IST 707: DATA MINING

Unit 3

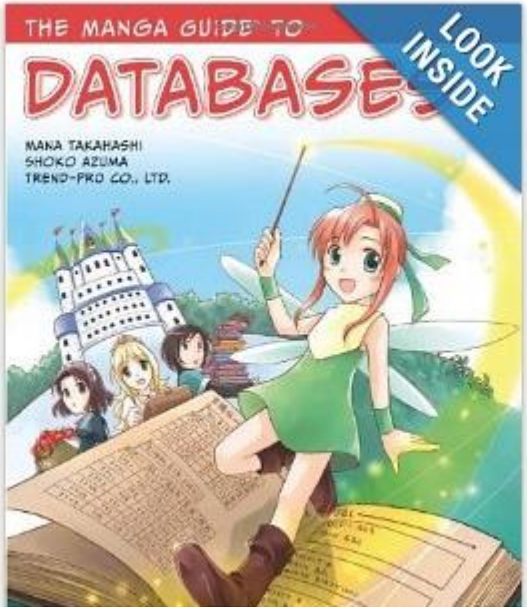
Association Rules

Syracuse University
School of Information Studies

What Is Frequent Pattern Analysis?

Frequent pattern

- What products do people frequently buy together?
- What other products would people buy if they bought a laptop?



The Manga Guide to Databases Paperback

by Mana Takahashi (Author) , Shoko Azuma (Author) , Trend-Pro Co. Ltd. (Author)

31 customer reviews

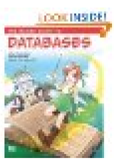
See all 3 formats and editions

Kindle	Library Binding	Paperback
$9.99	$26.06	$13.87
	1 New from $26.06	47 Used from $6.39 44 New from $10.64

Want to learn about databases without the tedium? With its unique combination of Japanese-style comics and serious educational content, *The Manga Guide to Databases* is just the book for you.

Princess Ruruna is stressed out. With the king and queen away, she has to manage the Kingdom of

Frequently Bought Together



+


+


Price for all three: $44.14

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

- [x] **This item:** The Manga Guide to Databases by Mana Takahashi Paperback $13.87
- [x] The Manga Guide to Statistics by Shin Takahashi Paperback $14.76
- [x] The Manga Guide to Linear Algebra by Shin Takahashi Paperback $15.51

Frequently Bought Together

Price for all three: $46.83

Add all three to Cart | Add all three to Wish List

Some of these items ship sooner than the others. Show details

- **This item:** The Manga Guide to Databases by Mana Takahashi Paperback $14.49
- The Manga Guide to Statistics by Shin Takahashi Paperback $15.80
- The Manga Guide to Linear Algebra by Shin Takahashi Paperback $16.54

Customers Who Bought This Item Also Bought

The Manga Guide to Electricity
› Kazuhiro Fujitaki
(24)
Paperback
$14.29 Prime

The Manga Guide to Calculus
› Hiroyuki Kojima
(28)
Paperback
$14.82 Prime

The Manga Guide to Physics
› Hideo Nitta
(31)
Paperback
$14.59 Prime

The Manga Guide to Statistics
› Shin Takahashi
(38)
Paperback
$15.80 Prime

Association Rule Mining

Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Association Rule (AR) Mining

Chapter 6 in the Tan textbook provides some background knowledge but may require some computer science background.

Requirement for this class: learn the basic concepts about AR and the main idea of the Apriori algorithm.

More applications

Product recommendation

- E.g. Amazon.com

Catalog design

Web log (click stream) analysis

DNA sequence analysis

Basic concepts in AR mining

Frequent itemset

Transaction-id	Items bought *
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

Can you answer the following questions?
- Which two items are frequently bought together?
- Which three items are often bought together?
- …

Definition: Frequent Itemset

Itemset

- a collection of one or more items
- k-itemset contains k items

1-itemset:

- {A}:3, {B}:3, {C}:2, {D}:4, {E}:3, {F}:2

2-itemset:

- {A,B}:1; {A,D}:3

3-itemset:

- {A,B,C}:0, {B, E, F}:2

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

Metrics to evaluate frequent level of itemsets

How frequent is an itemset?

Support count

- Number of transactions that contain an itemset
  - support_count({ D, E }) = 2

Support percentage

- Fraction of transactions that contain an itemset
  - support({D, E}) = 2/5

Frequent Itemset

- an itemset with support >= threshold

Definition: Association Rule

Association Rule

an implication of the form X → Y,

where X and Y are itemsets,

e.g. { E, F } → { B }

Example Rules:

LHS Left-Hand Side

{B, E} → {F}
{E, F} → {B}
{B, F} → {E}
{B} → {E, F}
{E} → {B, F}
{F} → {B, E}

RHS Right-Hand Side

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

Metrics to evaluate the rule's strength

Rule Evaluation Metrics

Support = P(X, Y)

Fraction of transactions that contain both X and Y

Support({E, F} → {B}) = support_count({B,E,F}) / N = 2/5

Confidence = P(Y|X)=P(X, Y)/P(X)

How frequently items in Y appear in transactions that contain X

confidence({E,F} → {B})
= support({B,E,F}} / support({E,F})
= support_count({B,E,F}) / support_count({E,F})
= 2/2 = 1

Confidence

confidence({E,F} → {B})

= P({B}|{E,F})

= support({B,E,F}} / support({E,F})

= support_count({B,E,F}) /support_count({E,F})

= 2/2 = 1

confidence({B}→ {E,F})

= P({E,F}|{B})

= support({B,E,F}} / support({B})

= support_count({B,E,F}) /support_count({B})

= 2/3 = .67

Switching LHS and RHS results in different rules with different confidence

Exercise: AR metrics

Calculate the support and confidence of association rules {A}->{D} and {D}->{A}

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

Apriori algorithm

How to mine association rules?

Given a set of transactions T, the goal of association rule mining is to find all rules having

support ≥ *minsup* threshold

confidence ≥ *minconf* threshold

Brute-force approach:

List all possible association rules

Compute the support and confidence for each rule

Prune rules that fail the *minsup* and *minconf* thresholds

⇒ Computationally prohibitive! (Why? Hint: calculate the number of subsets!)

Mining Association Rules

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} $\rightarrow$ {Beer} (s=0.4, c=0.67)
{Milk,Beer} $\rightarrow$ {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} $\rightarrow$ {Milk} (s=0.4, c=0.67)
{Beer} $\rightarrow$ {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} $\rightarrow$ {Milk,Beer} (s=0.4, c=0.5)
{Milk} $\rightarrow$ {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset:
  {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Mining Association Rules

Two-step approach:

1. Frequent Itemset Generation
   – Generate all itemsets whose support $\geq$ minsup

2. Rule Generation
   – Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

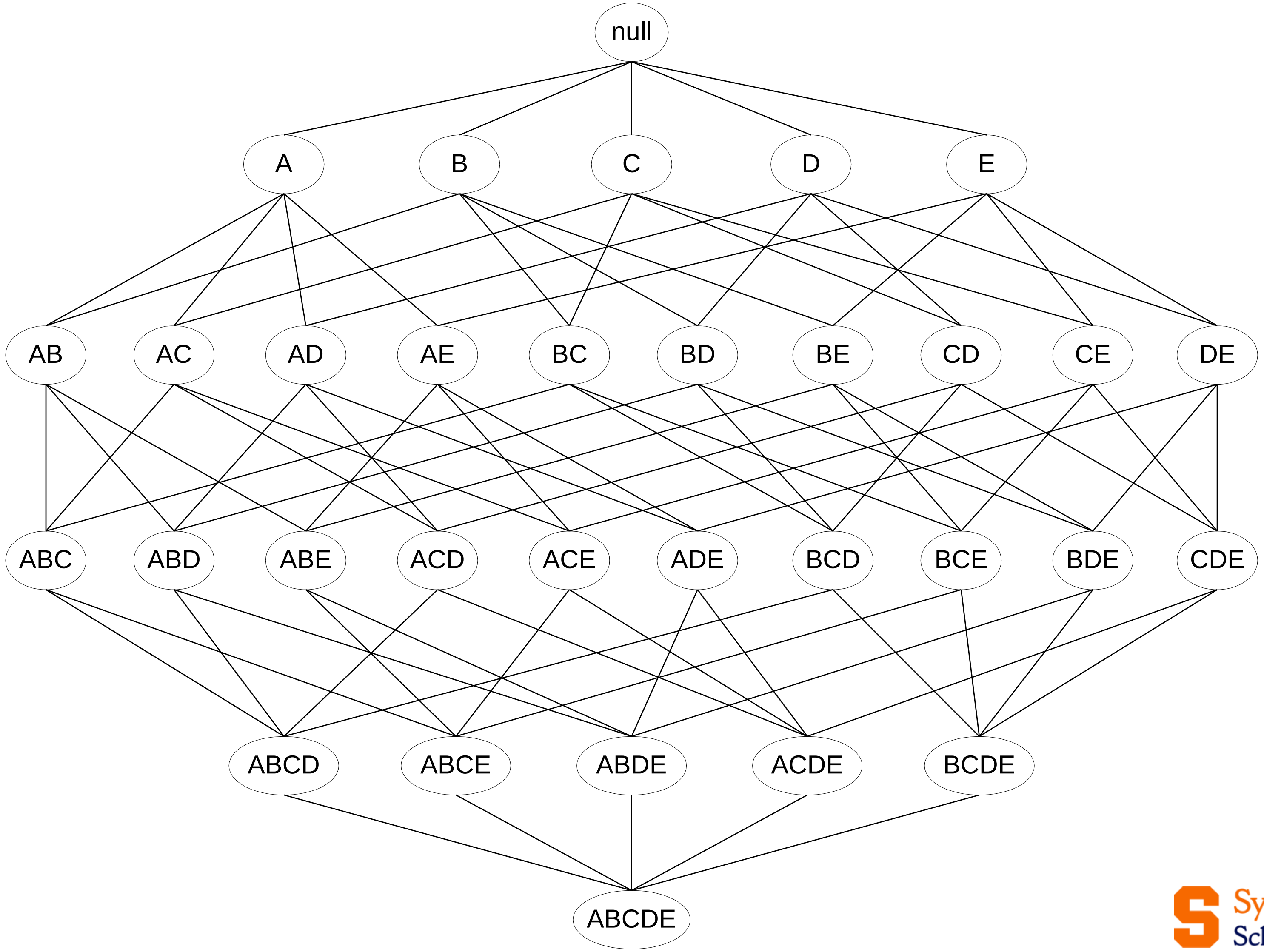
Frequent itemset generation is still computationally expensive

Scalable Methods for Mining Frequent Patterns

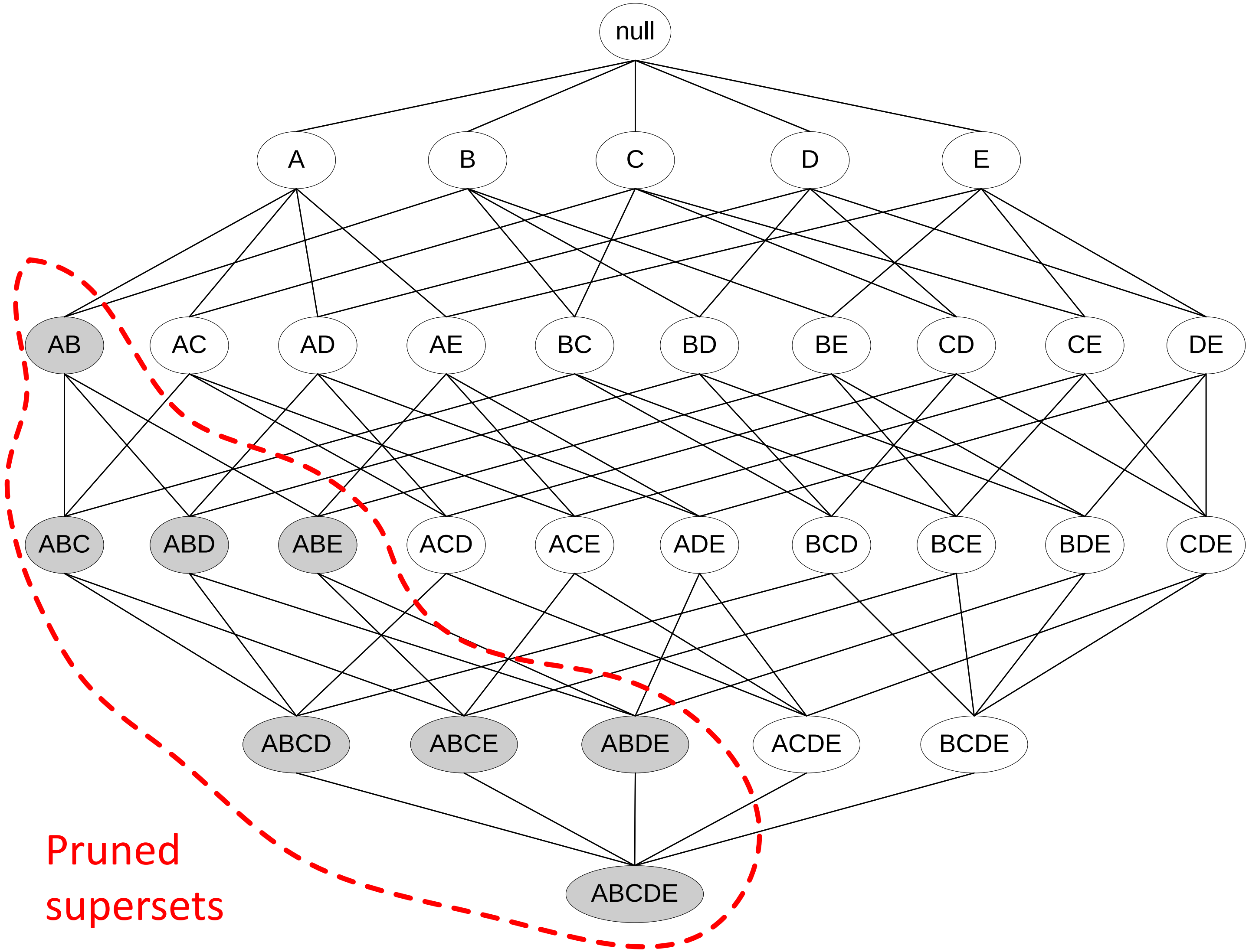
Scalable mining methods: Three major approaches

- Apriori (Agrawal & Srikant@VLDB’94)
- Freq. pattern growth (FPgrowth—Han, Pei & Yin @SIGMOD’00)
- Vertical data format approach (Charm—Zaki & Hsiao @SDM’02)

Frequent Itemset Generation



Illustrating Apriori Principle



Found to be Infrequent

Pruned supersets

Apriori: A Candidate Generation-and-Test Approach

Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!

Method:

- Initially, scan DB once to get frequent 1-itemset
- Generate length (k+1) candidate itemsets from length k frequent itemsets
- Test the candidates against DB
- Terminate when no frequent or candidate set can be generated

The Apriori Algorithm—Generate Frequent Itemset

$Sup_{min} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Rule Generation

Given a frequent itemset L, find all non-empty subsets f, such that f → (L – f) satisfies the minimum confidence requirement

- If {A, B, C, D} is a frequent itemset, candidate rules:
  - ABC → D,   ABD → C, ACD → B, BCD → A
  - AB → CD, AC → BD, …
  - A → BCD, B → ACD, C→ ABD, D → ABC
- Compute the confidence for each rule, and keep the ones that are greater than min_conf

Rule Generation

How to efficiently generate rules from frequent itemsets?

Start from long LHS

- For itemset {ABCD}, c(x) means confidence of rule x
- c(ABC → D) >= c(AB → CD) >= c(A → BCD)

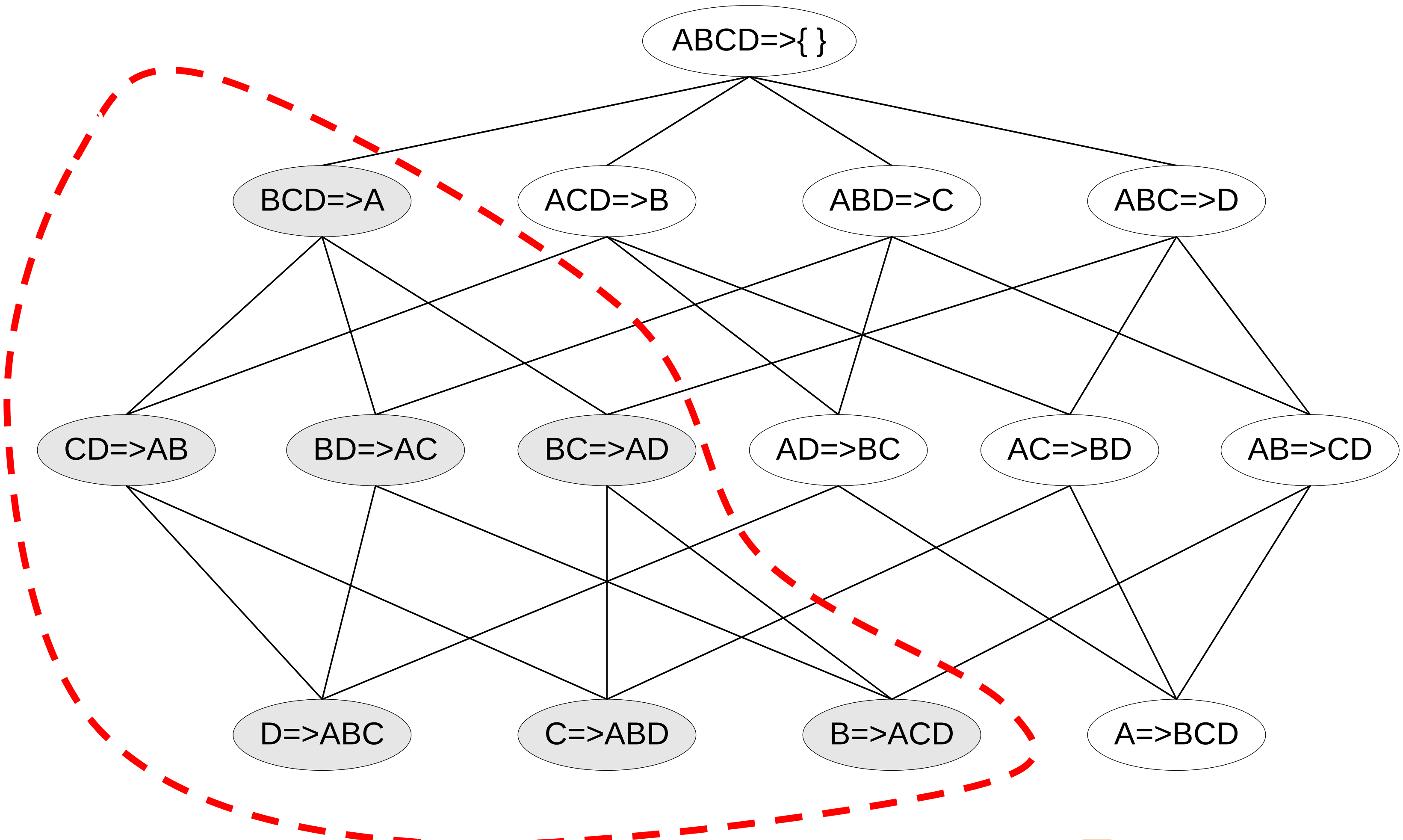
Proof

- C(ABC->D) = support(ABCD)/support(ABC)
- C(AB->CD) = support (ABCD)/support (AB)
- support(AB) >= support (ABC)
- So C(ABC->D) >= C(AB->CD)

if min_conf is not satisfied, no need to generate rules with larger right-hand-side (RHS)

The Apriori Algorithm: Rule pruning

Lattice of rules



Exercise: Apriori algorithm

Use your own words to explain why starting from longest LHS is an efficient method to generate association rules.

Limitation of confidence measure

100 transactions

75 bought movies

60 bought games

40 bought both

{movies}->{games}
support 40/100=0.4
confidence 40/75=0.53

{games}->{movies}
support 40/100=0.4
confidence 40/60=0.67

Both seem to be strong rules

However,

100 transactions

75 bought movies

60 bought games

40 bought both

P(movies) = 75/100=0.75

P(games) = 60/100=0.6

Expected: P(movies & games) = .6 * .75 = .45

Observed: S(movies & games) = .4

So the actual support for movies & games is less than would be expected if the two events were independent – ergo, people tend *not* to buy them together!

Metric: Lift

Measure of dependent/correlated events: lift

Lift is the ratio of the confidence of the rule to the expected confidence

OR

The ratio of the observed percentage to the unconditioned joint probability

$$\text{Lift } (A \Rightarrow B) = S(A \Rightarrow B) / S(A)*S(B)$$

Association rules should have >1 lift to be meaningful

The Lift Measure

	Game	Not game	total
Movie	40	35	75
Not movie	20	5	25
total	60	40	100

S(buy game)=0.6
S(not buy movie) =0.25
S(buy game -> not buy movie) =0.20
Lift (buy game -> not buy movie)
=0.20/(0.6*0.25)=1.33 > 1

Strong rule!

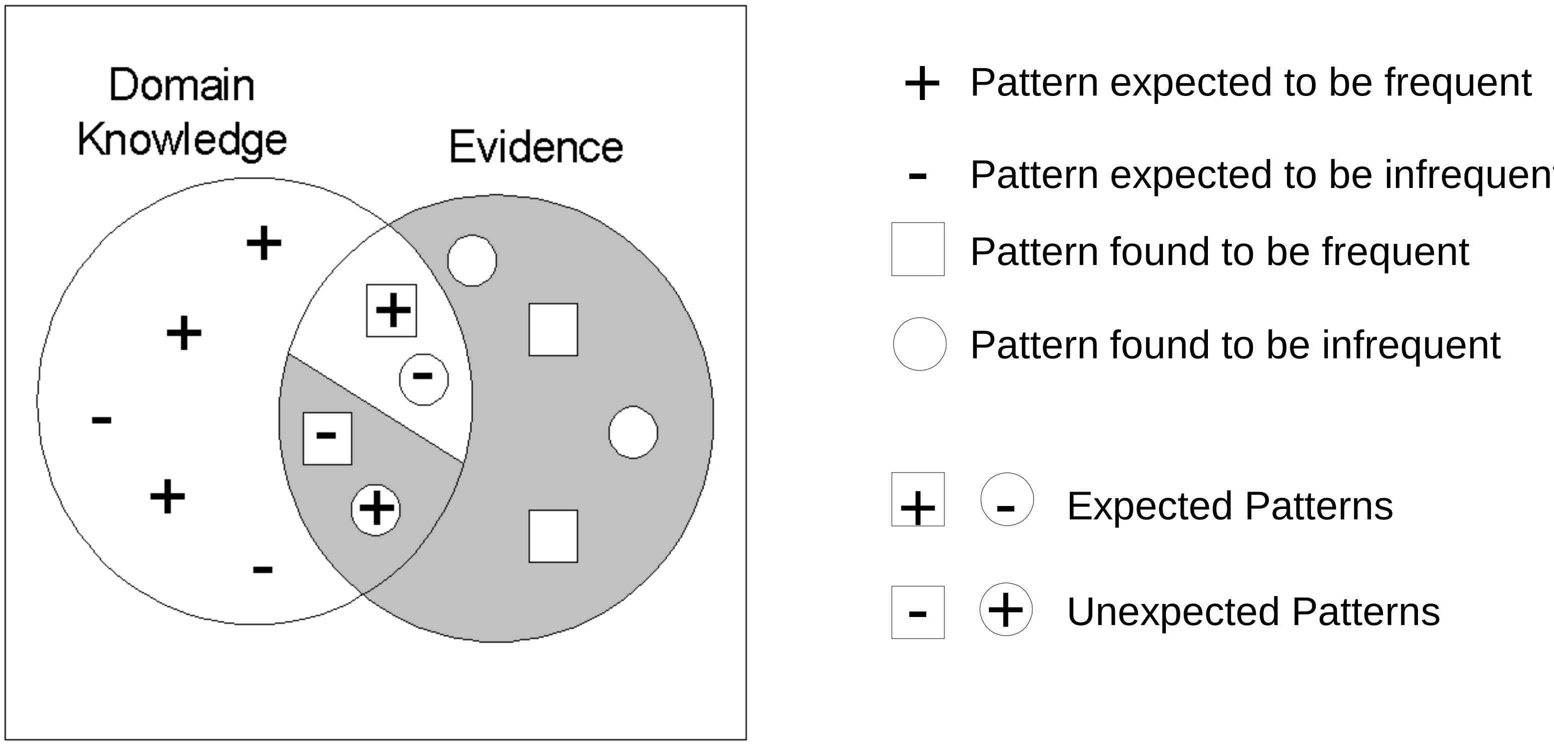
Alternative measures

Association rule algorithms tend to produce too many rules

- many of them are uninteresting or redundant
- Uninteresting if it is known knowledge
- Redundant if $\{A,B,C\} \rightarrow \{D\}$ and $\{A,B\} \rightarrow \{D\}$ have same support & confidence

Interestingness via Unexpectedness

Need to model expectation of users (domain knowledge)



Need to combine expectation of users with evidence from data (i.e., extracted patterns)

Association Rule Measures

In practice, what levels of support, confidence and lift should we aim for?

Support

- Depends on dataset and business problem
- Common setting: 20-40% of the transactions

Confidence

- Strong confidence rules >=.9, but .6 to .8 range might be o.k.

Lift

- should be above 1.0, the higher the better
- Levels of 2 and above can occasionally be seen, but more likely to see around 1.3 – 1.5

Exercise: calculate lift

Calculate the lift of rules {A}->{D} and {D}->{A}

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F