

# Discovering Spatial Co-location Patterns: A Summary of Results

---

Yan Huang  
Spatial Database Lab  
Department of Computer Science  
University of Minnesota  
huangyan@cs.umn.edu

Full paper by Shashi Shekhar and Yan Huang at:  
<http://www.cs.umn.edu/research/shashi-group>

# Application Domains

---

★ Ecology

★ *Lansing woods tree data [Diggle 83]*

★ *Predator-prey species, symbiosis*

★ Immunogold Labeling

★ Epidemiology

★ *food-types, obesity, heart disease*

★ Examples from climatology [Potter01]

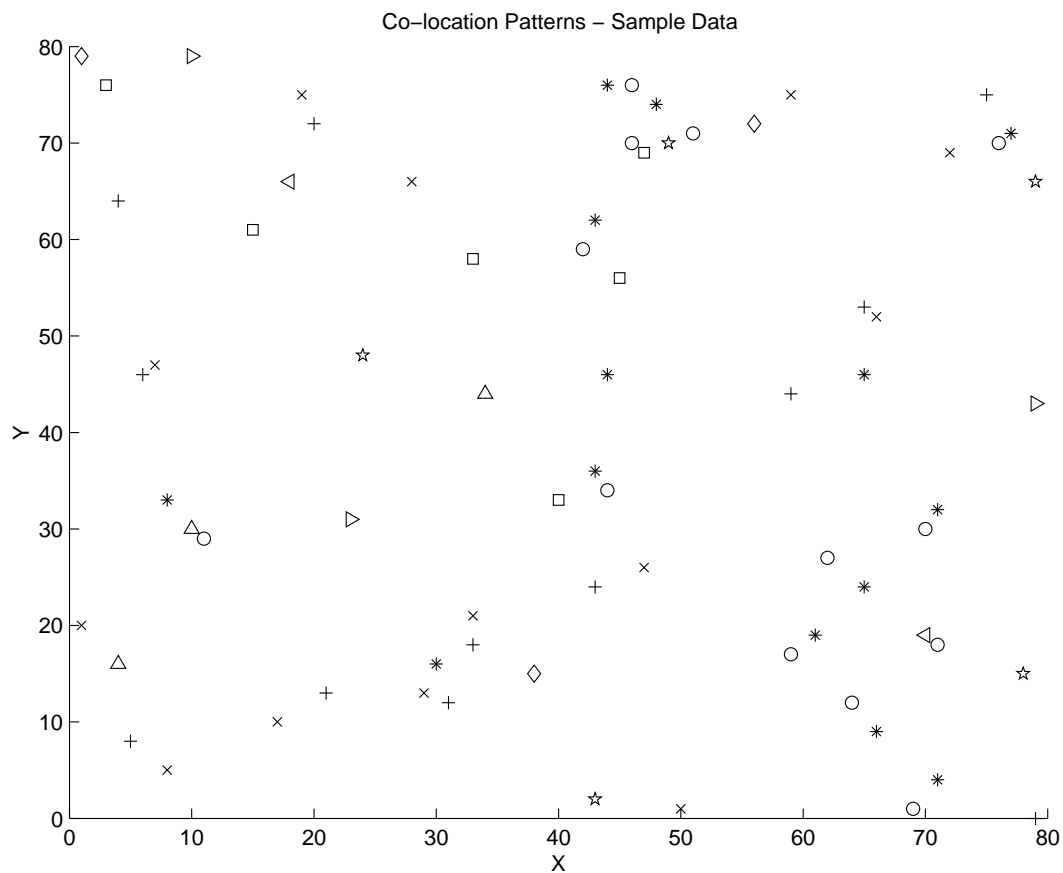
Pattern #	Variable $A$	variable $B$	Examples of interesting patterns
P1	Cropland Area	Vegetation	Higher cropland area alters NPP
P2	Smoke Aerosol Index	Precipitation	Smoke aerosols alter the likelihood of rainfall in a nearby region
P3	Sea Surface Temperature	Land Surface Climate and NPP	Surface ocean heating affects regional terrestrial climate and NPP

# Motivating Example

★ Given:

★ *A collection of different types of spatial events*

★ Illustration

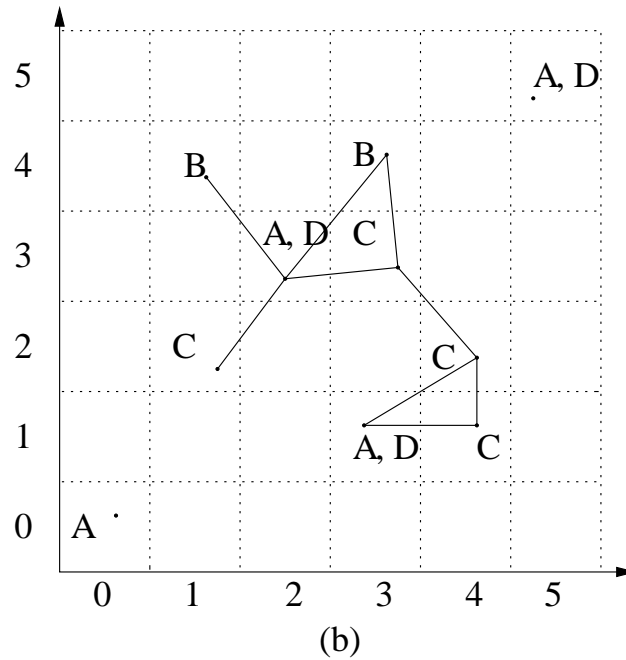


★

★ *Find: Co-located subsets of event types*

# Event Centric Model






- ★ Non-transaction approach



- ★ An example:  $A$  and  $B$  happen in a neighborhood  $\rightarrow C$  happens in their neighborhood with 80% conditional probability
- ★ Application domain: Ecology

## Association Rules - An Analogy

★ Association rule e.g. (Diaper in T  $\Rightarrow$  Beer in T)

Trans.	Items Bought
1	{socks,  milk,  , beef, egg, ... }
2	{ pillow,  , toothbrush, ice-cream, muffin, ... }
3	{  ,  , pacifier, formula, blanket, ... }
...	...
n	{battery, juice, beef, egg, chicken, ... }

★ Support:  $\text{pr}(\text{Diaper and Beer in T})$

★ Confidence:  $\text{pr}(\text{Beer in T} | \text{Diaper in T})$

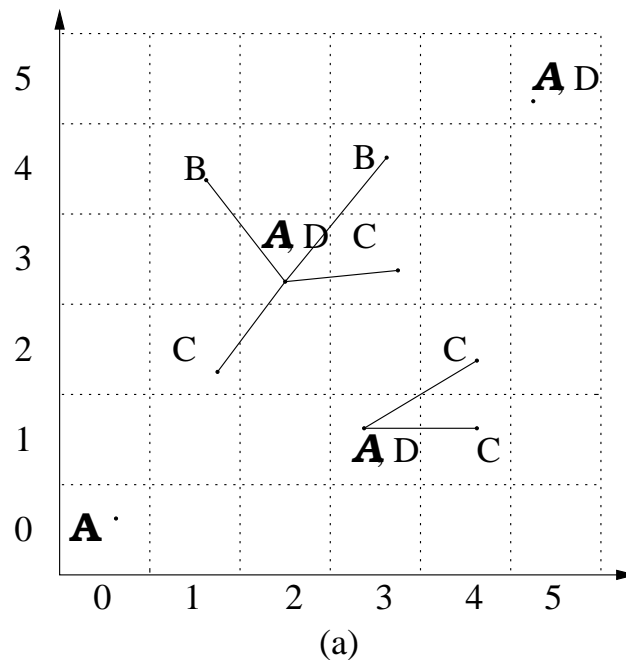
★ Algorithm Apriori [Agrawal, Srikant, VLDB94]

★ Support based pruning using monotonicity

★ Note: **Transaction is a core concept!**

## Related Work: Spatial Association Rules

- \* Force-fit notion of transaction
- \* Reference feature centric model [Koperski, Han, SSD95]
- \* *All relevant co-locations reference to one feature*



- \* *Item types = boolean spatial features*
- \* *Transactions = Instances of reference feature*
- \* *An example: B and C are close to A  $\Rightarrow$  D is close to A with 60% conditional probability*
- \* *Application domain:*
  - Focus on a specific boolean spatial feature, e.g. cancer
  - Need a reference feature
- \* *Q: will it discover co-location rules?*

## New Challenges

---

### ★ Association Rules Vs. Co-location Rules

Criteria	Association Rule	Co-location Rule
Underlying Space	Discrete Sets	Continuous Space
Item Types	Product types	Spatial Features(Boolean)
Item Collections	Transactions $\{T_i\}$	<b>Neighborhoods</b>
Prevalence ( $A \rightarrow B$ )	Support: $p(A \cup B \in T_i)$	Participation Index
Conditional Probability ( $A \rightarrow B$ )	$p(B \in T_i   A \in T_i)$	$p(B \in \text{Nbr}(L)   A \text{ at } L)$

### ★ Items? transactions?

- ★ *Spatial transactions may not be nature!*
- ★ *Support is not defined*
- ★ *Support based pruning (Apriori) not defined*

# Overview

---

- ★ Introduction
- ⇒ Problem Formulation
- ★ Co-location Miner Algorithm
- ★ Conclusions

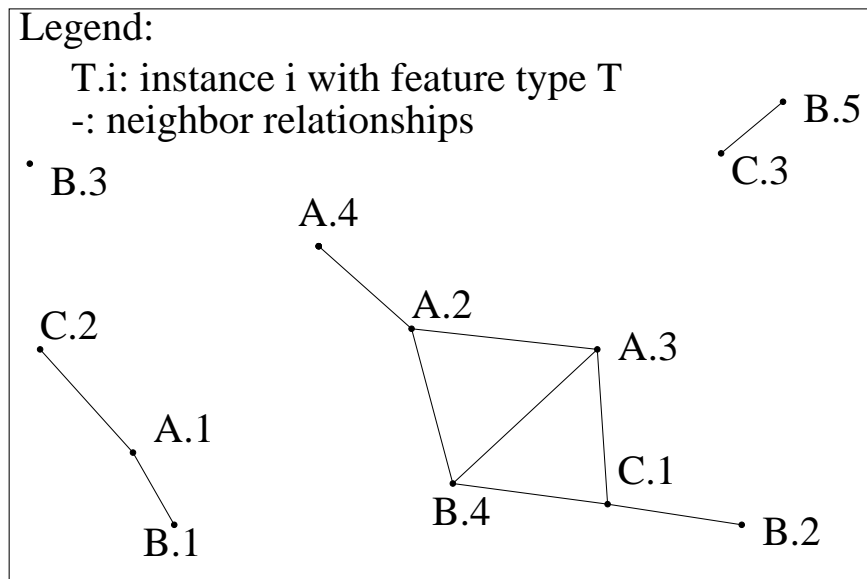
## Problem Formulation

---

- \* Given:
  - \*  $K$  Boolean spatial feature types
  - \* Instances  $\langle id, \text{feature type } t, \text{location } l \rangle$
  - \* A neighbor relation  $R$  over locations
  - \* Thresholds: prevalence and conditional probability
- \* Find:
  - \* Co-location rules with high prevalence and high conditional probability
- \* Objectives:
  - \* Completeness, Correctness, Efficiency
- \* Constraints:
  - \* Symmetric  $R$
  - \* Monotonic prevalence measure
  - \* Event centric model
  - \* Sparse data set

## Example of Key Concepts

- \* Co-locations of size 2
  - \*  $(A,B)$ ,  $(A,C)$ ,  $(B,C)$
- \* Table instances of co-locations of size 2
  - \*  $(A,B)$ :  $(1,1)$ ,  $(2,4)$ ,  $(3,4)$
  - \*  $(A,C)$ :  $(1,2)$ ,  $(3,1)$
  - \*  $(B,C)$ :  $(2,1)$ ,  $(4,1)$ ,  $(5,3)$
- \* Participation Indexes of co-locations of size 2
  - \*  $(A,B)$ :  $\min(3/4, 2/5) = 2/5$
  - \*  $(A,C)$ :  $\min(2/4, 2/3) = 2/4$
  - \*  $(B,C)$ :  $\min(3/5, 2/3) = 3/5$
- \* Example Dataset:



## Key Concepts

---

- \* **A Co-location  $C$ :**
  - \* *A subset of boolean spatial features*
- \* **A co-location rule  $C_1 \rightarrow C_2(p, cp)$ :**
  - \*  $C_1$  and  $C_2$  are co-locations
  - \*  $p$  = prevalence measure
  - \*  $cp = Pr[C_2 \in N(L) \mid C_1 @ L]$
- \* **A Neighborhood:**
  - \* *A clique in a graph of neighbor relation  $R$*
- \* **A row instance  $I$  of a co-location  $C = \{f_1, \dots, f_k\}$ :**
  - \*  $I = \{i_1, \dots, i_k\}$
  - \*  $i_j$ : instance of  $f_j (\forall j \in 1, \dots, k)$
  - \*  $I$  is a neighborhood
- \* **Table instance(co-location  $C = \{f_1, \dots, f_k\}$ ):**
  - \* *Collection of all its row instances*
  - \* *Spatial join interpretation*

## Key Concepts *cont ...*

---

- ★ **Participation ratio**

- ★  $pr(C, f_i) = |\pi_{f_i} table - instance(C)| / |instances(f_i)|$

- ★  $C = \{f_1, f_2, \dots, f_k\}$

- ★ *Monotonically decreasing*

- ★ **The participation index**

- ★  $pi(C) = \min_{i=1}^k pr(C, f_i)$

- ★ Statistical interpretation: ongoing work

# Overview

---

- ★ Introduction
- ★ Problem Formulation
- ⇒ Co-location Miner Algorithm
- ★ Conclusions

# Co-location Miner Algorithm

---

★ Co-location Miner

★ *Initialization*

★ *Generate size 2 co-location rules*

★ **for**  $k$  *in*  $(2, 3, \dots, K - 1)$  **do**

- 1. Generate size  $k$  candidate co-locations (*apriori\_gen*)

- 2. Generate table instances/prune based on neighborhood

- 3. Prune based on prevalence of co-locations

- 4. Generate co-location rules of size  $k$

★ **end**

★ Note: Step 2 not needed in mining association rules

★ *because item collections (i.e. transactions) are given*

★ Execution Trace

Size	Cand. Co-locs	Tables	Table Instances	Par. Ind.	Prev. ?
1	(A)	T1	(1),(2),(3),(4)	1	Y
	(B)	T2	(1),(2),(3),(4),(5)	1	Y
	(C)	T3	(1),(2),(3)	1	Y
2	(A,B)	T4=T1⋈T2	(1,1),(2,4),(3,4)	$\min(3/4, 2/5)=2/5$	Y
	(A,C)	T5=T1⋈T3	(1,2),(3,1)	$\min(2/4, 2/3)=2/4$	Y
	(B,C)	T6=T2⋈T3	(2,1),(4,1),(5,3)	$\min(3/5, 2/3)=3/5$	Y
3	(A,B,C)	T7=T4⋈T5	(3,4,1)	$\min(1/4, 1/5, 1/3)=1/5$	?

## Details of Co-location Miner

---

\* Apriori-gen

\* *Join step:*

```
insert into  $C_{k+1}$ 
select  $p.f_1, \dots, p.f_k, q.f_k, p.table\_id, q.table\_id$ 
from  $C_k$   $p, C_k$   $q$ 
where  $p.f_1 = q.f_1$  and  $\dots$  and  $p.f_{k-1} = q.f_{k-1}$ 
      and  $p.f_k < q.f_k$ 
```

\* *Prune step:*

```
forall co-locations  $c \in C_{k+1}$  do
  forall size  $k$  subset  $s$  of  $c$  do
    if ( $s \notin C_k$ ) then
      delete  $c$  from  $C_{k+1}$ ;
```

## Details of Co-location Miner cont...

---

- \* Generate table instance
  - \* forall co-location  $c \in C_{k+1}$ 
    - insert into  $T_c$
    - select  $p.i_1, p.i_2, \dots, p.i_k, q.i_k$
    - from  $c.table\_id_1 p, c.table\_id_2 q$
    - where  $p.i_1 = q.i_1$  and  $\dots$  and  $p.i_{k-1} = q.i_{k-1}$
    - and  $(p.i_k, q.i_k) \in R$ ;
  - end;
  
- \* Participation Indexes Calculation
  - \* *Bitmap index based*
  - \* *One scan of table instances in current iteration*
  
- \* Co-location rule generation:
  - \* *See paper ...*

## Completeness and Correctness

---

- \* Definition:
  - \* *Completeness:*  
*Find all rules with prevalence and conditional probability > thresholds*
  - \* *Correctness:*  
*Any rules found have prevalence and conditional probability > thresholds*
- \* Theorem
  - \* *Co-location Miner is complete and correct*
- \* Proof:
  - \* *Monotonic participation index*
  - \* *Any prevalent co-location's subset is prevalent*
  - \* *Table join will not miss any instance*

## Conclusions

---

- ★ Our Contributions

- ★ *A NATURAL spatial association model, i.e. co-location rules*
- ★ *Eliminates need to transactionize spatial data*
- ★ *Co-location Miner Algorithm*
- ★ *Proof of correctness and completeness*

- ★ Future Work

- ★ *Statistical interpretation*
- ★ *Performance evaluation*
- ★ *Other spatial data types: polygons, lines, etc.*
- ★ *Spatio-temporal datasets*
- ★ *Shameless plug for “Spatial Database: A Tour” book :)*
- ★ *Questions?*