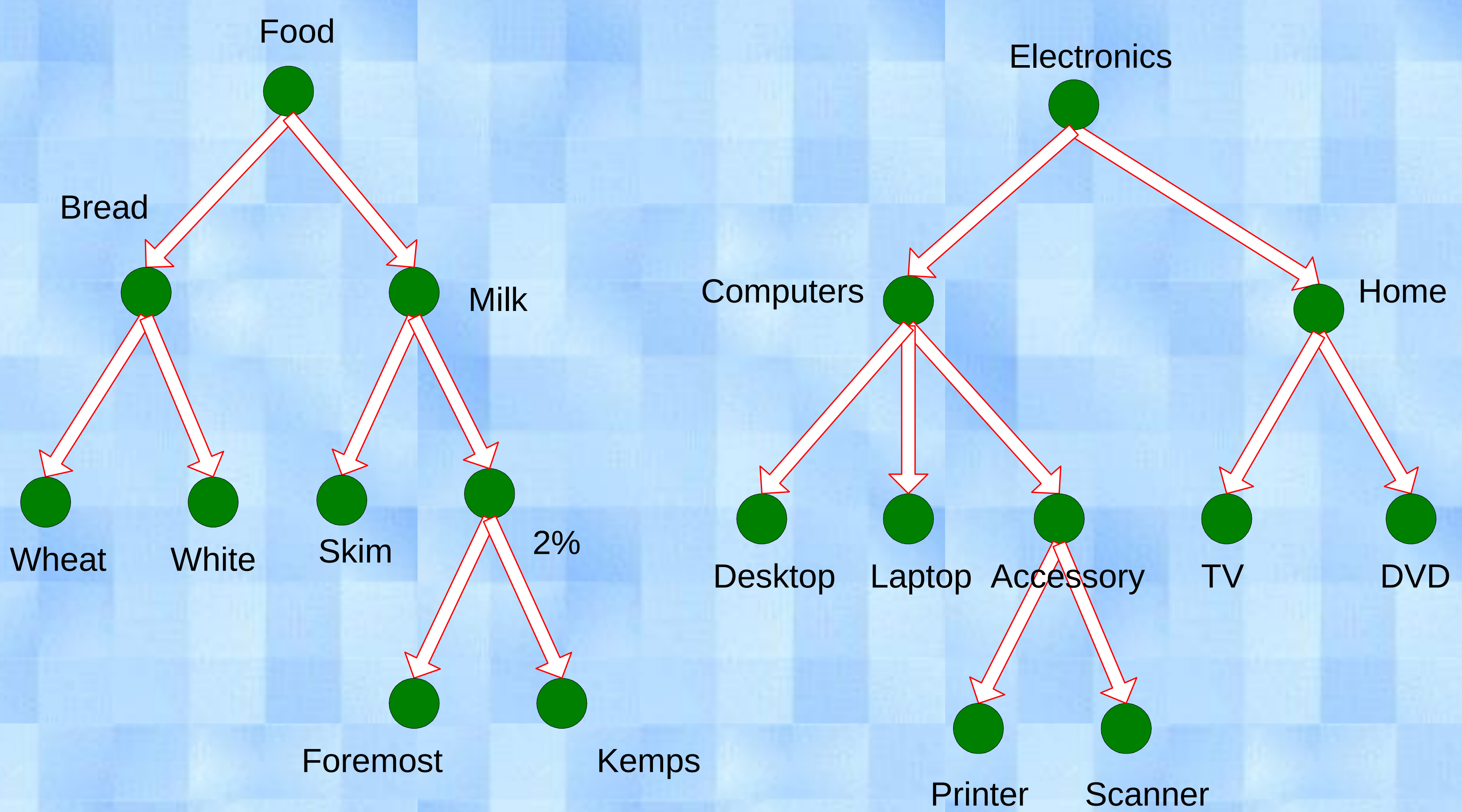
Association Analysis
Part 3

Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida

Multi-level Association Rules



Negative Association

- When do infrequent patterns become interesting?
  - Negative correlation:
    - $P(A,B) < P(A)P(B)$
    - e.g: Windows vs Linux
  - Negative association rules: ($\neg A \rightarrow B$):
    - $P(\neg A, B) = P(B) - P(A,B)$
    - e.g: $\neg$Regular $\rightarrow$ Diet (s=0.17, c=0.25)

Coke	Diet	¬Diet	
Regular	1	32	33
¬Regular	17	50	67
	18	82	100

Approach 1: Using Negative Items

Tid	A	¬A	B	¬B	C	¬C	D	¬D
1	1	0	0	1	1	0	0	1
2	1	0	0	1	0	1	0	1
3	1	0	0	1	1	0	0	1
4	1	0	1	0	0	1	0	1
5	1	0	0	1	0	1	1	0

- Computationally expensive
- Tends to produce many uninteresting negative associations

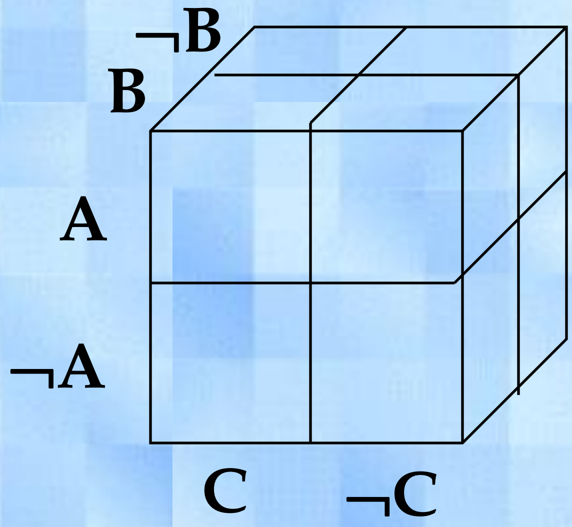
Approach 1: Using Negative Items

Size 2:

	B	¬B
A	10	320
¬A	170	500

Support of
{A,B}, {A,¬B} and {¬A,B}
can be large

Size 3:



Approach 2: Using Positive Itemsets

- Boulicaut et al [2000]:
  - Compute support of negative itemsets based on the support of positive itemsets
    - e.g. X = Y ∪ ¬Z

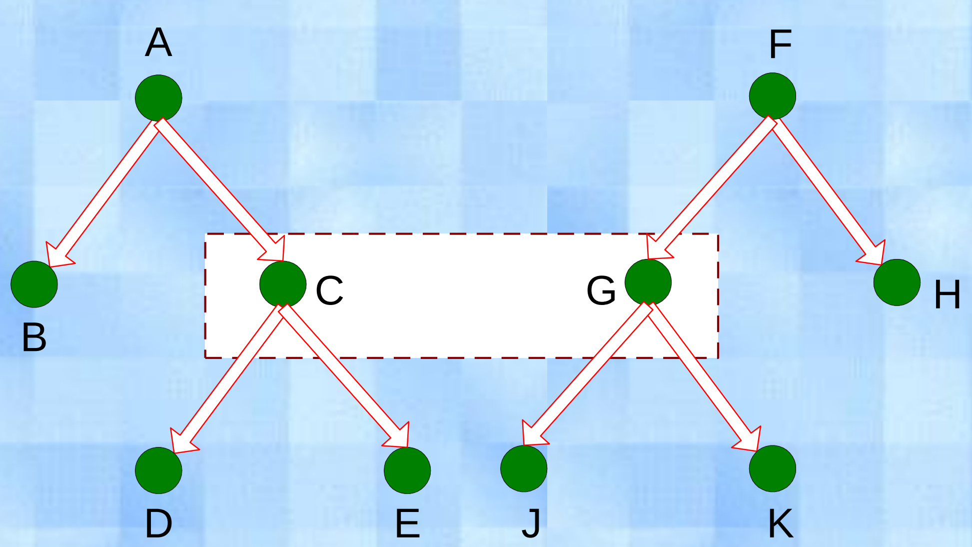
$$P(X) = \sum_{Y \subseteq I \subseteq (Y \cup Z)} (-1)^{|I|-|Y|} P(I)$$

    - e.g.: $P(AB\overline{CD}) = P(AB) - P(ABC) - P(ABD) + P(ABCD)$

  - To use this formula:
    - Need to use a very low support threshold, or
    - Use approximation

Approach 3: Using Domain Knowledge

- Compute expected support using item taxonomy



Suppose C and G are frequent:

$$Exp(\sup(EJ)) = \frac{\sup(CG) \times \sup(E) \times \sup(J)}{\sup(C) \times \sup(G)}$$

$$Exp(\sup(CJ)) = \frac{\sup(CG) \times \sup(J)}{\sup(G)}$$

$$Exp(\sup(CH)) = \frac{\sup(CG) \times \sup(H)}{\sup(G)}$$

- There could be multiple taxonomies defined (based on type, brand, size, etc.)
- Limited to the nodes that are directly connected to the frequent itemsets

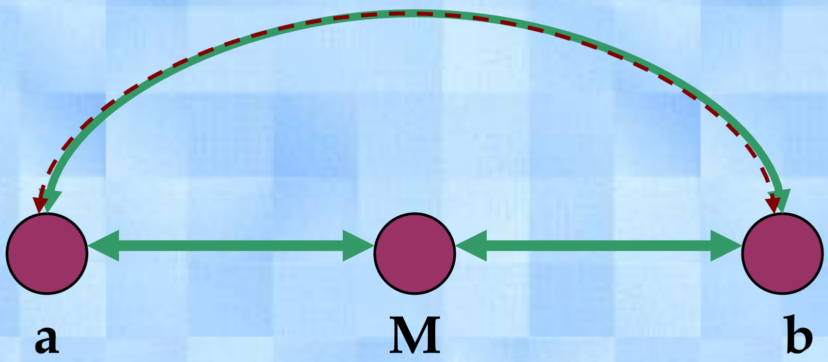
Approach 3: Using Domain Knowledge

- A negative itemset is a set of items whose actual support is significantly lower than its expected support
- Negative association rule: $X \Rightarrow Y$
- Rule interest measure:

$$RI = \frac{Exp(P(X \cup Y)) - P(X \cup Y)}{P(X)}$$

- Approach:
  - Find frequent itemsets at each level of the taxonomy
  - Identify candidate negative itemsets based on the frequent itemsets found and their item taxonomy
  - Count actual support of candidate itemsets and retain only the negative itemsets
  - Generate negative association rules from negative itemsets

Approach 4: Indirect Association



IF: a and M are frequent and highly dependent on each other
b and M are frequent and highly dependent on each other

THEN: a and b are expected to be frequent

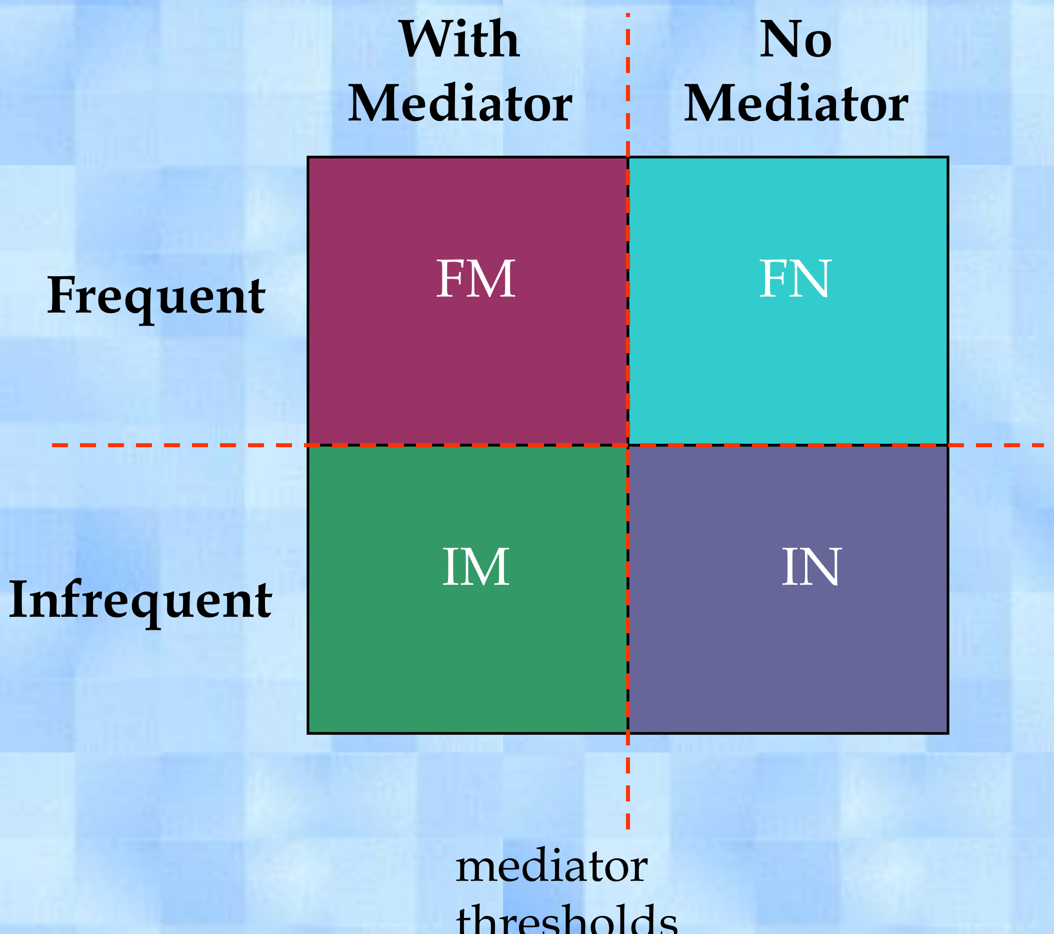
If a and b are infrequent, there is an interesting negative association

⇒ a and b are indirectly associated via mediator M

⇒ M identifies the context of negative association

Finding Interesting Negative Association

For all pairs of items:



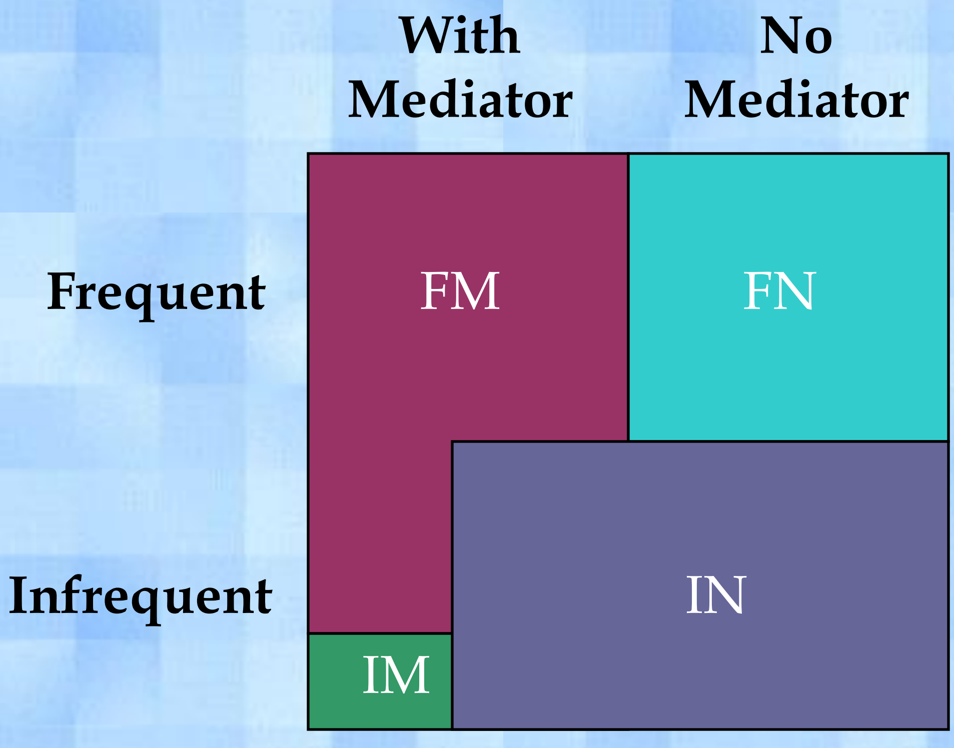
Minimum itempair support

If

$$IM/FM = IN/FN$$

then Indirect Association is not surprising

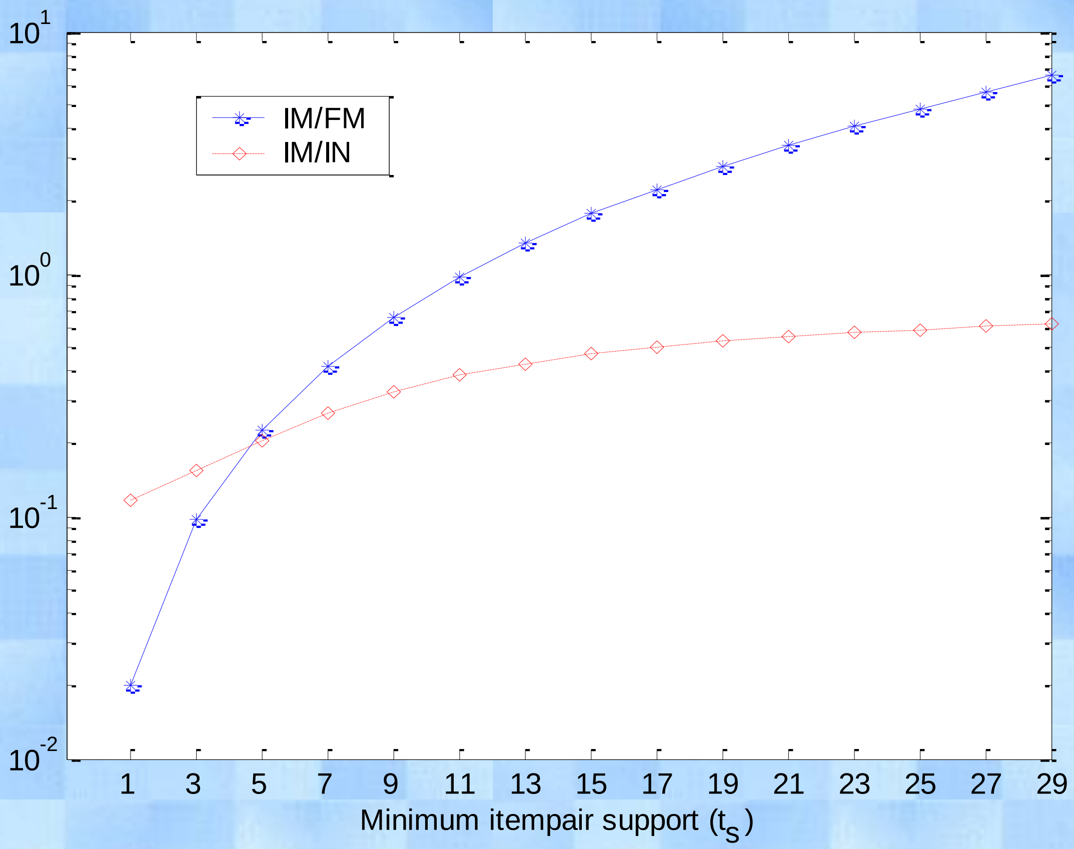
Finding Interesting Negative Associations



- IM/FM is small
- IM/IN is small

$\Rightarrow$ Indirect Association is interesting

Finding Interesting Negative Association

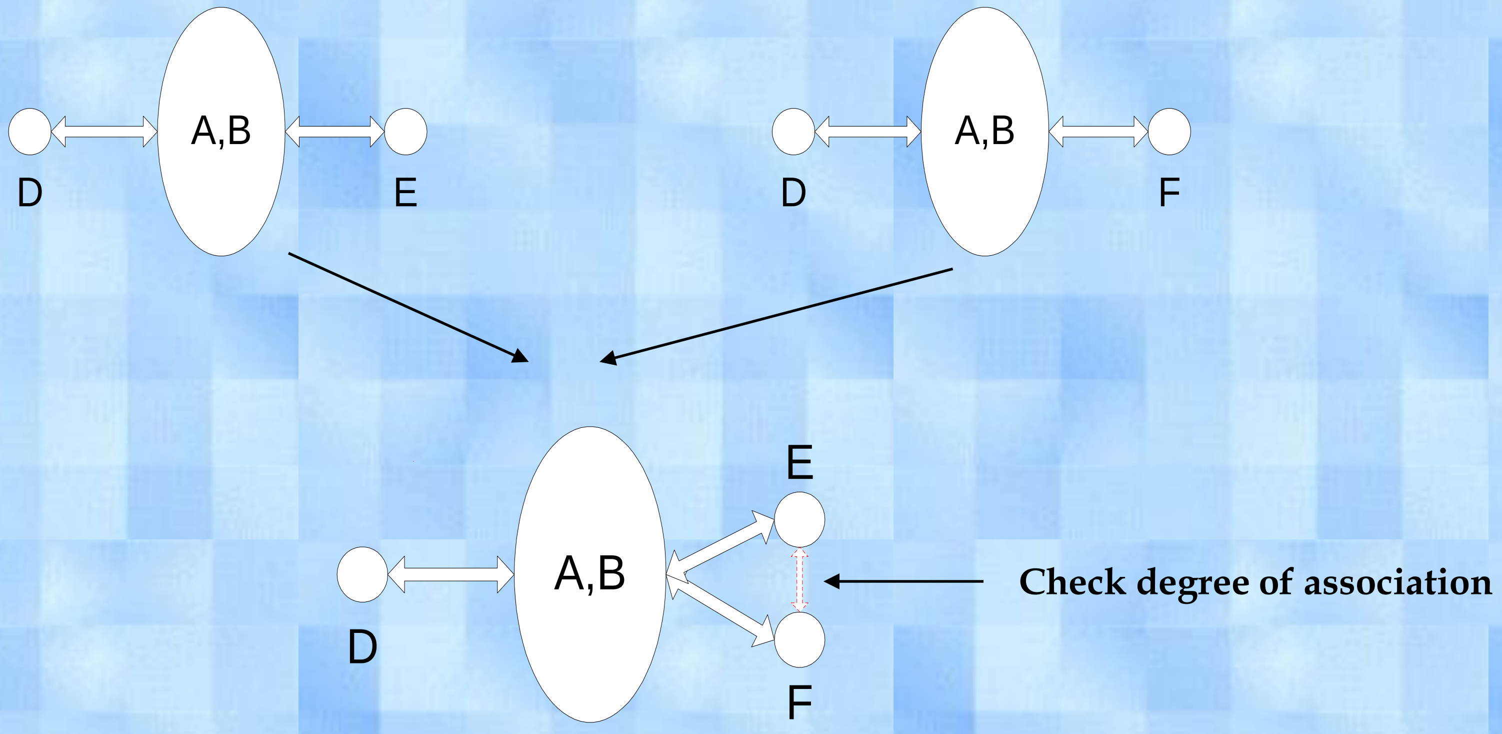


Indirect Association is interesting when minimum itempair support threshold is small.

But, if threshold is too low, very few indirect associations are obtained.

Grouping Indirect Association

- Indirect associations can be grouped together into more compact structures if they have same mediator



Mining Indirect Associations

Market-basket Data

Transaction Id	Items
1	{A,B,C,D}
2	{A,B,E}
3	{B,C}
4	{A,B,D,E}
5	{B,C,D}

Itempair Support Matrix (Frequent pairs are shaded)

	A	B	C	D	E
A	3	3	1	2	2
B	3	5	3	3	2
C	1	3	3	2	0
D	2	3	2	3	1
E	2	2	0	1	2

Frequent 3-itemset

Pattern	Support
{A,B,D}	2
{A,B,E}	2
{B,C,D}	2

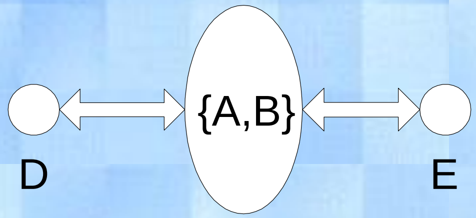
Join step

Candidate Indirect Associations

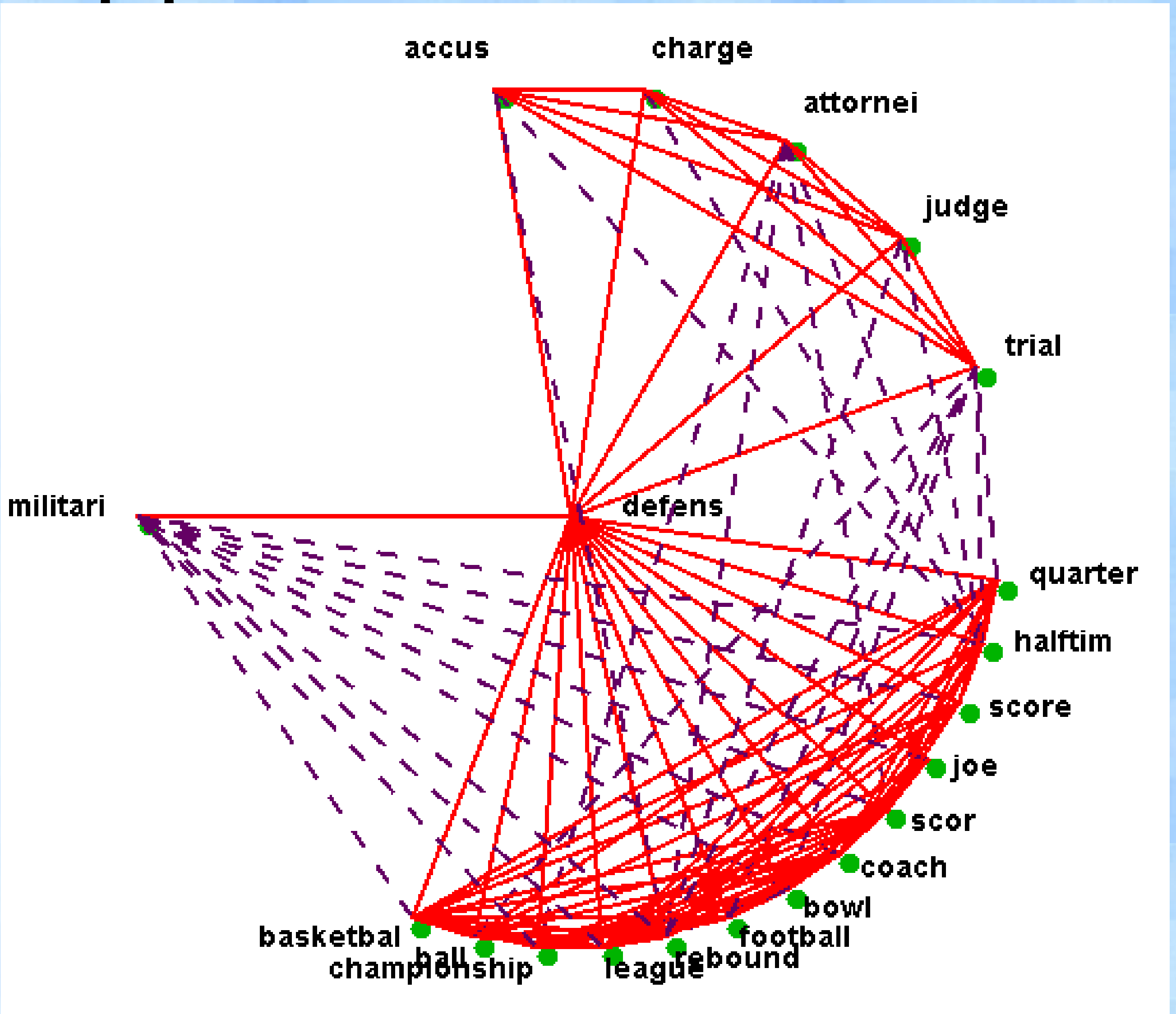
Itempair	Mediator
{D,E}	{A,B}
{A,C}	{B,D}

Prune step

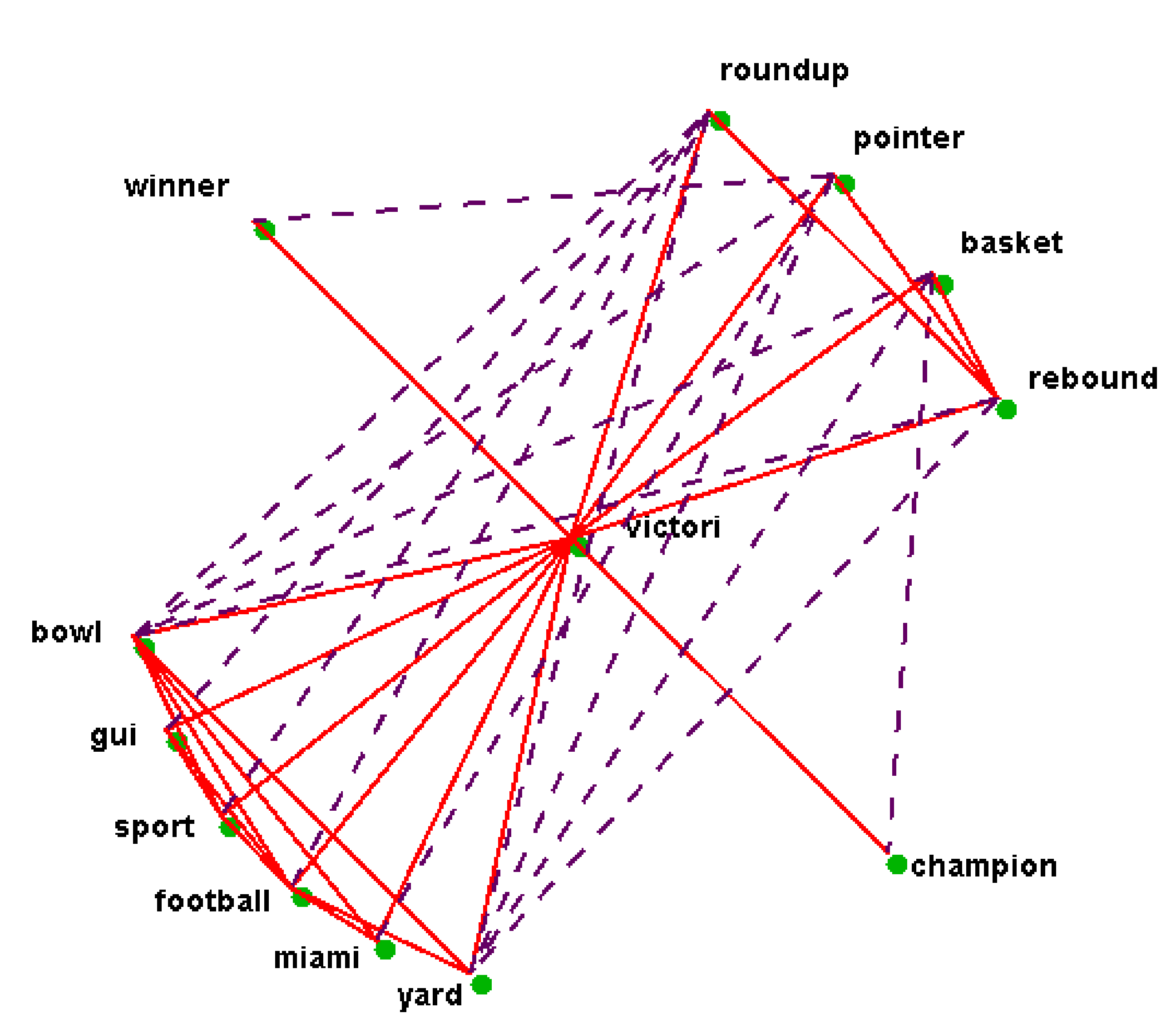
Indirect Association



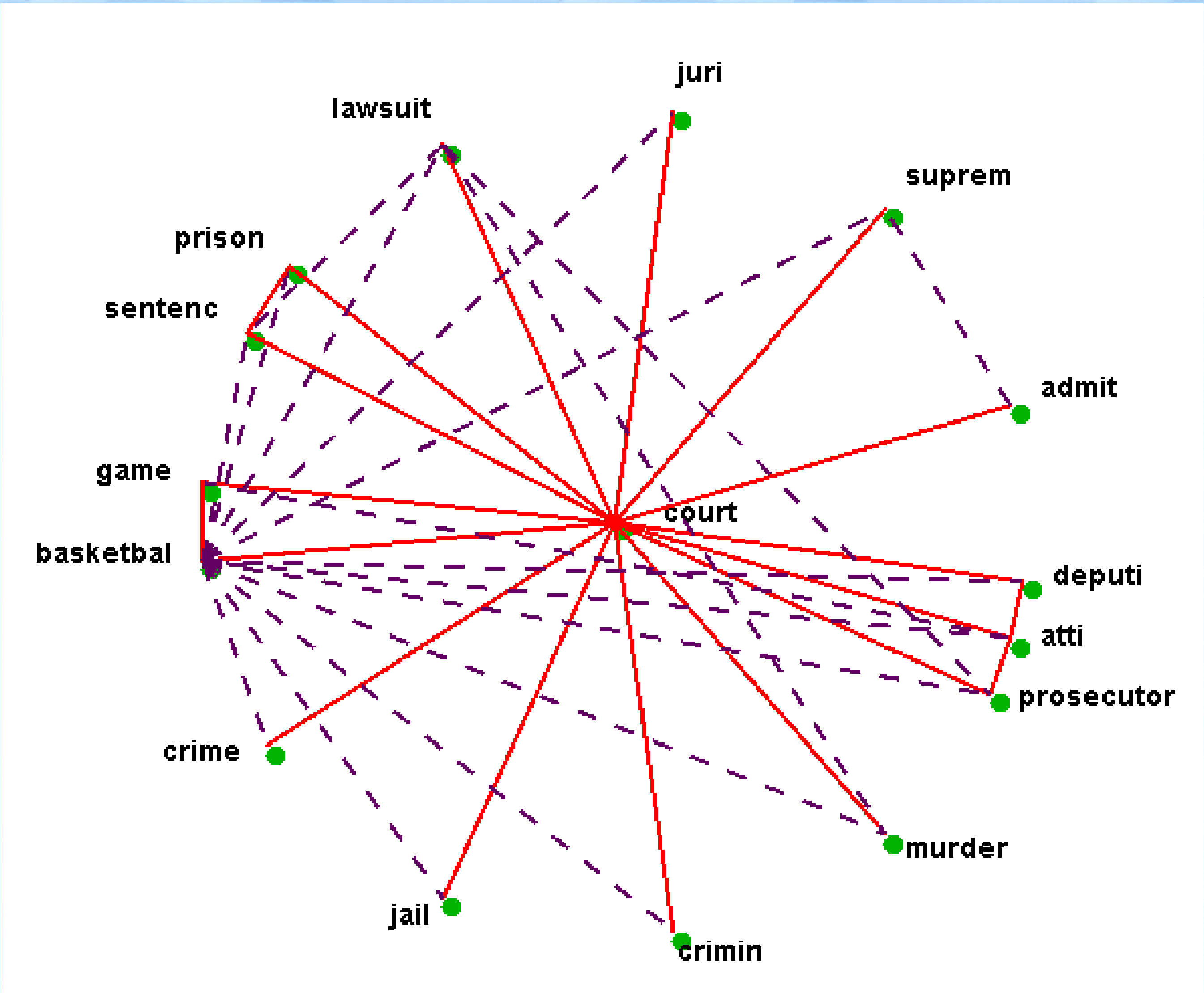
Applications: LA Times



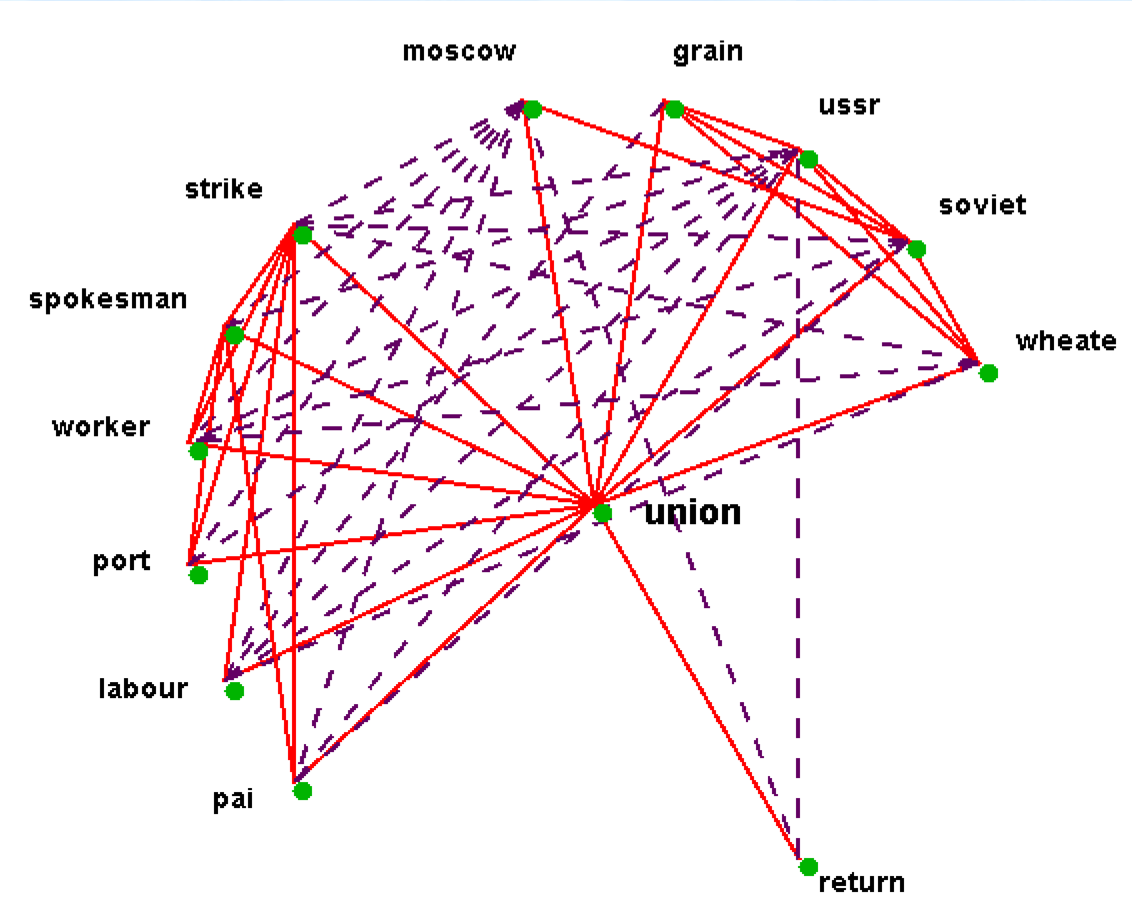
Application: LA Times



Application: LA Times

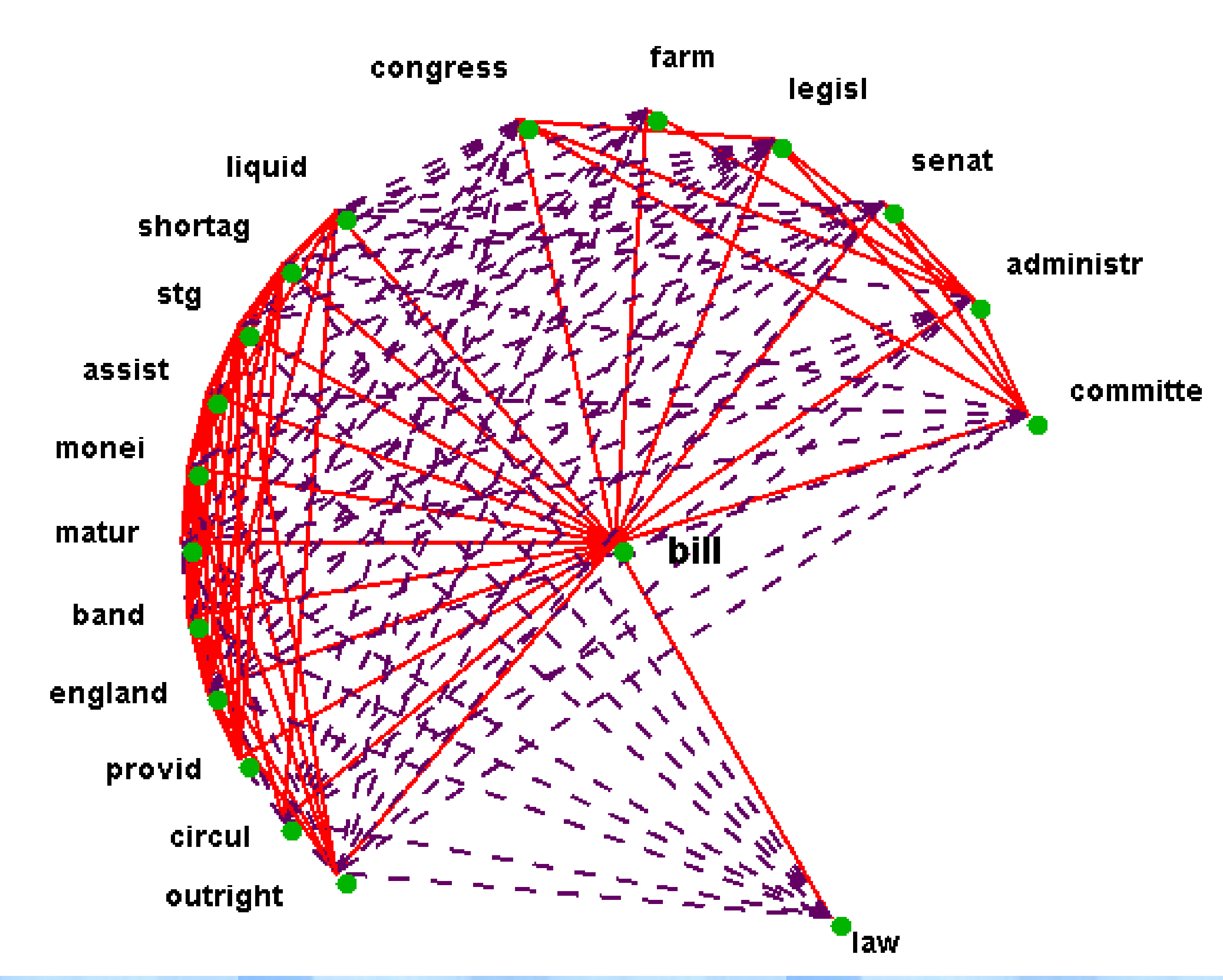


Application: Reuter-21758 news

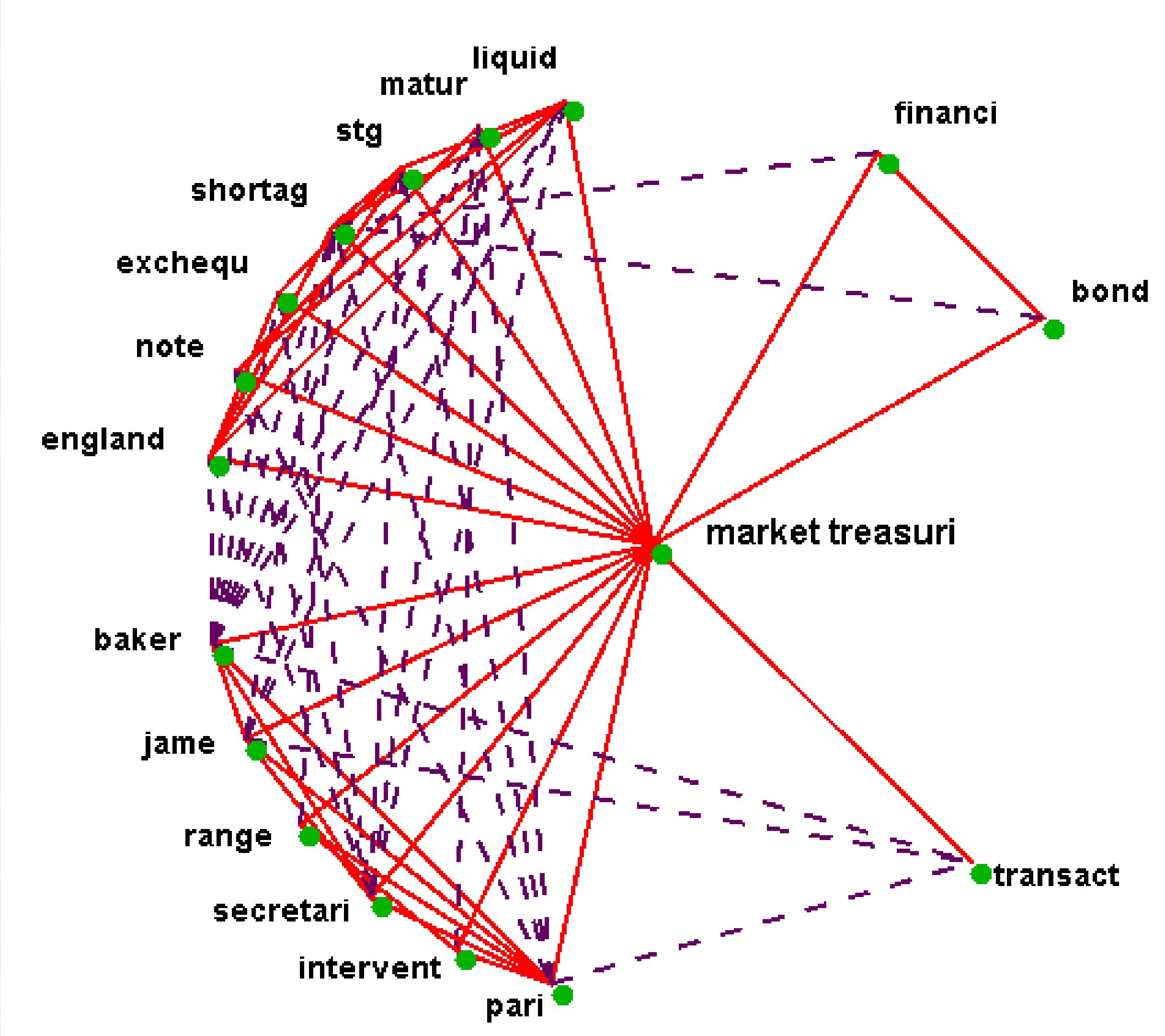


Indirect association can identify different contexts of a word

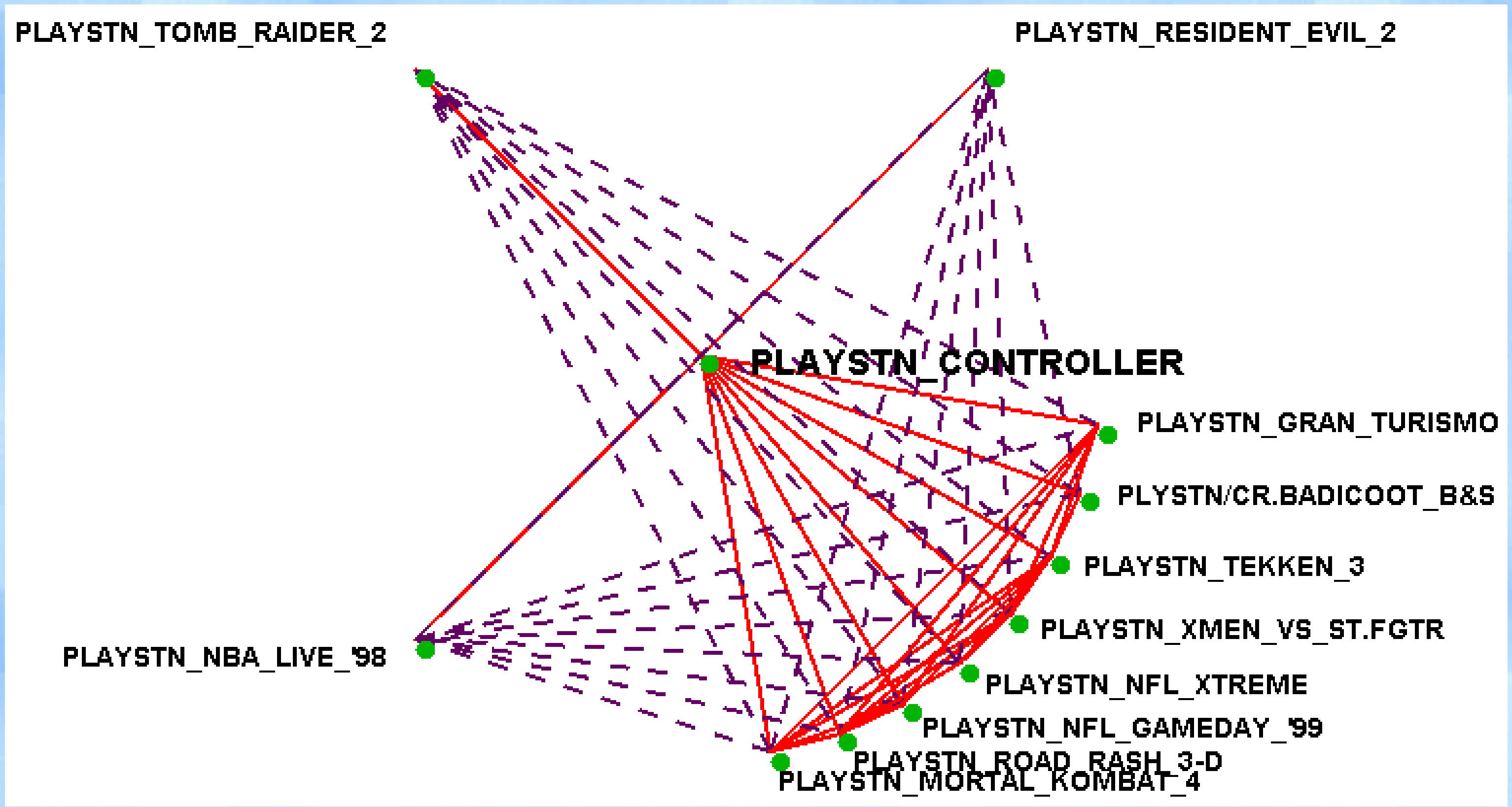
Application: Reuter-21758 news



Application: Reuter-21758 news

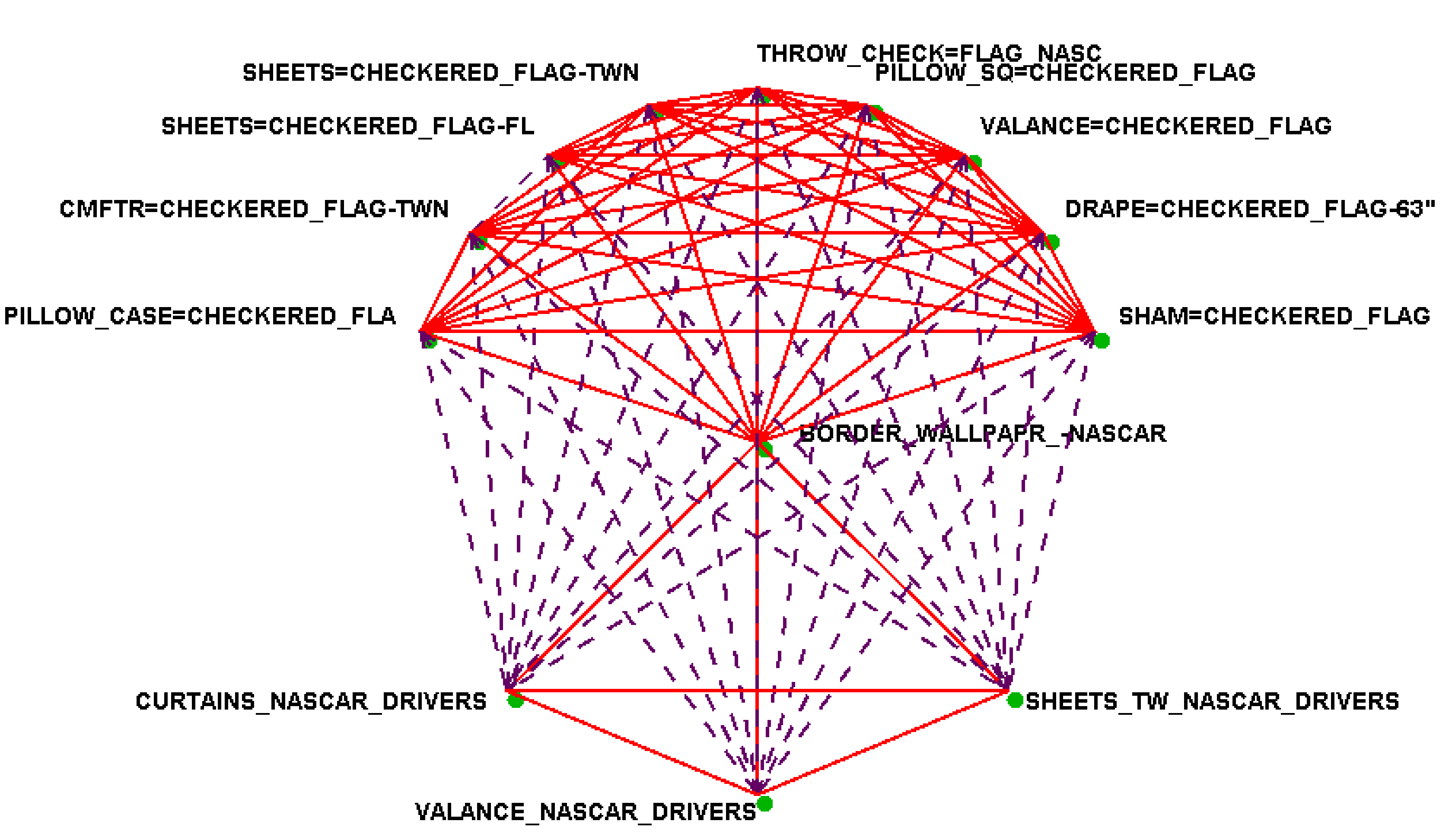


Application: Retail Data



Indirect association can identify competing and (sometimes) complimentary items

Application: Retail Data



Note: There is no checked-flag border wallpaper

Mining Continuous Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

?

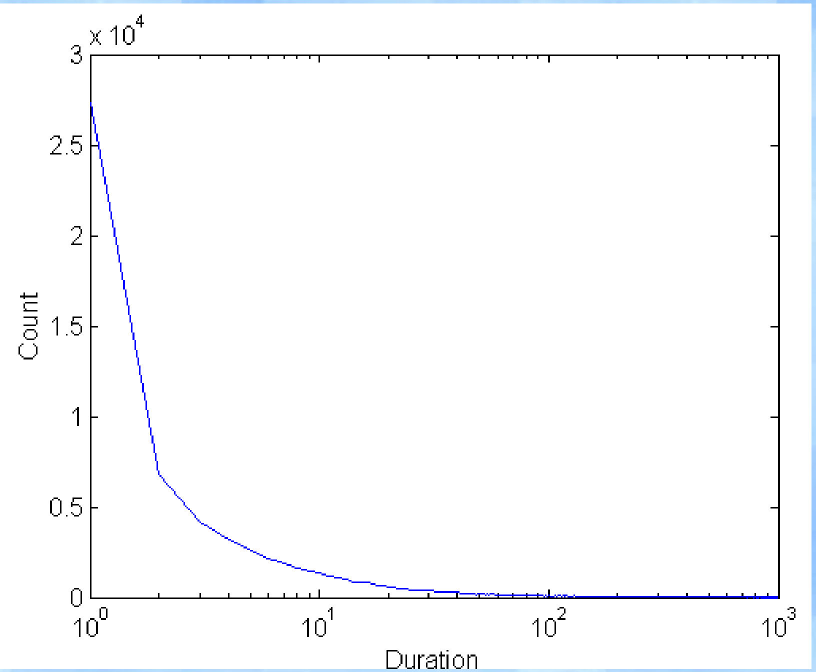
Tid	A	B	C	D	E
1	1	0	0	1	1
2	1	0	0	1	0
3	1	0	0	1	1
4	1	0	1	0	0
5	1	0	0	1	0

Example:

$$\{\text{Refund} = \text{No}, (60K \leq \text{Income} \leq 80K)\} \rightarrow \{\text{Cheat} = \text{No}\}$$

Discretize Continuous Attributes

- Unsupervised:
  - Equal-width binning
  - Equal-depth binning
  - Clustering



- Supervised:

Attribute values, v

Class	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
Anomalous	0	0	20	10	20	0	0	0	0
Normal	150	100	0	0	0	100	100	150	100

$\underbrace{v_1\ v_2}_{bin_1}$ $\underbrace{v_3\ v_4\ v_5}_{bin_2}$ $\underbrace{v_6\ v_7\ v_8\ v_9}_{bin_3}$

Discretization Issues

- Size of the discretized intervals affect support & confidence

$$\{Refund = No, (Income = \$51,250)\} \rightarrow \{Cheat = No\}$$

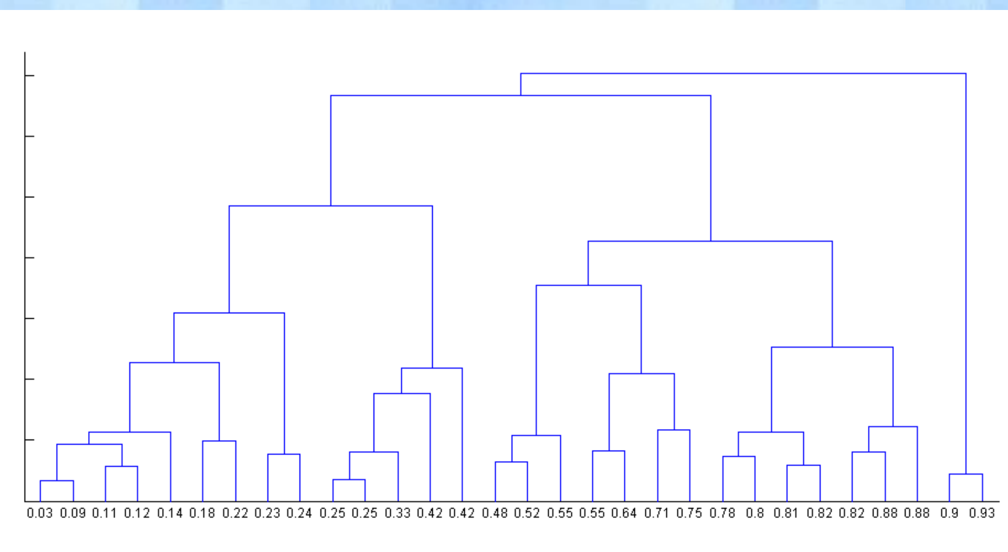
$$\{Refund = No, (60K \leq Income \leq 80K)\} \rightarrow \{Cheat = No\}$$

$$\{Refund = No, (0K \leq Income \leq 1B)\} \rightarrow \{Cheat = No\}$$

  - If intervals too small
    - may not have enough support
  - If intervals too large
    - may not have enough confidence

Discretization Issues

- Execution time
  - If intervals contain n values, there are on average $O(n^2)$ possible ranges



- Too many rules

{Refund = No, (Income = $51,250)} → {Cheat = No}

{Refund = No, (51K ≤ Income ≤ 52K)} → {Cheat = No}

{Refund = No, (50K ≤ Income ≤ 60K)} → {Cheat = No}

Approach by Srikant & Agrawal

- Discretize attribute using equi-depth partitioning
  - Use *partial completeness measure* to determine number of partitions

C: frequent itemsets obtained by considering all ranges of attribute values
P: frequent itemsets obtained by considering all ranges over the partitions

P is *K-complete* w.r.t C if $P \subseteq C$,and $\forall X \in C$, $\exists X' \in P$ such that:

1. X' is a generalization of X and support $(X') \leq K \times$ support(X) $(K \geq 1)$
2. $\forall Y \subseteq X$, $\exists Y' \subseteq X'$ such that support $(Y') \leq K \times$ support(Y)

Given *K (partial completeness level),* can determine number of intervals (N)

- Merge adjacent intervals as long as support is less than max-support
- Apply existing association rule mining algorithms
- Determine interesting rules in the output

Interestingness Measure

$\{\text{Refund} = \text{No}, (\text{Income} = \$51{,}250)\} \rightarrow \{\text{Cheat} = \text{No}\}$

$\{\text{Refund} = \text{No}, (51\text{K} \leq \text{Income} \leq 52\text{K})\} \rightarrow \{\text{Cheat} = \text{No}\}$

$\{\text{Refund} = \text{No}, (50\text{K} \leq \text{Income} \leq 60\text{K})\} \rightarrow \{\text{Cheat} = \text{No}\}$

- Given an itemset: $Z = \{z_1, z_2, \ldots, z_k\}$ and its generalization $Z' = \{z_1', z_2', \ldots, z_k'\}$

  P(Z): support of Z

  $E_{Z'}(Z)$: expected support of Z based on Z′

$$E_{Z'}(Z) = \frac{P(z_1)}{P(z_1')} \times \frac{P(z_2)}{P(z_2')} \times \cdots \times \frac{P(z_k)}{P(z_k')} \times P(Z')$$

  - Z is R-interesting w.r.t. Z′ if $P(Z) \geq R \times E_{Z'}(Z)$

Interestingness Measure

- For S: $X \rightarrow Y$, and its generalization S′: $X' \rightarrow Y'$

  $P(Y \mid X)$: confidence of $X \rightarrow Y$
  $P(Y' \mid X')$: confidence of $X' \rightarrow Y'$
  $E_{S'}(Y \mid X)$: expected support of Z based on Z′

$$E(Y \mid X) = \frac{P(y_1)}{P(y_1')} \times \frac{P(y_2)}{P(y_2')} \times \cdots \times \frac{P(y_k)}{P(y_k')} \times P(Y' \mid X')$$

- Rule S is R-interesting w.r.t its ancestor rule S′ if
  - Support, $P(S) \geq R \times E_{S'}(S)$ or
  - Confidence, $P(Y \mid X) \geq R \times E_{S'}(Y \mid X)$

Min-Apriori (Han et al)

Document-term matrix:

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

Example:

W1 and W2 tends to appear together in the same document

Min-Apriori

- Data contains only continuous attributes of the same "type"
  - e.g., frequency of words in a document

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

Normalize

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

- Discretization does not apply as users want association among words not ranges of words

Min-Apriori

- Why normalize?

TID	W1	W2
D1	0	10
D2	0	10
D3	0	10
D4	0	10
D5	1	1
D6	1	1
D7	10	0
D8	10	0
D9	10	0
D10	10	0

versus

TID	W3	W4
D1	0	0
D2	0	0
D3	0	0
D4	0	0
D5	1	1
D6	1	1
D7	0	0
D8	0	0
D9	0	0
D10	0	0

Min-Apriori

- New definition of support:

$$\sup(C) = \sum_{i \in T} \min_{j \in C} D(i, j)$$

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Example:

Sup(W1,W2,W3)

= 0 + 0 + 0 + 0 + 0.17

= 0.17

Anti-monotone property of Support

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Example:

Sup(W1) = 0.4 + 0 + 0.4 + 0 + 0.2 = 1

Sup(W1, W2) = 0.33 + 0 + 0.4 + 0 + 0.17 = 0.9

Sup(W1, W2, W3) = 0 + 0 + 0 + 0 + 0.17 = 0.17