CS57300
PURDUE UNIVERSITY

APRIL 11, 2019

DATA MINING

ANNOUNCEMENT

- Next class: April 16 (Tuesday)
  - Guest lecture: Graphic models and casual inference (Professor Elias Bareinboim)
- April 18 (Thursday): No class; work on your final project
- April 23 & 25: Final project presentations
  - 5 minute presentation + 1 minute Q&A; order will be out soon
  - Slides will be due on April 21, 11:59pm

PATTERN MINING

DATA MINING COMPONENTS

- Task specification: **Pattern discovery**
- Knowledge representation
- Learning technique
- Evaluation

PATTERN DISCOVERY

- Models describe entire dataset (or large part of it)
- Pattern characterizes local aspects of data
- Pattern: predicate/statement that returns “true” for the instances in the data where the pattern occurs and “false” otherwise

PATTERN IN TABULAR DATA

- Primitive pattern: subset of all possible observations over variables $X_1,\ldots,X_p$
  - If X_k is categorical then $X_k=c$ is a primitive pattern
  - If X_k is ordinal then $X_k \leq c$ is a primitive pattern
- Start from primitive patterns and combine using logical connectives such as AND and OR
  - age<40 AND income<100,000
  - chips=1 AND (beer=1 OR soda=1)

PATTERN DISCOVERY TASK

- Find all “interesting” patterns in the data
    - Find a pattern that is frequently true
    - Find associative property between patterns

EXAMPLES

- Supermarket transaction database
  - 10% of the customers buy wine and cheese
- Telecommunications alarms database
  - If alarms A and B occur within 30 seconds of each other then alarm C occurs within 60 seconds with p=0.5
- Web log dataset
  - If a person visits the CNN website, there is a 60% chance the person will visit the ABC News website in the same month

KNOWLEDGE REPRESENTATION

RULE

- A rule is an expression of the form $\theta \rightarrow \varphi$
- A statement about the co-occurrence of events/patterns
- **Support** (aka frequency)
  - $s(\theta \rightarrow \varphi) = fr(\theta \wedge \varphi) / N$
  - Proportion of N items with antecedent θ and consequent φ
- **Confidence** (aka accuracy)
  - $c(\theta \rightarrow \varphi) = p(\varphi \mid \theta) = fr(\theta \wedge \varphi) / fr(\theta)$
  - Proportion of items which have antecedent θ that also have consequent φ

ASSOCIATION RULES

- Find all rules of the form $\theta \rightarrow \varphi$ that satisfy the following constraints:
  - Support of the rule is greater than threshold *s*
  - Confidence of the rule is greater than threshold *c*

ASSOCIATION RULE EXAMPLE

- Support threshold: 30%, confidence threshold: 70%
- Flour –> Eggs
- Eggs –> Milk
- Milk –> Eggs
- Flour –> Milk
- Eggs, Flour –> Milk
- Flour, Milk –> Eggs

Transaction ID	beer	eggs	flour	milk
1	0	1	1	1
2	1	1	0	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1

LEARNING

MODEL SPACE AND SEARCH

- Model space: All possible rules
- Suppose there are N binary variables
- Even if we only consider rules where θ and φ are conjunctions of X_k=1
  - We still have $\binom{N}{2}\binom{2}{1} + \binom{N}{3}(\binom{3}{1} + \binom{3}{2}) + \ldots + \binom{N}{N} \times (\binom{N}{1} + \binom{N}{2} + \ldots + \binom{N}{N-1})$ rules
- Searching for all patterns is computationally intractable

SOLUTION: THE APRIORI ALGORITHM

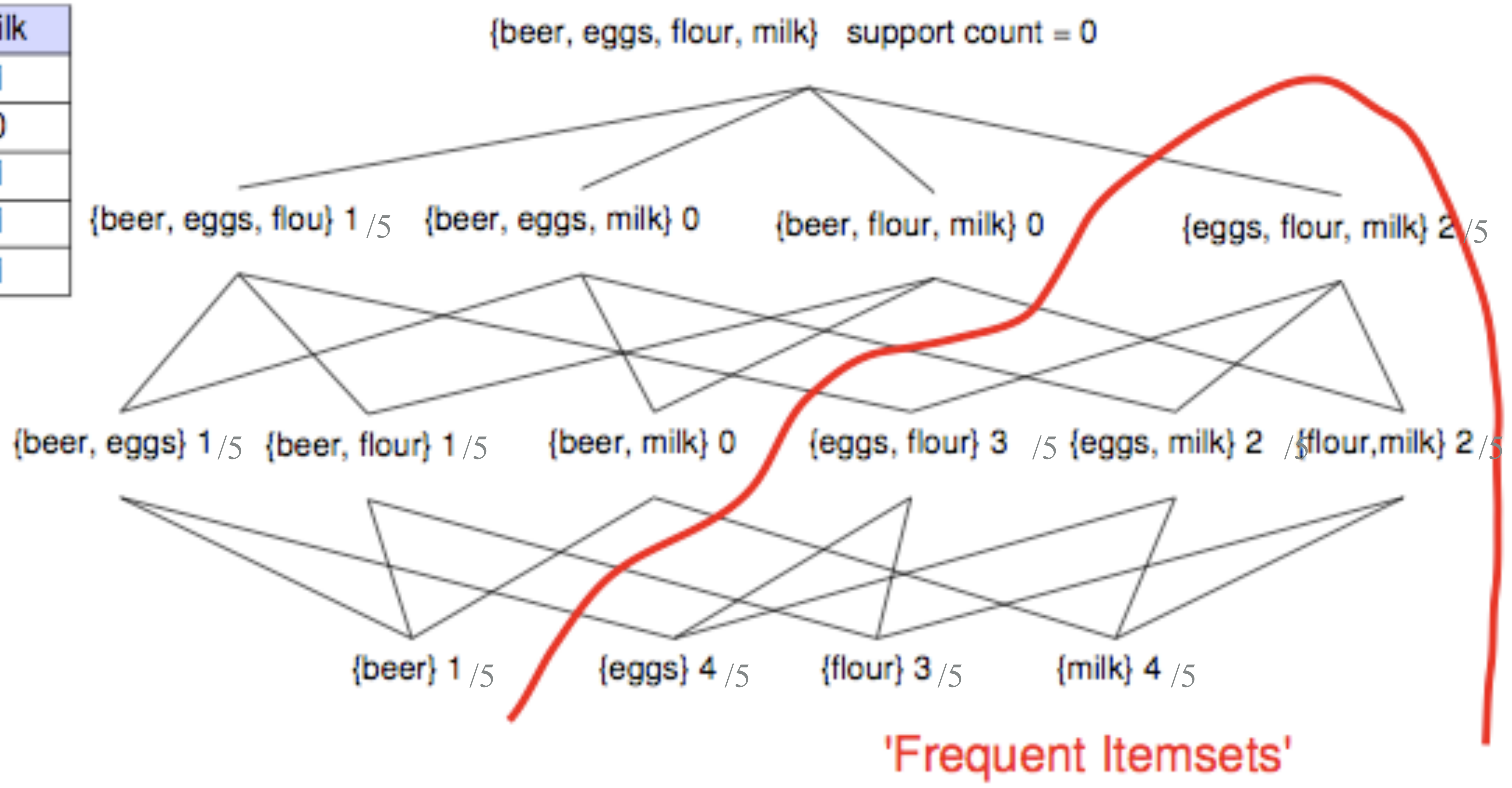
- Key idea: Decompose the search process into two steps
- First search for “frequent itemset”: combinations of predicate whose support is above the threshold
- Then search among frequent items to prune rules whose confidence is below threshold

FINDING FREQUENT ITEMSETS

- Find sets of items with minimum support
- Support is ***monotonic***
  - A subset of a frequent itemset must also be frequent
  - Eg. If {A,B} is a frequent itemset then both {A} and {B} are frequent itemsets as well
  - That is, if {A} is not a frequent itemset, then {A, B} can't be a frequent itemset either
- Approach
  - Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
  - Prune any sets of size k that are not frequent

EXAMPLE

Transaction ID	beer	eggs	flour	milk
1	0	1	1	1
2	1	1	1	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1



support threshold = 0.2

ALGORITHM TO FIND FREQUENT ITEMSETS

FrequentItemsetGeneration (D, minsup)

% C_k: candidate itemsets of size k; L_k: frequent itemsets of size k

L_1 = {frequent single items}

for (k=1; L_k!=∅; k++)

C_{k+1} = CandidateItemsetGeneration (L_k, minsup)

for each transaction t in D

increment the count of all candidates in C_{k+1} contained in t

L_{k+1} = candidates in C_{k+1} with minsup

Return $\bigcup_k L_k$

GENERATING CANDIDATES

CandidateItemsetGeneration (L_k, minsup)

% step 1: self-joining L_k

C_{k+1} = {}

For p in L_k, q in L_k ,p!=q:

Add p∪q in C_{k+1} if | p∪q |=k+1

% step 2: pruning

For c in C_{k+1}

For all k-item subsets s of c

If s not in L_k then delete c from C_{k+1}

EXAMPLE

support threshold = 0.3

Database D

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

Scan D

C_1

itemset	sup.
{1}	2/4
{2}	3/4
{3}	3/4
{4}	1/4
{5}	3/4

L_1

itemset	sup.
{1}	2/4
{2}	3/4
{3}	3/4
{5}	3/4

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1/4
{1 3}	2/4
{1 5}	1/4
{2 3}	2/4
{2 5}	3/4
{3 5}	2/4

L_2

itemset	sup
{1 3}	2/4
{2 3}	2/4
{2 5}	3/4
{3 5}	2/4

C_3

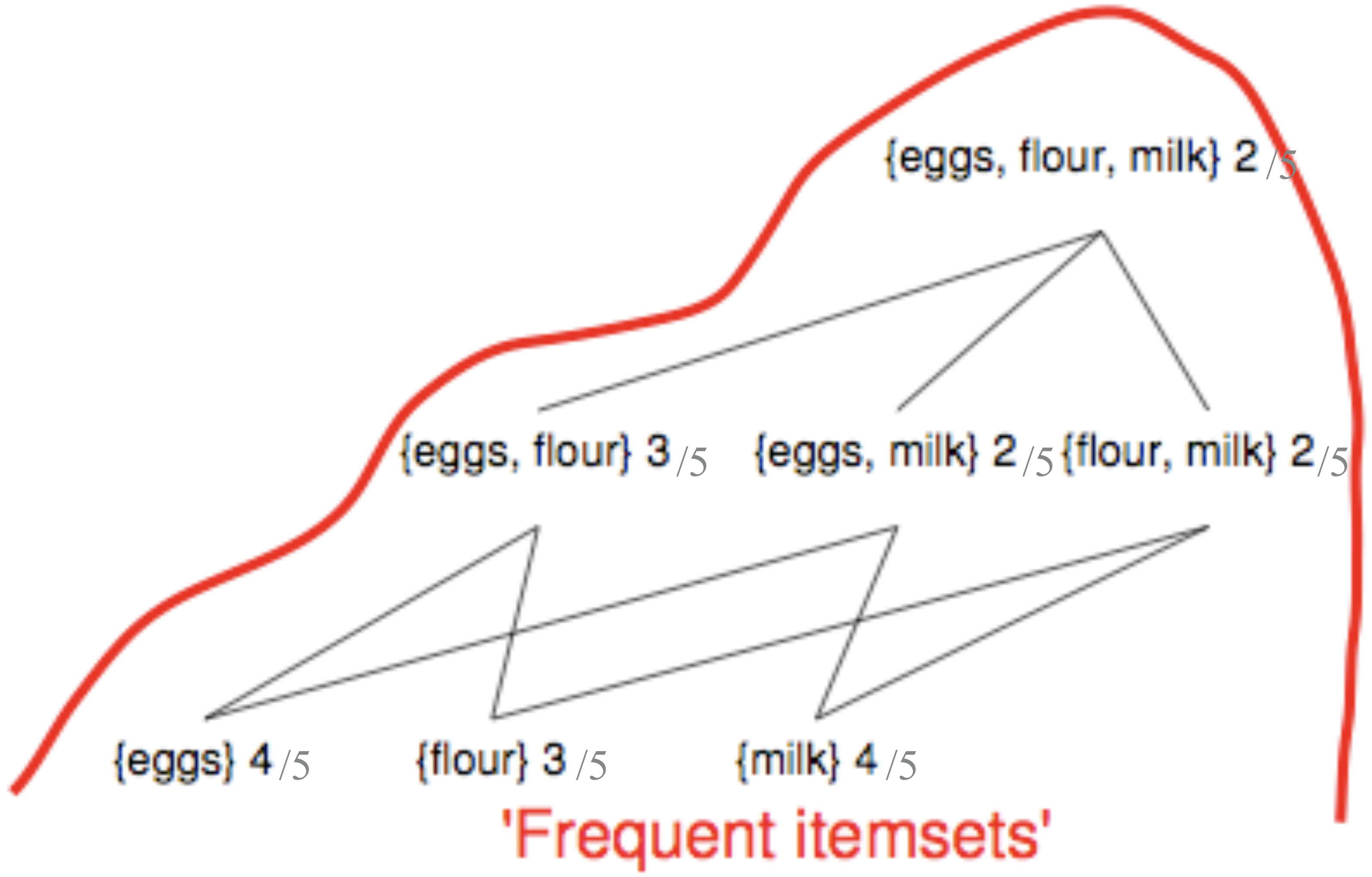
itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2/4

EXAMPLE



			Confidence
{eggs}	$\rightarrow$	{flour}	$3/4 = 0.75$
{flour}	$\rightarrow$	{eggs}	$3/3 = 1$
{eggs}	$\rightarrow$	{milk}	$2/4 = 0.5$
{milk}	$\rightarrow$	{eggs}	$2/4 = 0.5$
{flour}	$\rightarrow$	{milk}	$2/3 = 0.67$
{milk}	$\rightarrow$	{flour}	$2/4 = 0.5$
{eggs, flour}	$\rightarrow$	{milk}	$2/3 = 0.67$
{eggs, milk}	$\rightarrow$	{flour}	$2/2 = 1$
{flour, milk}	$\rightarrow$	{eggs}	$2/2 = 1$
{eggs}	$\rightarrow$	{flour, milk}	$2/4 = 0.5$
{flour}	$\rightarrow$	{eggs, milk}	$2/3 = 0.67$
{milk}	$\rightarrow$	{eggs, flour}	$2/4 = 0.5$

RULE GENERATION

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow (L - f)$ satisfies the minimum confidence requirement

If {A,B,C,D} is a frequent itemset, candidate rules:

$ABC \rightarrow D$,	$ABD \rightarrow C$,	$ACD \rightarrow B$,	$BCD \rightarrow A$,
$A \rightarrow BCD$,	$B \rightarrow ACD$,	$C \rightarrow ABD$,	$D \rightarrow ABC$
$AB \rightarrow CD$,	$AC \rightarrow BD$,	$AD \rightarrow BC$,	$BC \rightarrow AD$,
$BD \rightarrow AC$,	$CD \rightarrow AB$,		

- If $|L|=k$ then there are 2^k-2 candidate association rules (ignoring $L \rightarrow \varnothing$ and $\varnothing \rightarrow L$)

EFFICIENT RULE GENERATION

- Key insight: the confidence of rules generated from the same itemset is monotonic with respect to the number of items in the consequent
  - Recall that:
    $c(\theta \rightarrow \varphi) = p(\varphi \mid \theta)$
  - Consider frequent itemset L={A,B,C,D}:

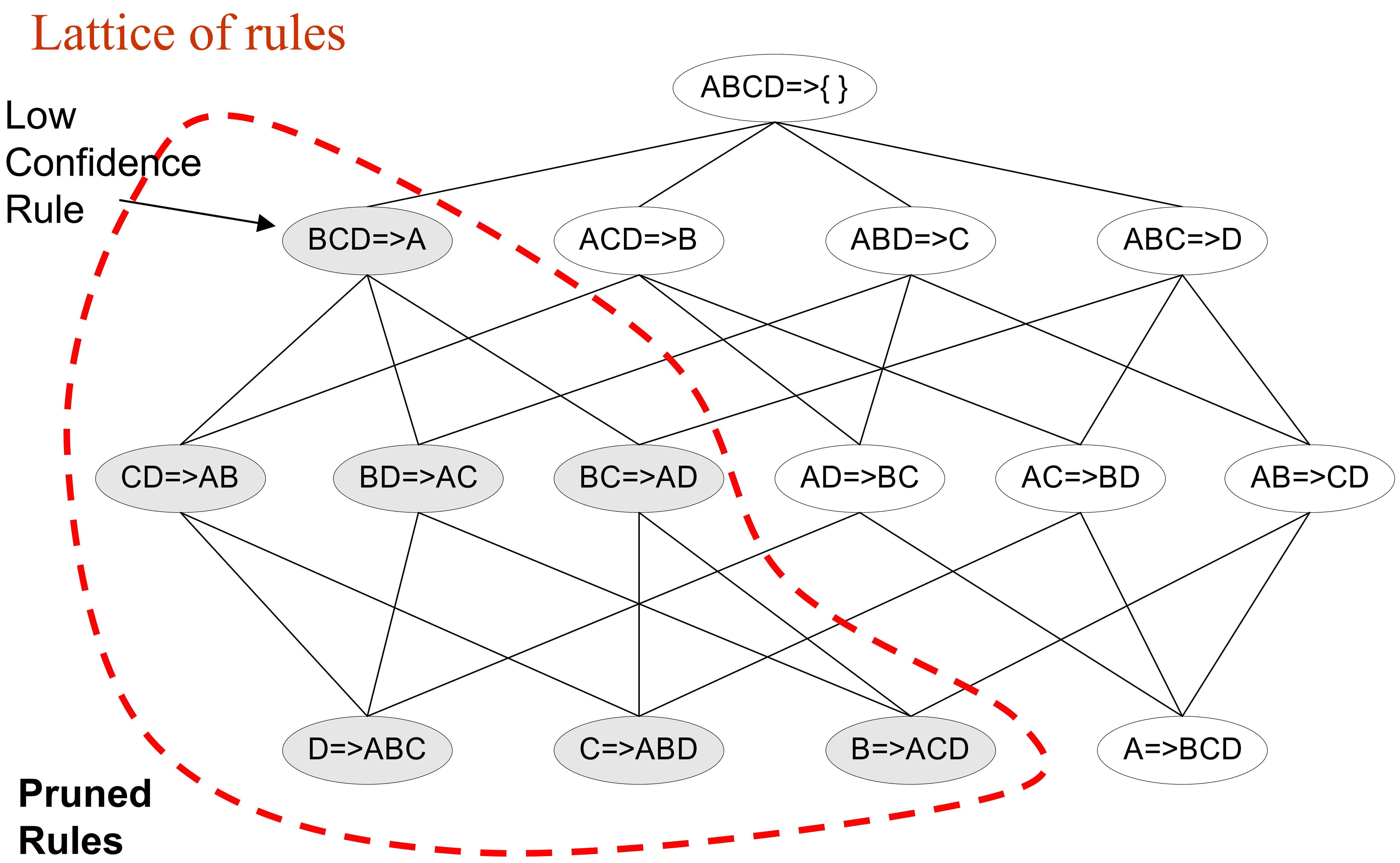
$$c(ABC \rightarrow D) = P(D|ABC) = \frac{fr(ABCD)}{fr(ABC)}$$

$$c(AB \rightarrow CD) = P(CD|AB) = \frac{fr(ABCD)}{fr(AB)}$$

$$\text{We know: } fr(ABC) \leq fr(AB) \text{ and } \frac{1}{fr(ABC)} \geq \frac{1}{fr(AB)}$$

$$\text{thus: } c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

PRUNING RULES



ALGORITHM TO FIND RULES WITH HIGH CONFIDENCE

Let R_m=confident rules with m variable consequents

Let H_m=candidate rules with m variable consequents

RuleGeneration (**L**, minconf)

for (k=1; L_k!=∅; k++)

H_1=candidate rules with single variable consequent from L_k

for (m=1; H_m!=∅; m++)

If k > m + 1:

H_{m+1} = generate candidate rules from R_m

R_{m+1} = select candidates in H_{m+1} with minconf

Return $\bigcup_m R_m$

APRIORI ALGORITHM

- Input: data (D), minsup, minconf
- Output: All rules (**R**) with support ≥ minsup and confidence ≥ minconf

Apriori Algorithm (D, minsup, minconf)

% Find all itemsets with support ≥ minsup
L = FrequentItemsetGeneration (D, minsup)

% Find all rules with confidence ≥ minconf
R = RuleGeneration (**L**, minconf)

Return R

EVALUATION

EVALUATION

- Association rules algorithms usually return many, many rules
  - Many are uninteresting or redundant
    (e.g., $ABC \rightarrow D$ and $AB \rightarrow D$ may have same support and confidence)
- How to quantify interestingness?
  - Objective: statistical measures
  - Subjective: *unexpected* and/or *actionable* patterns (requires domain knowledge)

OBJECTIVE MEASURES

- Given a rule X→Y, can compute statistics based on contingency tables

Contingency table for $X \rightarrow Y$

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	$\lvert T \rvert$

f_{11}: support of X and Y
f_{10}: support of X and $\overline{Y}$
f_{01}: support of $\overline{X}$ and Y
f_{00}: support of $\overline{X}$ and $\overline{Y}$

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

DRAWBACK OF SUPPORT

- Support suffers from the **rare item problem** (Liu et al.,1999)
  - Infrequent items not meeting minimum support are ignored which is problematic if rare items are important
  - E.g. rarely sold products which account for a large part of revenue or profit
- Support falls rapidly with itemset size. A threshold on support favors short itemsets

DRAWBACK OF CONFIDENCE

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

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Although confidence is high, rule is misleading

$\Rightarrow$ P(Coffee|$\overline{\text{Tea}}$) = 0.9375

LIFT EXAMPLE

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

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Lift = 0.75/0.9= 0.8333 (< 1, therefore is negatively associated)