# Mining Co-location Patterns with Rare Events from Spatial Data Sets

Yan Huang, Jian Pei, Hui Xiong

Petr Glos

**Masaryk University**

**Faulty of Informatics, Knowledge Discovery Group**

**Institute of Computer Science, Department of Intelligent Building Systems**

**Botanická 68a, 602 00 Brno**

**glos@ics.muni.cz**

**http://maps.muni.cz, http//:www.muni.cz, http://ics.muni.cz**

**Seminar on Knowledge Discovery, April 2011**

# Outline

- Co-location Patterns
- Participation Index
- Participation Ratio
- MinMax Algorithm
- Algorithm maxPrune
- Q&A

# Co-Location Patterns

- **Co-Location Pattern** – group of spatial features/events that are frequently co-located in the same region.
- **Co-Location Pattern** – set of spatial features that are frequently located together in spatial proximity.
- Location based services,
- Ecology mapping,
- Road works, Closures, Accidents,
- **Spatial feature is rare** if its instances are substantially less than those of other features in a co-location.

# Questions and tasks

- How to identify and measure spatial co-location patterns involving rare spatial features?
  - Measure called maximal participation ratio
- How to mine the patterns involving rare spatial feature effieciently?
  - Extension of apriori-like solution to do post-procesing
    - Very low participation index treshold to prune
    - Maximal participation ratio treshold to do a postprocessing
  - Algorithm using weak monotonic property of the maximal participation ratio to push the maximal participation ratio treshold deep into the mining.

# Frequent pattern x Co-location pattern Mining

| | |
|---|---|
| Item | Spatial feature |
| Item set | Spatial feature set |
| Frequent pattern | Co-location pattern |
| Support | Spatial interestigness measures |
| Transactional database | Spatial database |

# Neighbor-set

- S – **spatial dataset**
- $F = \{f_1, \dots, f_k\}$ - set of **boolean spatial features**
- $i = \{i_1, \dots, i_n\}$ - set of n **instances** in S,
- Each instance is a vector (instance-id, location, spatial feature)
- i.f – **spatial feature f of instance i**
- R is **neighborhood realation** over pairwise instances in S.
- **Neighbor-set** L is a set of instances such that all pairwise locations in L are neighbors.

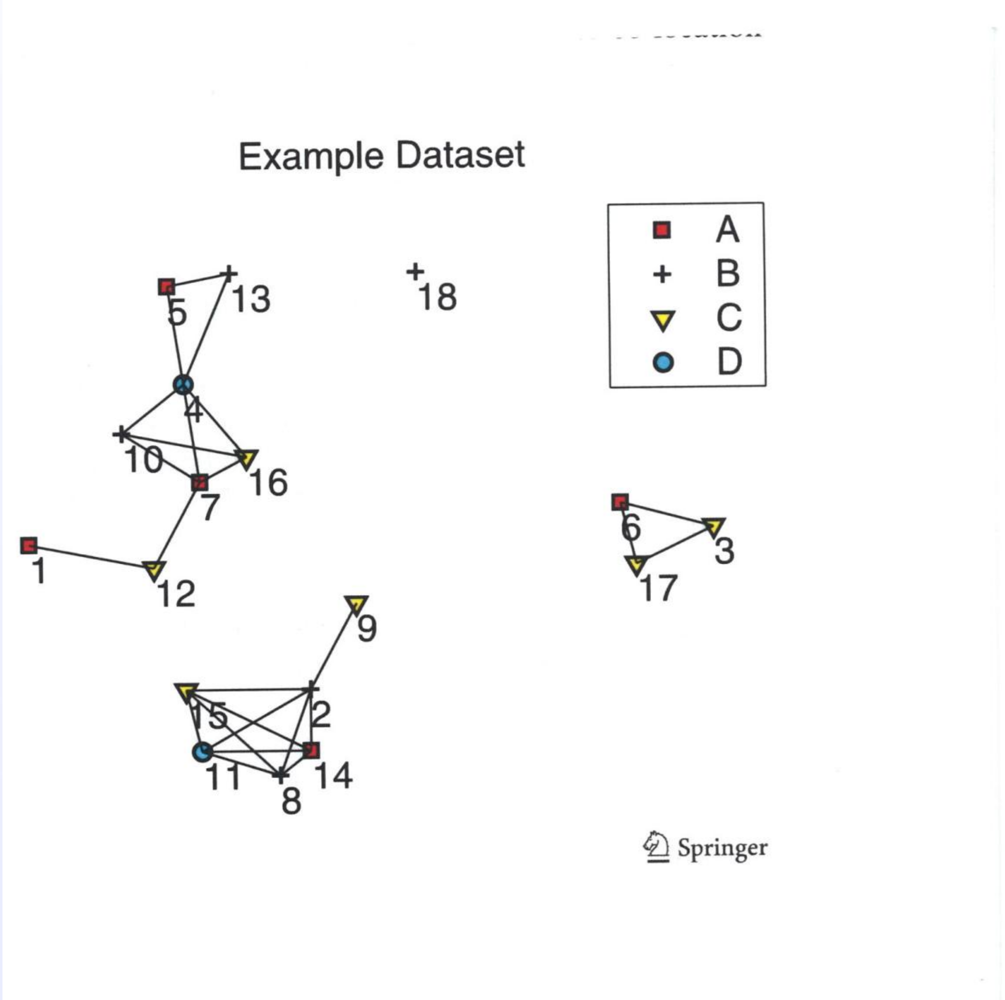
# Example

Co-location pattern
{A,B,C,D}

Neighbor sets

{3,6,17}
{6,17}
{3,6}

{4,5,13}
{4,7,10,16}
...



# Row instance, Participation ratio

- **Co-location pattern** C is a set of spatial features, $C \leq F$.
- A neighbor-set L is said to be a **row instance** of co-location pattern C if every feature in C appears as a feature of an instance in L and there exists no proper subset of L does so.
- **rowset(C)** – all row instances of co-location pattern C
- **Participation ratio**

$$\mathbf{pr(C,f)} = \frac{|\{r \mid (r \in S) \text{ and } (r.f=f) \text{ and } (r \text{ is a row instance of } C)\}|}{|\{r \mid (r \in S) \text{ and } (r.f=f)\}|}$$

- Wherever the feature f is observed, with probability pr(C,f), all other features in C are also observed in neighbor-set.

# Example

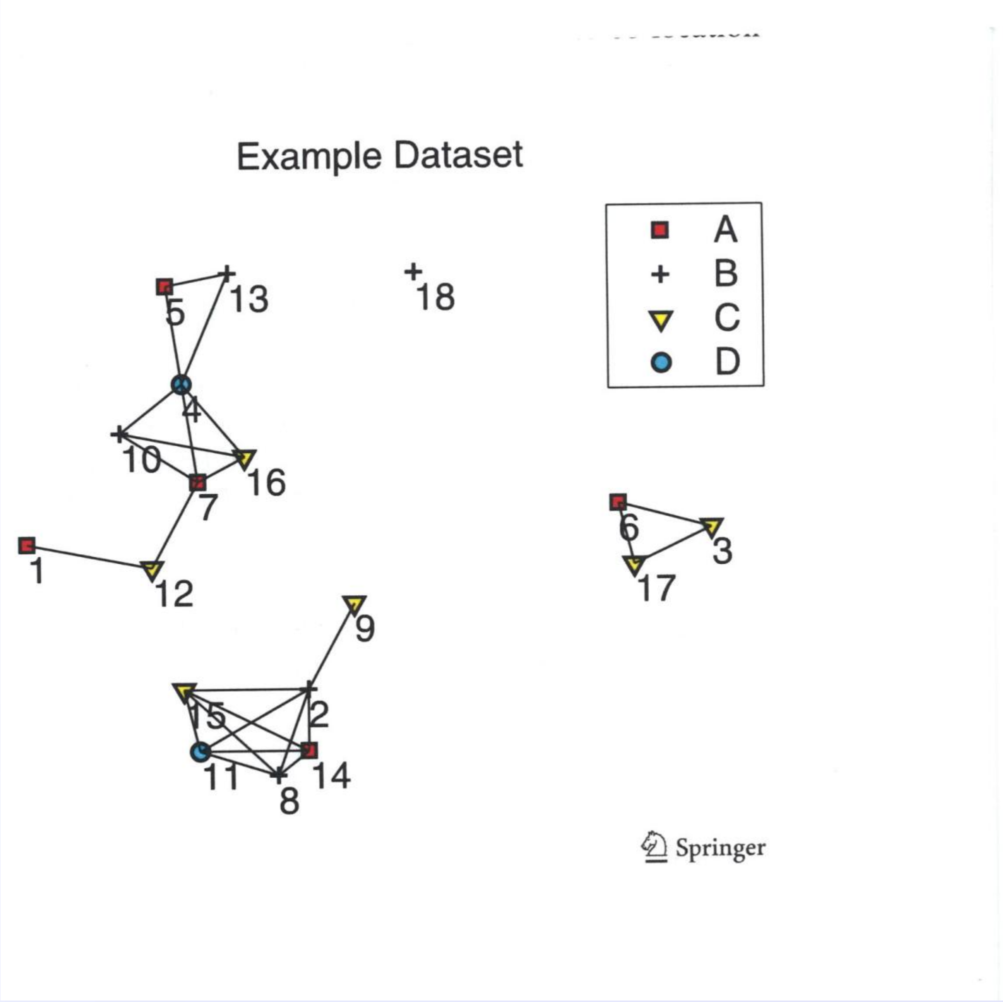
Row instances for ({A,B,C,D})

{2,11,14,15}

~~{2,8,11,14,15}~~

rowset({A,B,C,D})
=
{{4,7,10,16}
{2,11,14,15}
{8,11,14,15}}

rowset({A,B})
=
{{7,10}
{2,14}
{5,13}
{8,14}}



# Example

Participation ratios

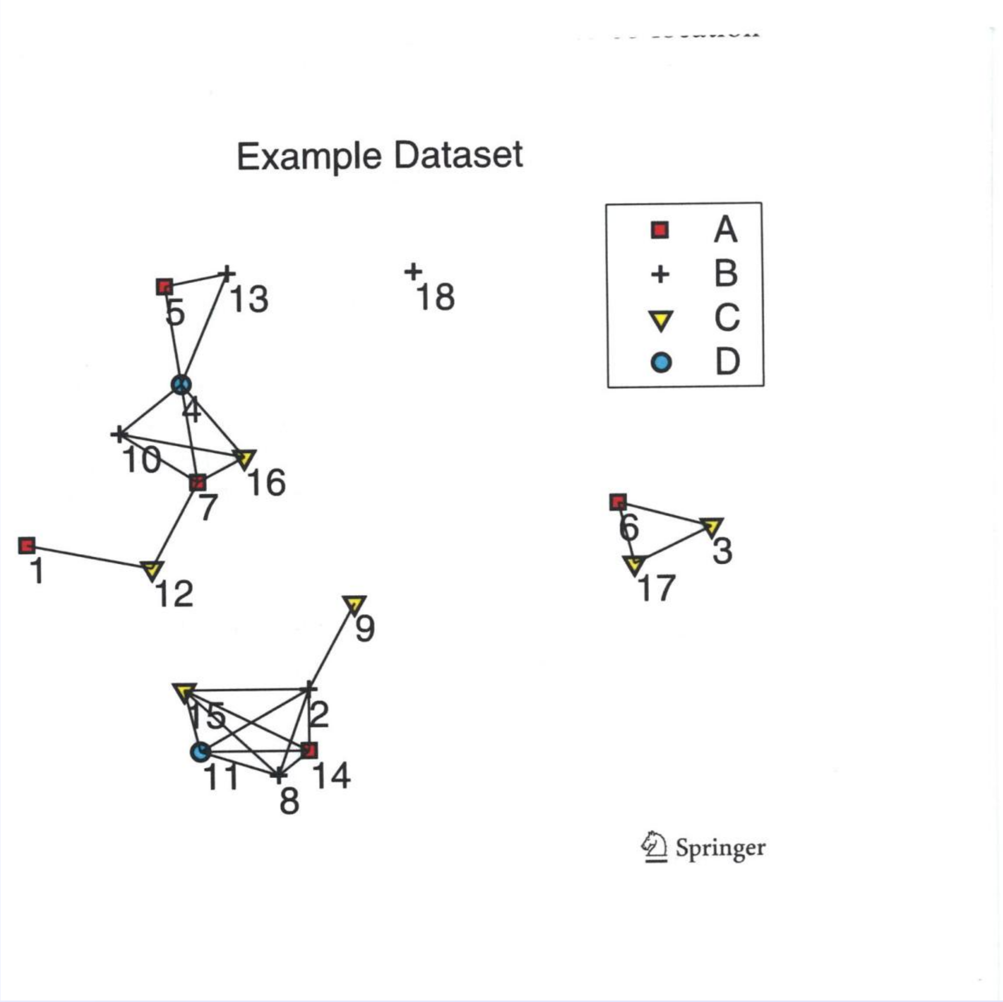
pr({A,B,C,D},A)=2/5

pr({A,B,C,D},B)=3/5

pr({A,B,C,D},C)=1/3

pr({A,B,C,D},D)=1

PI ({A,B,C,D},A)=1/3



# Participation index and monotonicity of participation ratio and index

- **$PI(C) = \min_{f \in C} \{pr(C,f)\}$**
- Wherever any feature from C is observed, with probability of at least PI(C), all other features in C can be observed in neighbor-set.
- A high participation index value indicates that the spatial features in a co-location pattern likely occur together.
- Given a user-specified **participation index treshold min_prev**, a co-location pattern C is called **prevalent** if PI(C) >= min_prev.
- Let C and C' be two co-location patterns such that C is subset of C'. Then, for each feature f∈C, pr(C,f) >= pr(C',f).
- Furthemore, PI(C) >= PI(C')

# Maximal participation ratio

- **Maximal participation ratio** $maxPR(C) = \max_{f \in C} \{pr(C,f)\}$
- A high maximal participation ratio value indicates that there are some spatial features strongly imply the pattern.
- $C = \{f_1, \dots, f_k\}$ co-location pattern,
- Minimum maximal participation ratio treshold min_maxPR
- $pr(C,f_1) \Rightarrow \dots \Rightarrow pr(C,f_l) \Rightarrow \dots \Rightarrow pr(C,fk)$ ,
- $f_l$ is the last spatial feature that has participation ration above min_maxPR
- If spatial feature $f_i$ $(1 <= i <= l)$ is observed in some location, then the probability of observing all other spatial feature in $C - \{f_i\}$ in neighbor set is at least min_maxPR.

# Example

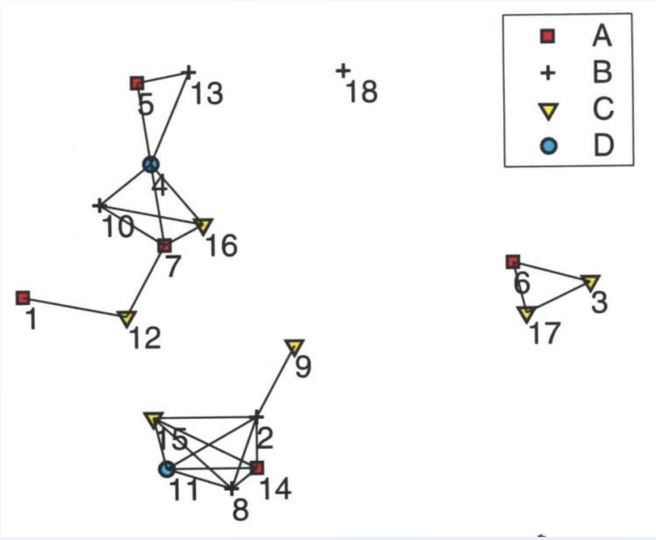



**Table 2** Rowsets, *PI*s and *maxPR*s of co-locations of dataset in Fig. 2

| ID | Co-loc | Rowset | pr | PI | max PI |
|---|---|---|---|---|---|
| 1 | {A} | {{1},{5},{6},{7},{14}} | {1} | 1 | 1 |
| 2 | {B} | {{2},{8},{10}, {13},{18}} | {1} | 1 | 1 |
| 3 | {C} | {{3},{9},{12}, {15},{16},{17}} | {1} | 1 | 1 |
| 4 | {D} | {{4},{11}} | {1} | 1 | 1 |
| 5 | {A,B} | {{5,13},{7,10},{14,2},{14,8}} | {4/5,4/5} | 4/5 | 4/5 |
| 6 | {A,C} | {{1,12},{6,3},{6,17},{14,15},{7,16}} | {4/5,5/6} | 4/5 | 5/6 |
| 7 | {A,D} | {{5,4},{14,1},{7,4}} | {3/5,2/2} | 3/5 | 1 |
| 8 | {B,C} | {{2,9},{2,15},{8,15},{10,16}} | {3/5,3/6} | 1/2 | 3/5 |
| 9 | {B,D} | {{2,11},{8,11},{10,4},{13,4}} | {4/5,2/2} | 4/5 | 1 |
| 10 | {C,D} | {{15,11},{16,4}} | {2/6,2/2} | 1/3 | 1 |
| 11 | {A,B,C} | {{7,10,16},{14,2,15},{14,8,15}} | {2/5,3/5,2/6} | 1/3 | 3/5 |
| 12 | {A,B,D} | {{5,13,4},{7,10,4},{14,2,11},{14,8,11}} | {3/5,4/5,2/2} | 3/5 | 1 |
| 13 | {A,C,D} | {{7,16,4},{14,15,11}} | {2/5,2/6,2/2} | 2/5 | 1 |
| 14 | {B,C,D} | {{2,15,11},{10,16,4},{8,15,11}} | {3/5,2/6,2/2} | 1/3 | 1 |
| 15 | {A,B,C,D} | {{7,10,16,4},{14,2,15,11},{14,8,15,11}} | {2/5,3/5,2/6,2/2} | 1/3 | 1 |

# Rundimentary Algorithm

**Input:** A spatial database $S$, a neighborhood relation $\mathcal{R}$, a minimum prevalent threshold $min\_prev$, and a minimum maximal participation index threshold $min\_maxPR$.

**Output:** Co-location patterns $P$ such that $PI(P) \geq min\_prev$ and $maxPR(P) \geq min\_maxPR$.

**Method:**

1. let $k = 2$; generate $C_2$, the set of candidate 2-patterns and their rowsets, by geometric methods;
2. for each $C \in C_k$ calculate $PI(C)$ and $maxPR(C)$ from $C$'s rowset $rowset(C)$;
3. let $P'_k$ be the subset of $C_k$ such that for each $P \in P'_k$, $PI(P) \geq min\_prev$;
4. let $P_k$ be the subset of $P'_k$ such that for each $P \in P_k$, $maxPR(P) \geq min\_maxPR$;
5. generate the set $C_{k+1}$ of candidate $(k+1)$-patterns, a co-location pattern $P$ with $(k+1)$ spatial features is in $C_{k+1}$ if and only if for each feature $f \in P$, $(P - \{f\}) \in P'_k$;
6. if $C_{k+1} \neq \emptyset$, let $k = k + 1$, go to Step 2;
7. output $\cup_i P_i$ ■

**Fig. 3** Algorithm Min–Max

# Rundimentary Algorithm

- If min_prev = 0 then algorithm can find the complete set of patterns.
- If min_prev > 0 then some patterns with high maximal participation ratio but low prevalence may be missed.
- Major disadvantage – If user wants to find the complete answer, the algorithm has to generate a huge number of candidates and test them, even though the maximal participation ration treshold min_maxPR is high.

# Weak monotonocity of maximal participation ratio

- Let P be a k-co-location pattern.
  Then, there exists at most one (k-1) – subpattern P' such that P' is subset of P and maxPR(P') < maxPR(P)

- If a k-pattern is above the maximal participation ratio treshold, then at least (k-1) out of its k subpatterns with (k-1) features are above the maximal participation ratio treshold.

# Algorithm maxPrune

**Example 8:** (Candidate generation using weak monotonicity) Suppose the maximal participation ratio values of $\{A, B, C\}$, $\{A, C, D\}$ and $\{B, C, D\}$ are all over the threshold $min\_maxPR$, but that of $\{A, B, D\}$ is not. We still should generate a candidate $P = \{A, B, C, D\}$, since it is possible that $maxPR(P)$ passes the threshold.

To achieve this, we need a systematic way to generate the candidates. Please note that, in apriori, for the above example, $\{A, B, C, D\}$ is generated only if $\{A, B, C\}$ and $\{A, B, D\}$ (differ only in their last spatial feature) are both frequent. However, in the co-location pattern mining with rare spatial features using maximal participation ratio measure, it is possible that $\{A, B, D\}$ is below the given threshold min_maxPR while $\{A, B, C, D\}$ is above the threshold $min\_maxPR$.

In general, for two co-location patterns $P$ and $P'$ from the set $P_k$ of $k$-patterns above threshold $min\_maxPR$, i.e., $P \in P_k$ and $P' \in P_k$, $P$ and $P'$ can be joined to generate a candidate $(k + 1)$-pattern in $C_{k+1}$ if and only if $P$ and $P'$ have one different feature in the last two features. For example, even $\{A, B, D\}$ is below threshold $min\_maxPR$, candidate $\{A, B, C, D\}$ can be generated by $\{A, B, C\}$ and $\{A, C, D\}$ since they have the common feature $C$ in their last two features, i.e., they differ one spatial feature in their last two spatial features. ■

We will illustrate the correctness of the above candidate generation method in Lemma 3 and Example 9. Also, with the revised candidate generator, the mining algorithm is presented in Fig. 4.

The algorithm does not need a minimum prevalence threshold but still finds all co-location patterns with maximal participation index above threshold $min\_maxPR$.

To make sure the candidate generation does not miss any co-location, we need to prove that the candidate $(k + 1)$-patterns $C_{k+1}$ generated by the maxPrune algorithm

# Algorithm maxPrune

**Input:** A spatial database $S$, a neighborhood relation $\mathcal{R}$, a minimum maximal participation ratio $min\_maxPR$.

**Output:** Co-location patterns $P$ such that $maxPR(P) \geq min_$maxPR.

**Method:**

1. let $k = 2$; generate $C_2$, the set of candidate 2-patterns and their rowsets, by geometric methods;
2. For each $C \in C_k$ calculate $maxPR(C)$ from C's rowset $rowset(C)$; Let $P_k$ be the subset of $C_k$ such that for each $P \in P_k$, $maxPR(P) \geq min\_maxPR$;
3. generate $C_{k+1}$, the set of candidates $(k+1)$-patterns, as illustrated in Example 8 ; if $C_{k+1} \neq \emptyset$, let $k = k + 1$, go to Step 2;
4. output $\cup_i P_i$ ■

**Fig. 4** Algorithm maxPrune

Springer

Thank you for your attention.