Mining association patterns (Part 3)

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

Contents

1. Recap: 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 $\mathcal{F}_i \neq \emptyset$:

- Generate candidates C_{i+1} from $\mathcal{F}_i$
- Prune $\mathbf{X} \in C_{i+1}$ if $\exists \mathbf{Y} \subsetneq \mathbf{X}, |\mathbf{Y}| = i, \mathbf{Y} \notin \mathcal{F}_i$

} (monotonicity)

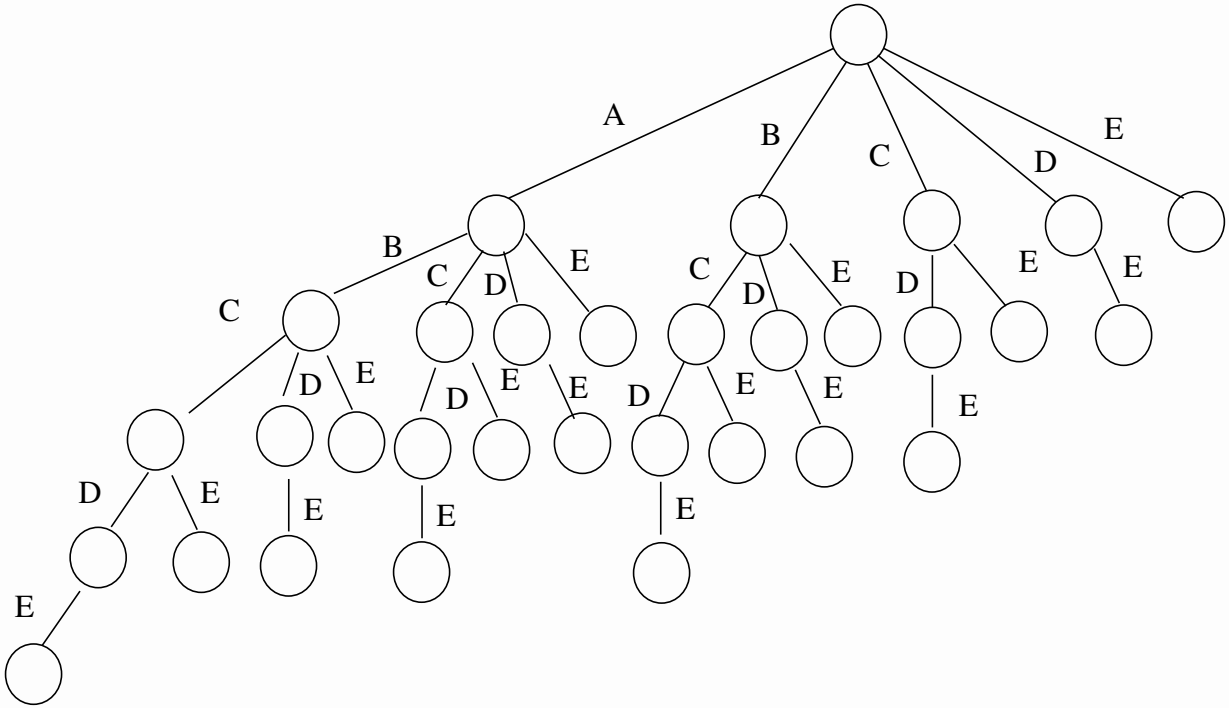
- Count frequencies $fr(\mathbf{X}), \mathbf{X} \in C_{i+1}$
- Set $\mathcal{F}_{i+1} = \{\mathbf{X} \in C_{i+1} \mid P(\mathbf{X}) \geq min_{fr}\}$
- $i = i + 1$

Return $\cup_i \mathcal{F}_i$

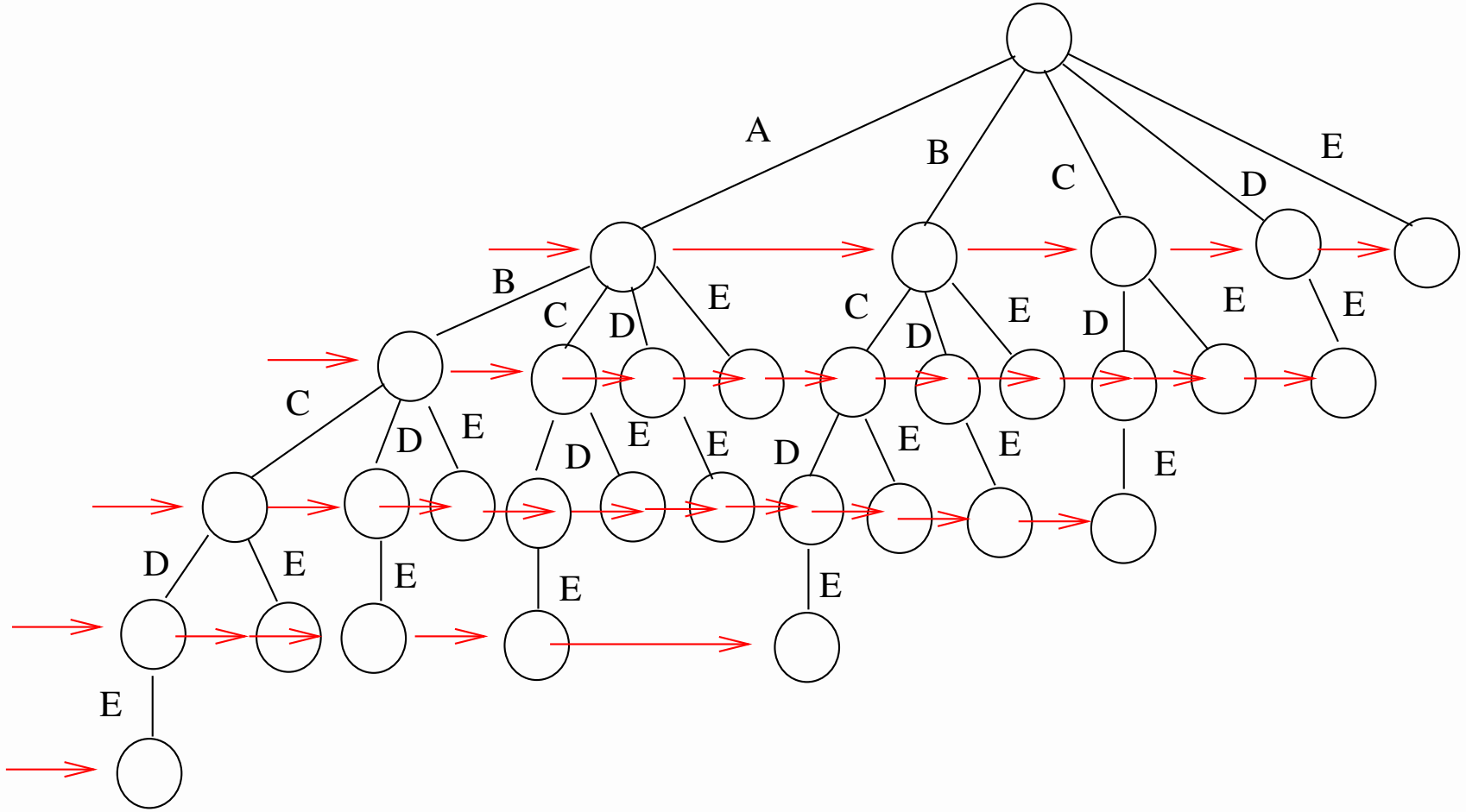
2.1 Enumeration tree can be large! ($\leq 2^k$ nodes, $k = |R|$)

Keep the tree as small as possible!

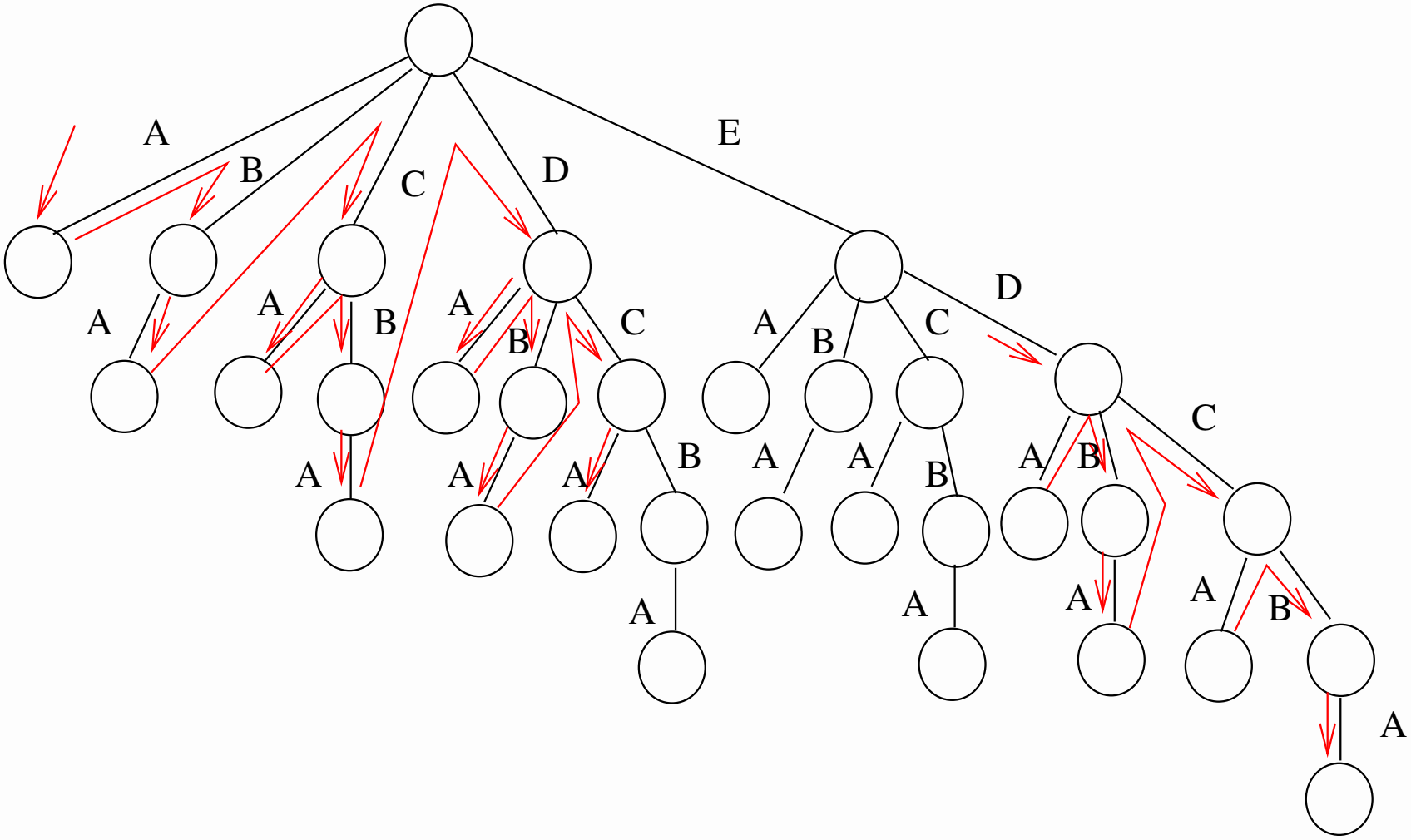
Trick: Order the main branches by ascending frequency
e.g., if $fr(A) \leq fr(B) \leq fr(C) \leq fr(D) \leq fr(E)$, put the largest branch under A:



Breadth-first traversal of the tree (like Apriori)

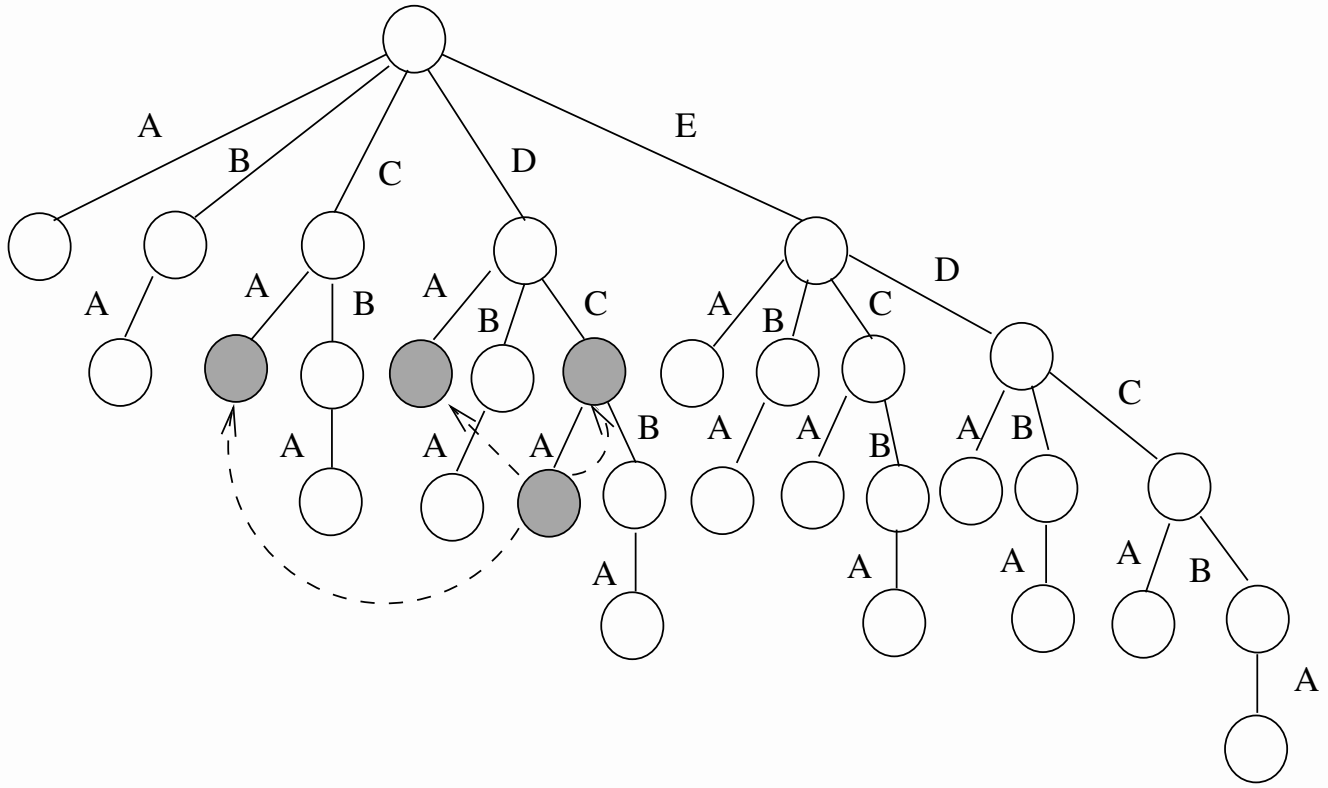


Depth-first traversal of the tree



Depth-first traversal of the tree

Construct the tree such that all parent sets are processed before child sets! e.g., parents of DCA:



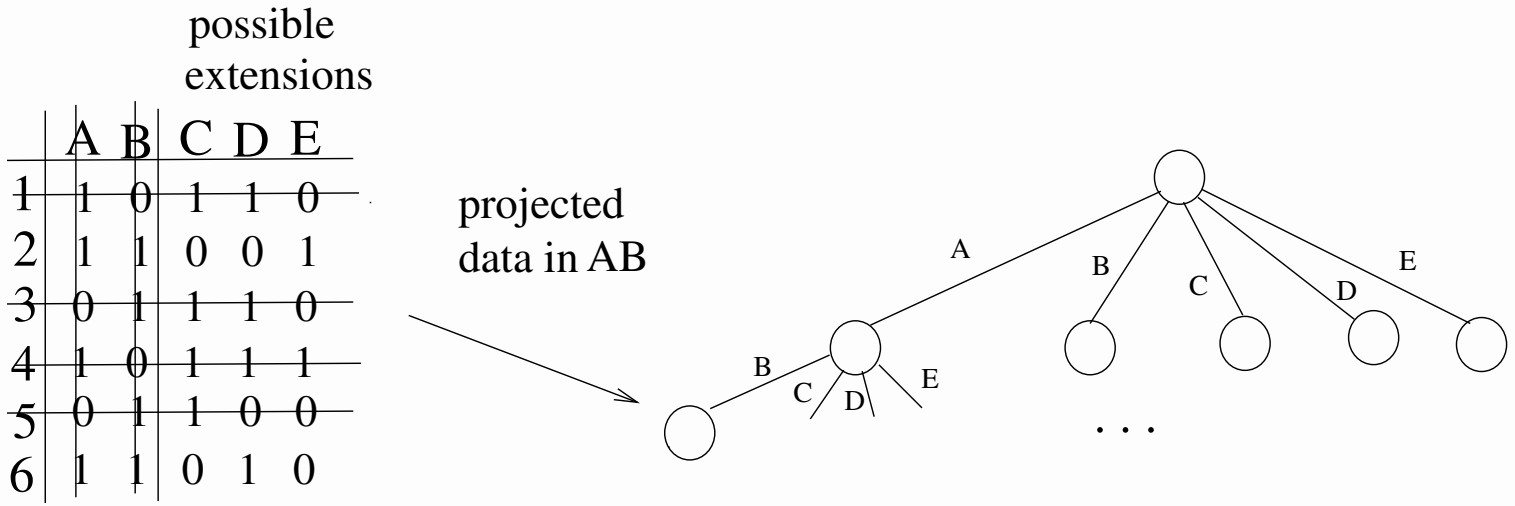
2.2 Cost of frequency counting

- **breadth-first search**: frequencies of the entire level can be checked at once
- **depth-first search**: check frequency of each pattern before continuing $\Rightarrow$ more database scans
- scanning database always costly!
- if enough memory, you can speed up frequency counting by **auxiliary structures** in the tree nodes, like
  - database projections
  - transaction id (tid) lists
- or construct a FP-tree

Database projections

Idea: Each node contains a projection of data onto the needed transactions and items.

- transactions covering $\mathbf{X}$ and items that can extend $\mathbf{X}$
- child nodes inherit the projection and update it



Tid lists (vertical counting methods)

Idea: Store into node $\mathbf{X}$ ids of transactions that cover $\mathbf{X}$. When creating a new child, take an intersection of parents' tid lists.

- $tids(A) = \{1, 2, 4, 6\}$ and $tids(B) = \{2, 3, 5, 6\} \Rightarrow$ $tids(AB) = \{2, 6\}$
- $tids(AB) = \{2, 6\}$ and $tids(AD) = \{1, 4, 6\} \Rightarrow$ $tids(ABD) = \{6\}$
  - no need to intersect with $tids(BD)$
- can be implemented with bit-vectors $\Rightarrow$ logical bit-and operation + count number of 1-bits

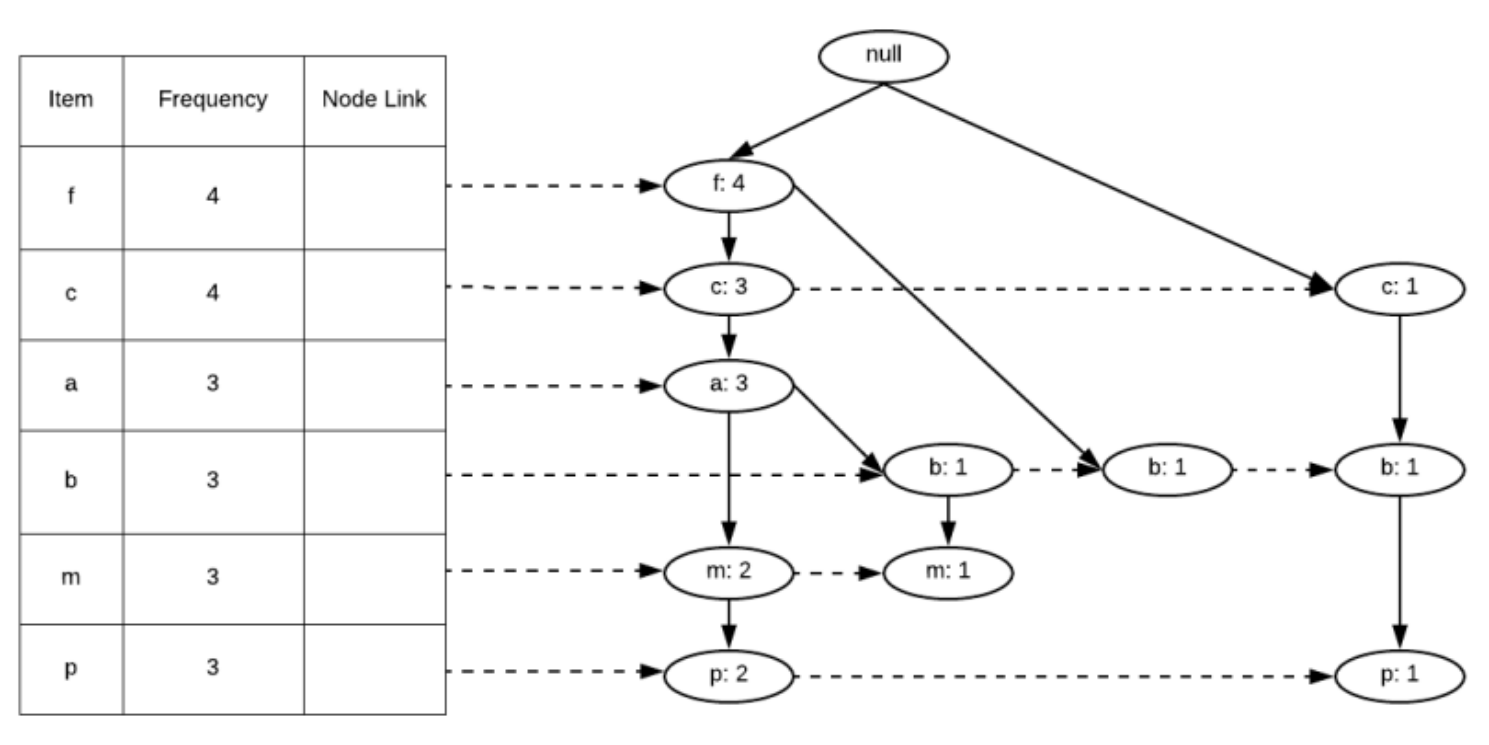
Extra: FP-tree

Idea: store data into a tree!

- present transaction database as a trie, FP-tree $\mathcal{T}$
- root–node path = prefix of a transaction
- each node contains a frequency counter = number of transactions having the prefix
- can be utilized in the FP-growth algorithm

See e.g., Aggarwal Sec 4.4.4

Extra: FP-tree example



Transactions
f,c,a,m,p
f,c,a,b,m
f,b
c,b,p
f,c,a,m,p

(image from Jain 2018)

3. Apriori for other pattern types and properties?

Φ = monotonic property (like frequency)
α = pattern (like set, graph, sequence)
$\beta \subseteq \alpha$ = subpattern (subset, subgraph, subsequence)
$|\alpha|$ = complexity of α
i-pattern = pattern of complexity i

Idea: Begin from 1-patterns and search patterns incrementally utilizing monotonicity of Φ

Generic Apriori given monotonic property Φ

$\mathcal{F}_i$ = i-patterns having property Φ
C_i = candidate i-patterns

i=1; $\mathcal{F}_1$ = {1-patterns with property Φ}
while $\mathcal{F}_i \neq \emptyset$

- Generate candidates C_{i+1} from $\mathcal{F}_i$
- Prune $\alpha \in C_{i+1}$ if $\exists\, \beta \subseteq \alpha$, $|\beta| = i$, $\beta \notin \mathcal{F}_i$
- Evaluate Φ for all $\alpha \in C_{i+1}$
- Set $\mathcal{F}_{i+1} = \{\alpha \in C_{i+1} \mid \alpha \text{ has property } \Phi\}$
- $i = i + 1$

Return $\cup_i \mathcal{F}_i$

4. Yule-Simpson's paradox and other specious associations

Statistical dependence is a necessary but not a sufficient condition of causal relation!

- Often a **majority** of dependencies are specious (illusory, spurious, apparent) associations
- e.g., cake $\rightarrow$ exam failure was a sideproduct of alcohol $\rightarrow$ exam failure and alcohol $\rightarrow$ cake
- Don't make too fast conclusions!

Introduction: Yule-Simpson's paradox

Example: Does a new treatment kill or cure? [a]

T=treatment, R=recovery

	R	$\neg R$	$\sum$
T	20	20	40
$\neg T$	16	24	40
$\sum$	36	44	80

$P(R|T) = 0.50 > 0.475 = P(R)$

positive association **Treatment → Recovery**

[a] Lindley and Novick 1981

Let's analyze female and male separately

Among female patients (F)

	R	$\neg R$	$\sum$
T	2	8	10
$\neg T$	9	21	30
$\sum$	11	29	40

$P(R|T,F) = 0.20 < 0.275 = P(R|F)$ negative association

i.e., **Treatment** $\rightarrow \neg$ **Recovery** in the female subgroup!

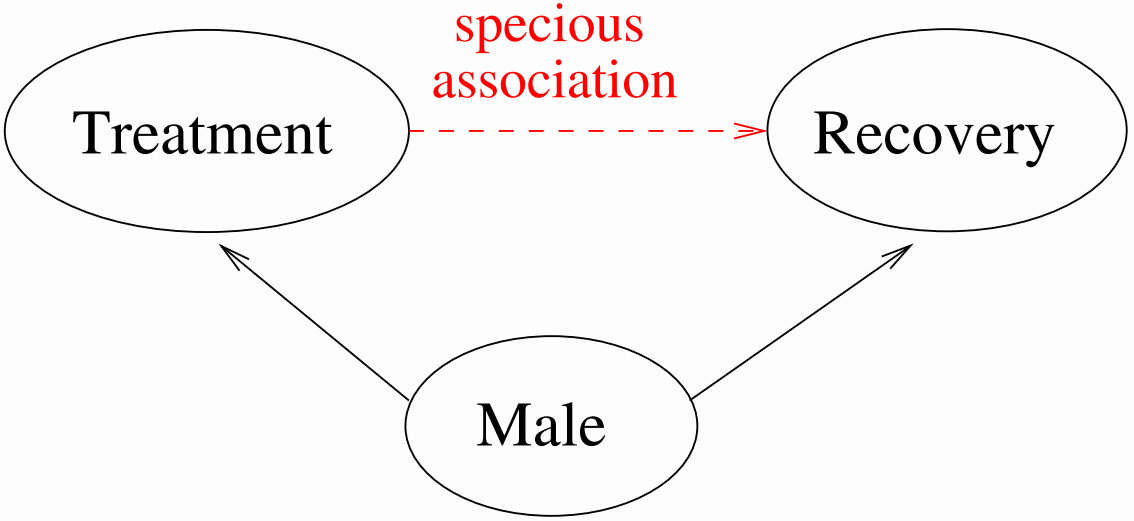
Among male patients (M)

	R	$\neg R$	$\sum$
T	18	12	30
$\neg T$	7	3	10
$\sum$	25	15	40

$P(R|T,M) = 0.60 < 0.625 = P(R|M)$ negative association

i.e., **Treatment** $\rightarrow \neg$ **Recovery** in the male subgroup!

Explanation: a specious sideproduct of other rules



male $\rightarrow$ treatment $\phi = 0.75, \gamma = 1.50, \delta = 0.125, p_F = 7.44e{-}6$
male $\rightarrow$ recovery $\phi = 0.63, \gamma = 1.32, \delta = 0.088, p_F = 1.61e{-}3$

Expected freq. given $H_{01} : T \perp\!\!\!\perp R \mid M$ and $H_{02} : T \perp\!\!\!\perp R \mid F$ is
$E(fr(TR) \mid H_{01}, H_{02}) = fr(TM)P(R|M) + fr(TF)P(R|F) =$
$30 \cdot \frac{25}{40} + 10 \cdot \frac{11}{40} = 21.5 > 20 = fr(TR)$. Conditionally negative dependence!

Types of specious associations $\mathbf{Q} \rightarrow C{=}c$

Here association **disappears** or **reverses its sign** when conditioned on some confounding factor $\mathbf{X}$:

- **Yule-Simpson's paradox**: $\mathbf{Q} \rightarrow C{=}c$ positive association, but $\mathbf{Q}$ and $C{=}c$ either conditionally independent or negatively dependent given $\mathbf{X}$ and given $\neg\mathbf{X}$
- **Specious generalization**: some $\mathbf{QZ} \rightarrow C{=}c$ completely explains $\mathbf{Q} \rightarrow C{=}c$ ($\mathbf{X} = \mathbf{QZ}$)
- **Specious specialization**: some $\mathbf{X} \rightarrow C{=}c$ completely explains $\mathbf{XZ} \rightarrow C{=}c$ ($\mathbf{Q} = \mathbf{XZ}$)
- **Equivalence between $\mathbf{Q}$ and $\mathbf{X}$ or $\neg\mathbf{X}$** (not specious per se)

Definition: Conditional leverage δ_c

Conditional leverage of $\mathbf{Q} \rightarrow C{=}c$ given $\mathbf{X}$ or $\neg\mathbf{X}$

$$\delta_1 = \delta_c(\mathbf{Q}, C{=}c|\mathbf{X}) = P(\mathbf{X}, \mathbf{Q}, C{=}c) - P(\mathbf{X}, \mathbf{Q})P(C{=}c|\mathbf{X})$$

$$\delta_2 = \delta_c(\mathbf{Q}, C{=}c|\neg\mathbf{X}) = P(\neg\mathbf{X}, \mathbf{Q}, C{=}c) - P(\neg\mathbf{X}, \mathbf{Q})P(C{=}c|\neg\mathbf{X})$$

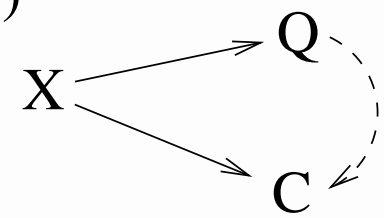
Recall: $\mathbf{Q}$ and C are conditionally independent given $\mathbf{X}$ if

$P(\mathbf{XQ}C) = P(\mathbf{XQ})P(C|\mathbf{X})$

Definition: Specious association rule

Association rule $\mathbf{Q} \rightarrow C{=}c$ $(c \in \{0,1\})$ is specious if there is another rule $\mathbf{X} \rightarrow C{=}c_x$ $(c_x \in \{0,1\})$ such that $\delta_c(\mathbf{Q}, C{=}c|\mathbf{X}) \leq 0$ and $\delta_c(\mathbf{Q}, C{=}c|\neg\mathbf{X}) \leq 0$ in the **population**.

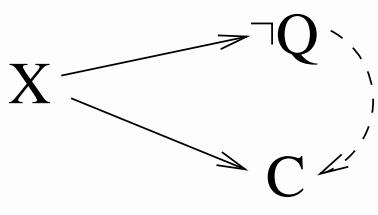
1)



Q - - > C

X ⟶ C

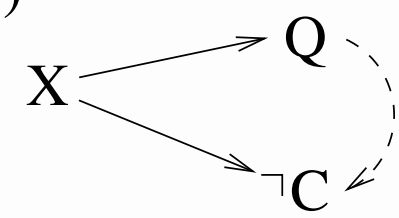
2)



Q - - > ¬C

X ⟶ C

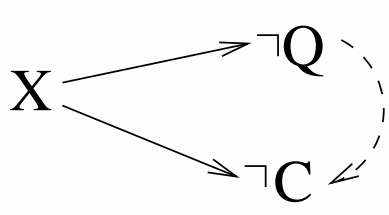
3)



Q - - > ¬C

X ⟶ ¬C

4)

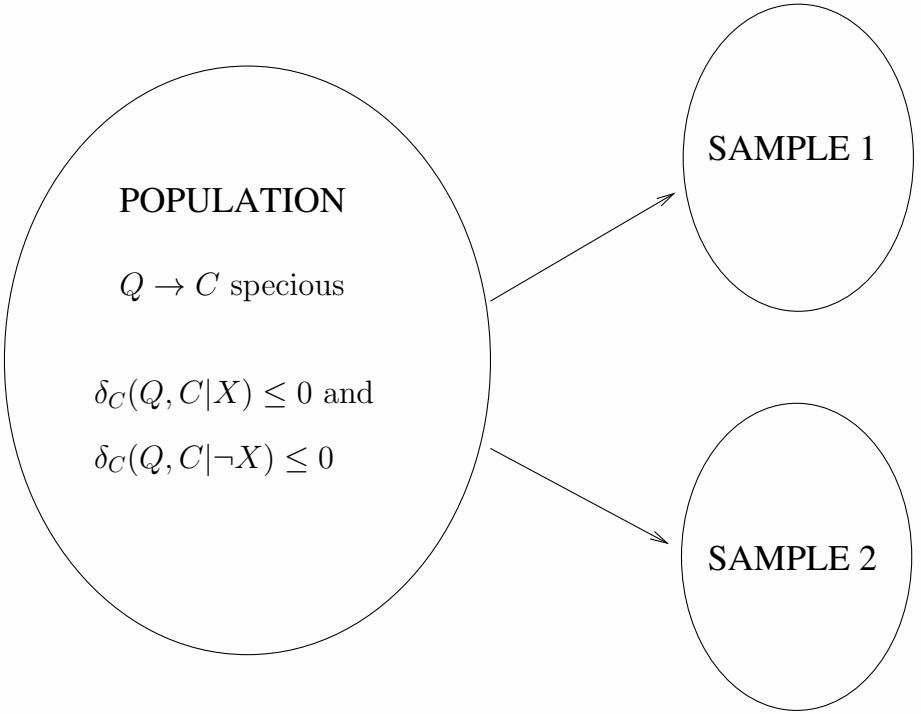


Q - - > C

X ⟶ ¬C

Detecting speciousness in the sample

E.g., Is $\mathbf{Q} \rightarrow C$ specious by $\mathbf{X} \rightarrow C$ or $\mathbf{X} \rightarrow \neg C$?



$\delta_C(Q, C|X) \leq 0$ and

$\delta_C(Q, C|\neg X) \leq 0$

Easy to detect!

$\delta_C(Q, C|X) > 0$ or

$\delta_C(Q, C|\neg X) > 0$

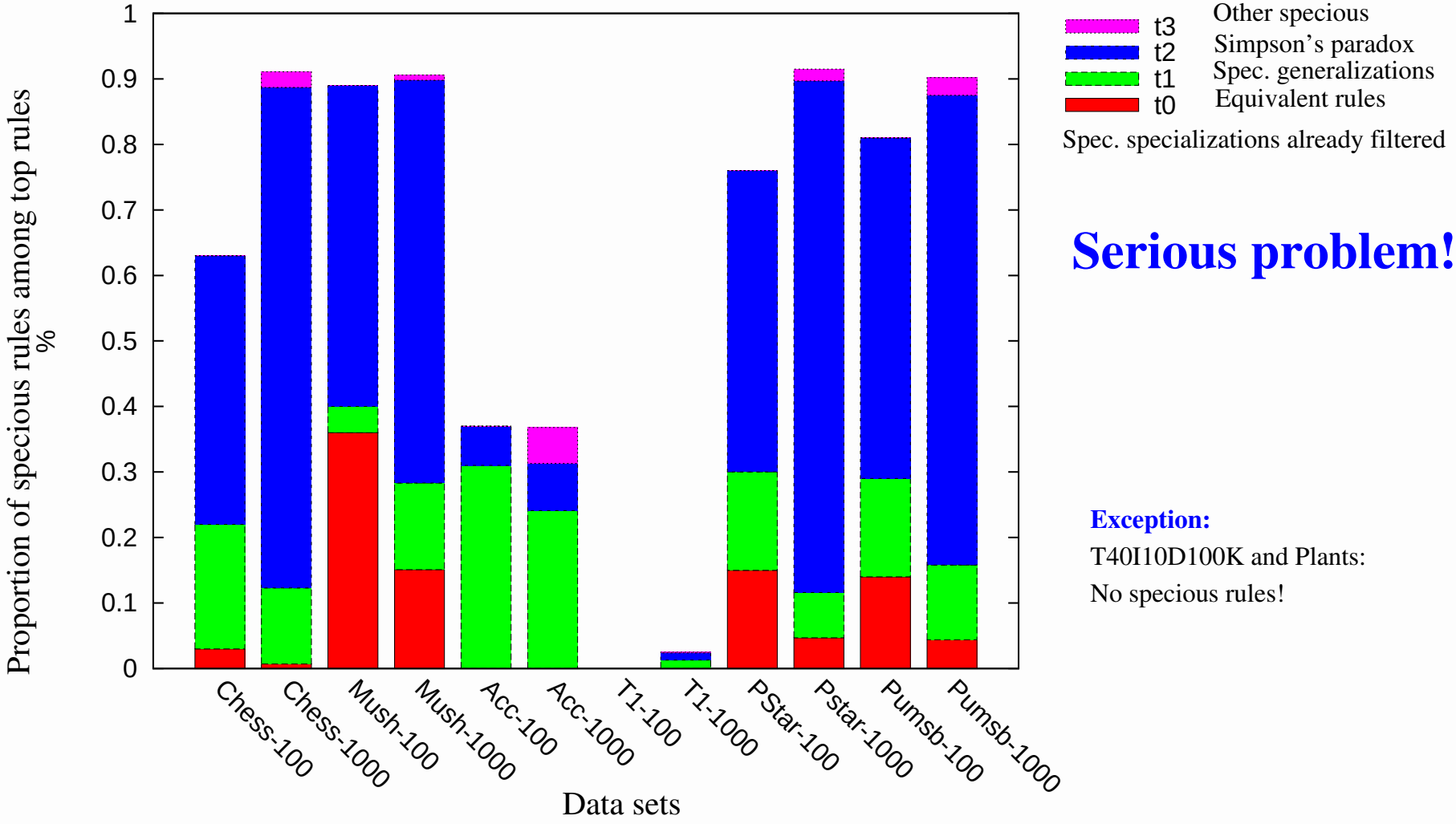
$\Rightarrow$ Test significance!

(conditional independence of Q and C given X and given $\neg X$)

Problem: How to find confounding $\mathbf{X}$ among exponentially many possibilities?? For an efficient solution, see Hämäläinen and Webb 2017!

Experiments: Specious rules and Simpson's paradox are very common!

Proportion and types of specious rules among top−100 or top−1000 best rules



Spec. specializations already filtered

Serious problem!

Exception:

T40I10D100K and Plants:

No specious rules!

Summary

- **Computational problems**:
  - complete enumeration tree has 2^k nodes – not feasible!
  - frequency counting done for each generated node – costly!

  $\Rightarrow$ strategies
- **Generic Apriori** given monotonic property Φ
- **Specious association** $\mathbf{Q} \rightarrow C{=}c$ disappears or reverses its sign when conditioned on confounding factor $\mathbf{X}$
  - sideproduct of $\mathbf{X} \rightarrow C{=}c$ and $\mathbf{X} \rightarrow \mathbf{Q}$
  - or $\mathbf{X} \rightarrow C{\neq}c$ and $\mathbf{X} \rightarrow \neg\mathbf{Q}$

References

- Hämäläinen and Webb: Specious rules: an efficient and effective unifying method for removing misleading and uninformative patterns in association rule mining. SIAM Int. Conf. Data Mining, 2017. `https://arxiv.org/pdf/1709.03915.pdf`
- Lindley and Novick: The Role of Exchangeability in Inference, Annals of Statistics 9(1):45-58, 1981.