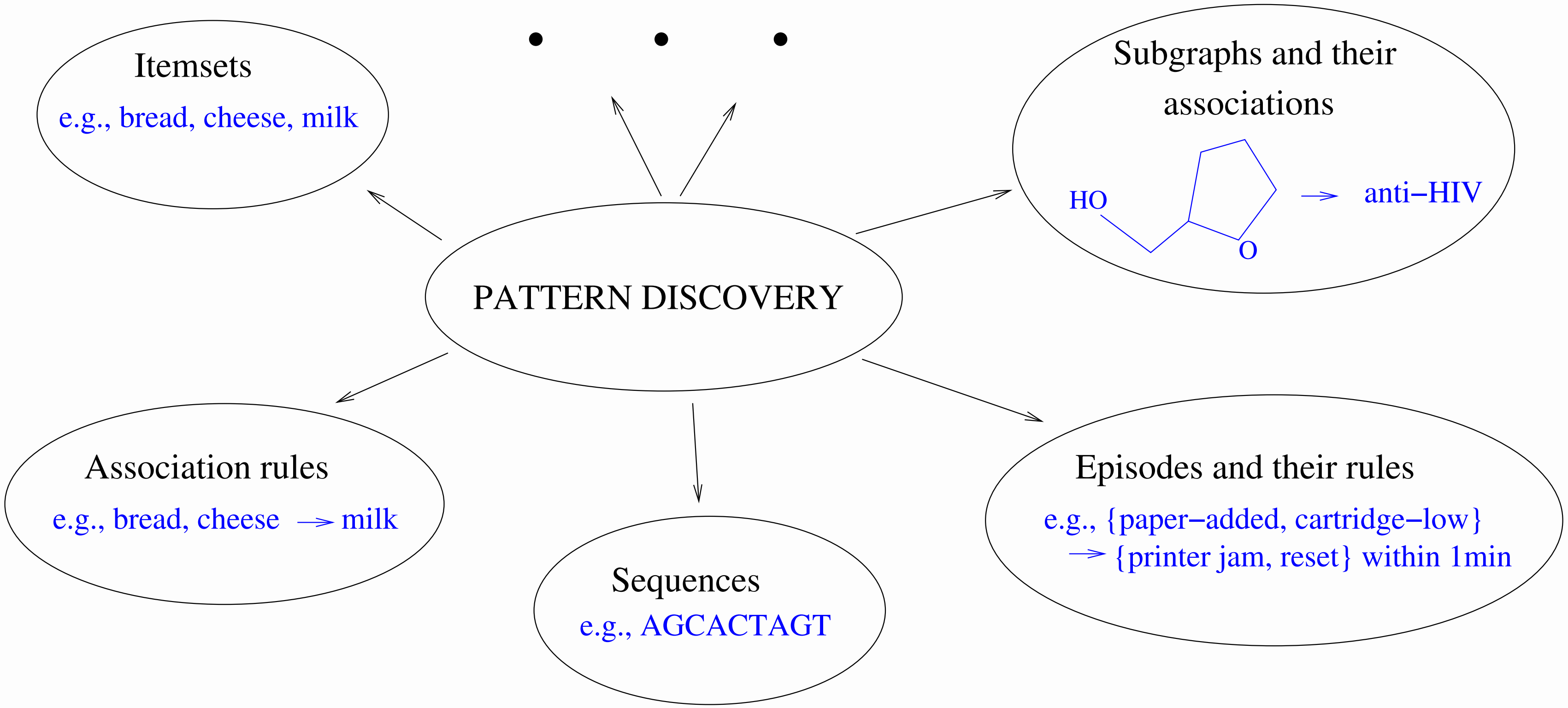
Mining association patterns (Part 1)

milk, cheese and bread
are often bought together

genes g1, g2, g3 and g4
are often over−expressed
in DLBC lymphomas

occurrence of certain insect species
makes it more likely to meet the
threatened white−backed woodpecker

Pattern discovery: search for all sufficiently good/top-K patterns of a certain type



Many variants of pattern types!

Contents for the next lectures

- Overview
  - pattern types, notions of association
- Frequent sets and association rules
  - Pruning the search space (monotonicity, Apriori)
  - Condensed representations
- Search for statistically significant association rules
  - Measures and algorithms, filtering redundant associations
- Advanced topics
  - Computational strategies, generic Apriori, etc.

1. Why associative pattern mining?

- **really efficient** way of discovering associations in large data sets!
  - sometimes globally optimal solutions to *NP*-hard optimization problems!
  - can handle even 20 000 attributes and millions of samples in a few minutes
  - for binary data – other data types should be binarized
- numerous applications!
- dependency analysis often a first step of data modelling ⇒ helps to choose methods (and features)
- as a subroutine of other methods
  - e.g., associative classifiers, clustering [a]

[a] e.g., Li & Zaiane 2017, Zimek et al. 2014, Zimmermann & Nijssen 2014

2. Data: typically occurrence data

	milk	juice	bread	cheese	oranges
basket1	1	0	1	1	0
basket2	1	1	0	0	1
basket3	0	1	1	1	0
basket4	1	0	1	1	1
basket5			⋮		

- items in market baskets
- species in ecological sites
- over-expressed or under-expressed genes in samples
- feature extraction for other data types

Data presented as **transactions** listing only 1-valued attributes: {milk, bread, cheese}, {milk, juice, oranges}, {juice, bread, cheese}, {milk, bread, cheese, oranges}, . . .

Data: Binarization and discretization

Categorical: Create a new binary attribute for each value

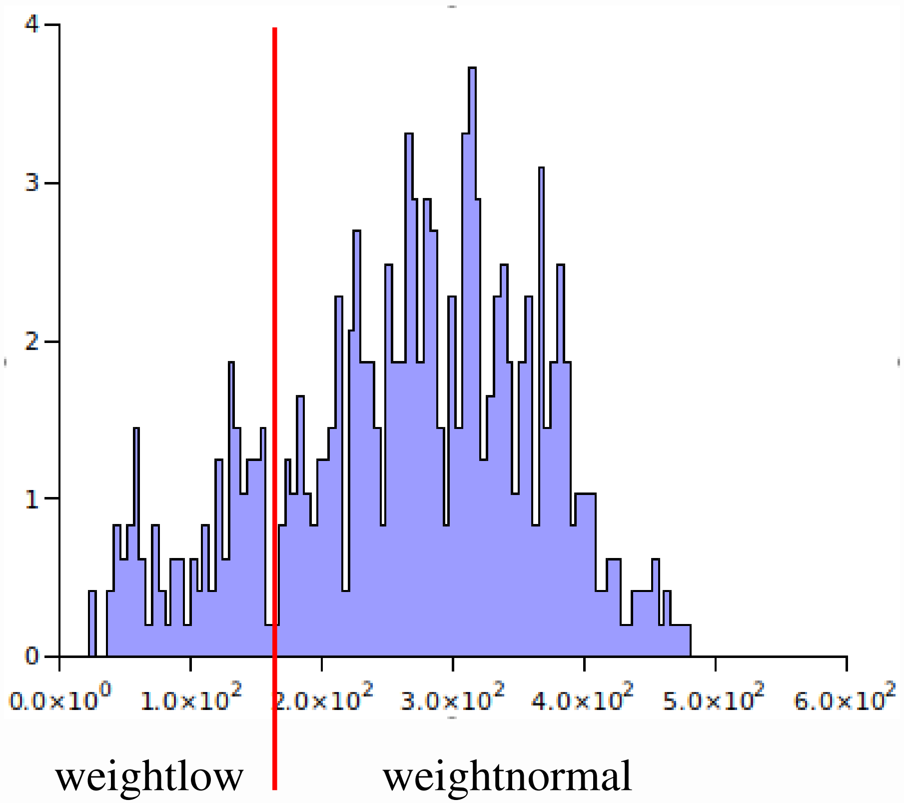
- *Colour*={red, blue, green} ⇒ attributes C_{red}, C_{blue}, C_{green}
- usually needed also for binary features!
  Gender={F, M} ⇒ attributes F, M
- no information loss!

Numerical: Discretize into bins and create a new attribute for each (interesting) bin

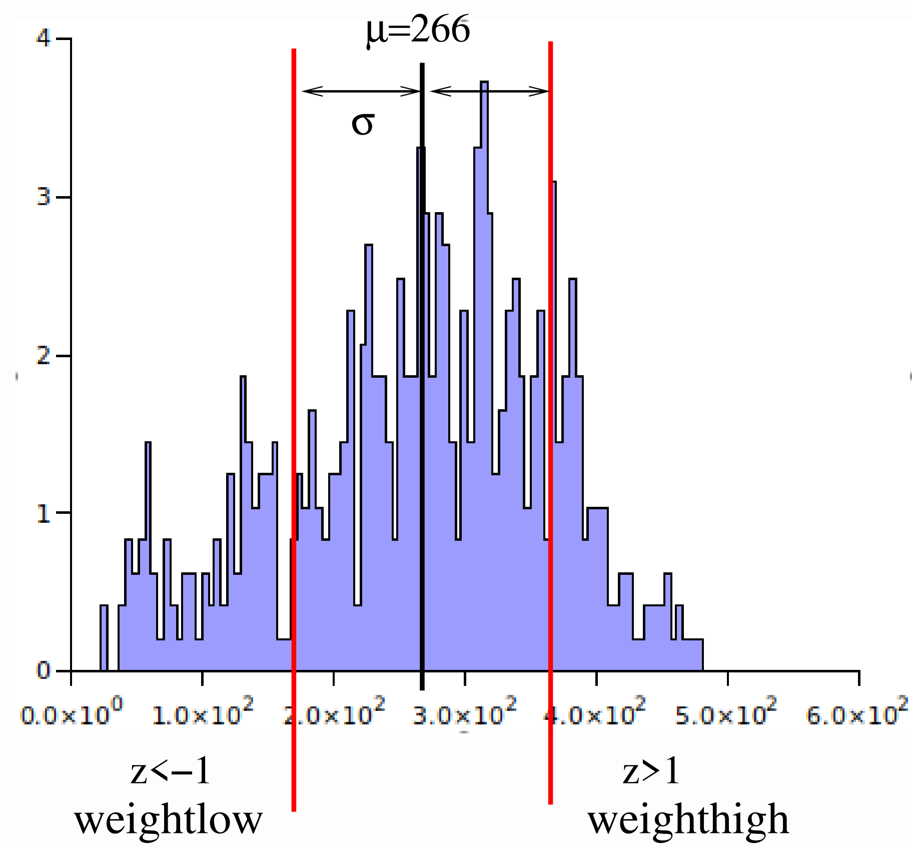
- Many approaches!
- static vs. dynamic, disjoint vs. overlapping, entire range vs. distribution tails
- loses some information, but also reduces noise

Example: 4 ways to discretize the rat's weight

1. One visually determined cut−off
entire range, two attributes



2. z−score discretization
attributes only for extreme values



3. Equi-width or 4. Equi-depth discretization of range $[20, 500]$

3. Types of association patterns

Given binary occurrence data
$\mathbf{R} = \{A_1, \ldots, A_k\}$ set of binary attributes ($Dom(A_i) = \{0, 1\}$)
$\mathcal{D} = \{d_1, \ldots, d_n\}$, $d_i \in Dom(A_1) \times \ldots \times Dom(A_k)$, data

Association patterns can be

1. **sets** $\bigwedge A_{i_j}{=}a_{i_j}$ ($A_{i_j} \in \mathbf{R}$, $a_{i_j} \in \{0, 1\}$) such that all $A_{i_j}{=}a_{i_j}$ are associated
   e.g., *milk=1, cheese=1, bread=1* (these items occur often together in a market basket), or
2. **rules** $\bigwedge A_{i_j}{=}a_{i_j} \rightarrow C{=}c$, $A_{i_j}, C \in \mathbf{R}$, $a_{i_j}, c \in \{0, 1\}$, such that rule condition and consequence are associated
   e.g., *cheese=1 → bread=1* (people who buy cheese tend to buy bread, too)

Simplification

Usually we are interested in patterns containing only 1-valued attributes $\Rightarrow$ Simplified notations:

- sets $\mathbf{X} \subseteq \mathbf{R}$ (simply list elements of set)
  e.g., *milk, cheese, bread*
- rules $\mathbf{X} \to C$, where $A_i = 1$ for all $A_i \in \mathbf{X}$
  e.g., *cheese* $\to$ *bread*

Notes:

- if needed, you can create new attributes A_{neg} for $\neg A$
- we concentrate on single attribute consequences (could be a set, too)

What association means??

Statistics

- **= statistical dependence** between categorical variables
- defined by the opposite, statistical independence
- measures for the strength of association
- statistical significance: is the observed association spurious?

Frequent pattern mining

- **= frequent co-occurrence** of attributes (given some minimum frequency)
- extra criteria to filter interesting patterns
- statistical measures and tests can also be applied (post-processing)

Statistical independence and dependence

- **Events** $A{=}a$ and $B{=}b$ are **statistically independent**, if $P(A{=}a, B{=}b) = P(A{=}a)P(B{=}b)$
- **Variables** A and B are **statistically independent**, if for all $a \in Dom(A)$, $b \in Dom(B)$, $P(A{=}a, B{=}b) = P(A{=}a)P(B{=}b)$
- If A and B binary, these conditions are equivalent!
- **Leverage** δ and **lift** γ measure the strength of dependence [a]

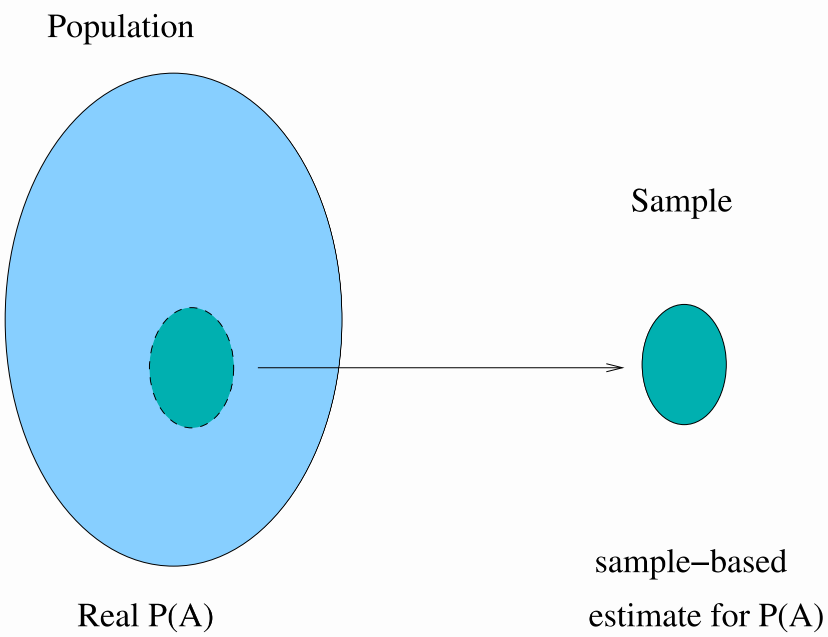
$$\delta(A{=}a, B{=}b) = P(A{=}a, B{=}b) - P(A{=}a)P(B{=}b)$$

$$\gamma(A{=}a, B{=}b) = \frac{P(A{=}a, B{=}b)}{P(A{=}a)P(B{=}b)}$$

[a] Also notations $\delta(\mathbf{X} \rightarrow C)$, $\gamma(\mathbf{X} \rightarrow C)$

Note: Simplified notation $P = \tilde{P}$

- in the definition of statistical independence, P refers to real (but unknown) probability
- its maximum likelihood estimate is **relative frequency** in data, $\tilde{P}(A) = \frac{fr(A)}{n}$, where $fr(A)$=number of rows containing A, n=number of rows in data
- Here we use $P = \tilde{P}$



Statistical (in-)dependence in rules $\mathbf{X} \rightarrow C{=}c$

$\mathbf{X}$ lists true-valued attributes, e.g., $\mathbf{X} = A, B, D$.
Notate $C{=}1$ by C and $C{=}0$ by $\neg C$.

- $\mathbf{X}$ and C are **independent**, if $P(\mathbf{X}, C) = P(\mathbf{X})P(C)$ ($\delta = 0$, $\gamma = 1$) ("independence rule")
- $\mathbf{X}$ and C are **positively associated**, if $P(\mathbf{X}, C) > P(\mathbf{X})P(C)$ ($\delta > 0$, $\gamma > 1$) (rule $\mathbf{X} \rightarrow C$)
- $\mathbf{X}$ and C are **negatively associated**, if $P(\mathbf{X}, C) < P(\mathbf{X})P(C)$ ($\delta < 0$, $\gamma < 1$). Now $\mathbf{X}$ and $\neg C$ positively associated! (rule $\mathbf{X} \rightarrow \neg C$ [a])

---

[a] Customary to present positive associations, if both C and $\neg C$ are allowed

Note on confidence (precision)

Strength of frequent association rules is often measured with “confidence” (precision) ϕ (*cf*) + required $\phi \geq \min_{cf}$

$$\phi(\mathbf{X} \rightarrow C) = P(C|\mathbf{X}) = \frac{P(\mathbf{X}C)}{P(\mathbf{X})}$$

But high ϕ does not guarantee statistical dependence!

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

(image from Tan et al. 2018)

e.g., $\min_{cf} = 0.70$
$\phi(\text{tea} \rightarrow \text{coffee}) = 0.75$
$P(\text{coffee}) = 0.90$
$\gamma(\text{tea} \rightarrow \text{coffee}) = 0.75/0.90$
$= 0.83 < 1$ **neg. association**

Statistical independence in sets $\mathbf{X}$

Let $\mathbf{X} = A_1, \ldots, A_m$ and $a_i \in \{0, 1\}$.

Variables $A_i \in \mathbf{X}$ are **mutually independent**, if for all $a_i \in \{0, 1\}$

$$P\left(\bigwedge_{A_i \in \mathbf{X}} A_i{=}a_i\right) = \prod_{A_i \in \mathbf{X}} P(A_i{=}a_i)$$

Equivalently, $\mathbf{X}$ is an "independence set", if for all $\mathbf{Y} \subseteq \mathbf{X}$

$$P(\mathbf{Y}) = \prod_{A_i \in \mathbf{Y}} P(A_i{=}1)$$

Note

It is possible that $P(\mathbf{X}) = \prod_{A_i \in \mathbf{X}} P(A_i{=}1)$, even if $\mathbf{X}$ is not an independence set!

E.g., *bread, cheese, juice*:
$P(B, C, J) = 0.06 = 0.4 \cdot 0.3 \cdot 0.5 = P(B)P(C)P(J)$, but
$P(B, C) = 0.25 > 0.4 \cdot 0.3$ (positive association)
$P(J|BC) = 0.24 < P(J)$ (negative association)

$\Rightarrow$ less misleading to report only *bread, cheese*?

Set dependencies

Many choices for dependency sets (="correlated" sets):

- **Minimal** sets $\mathbf{X}$ that express positive dependence between true-valued attributes: $P(\mathbf{X}) > \prod_{A_i \in \mathbf{X}} P(A_i{=}1)$
- Sets $\mathbf{X}$ that express bipartition dependence: for **some** $\mathbf{Q} \subsetneq \mathbf{X}$: $P(\mathbf{X}) > P(\mathbf{X} \setminus \mathbf{Q})P(\mathbf{Q})$

  i.e., association $\mathbf{X} \setminus \mathbf{Q} \rightarrow \mathbf{Q}$
- Self-sufficient sets [a] $\mathbf{X}$ where for **all** $\mathbf{Q} \subsetneq \mathbf{X}$, $\mathbf{Q} \neq \emptyset$: $P(\mathbf{X}) > P(\mathbf{X} \setminus \mathbf{Q})P(\mathbf{Q})$ (+ some extra criteria)

---

[a] Webb and Vreeken, 2014

Extra: Extensions of leverage and lift to sets

- Leverage

$$\delta_S(\mathbf{X}) = P(\mathbf{X}) - \prod_{A_i \in \mathbf{X}} P(A_i)$$

- Lift

$$\gamma_S(\mathbf{X}) = \frac{P(\mathbf{X})}{\prod_{A_i \in \mathbf{X}} P(A_i)}$$

Problem: $\mathbf{Y} \subsetneq \mathbf{X}$ may express association (i.e., $\mathbf{X}$ is not independence set), even if $\delta_S(\mathbf{X}) = 0$ and $\gamma_S(\mathbf{X}) = 1$!

4. Search problems

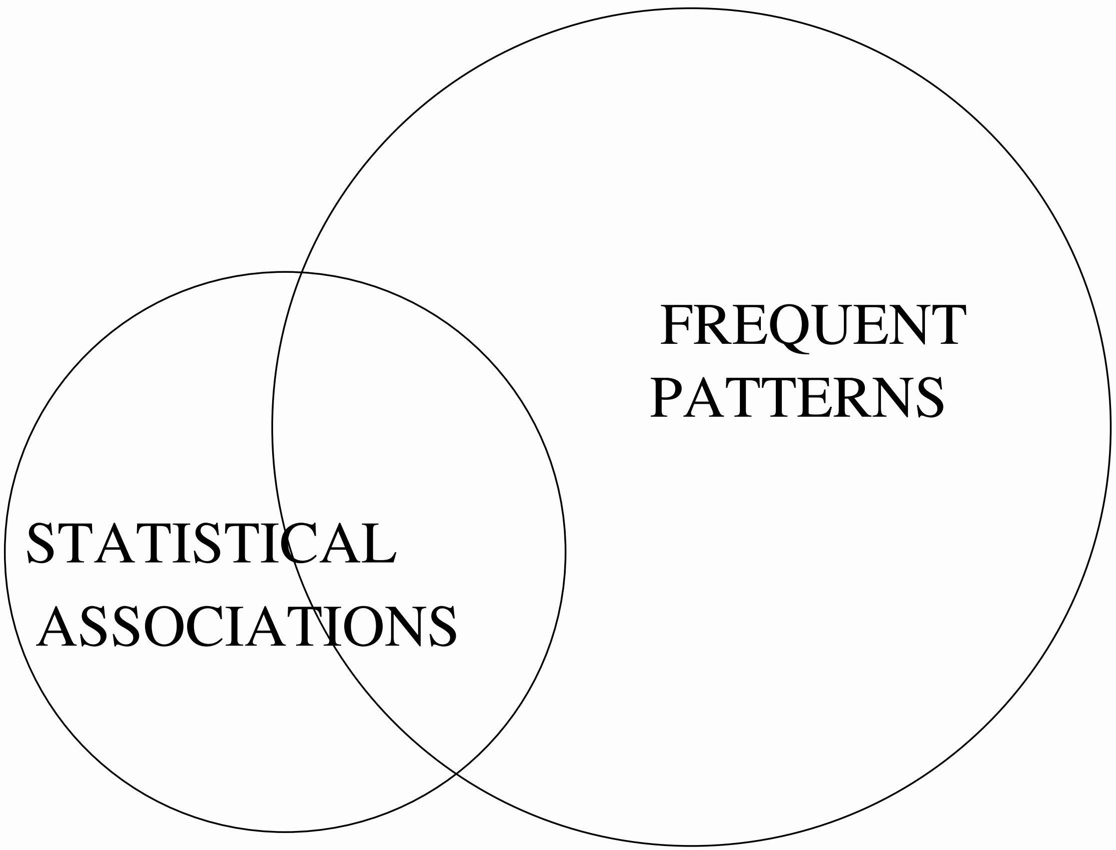
1. **Enumeration problem**: Search **all** patterns whose measure values are **sufficiently good** (given thresholds), e.g.,
   - all frequent sets given minimum frequency $\min_{fr}$
   - all frequent and confident rules, given $\min_{fr}$ and $\min_{cf}$
   - all rules whose mutual information is at least $\min_{MI}$
2. **Optimization problem**: Find the **best** K patterns with a given goodness measure (+ possible extra constraints)
   - top-100 frequent rules with largest lift (given $\min_{fr}$)
   - top-100 rules with mutual information

5. Main approaches

1. **Frequency-based search**
   i) search frequent sets
   ii) construct rules from sets (if wanted)
   iii) filter with statistical measures
   iv) possibly test statistical significance
2. **Direct search of statistical associations** with measures for the strength and/or statistical significance of association (for rules or sets)

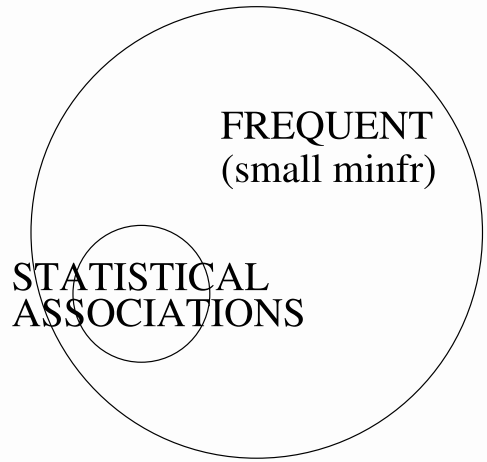
Approach 1 is easier to implement, but often fails to find everything we want (and finds a lot that we don't want!)

Two approaches find generally different patterns!

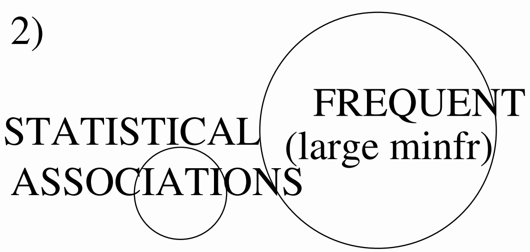


TWO EXTREMES:

1)



2)



If sufficiently small min_{fr} is feasible, you can filter statistical associations afterwards (but this can be heavy!)

Empirical comparison: frequent vs. statistical associations

Proportions of top-100 statistically significant rules with χ^2 that could be found with Apriori, when $\min_{fr}$ as small as computationally possible (given 256GB memory!):

data	n	k	tlen	$\min_{fr}$	discovered %
Mushroom	8124	119	23.0	0.01	100%
Chess	3196	75	37.0	0.20	1%
T10I4D100K	100000	870	10.1	0.00007	100%
T40I10D100K	100000	942	39.6	0.01	0%
Accidents	340183	468	33.8	0.25	14%
Pumsb	49046	2113	74.0	0.45	27%
Retail	88162	16470	10.3	0.000085	100%

Lesson to learn

Check always what is the **definition of association** of a given algorithm!

- What kind of patterns do you find precisely?
- Do you find what you want?
- If not, consider alternatives or try to tailor the algorithm to find your target patterns

6. Algorithm for frequent association mining

Let $\mathbf{R} = \{A_1, \ldots, A_k\}$ binary attributes.
Focus: How to find **frequent sets** i.e., $\mathbf{X} \subseteq \mathbf{R}$ such that $P(\mathbf{X}) \geq min_{fr}$?

- rules are easy to derive: check $\mathbf{X} \setminus \{C\} \rightarrow C$ for all $C \in \mathbf{X}$
- postprocessing may still take time...

Contents:

6.1 Pruning the search space

Problem: Search space has **exponential size!**

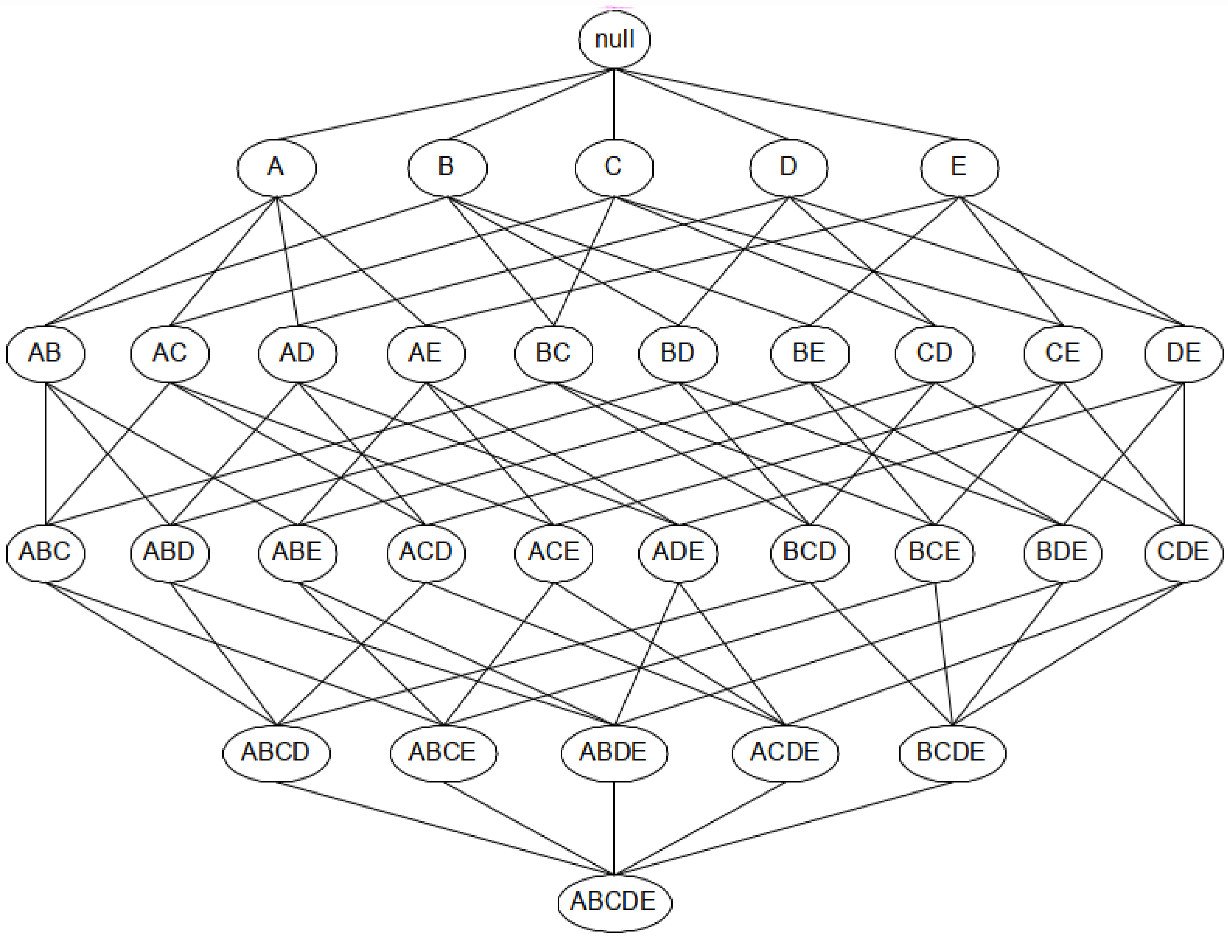
Given $\mathbf{R} = \{A_1, \ldots, A_k\}$, there are

- $\sum_{i=1}^{k} \binom{k}{i} = 2^k - 1$ non-empty sets $\mathbf{X} \subseteq \mathbf{R}$
- $\sum_{i=2}^{k} i \binom{k}{i} = \sum_{i=2}^{k} \frac{i \cdot k!}{i!(k-i)!} = \sum_{i=2}^{k} \frac{k \cdot (k-1)!}{(i-1)!(k-i)!} = k(2^{k-1} - 1)$ possible rules $\mathbf{X} \setminus \{C\} \rightarrow C$

(e.g., $2^{20} \approx 10^6$, $2^{100} \approx 10^{30}$)

How to find frequent ones?

Search space as a grid for $\mathbf{R} = \{A, B, C, D, E\}$



(image from Tan et al. 2018)

Key idea: Monotonicity of frequency

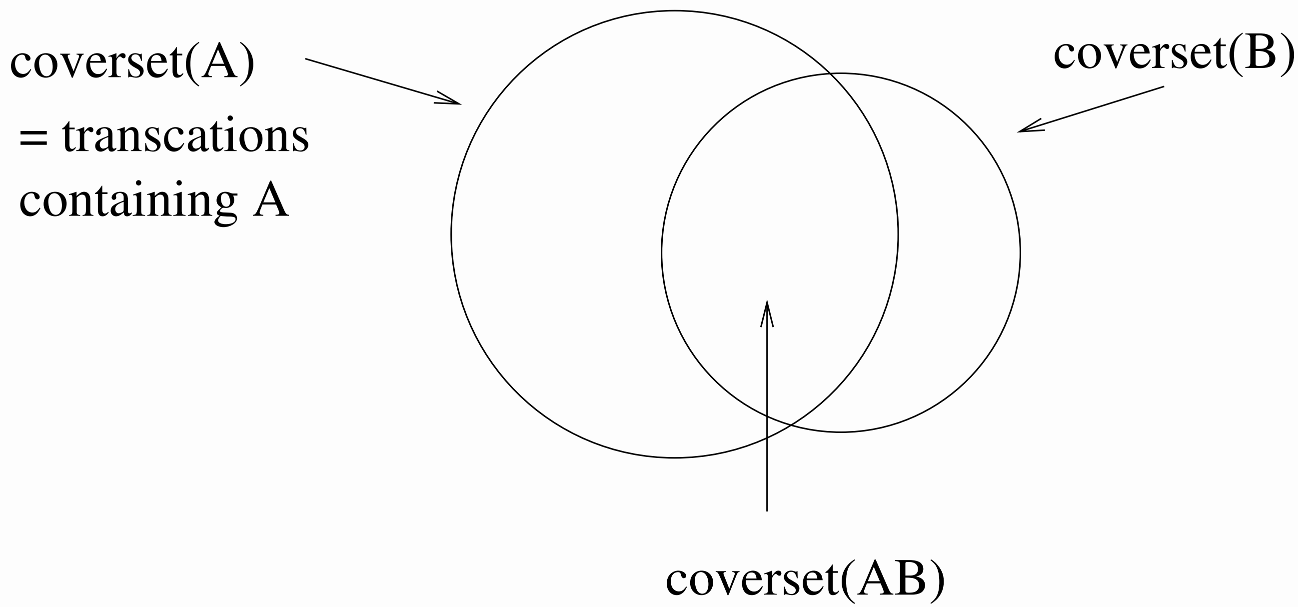
$fr(\mathbf{X})$= absolute frequency of $\mathbf{X}$
$P(\mathbf{X})$= relative frequency of $\mathbf{X}$
(both can be called “support”)

Frequency is **monotone** property: For all $\mathbf{X}$, $\mathbf{Y}$:

$$\mathbf{Y} \subseteq \mathbf{X} \Rightarrow fr(\mathbf{Y}) \geq fr(\mathbf{X})$$

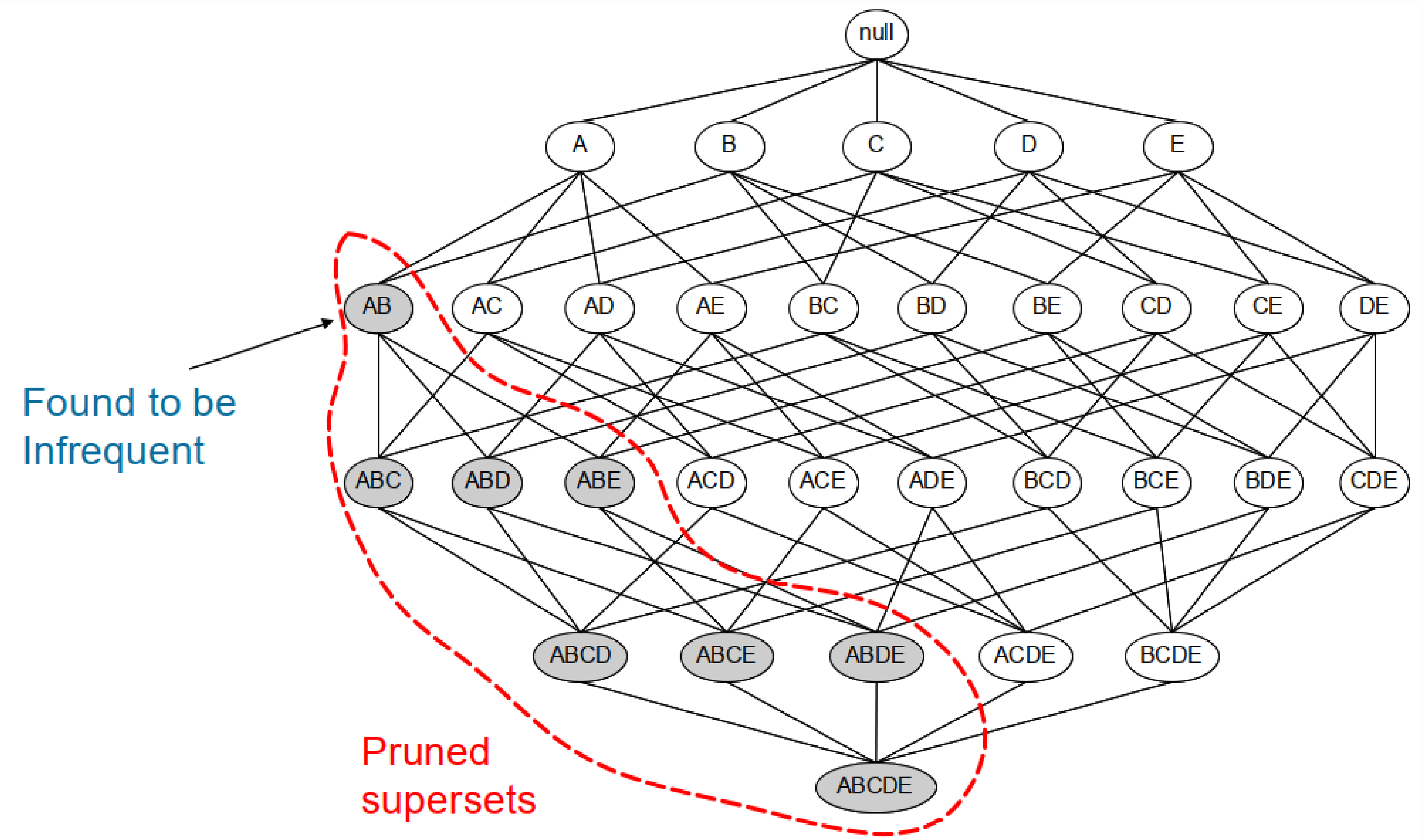
Key idea: Monotonicity of frequency

E.g., $\mathbf{X} = \{A, B\}$ and $\mathbf{Y} = \{A\}$. $fr(A) \geq fr(AB)$:



Consequence: If $\mathbf{Y}$ is infrequent ($P(\mathbf{Y}) < min_{fr}$), then all $\mathbf{X} \supsetneq \mathbf{Y}$ are infrequent ($P(\mathbf{X}) < min_{fr}$)

Pruning by monotonicity



(image from Tan et al. 2018)

6.2 Apriori algorithm (given $\mathbf{R}$, $\mathcal{D}$ and min_{fr})

$\mathcal{F}_i$ = frequent i-itemsets, C_i = candidate i-itemsets

i=1
$\mathcal{F}_1 = \{A_i \in \mathbf{R} \mid P(A_i) \geq min_{fr}\}$
while $\mathcal{F}_i \neq \emptyset$:

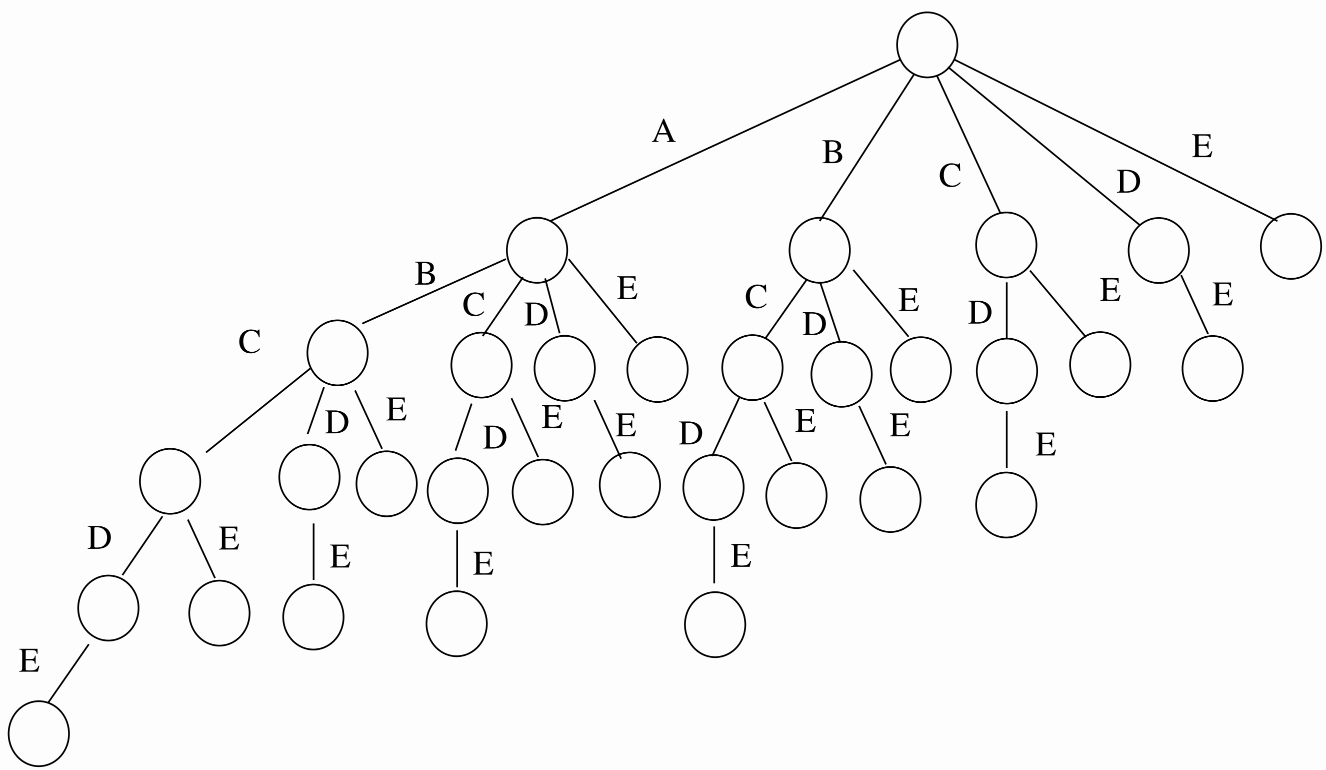
$$\left.\begin{array}{l} \bullet \text{ Generate candidates } C_{i+1} \text{ from } \mathcal{F}_i \\ \bullet \text{ Prune } \mathbf{X} \in C_{i+1} \text{ if } \exists \mathbf{Y} \subsetneq \mathbf{X}, |\mathbf{Y}| = i, \mathbf{Y} \notin \mathcal{F}_i \end{array}\right\} \text{(monotonicity)}$$

- Count frequencies $fr(\mathbf{X})$, $\mathbf{X} \in C_{i+1}$
- Set $\mathcal{F}_{i+1} = \{\mathbf{X} \in C_{i+1} \mid P(\mathbf{X}) \geq min_{fr}\}$
- $i = i + 1$

Return $\cup_i \mathcal{F}_i$

Useful data structure: enumeration tree

- Idea: each root–node path corresponds an itemset
- A complete tree has 2^k nodes ($k = |R|$) $\Rightarrow$ construct only as much as you need!

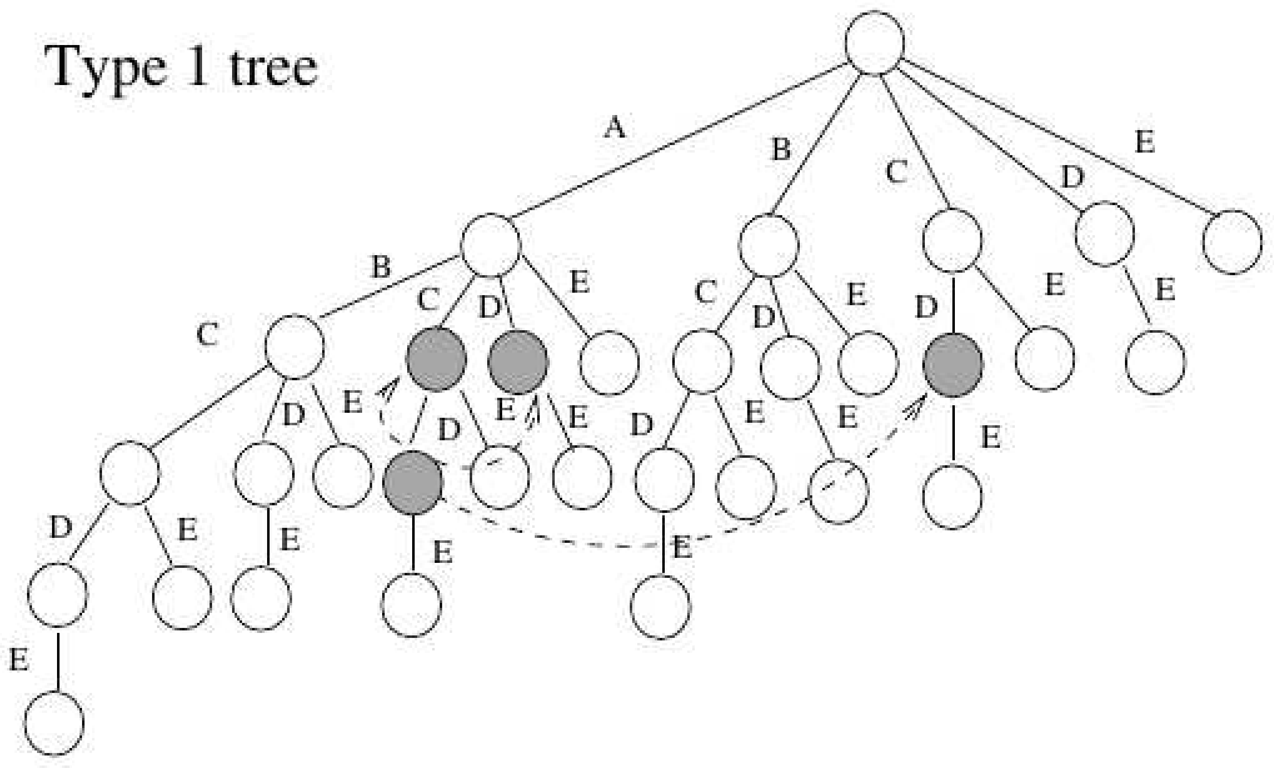


Terminology: a node may have many "parents"

parent of node $\mathbf{X}$ = node presenting $\mathbf{Y} \subsetneq \mathbf{X}$, $|\mathbf{X}| = |\mathbf{Y}| + 1$

Monotonicity: if any parent is infrequent, the child is infrequent

e.g., ACD has 3 parents:



Simulating Apriori

$\mathbf{R} = \{A, B, C, D, E\}$, $n = 6$, $min_{fr} = 2/6 = 0.33$

Transactions:

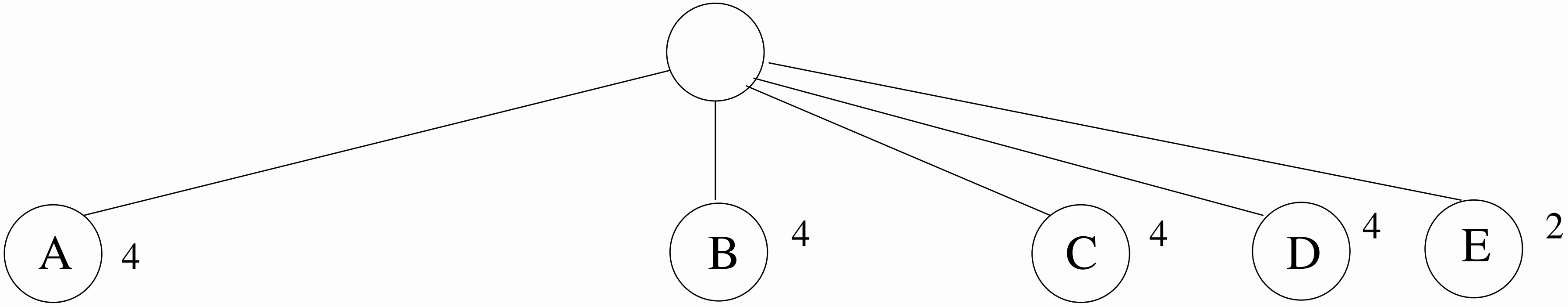
A, C, D
A, B, E
B, C, D
A, C, D, E
B, C
A, B, D

$fr(A) = fr(B) = fr(C) = fr(D) = 4$, $fr(E) = 2$

A=milk, B=juice, C=bread, D=cheese, E=oranges

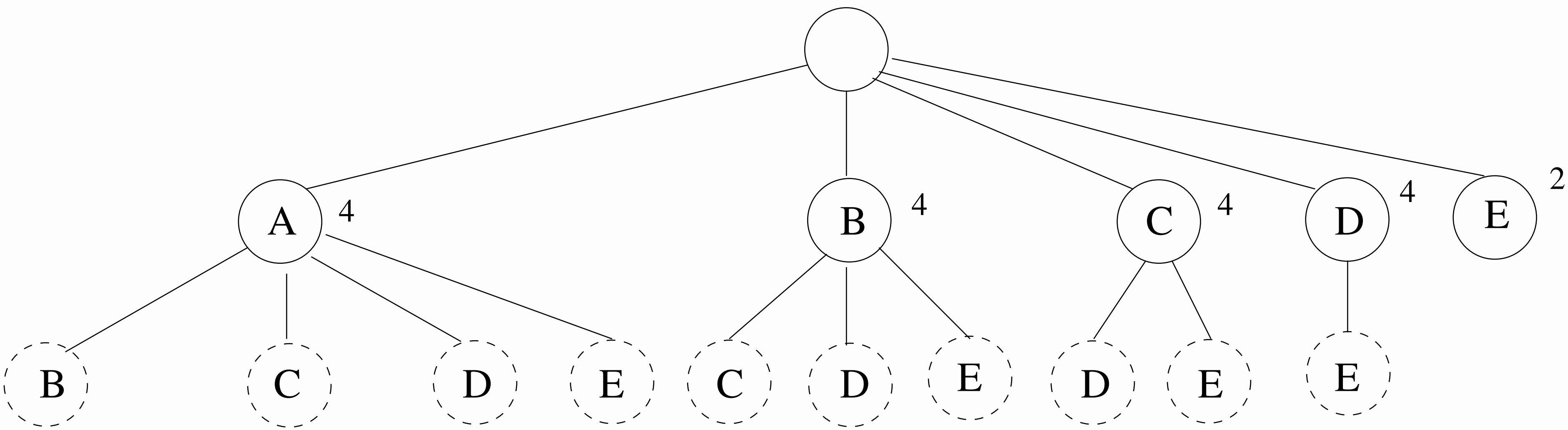
Simulation: Level $i = 1$

Check frequencies of all 1-sets and add to tree if frequent



All 1–sets frequent

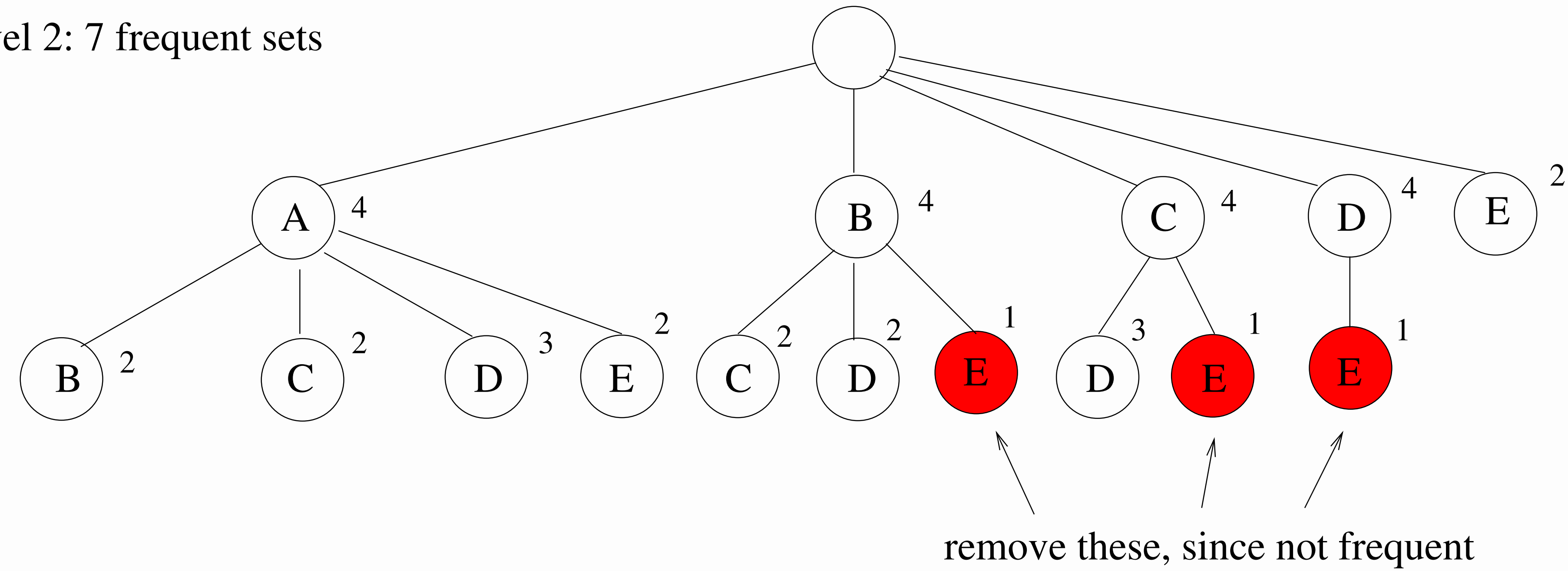
Simulation: $i = 2$ Candidate generation



Since all 1−sets frequent, check all possible 2−sets

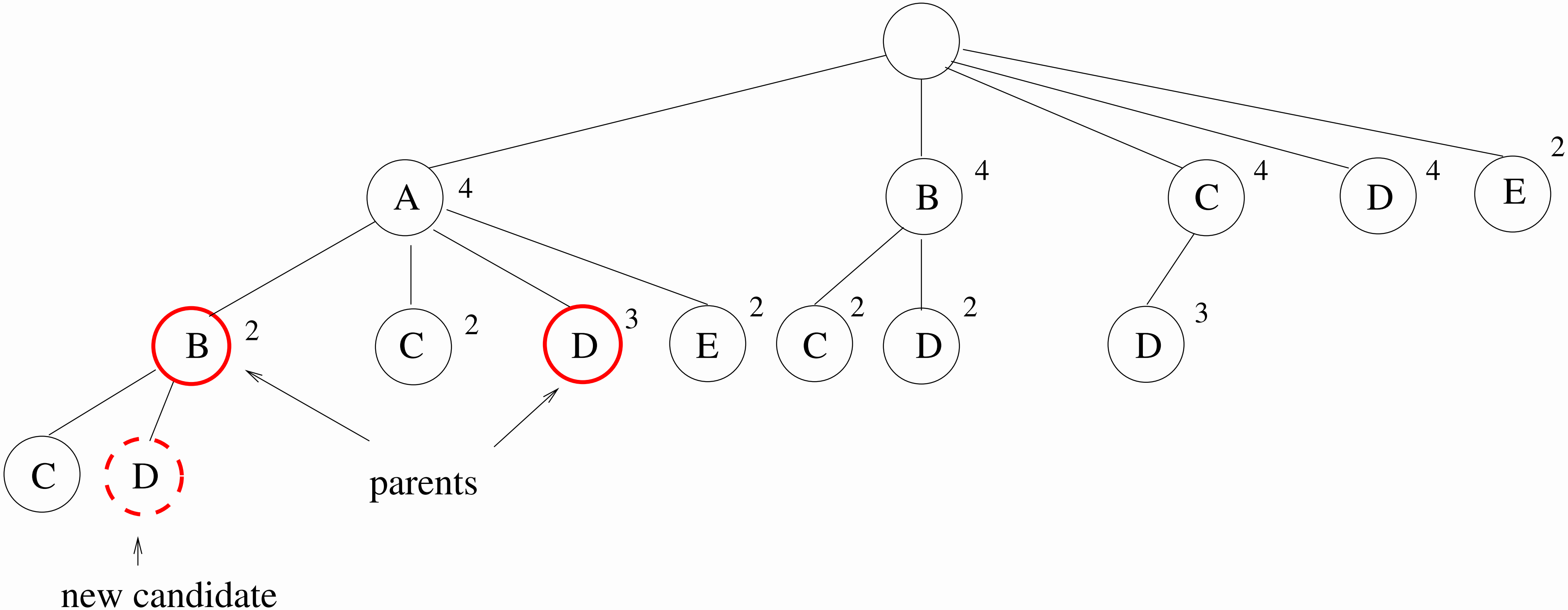
Simulation: $i = 2$ *Frequency counting*

Level 2: 7 frequent sets



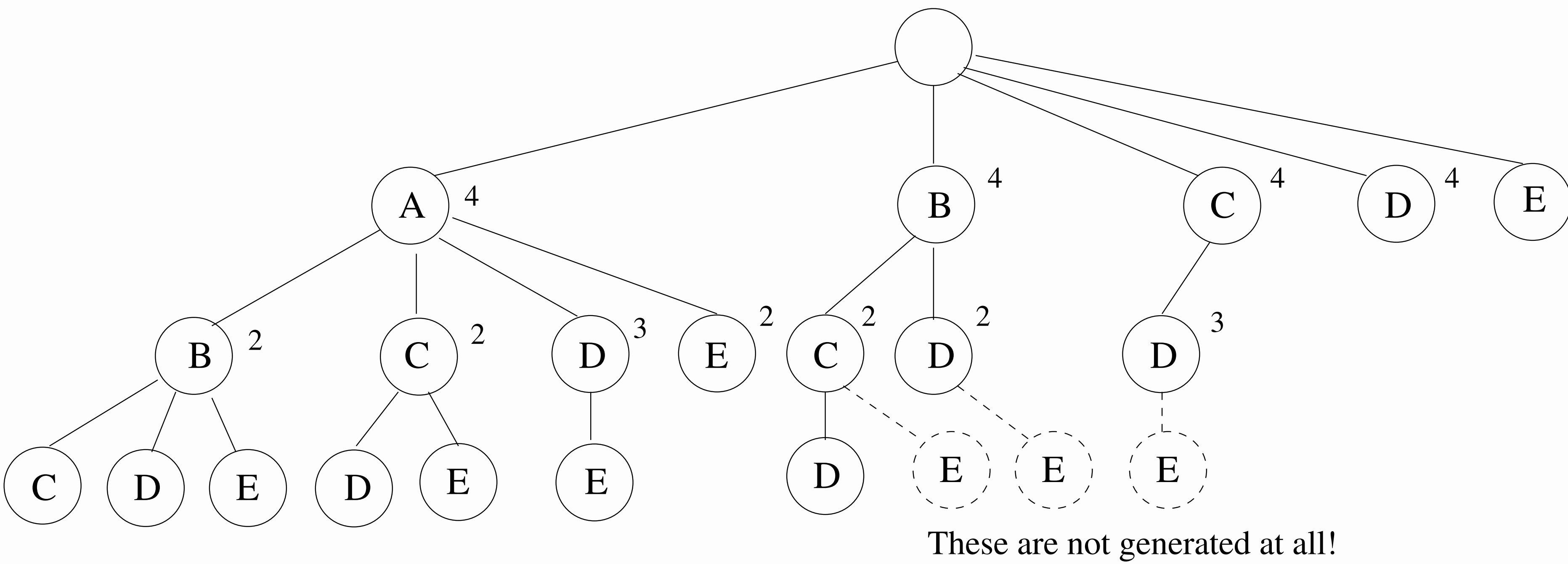
Simulation: $i = 3$ *Candidate generation*

Idea: Given parents $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathcal{F}_i$, such that $|\mathbf{Y}_1 \cap \mathbf{Y}_2| = i - 1$, generate (i+1)-candidate $\mathbf{X} = \mathbf{Y}_1 \cup \mathbf{Y}_2$. E.g.,



Simulation: $i = 3$ *Candidate generation*

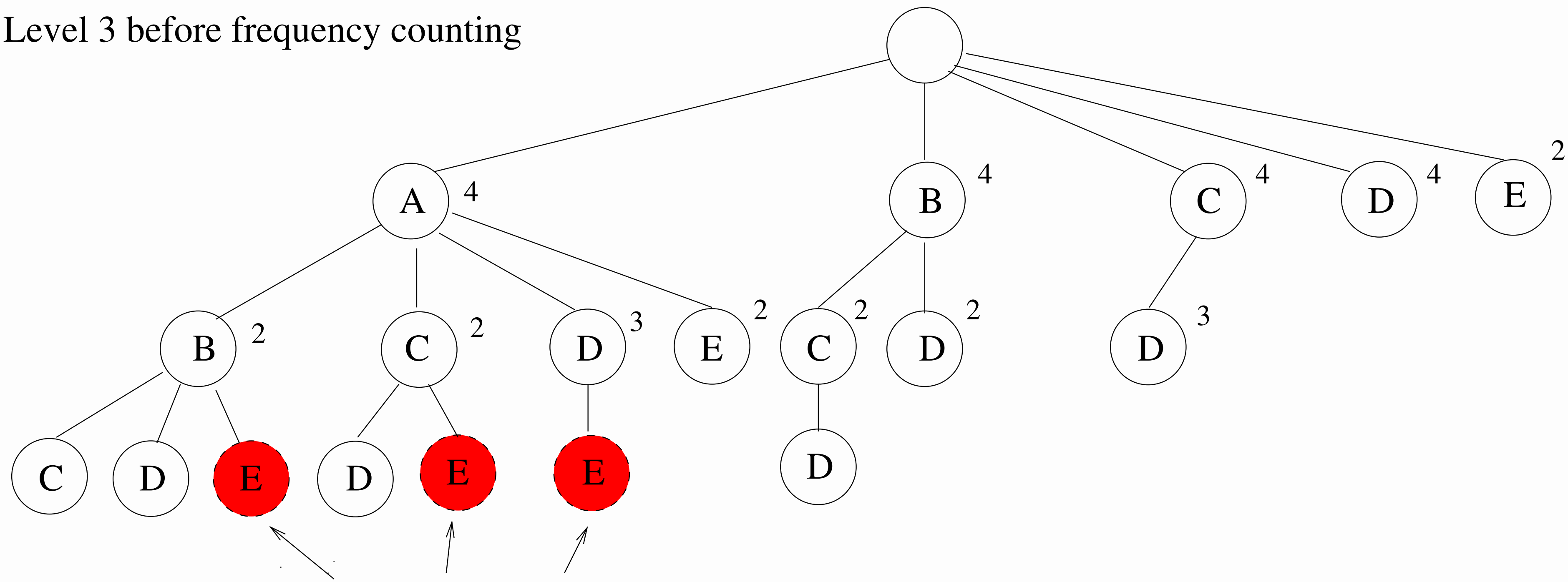
Possible 3-candidates (no pruning yet)



Simulation: $i = 3$ *Candidate generation*

Prune with monotonicity

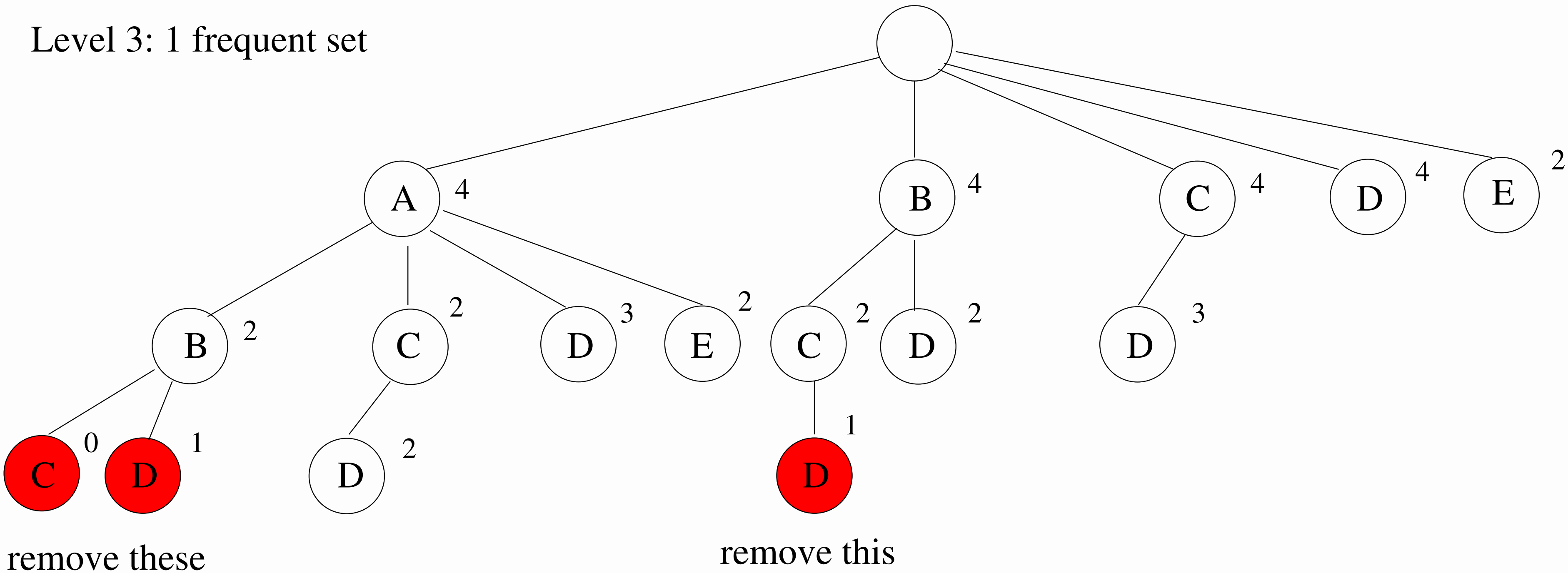
Level 3 before frequency counting



These are immediately removed because some parents not frequent

Simulation: $i = 3$ ***Frequency counting***

Level 3: 1 frequent set



Simulation: $i = 4$ *algorithm stops*

Cannot create any 4-itemset candidates!

Return frequent sets (abs freq. in parenthesis):
A (4), B (4), C (4), D (4), E (2),
AB (2), AC (2), AD (3), AE (2),
BC (2), BD (2), CD (3), ACD (2)

6.3 Constructing rules from frequent sets

Given frequent sets $\mathcal{F} = \cup_i \mathcal{F}_i$. For all $\mathbf{X} \in \mathcal{F}$, $|\mathbf{X}| \geq 2$, and all $C \in \mathbf{X}$, evaluate $\mathbf{X} \setminus \{C\} \rightarrow C$

- confidence/precision: $\phi(\mathbf{X} \setminus \{C\} \rightarrow C) = P(C|\mathbf{X} \setminus \{C\})$
- statistical dependence:
  - lift $\gamma(\mathbf{X} \setminus \{C\} \rightarrow C) = \frac{P(\mathbf{X})}{P(\mathbf{X}\setminus\{C\})P(C)}$ or
  - leverage $\delta(\mathbf{X} \setminus \{C\} \rightarrow C) = P(\mathbf{X}) - P(\mathbf{X} \setminus \{C\})P(C)$
  - minimum requirement: $\gamma > 1$ or $\delta > 0$
- statistical significance p_F, χ^2, mutual information, ...
- optional: prune out redundant/overfitted rules
  Does $\mathbf{X} \setminus \{C\} \rightarrow C$ improve $\mathbf{Y} \setminus \{C\} \rightarrow C$ for all $\mathbf{Y} \subsetneq \mathbf{X}$?

Task: Construct rules from frequent sets

What are confident rules, if $min_{cf} = 0.6$? Which of them express positive statistical dependence? ($n = 6$)

A (4), B (4), C (4), D (4), E (2),
AB (2), AC (2), AD (3), AE (2),
BC (2), BD (2), CD (3), ACD (2)

$\phi(\mathbf{X} \rightarrow C) = \frac{P(\mathbf{X},C)}{P(\mathbf{X})}$ $\qquad$ $\gamma(\mathbf{X}, C) = \frac{P(\mathbf{X},C)}{P(\mathbf{X})P(C)} = \frac{n \cdot fr(\mathbf{X},C)}{fr(\mathbf{X})fr(C)} = \frac{P(C|\mathbf{X})}{P(C)}$

Extra task (at home): What would be negative dependencies (form $\mathbf{X} \rightarrow \neg A_i$) with these constraints?

Task: Confident rules with $min_{cf} = 0.6$

A (4), B (4), C (4), D (4), E (2), AB (2), AC (2), AD (3), AE (2), BC (2), BD (2), CD (3), ACD (2)

$\phi(A \rightarrow B) = \frac{2}{4}$

$\phi(A \rightarrow C) = \frac{2}{4}$

$\phi(A \rightarrow D) = \frac{3}{4}$

$\phi(A \rightarrow E) = \frac{2}{4}$

$\phi(B \rightarrow C) = \frac{2}{4}$

$\phi(B \rightarrow D) = \frac{2}{4}$

$\phi(C \rightarrow D) = \frac{3}{4}$

$\phi(AC \rightarrow D) = \frac{2}{2}$

$\phi(CD \rightarrow A) = \frac{2}{3}$

$\phi(B \rightarrow A) = \frac{2}{4}$

$\phi(C \rightarrow A) = \frac{2}{4}$

$\phi(D \rightarrow A) = \frac{3}{4}$

$\phi(E \rightarrow A) = \frac{2}{2}$

$\phi(C \rightarrow B) = \frac{2}{4}$

$\phi(D \rightarrow B) = \frac{2}{4}$

$\phi(D \rightarrow C) = \frac{3}{4}$

$\phi(AD \rightarrow C) = \frac{2}{3}$

Task: Confident rules with $min_{cf} = 0.6$

rule	ϕ
$A \rightarrow D$ (or $D \rightarrow A$)	0.75
$E \rightarrow A$	1.00
$C \rightarrow D$ (or $D \rightarrow C$)	0.75
$AD \rightarrow C$	0.67
$CD \rightarrow A$	0.67
$AC \rightarrow D$	1.00

Note: $A \rightarrow D$ and $D \rightarrow A$ express the same association (here also ϕ happens to be the same).

Task: Statistical dependence, $\gamma > 1$?

A (4), B (4), C (4), D (4), E (2), AB (2), AC (2), AD (3), AE (2), BC (2), BD (2), CD (3), ACD (2)

$$\gamma(\mathbf{X}, C) = \frac{n \cdot fr(\mathbf{X}, C)}{fr(\mathbf{X})fr(C)} = \frac{P(C|\mathbf{X})}{P(C)}$$

$\gamma(A \to D) = \frac{6 \cdot 3}{4 \cdot 4} = 1.125$

$\gamma(C \to D) = \frac{6 \cdot 3}{4 \cdot 4} = 1.125$

$\gamma(AD \to C) = \frac{6 \cdot 2}{3 \cdot 4} = 1$

$\gamma(E \to A) = \frac{6 \cdot 2}{2 \cdot 4} = 1.5$

$\gamma(AC \to D) = \frac{6 \cdot 2}{2 \cdot 4} = 1.5$

$\gamma(CD \to A) = \frac{6 \cdot 2}{3 \cdot 4} = 1$

Task: Statistical dependence, $\gamma > 1$?

rule	ϕ	γ	
$A \rightarrow D$ (or $D \rightarrow A$)	0.75	1.13	
$E \rightarrow A$	1.00	1.50	
$C \rightarrow D$ (or $D \rightarrow C$)	0.75	1.13	
$AD \rightarrow C$	0.67	$\mathbf{1.00}$	$\rightarrow$ prune out!
$CD \rightarrow A$	0.67	$\mathbf{1.00}$	$\rightarrow$ prune out!
$AC \rightarrow D$	1.00	1.50	

Strongest positive rules with lift:
oranges (E) $\rightarrow$ milk (A)
milk (A), bread (C) $\rightarrow$ cheese (D)

Many negative rules! e.g., $\gamma(ACD \rightarrow \neg B) = 3.0$

7. Pattern explosion and condensed representations

Pattern explosion a big problem!

- small min_{fr} $\Rightarrow$ too many frequent patterns (worst case: $\mathcal{O}(2^k)$ patterns!)
- but large min_{fr} not good (trivial patterns, interesting missed)

$\Rightarrow$ condensed representations = **representatives** of all frequent sets

- faster to search
- all frequent sets can be derived from them (but **very costly!**)
- **bad shortcuts**: use only condensed representations $\Rightarrow$ How to find statistical or significant associations??

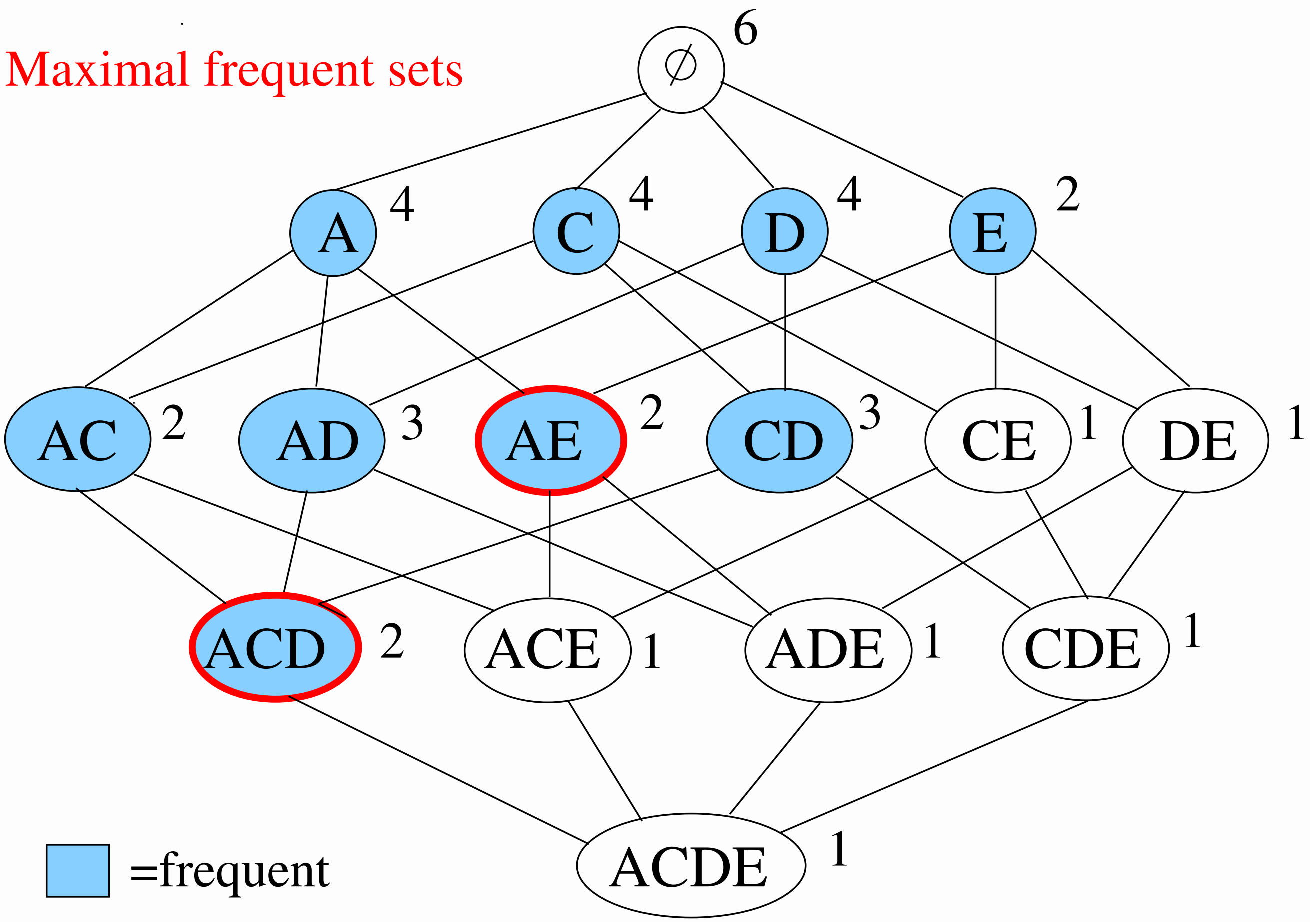
Maximal, closed and free sets

Frequent set $\mathbf{X}$, $P(\mathbf{X}) \geq min_{fr}$, is

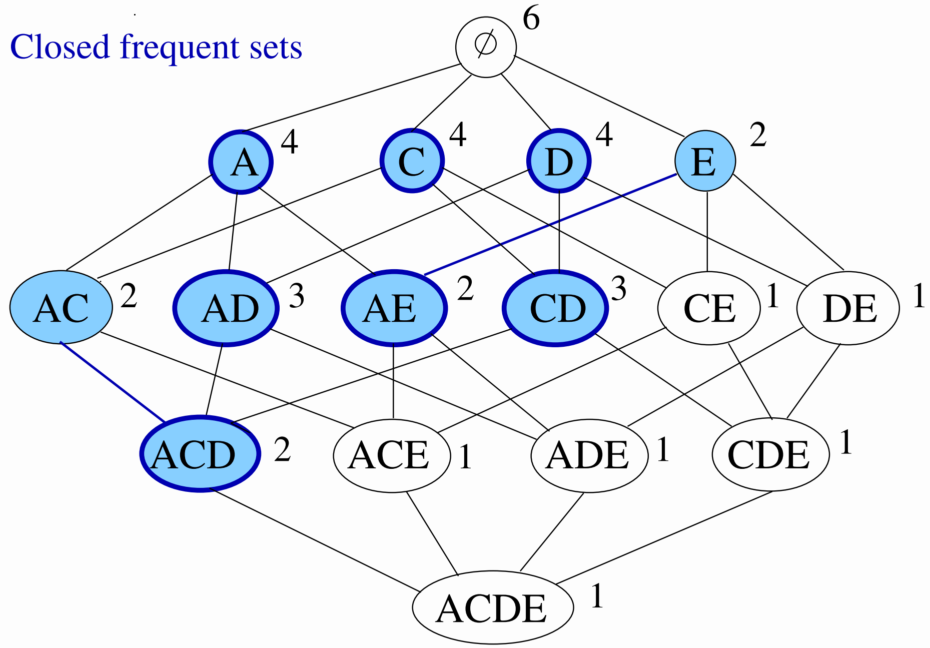
- **maximal frequent set**, if for all $\mathbf{Y} \supsetneq \mathbf{X}$: $P(\mathbf{Y}) < min_{fr}$ (most complex sets that are frequent)
- **closed set**, if for all $\mathbf{Y} \supsetneq \mathbf{X}$: $P(\mathbf{Y}) < P(\mathbf{X})$ (most specific set as a representative of **nested** sets with the same fr; e.g., $fr(A) = fr(AC) = fr(ACD) \Rightarrow ACD$ closed)
- **0-free set = minimal generator**, if for all $\mathbf{Y} \subsetneq \mathbf{X}$: $P(\mathbf{Y}) > P(\mathbf{X})$ (most general set as a representative, e.g., A above)

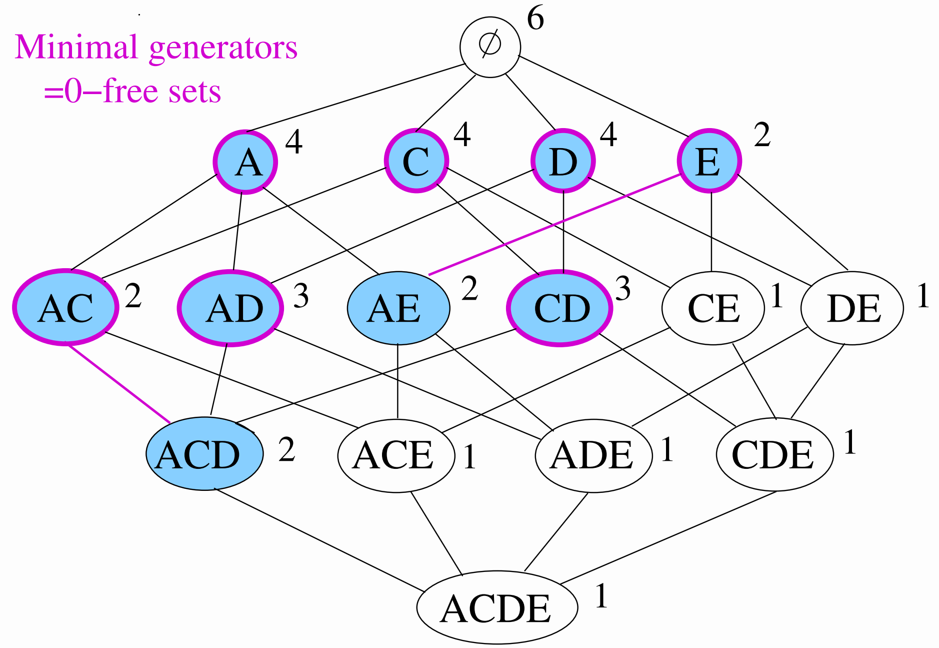
Worst case: all sets closed and free!

Example: maximal sets ($min_{fr} = 2/6 = 0.33$)



Example: closed and 0-free sets





Summary

- **know what you find!** (and what you miss)
  - frequent rules $\neq$ statistical associations
  - constraints and goodness measures
- exponential search space, but very **scalable algorithms**
- **monotonicity of frequency** very useful!
- dilemma: large $\min_{fr}$ misses significant associations, but small $\min_{fr}$ causes **pattern explosion**
- be **aware** when the algorithm uses **condensed representations**!

Question: How to utilize monotonicity when you search for statistical associations?

Further reading

- Aggarwal: Data mining – The textbook, Springer 2015, chapters 4–5.
- Leskovec et al.: Mining of Massive Datasets, Cambridge University Press 2014, chapter 6.
- Tan et al.: Introduction to Data Mining, Pearson, 2019, chapters 5–6.
- Hämäläinen and Webb: A tutorial on statistically sound pattern discovery. Data Mining and Knowledge Discovery 33(2):325-377, 2019. `https://doi.org/10.1007/s10618-018-0590-x`

Other references

- Li and Zaiane: Exploiting statistically significant dependent rules for associative classification. Intelligent Data Analysis, 21(5):1155-1172, 2017.
- Webb and Vreeken: Efficient discovery of the most interesting associations. Transactions on Knowledge Discovery from Data 8(3):15:1-15:31, 2014.
- Zimek et al.: Frequent pattern mining algorithms for data clustering. Chapter 16 in Frequent Pattern Mining, 2014.
- Zimmermann and Nijssen: Supervised pattern mining and applications to classification. Chapter 17 in Frequent Pattern Mining, 2014.

Image sources

- Tan et al. slides for the book, 2018, `https://www-users.cs.umn.edu/~kumar001/dmbook/`
- Animal pictures: `https://commons.wikimedia.org/wiki/File:Dendrocopos_leucotos_NAUMANN.jpg` and `https://www.upmmetsa.fi/tietoa-ja-tapahtumia/tietoartikkelit/metsiemme-kovakuoriaisia/`