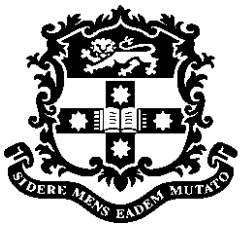

The University of Sydney

Complex Spatial Relationships

TECHNICAL REPORT NUMBER 539

August 2003

Robert Munro, Sanjay Chawla and Pei Sun

ISBN 1 86487 584 4

School of Information Technologies
University of Sydney NSW 2006

Complex Spatial Relationships

Robert Munro, Sanjay Chawla & Pei Sun
School of Information Technologies, University of Sydney
{rmunro, chawla, psun2712}@it.usyd.edu.au

Abstract

This paper describes the need for mining complex relationships in spatial data. Complex relationships are defined as those involving two or more of: multi-feature co-location, self-co-location, one-to-many relationships, self-exclusion and multi-feature exclusion. We demonstrate that even in the mining of simple relationships, knowledge of complex relationships is necessary to accurately calculate the significance of results. We implement a representation of spatial data such that it contains 'weak-monotonic' properties, which are exploited for the efficient mining of complex relationships, and discuss the strengths and limitations of this representation.

1 Introduction

A relationship in spatial data is a relationship between features in a Euclidean space, defined in terms of the co-locational trends of two or more features over that space. An example is determining the confidence of the '*where there's smoke there's fire*' with respect to a set of co-ordinates, each representing the feature *smoke* or *fire*. The task here would be to determine whether *fire* occurs in the neighborhood of *smoke* more than is randomly likely.

Neighborhoods are defined in terms of cliques (also known as neighbor-sets). A clique is defined as any set of items such that all items in that set co-locate. For example, in figure 1 and table 1, the co-locational pattern {A,C} occurs four times in five cliques, *iv, v/vi, vii & x*. In spatial data, two items are typically said to co-locate if they are positioned within a certain distance *d* of one another. As has been assumed in figure 1, *d* is usually a constant, but it may also be defined as varying locally within the space or with respect to a given feature.

Typically, the mining of information in a spatial domain involves representing the cliques as transactions, and undertaking association rule mining upon these transactions.

While association rule mining is a well-developed field (Han 2000), the mining of confident cliques as transactions fails to capture many spatial phenomena of interest, due to most association mining techniques being optimized for '*market-basket*' data.

Spatial data is fundamentally different from market-basket data, both in its basic nature and distributional tendencies.

One factor unique to spatial data is that the number of transactions a single item may participate in is potentially unbounded, while in a market-basket this is limited to one (obviously, two people may purchase the same toothpaste product, but not the same exact tube).

On the other hand, in spatial data, the number of features is explicitly bounded, as, typically, it involves the deliberate selection of features so as to mine the spatial relationships between them. In market-basket data, the potential number of features (products) is unbounded.

Self-co-location is also more likely in spatial data. The upper limit in a market-basket is multiple purchases of only one product, which is less likely than an equivalent spatial situation of an area of monoculture forest.

Similarly, there may be direct relationships between features that don't co-locate in spatial data, such between animals displaying territorial behavior, or between a certain virus and medication in a biological system, making such relationships intrinsically more interesting in spatial data.

A complex relationship is simply any combination of these. It is important to note that while the relationships are defined as complex, the phenomena they represent are often very simple, for example:

1. Crime is more likely to co-occur on streets with no lighting near a subway station.
2. Many feral (non-native) animals in an area of forest implies the absence of native ones.
3. Dogs only act as pack animals in non-urban environments.

Perhaps the most fundamental difference between spatial and other data in a transactional representation is the notion of significance. In general, a co-location is considered significant if it occurs more than is randomly likely. In transactions representing market-baskets, the transactions, by definition, represent the complete space of the data, that is, there are no empty baskets. In such cases, the significance of the data may be represented by frequency of the feature in relation to the number of transactions, such as the interest measure proposed by Piatetsky-Shapiro (1991). In spatial data, however, the random likelihood of a co-location depends on the volume of the space from which it was taken. This is discussed in more detail in section 6.1.

1.1 Our contribution

We describe the need for mining complex relationships in spatial data. To the authors' best knowledge, this problem has not previously been addressed.

We demonstrate that even when the goal is the mining of only simple relationships:

1. It is necessary to mine complex relationships to accurately calculate the significance of results.
2. Including the notion of exclusion can lead to a stronger definition of clique confidence.

We demonstrate that a representation of spatial data is possible such that it facilitates the efficient mining of complex relationships.

1.2 Outline

Sections 2 and 3: we give the problem definition and a discussion of related work.

Section 4: we describe and discuss the use of a participation threshold, rather than a support threshold, in the mining of spatial co-locations. It should be noted that we are *not* redefining or refining any association rule algorithms; rather than throwing out the baby with the bathwater, we explore new applications and interpretations of existing ones.

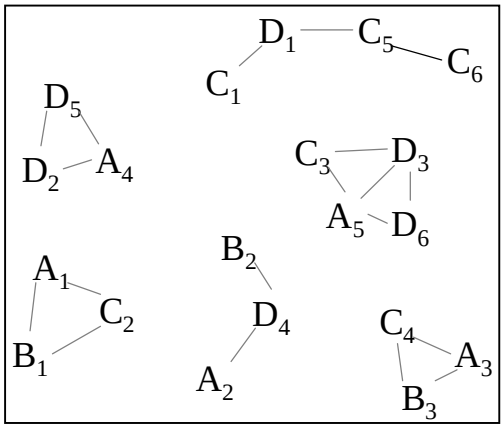

Figure 1: An example of spatial co-locational patterns of the features A,B,C and D

No	Clique
i.	C_1, D_1
ii.	C_5, D_1
iii.	C_5, C_6
iv.	A_4, D_2, D_5
v.	A_5, C_3, D_3
vi.	A_5, D_3, D_6
vii.	A_1, B_1, C_2
viii.	B_2, D_4
ix.	A_2, D_4
x.	A_3, B_3, C_4

Table 1: Cliques in figure 1

Section 5: we define and give examples of the various types of relationships in spatial data, including complex relationships.

Section 6: we demonstrate that the knowledge of complex relationships in spatial data may be desirable and/or necessary, even when the goal is the mining of only simple relationships.

Section 7: we implement a transactional representation of spatial data such that it contains weak-monotonic properties, which are exploited by the *maxPI* algorithm (Huang, 2003) for the efficient mining of complex relationships, and discuss the strengths and limitations of this representation.

Section 8: we conclude and discuss possible future directions.

2 Problem Definition

Given a set of items, each representing one feature at a given coordinate, the goal is to find all spatial relationships between items, with respect to the features they represent and their relative positions within the coordinate space.

One of the main concerns with the mining of spatial data is the appropriate choice regarding the representation of the data for the purposes of mining information. There can be a loss of information when transcribing coordinate information into continuous/discrete attributes for machine learning, or cliques for discovering co-locations. The latter, a transactional representation (as in table 1), is the most commonly used in mining spatial data, as it allows the inclusion and the discovery of the interrelation of non-spatial features.

For mining simple co-locations, this is a 3-step process:

1. Generate a set of the cliques in a representation that facilitates the mining of co-locations.
2. Apply a mining algorithm to the cliques, returning a set of co-locations and their confidences, the constituency of which is determined by given pruning and confidence thresholds.
3. Calculate the significance of the mined co-locations.

The first two steps are typically combined, so as to not to generate cliques already known to be below the given thresholds. In this paper, we assume that the first step has already taken place.

For mining complex relationships, the problem becomes:

1. Generate a representation of spatial data that facilitates the mining of multiple relationship types, including the interaction of such relationships (complex relationships).
2. Select a mining algorithm appropriate to the efficient mining of both simple and complex relationships, and apply it to the representation, returning a set of co-locations and their confidences.
3. Calculate the significance of the mined relationships.

3 Related Work

Koperski and Han (1995) proposed the first extension of the *Apriori* paradigm to spatial data. However in their method they materialized all the possible spatial relationships that they intended to mine. This is equivalent to determining the universe of candidate interesting relationships. Thus in some ways their technique was 'hypothesis driven' rather than 'hypothesis generating.'

Shekhar et al (2001) presented an efficient algorithm to mine a kind of spatial co-locations. The concepts of neighborhood, participation ratio, participation index were defined. Instead of support, the participation index was used as a pruning measure in the conventional *Apriori*-like technique.

The drawback of above method is that some confident co-location rules with low support are also pruned. In order to solve this problem, Huang et al (2003) proposed the concept of maximal participation index and it was used as pruning measure to replace participation index. We will discuss these measures in detail in the next section, as they are central to our approach.

Wu et al (2002) proposed an algorithm to mine both positive and negative association rules. Negative rules are generated from infrequent item sets and interest is used as a further pruning measure. Their algorithm involves no spatial component.

4 Maximal Participation Ratio

In this section we will briefly describe the notion of Maximal Participation Index as described in Huang (2003) where more details can be found.

4.1 Participation ratio

Given a co-location pattern L and a feature $f \in L$, the participation ratio of f, $pr(L, f)$, can be defined as the support of L divided by the support of f.

For example, in figure 1, the support of $\{A, B, C\}$ is 2 and the support of C is 6, so $pr(\{A,B,C\},C)=2/6$.

4.2 Maximal participation index

Given a co-location pattern L, the maximal participation index of L, $maxPI(L)$ can be define as the maximal participation ratio of all the features in L, that is:

$$maxPI(L) = max_{f \in L}\{pr(L, f)\}.$$

For example, in figure 1, *maxPI({ A, B, C}) = max(pr({ A, B, C}, A), pr({ A, B, C}, B), pr({ A, B, C}, C) = max(2/5, 2/3, 2/6) = 2/3* .

A high maximal participation index indicates that at least one spatial feature strongly implies the pattern. By using *maxPI*, we can find some low frequency confident rules, which may be pruned by a support threshold (Huang 2003).

4.3 The weak monotonic property of maxPI

Maximal participation index is not monotonic with respect to the pattern containment relations. For example, in the figure 1, *(maxPI({A,C}) = 3/5) < (maxPI({A, B, C}) = 2/3).*

Interestingly, the maximal participation index does have the following weak monotonic property:

If *P* is a k-co-location pattern, then there exists at most one *(k-1)* subpatterns *P'* of *P* such that *maxPI(P') < maxPI(P).*

Relying on this weak monotonic property, we can modify the *Apriori*-like algorithm to mine confident patterns by using a *maxPI* threshold.

5 Relationship Definitions

5.1 Notation used

Feature: In this paper, a feature is represented as a capital letter, for example *A*.

Item: An instance of a feature (one item) is represented as the feature followed by an id number unique for that feature, for example A_2.

Absence: The absence of an item is represented by negation, for example *–A*. (**Note**: this is *not* the equivalent of the set-theory, ¬*A*, meaning the presence of any item other than *A*).

Self-co-location: Multiple instances of a feature (multiple items) are represented by a '+' following the feature, for example *A+*.

The representations for the complex relationships described in the introduction would be:

1. Crime, *C*, is more likely to co-occur on streets with no lighting, *L*, near a subway station, *S:* $S, -L \rightarrow C+$
2. Many feral animals, *F*, in an area of national park implies the absence of native ones, *N:* $F+ \rightarrow -N$
3. Dogs, *D*, act as pack animals in urban environments, *U:* $D, U \rightarrow D+$

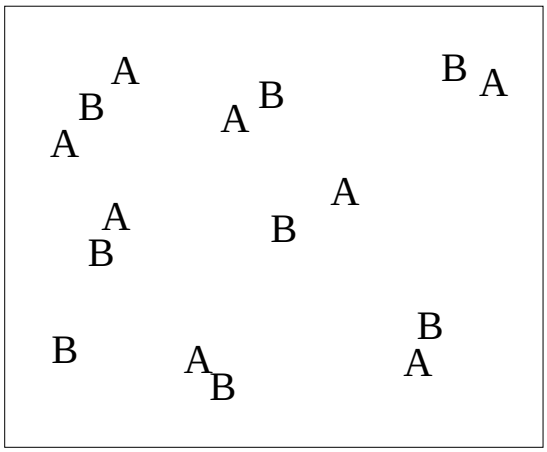

Figure 2: Example of a positive relationship $A \rightarrow B$

Further examples of each relationship type, taken from figure 1, are given in table 2.

5.2 Positive Relationships

This is the most common type of relationship mined.

Definition 1: A positive relationship (multi-feature co-location) in spatial data is a set of features that co-locate at a ratio greater than some predefined threshold. In spatial data, the confidence of a positive relationship $A \rightarrow B$, is given by the fraction of unique *A*'s that co-occur in a clique containing a the feature *B.*

5.3 Self-co-location / Self-exclusion

This is the measure of which a feature tends to co-locate with itself. Formally, it is the average cardinality of an item in a clique with respect to the expected cardinality of a random distribution. (In some domains 'exclusion' is referred to 'repulsion'. Here, 'exclusion' is used, as 'repulsion' implies a causal relationship, which will not always be the case). Extreme self-exclusion will be a perfectly uniform distribution with respect to the data space. In terms of confidence, self-exclusion is simply the compliment of self-co-location.

Definition 2: A feature is defined as self-collocating in spatial data if the items representing that feature co-locate with each other at a ratio greater than some predefined threshold.

Definition 3: A feature is defined as self-excluding in spatial data if the items representing that feature co-locate with each other at a ratio less than some predefined threshold.

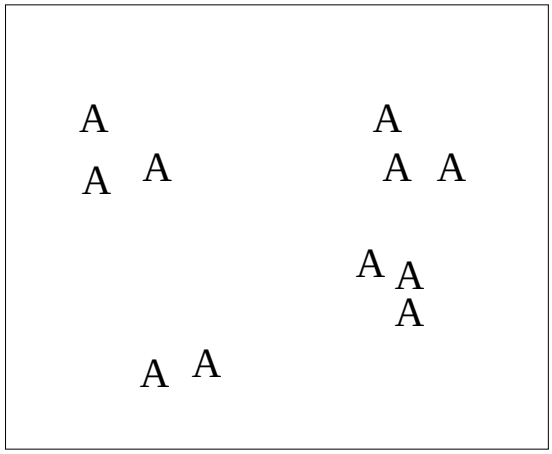

Figure 3: Example of self-co-location
A → A+

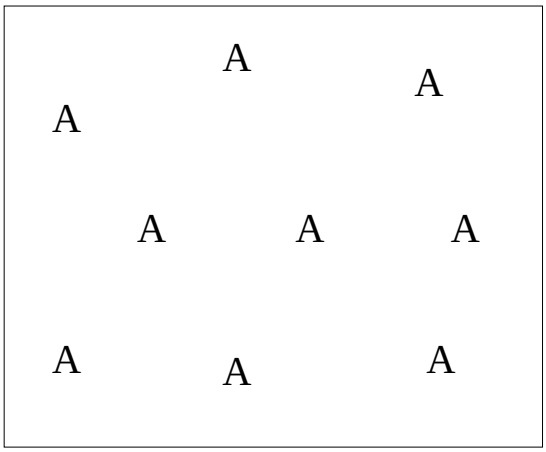

Figure 4: Example of self-exclusion: low confidence for
A → A+

5.4 One-to-Many relationships

Here, the user is specifically interested in the cardinality of a relationship between two features. As well as the possibility of a single item appearing in many co-locations, another property typical of spatial data is large variations in the frequency of items, such as the frequency of a certain microbe in forest compared to the frequency of certain bird. A result of this is that there may a large numbers of instances of one feature in a single clique. A one-to-many relationship, given by $A \rightarrow B+$ may be used to explicitly capture whether B is more likely to exhibit self-co-location in the presence of A, given by the situation where the confidence of $A \rightarrow B+$ is greater than the confidence of $B \rightarrow B+$.

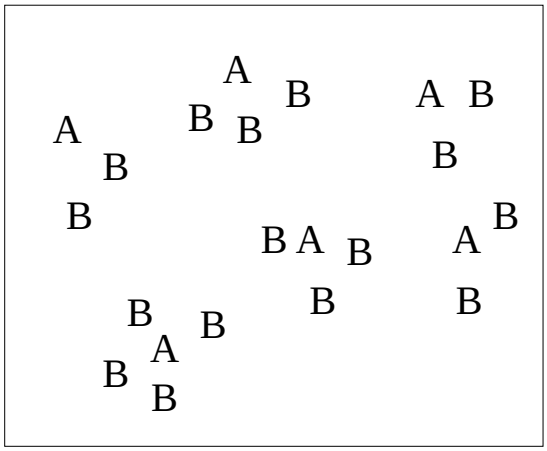

Figure 5: Example of a one-to-many relationship
A → B+

Definition 4: Two features are defined as having a one-to-many relationship in spatial data if one feature occurs multiple times in the presence of the other feature, greater than some pre-defined threshold. Included within this definition are two-way one-to-many (many-to-many) relationships.

5.5 Multi-feature exclusive relationships

These are exclusive relationships with respect to two or more features.

In terms of a transactional representation, they are negative rules, which are explored in Wu et al (2002).

There are two main problems associated with the mining of negative rules:

Firstly, with sparse data, negative sets will often be much larger than the positive ones. Wu et al (2002) addresses this problem by only generating negative rules from infrequent itemsets (a negative support threshold), with further pruning based on an interest measure adapted from Piatetsky-Shapiro (1991). While the former may be applied to spatial data, as discussed, a pruning strategy based on frequency will not always be appropriate. Furthermore, as described in the introduction, the interest measure of Piatetsky-Shapiro (1991) does not apply to spatial data. Section 6.1 describes how this problem should be addressed in spatial data.

Secondly, for every positive set of items of size k, there are 2^k-1 corresponding sets containing negative examples. For example, for the set *{A,B}*, there are also the negative sets: *{-A,B}*, *{-A,-B}* & *{A,-B}*. This is unavoidable if multi-feature exclusion relationships are to be mined, but with

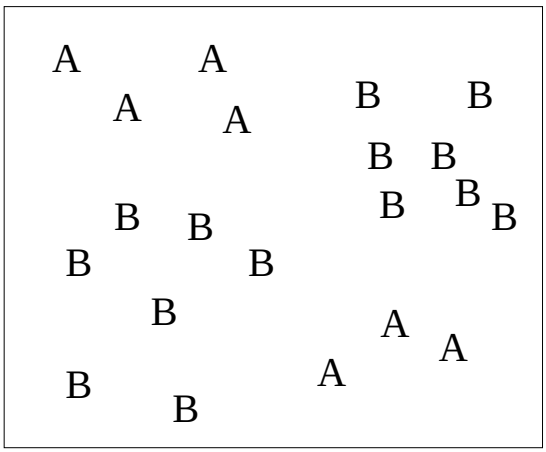

Figure 6: Example of a multi-feature exclusive relationship
$A \rightarrow -B$

the pruning of nonsensical rules such as $A \rightarrow -A$, the number of rules generated are less than that if there were an equal number of positive features added.

That an item is unbounded in the number of co-locations it may take place in also affects its positive/negative confidence. Given a confidence function *C()*, in market-basket data, there is a strict complementary relationship between a positive and negative rule given by:

$$C(A, B) = 1 - C(A, -B),$$

This will *not* always be the case in spatial data. For example, in figure 1, $C(C,D) = 3/6$, while $C(C,D) = 4/6 \neq 1-3/6$. This increases the inherent interest in explicitly mining multi-feature exclusive relationships.

Definition 5: A multi-feature exclusive relationship in spatial data is defined as where a feature is absent from a given co-location at a ratio greater than a predefined threshold.

Type	**Features**
Positive	A, B → C
Self-co-location/ self-exclusion	D → D+
One-to-many	A → D+
Multi-feature-exclusion	C → –A
Complex	A, –B → D+

Table 2: Examples of types of relationships in figure 1.

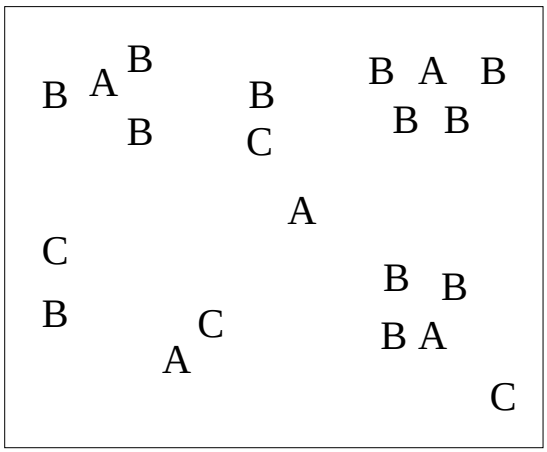

Figure 7: Example of a complex relationship
A, -C → B+

5.6 Complex relationships

These are any combination of two or more spatial relationship types.

Definition 6: A complex relation in spatial data is any relationship containing two or more of the properties defined in definitions 1-5.

The independent application of the above rules may produce complex relationships such as $A+ \rightarrow B$ *and* $A \rightarrow -C$, but will not produce complex relationships such as $\{A+, -C\} \rightarrow B+$.

5.7 Sparse data and the mining of absence

A participation index directly addresses the problem that a certain item may have low support resulting in it being absent from very many cliques, and hence having a high negative support, in that it with therefore have a low participation ratio for each of those cliques. For example, if *D* is an infrequent feature, then the rule $A \rightarrow -D$ will most likely be confident. However, the participation ratio of $-D$ in $\{A,-D\}$ will be very low, as $-D$ will occur in many cliques. In other words the participation ratio of $-D$ in $\{A,-D\}$ will only be high if *D* is *atypically* absent from cliques containing *A*. This also applies to dense data. If *D* is very frequent, then the participation ration of *D* in $\{A, D\}$ will also be low. Therefore, when using a participation ratio for sparse or dense data, there is, in fact, a gain in efficiency. The results in section 7 confirm this.

6 Statistical Applications of Complex Relationships

Complex relationships are not restricted to mining complex rules. Complex relationships can be used to provide stronger definitions and more accurate significance testing for simple relationships.

6.1 Significance as a Complex relationship

In terms of confidence, the significance of a rule is given by the extent to which the observed confidence of a rule differs from the expected confidence given by a random distribution.

Given a set of confident rules, the significance of these rules will dependent on the relative size of the space from which they were taken.

Lemma 1: The significance of a confidence rule of the form $A \rightarrow B$, is independent of the self-co-location/exclusion of *A*, but is dependent on the self-co-location/exclusion of *B*.

Proof Outline:

In general, given that *A* occurs with frequency *F(A)*, and *B* with frequency *F(B)*, in a two dimensional space with dimensions *x* and *y*, with cliques formed by a distance *d* the random chance of $A \rightarrow B$ can be given by the product of the fraction of the total volume that each features occupies:

$$\frac{F(A)\pi d^2}{xy} \times \frac{F(B)\pi d^2}{xy}$$

$$= \frac{F(A)F(B)\pi^2 d^4}{x^2 y^2}$$

The problem with the above, however, is that *B* may not be exclusively distributed with respect to itself.

Note that the random chance will not change with respect to the self-co-location/exclusion of *A*. The two extreme cases are:

1. *A* self-excludes such that no *A*'s are in the same clique. This is the equation given above.
2. *A* self-co-locates such that all *A*'s co-occur in one clique. In this case all *A*'s occupy an effective total space of $\pi^2 d$, giving *F(A)* an effective value of 1. However, if a *B* exists in that clique, then all *A*'s in that clique co-locate with it, so the equation must be multiplied by the number of *A's* in the clique - in this case *F(A)*. The equation is, therefore:

$$F(A)\frac{1 \times F(B)\pi^2 d^4}{x^2 y^2}$$

Which is obviously the equivalent of where *A* self-excludes.

The random chance will, however, change with respect to the self-co-location/exclusion of B. Assume the two extreme cases:

1. *B* self-excludes such that no *B*'s are in the same clique. Again, this is the equation given above.
2. *B* self-co-locates such that all *B*'s co-occur in one clique. In this case, all *B*'s occupy an effective total space of $\pi^2 d$, giving *F(B)* an effective value of 1. If one *A* exists in that clique, the number of *B*'s in that clique has no affect on the confidence as, by definition 1, only one unique *A* co-locates with a *B*. The expected value of $A \rightarrow B$ is therefore given by:

$$\frac{F(A) \times 1\pi^2 d^4}{x^2 y^2}$$

As the two extreme cases demonstrate, the expected value of $A \rightarrow B$ exists in a range with boundaries differing by a factor of *F(B)*. The consequence of this is that an accurate measure of the significance of a rule $A \rightarrow B$ must also include the measure of *B*'s self-co-location/exclusion. Therefore, by definition 6, in order to measure the significance of a simple relationship $A \rightarrow B$, it is necessary to know a complex relationship.

■

An approximation of self-exclusion may be given by the ratio of number of cliques containing *B* to the total number of *B*'s. Assuming *B* occurs in *cf(B)* cliques, this is given by:

$$\frac{F(B) - cf(B)}{F(B)}$$

This can underestimate the random chance, as it doesn't take into account the intersection of cliques in the data space where two or more *B*'s are greater than distance *d* but less than distance *2d* apart, or over-estimate, as it doesn't take into account the distance between items within a clique. The range

of possible underestimation to overestimation is dependent on the dimension of the data space. For one dimension, the range is a factor of *2*. In two and greater dimensions, the potential range increases, although the likelihood of intersection decreases.

Lemma 2: The potential range of confidence rules of the form $A \rightarrow B$, will depend on the self-co-location/exclusion of *both* *A* and *B*.

Proof Outline:

While the self-co-location/exclusion of *A* does not affect the significance of a confidence rule, it can limit the possible range of the observed confidence.

Assuming that all *A*'s and B's self-exclude, then, as in market-basket data, the maximum possible confidence for $A \rightarrow B$ is simply given by:

$$\min\left(\frac{F(B)}{F(A)},1\right)$$

Where $F(A) > F(B)$, this will obviously be less than *1*. However, if *A* self-co-locates such that $cf(A) \leq F(B)$, then the maximum possible confidence will be *1*.

Similarly, if *B* self-co-locates, then the maximum possible confidence may be lower. The exact measure for the maximum possible observed confidence is:

$$\min\left(\frac{cf(B)}{cf(A)},1\right)$$

■

A further factor that is not discussed here is where the potential size of some cliques extend beyond the boundaries of the measured space. Again, the random likelihood of this will relate to the ratio between *d* and the dimensions of the space. Here, it is simply assumed that it is very low.

6.2 Exclusion and maxPI

As a support threshold can prune low frequency confident rules, a maximal participation threshold can prune confident rules with low participation.

While maxPI will return the complete set of items that satisfy both thresholds of *maxPI* and the minimum confidence, there may be the case such that a high confidence rule will not have a high corresponding maximal participation index. For example, in figure 1, $C(A,B,C) = 1$ while $maxPI(A,B,C) = 2/3$.

An improved measure of participation includes the atypical exclusion of an item. We posit that by including the absence of items, we may discover a more robust measure for a participation index measure. In figure 1, $maxPI(A,B,C,-D) = 1$, indicating that $\{A,B,C\}$ is a strong co-location in that it atypically excludes *D*.

Whether or not such a measure is appropriate will depend on the nature of the phenomena the respective features represent.

7 A Representation of Spatial Data for Mining Complex Relationships

In this section we propose and test one simple representation of spatial data that facilitates the efficient mining of complex relationships.

7.1 Mining complex relationships using the maximal participation index

In terms of the steps in the problem definition, the steps taken are:

1. Generate all positive cliques in a transactional representation. To every clique we add features representing the absence of items and the presence of multiple items. (An example of this applied to the data in figure 1 / table 1 is given in table 3).
2. Apply the *maxPI* algorithm to the transactions, as described in section 4, automatically pruning trivial/nonsensical collocations such as $A \rightarrow -A$ and $A+ \rightarrow A$. (For an analysis of the efficiency and application across different spatial data sets, see Huang et al (2003)). Return a set of co-locations and their confidences. (Example results of different relationship types from figure 1 are given in table 4).
3. Calculate the significance of the confidences of the mined relationships, with respect to their significance, as described in section 6.1.

No	Clique
i.	C_1, D_1, –A, –B
ii.	C_5, D_1, –A, –B
iii.	C_5, C_6, –A, –B, –D, C+
iv.	A_4, D_2, D_5, –B, –C, D+
v.	A_5, C_3, D_3, –B
vi.	A_5, D_3, D_6, –B, –C, D+
vii.	A_1, B_1, C_2, –D
viii.	B_2, D_4, –A, –C
ix.	A_2, D_4, –B, –C
x.	A_3, B_3, C_4, –D

Table 3: Cliques in figure 1 with added features representing absent features (negative items), and the presence of multiple items.

Type	Features	Conf	MaxPI
Positive	A, B → C	1	2/3
Self-coll	D → D+	3/5	1
Self-excl	A → A+	0	n/a
One-to-many	A → D+	2/5	1
Multi-feature-excl	C → –A	1/2	3/4
Complex	A, –B → D+	1	1

Table 4: Relationship confidences in figure 1.

7.2 Test Sets

Data sets were created similar to those described in Agarwal et al (1994), but with the specific properties of spatial data, such the occurrence of a single item in many cliques, the occurrence of many items representing a single feature in one clique, and large variations in the relative frequency of items.

Set constituency was varied according to sparseness, the number of features, and the number of items. The mining of relationships was varied according to the participation and confidence thresholds. A comprehensive set of tests corresponding was completed across approximately different 100,000 set/parameter combinations. A summary of results is given below.

7.3 Results

Testing was undertaken to compare the efficiency of mining complex relationships to the mining of simple relationships with *maxPI*, and to investigate the relative frequencies of the different relationship types.

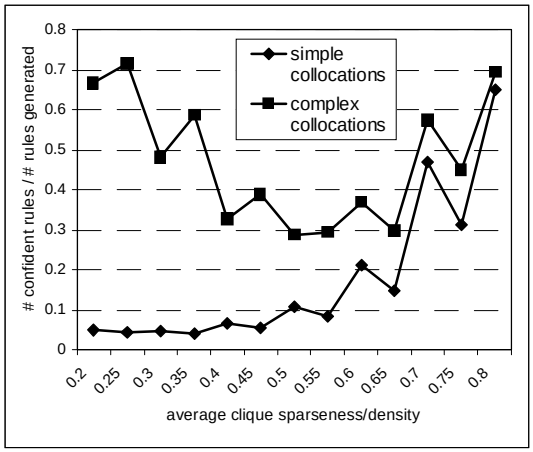

Figure 8: The relative efficiency of mining simple and complex co-locations.

Efficiency:

As Figure 8 shows, the ratio of rules generated to confident rules found is typically more efficient for the mining of complex rules, especially when the data is sparse. That the mining of complex rules in dense data is in all cases more efficient is due the automatic pruning of trivial/nonsensical collocations, as described above. Although it was never the case here, we do not rule out the possibility of the existence of a set such that the mining of simple relationships is more efficient than the mining of complex relationships.

The results in Figure 8 are the average ratios for approximately 10,000 randomly generated data sets, varied according to the average likelihood of a feature occurring in a given clique, given by the sparseness/density measure on the x-axis. The maximal participation index and confidence were held constant at *0.6* and *0.8*, respectively. Varying the maximal participation index had little impact on the respective ratios. Varying the confidence threshold varied the scale of ratio, but did not affect the scale of the two distributions with respect to each other.

A constant maintained across the generation of all sets were the inclusion of skews in the data such as: *'the probability of C appearing in a clique increases by 0.15 if A and B are present'*. These were originally generated randomly, than

maintained as averages about which all random sets were created. It is the interaction of such skews with the various thresholds that cause the unevenness in distributions in Figure 8.

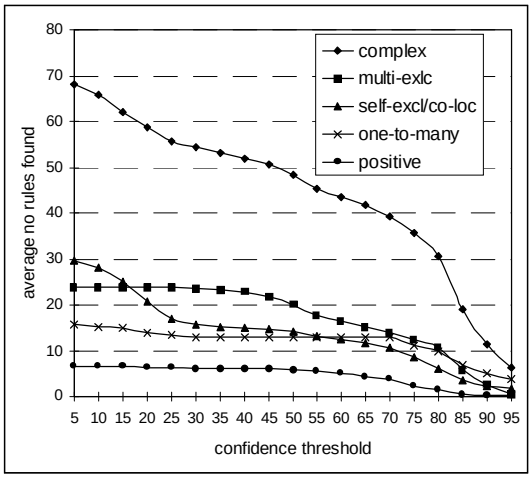

Figure 9: The relative frequency of different relationship types

Frequency of relationship types:

The results in Figure 9 are averages for approximately 1000 sets, each with 10 features. The number of features is the most sensitive variable in the relative frequencies, due to the fact that there is the possibility of exponentially more exclusive and therefore complex sets with respect to the number of features in a clique, as discussed in section 5.

Typically, the number of complex relationships found was greater than but correlated with the other relationship types found. As Figure 9 shows, the number of complex relationships at a given confidence threshold was sensitive to the variance in the number of the other relationship types.

Self-exclusion and self-co-location were modeled together in Figure 9 emphasize the complementary relationship between the two, as described in section 5.3. This is revealed in the corresponding steepness of the graph for self-exclusion/co-location at confidence < *0.3* and confidence > *0.7*.

7.4 The limitations and strengths of the representation

There are, of course, representational issues with any type of data. With market-basket transactions, there is a loss of information due to the granularity of the feature representation. For example, the well-known *'diapers → beer'* rule is a coarser granularity than mining specific types/brands of beer, but finer than mining simply *'diapers → alcoholic drink'*.

Limitation 1: in one-to-many relationships, this model doesn't capture interesting ranges or distributions in the 'many'. This is a task better suited for a mixture modeler, or the techniques described in Brin et al (2003).

Limitation 2: as pointed out in Shekhar and Chawla (2003), the cost of fully transcribing spatial data into a transactional representation can, in some cases, be more expensive than the mining of the co-locations themselves. As a full representation is necessary to accurately add the cliques with features representing absent and multiple items, a solution to this in the current representation may be problematic.

Limitation 3: as is typical in spatial data, it is assumed the number of features is unbounded. Where this is not the case, there would be problems in adding features representing the absence of items.

Strength 1: The most obvious strength of this representation is that, currently, it is the only model that allows the mining of complex relationships in spatial data.

Strength 2: A major strength of a transactional representation of spatial data not explored here is that it may be combined with non-spatial data. The addition of non-spatial data to the representation described here would be uncomplicated.

8 Conclusions / Future Work

We have defined the concept of complex relationships in spatial data.

We have described how, even in transactional representations, spatial data is fundamentally different from other forms of data, making the need to mine complex relationships of inherent interest.

We have demonstrated that even when simple relationships are the goal of mining spatial data, the mining of complex relationships is necessary for determining the significance of those relationships, and how the knowledge of complex relationships can lead to an improved measure of confidence for simple relationships.

We have implemented and demonstrated a transactional representation of spatial data that allows the efficient mining of complex

relationships, and discussed its limitations and strengths.

8.1 Future Work:

Apart from investigating improvements to the representation to address the limitations mentioned in 7.4, there are several future directions evident:

1. The potential use of other mining algorithms with the representation.
2. The combination of complex relationships with non-spatial features.
3. The application to other types of data with a spatial component, such as spatio-temporal data and to a lesser extent natural language and biological systems.

9 References

1. R. Agarwal and R. Srikant, 1994. *"Fast algorithms for Mining association rules"*, in proc of the 20th VLDB

2. S. Brin, R. Rastogi and K. Shim, *"Mining Optimized Gain Rules for Numeric Attributes"*, IEEE transactions on knowledge and data engineering, VOL.15, No.2, March/April 2003.

3. J. Han, J. Pei, and Y. Yin, *"Mining Frequent Patterns without Candidate Generation"*, in proc. 2000 ACM-SIGMOD Int. Conf. on Management of Data (SIGMOD' 00), Dallas, TX, May 2000.

4. Y. Huang, H. Xiong & S. Shekhar, 2003. *"Mining Confident Co-location Rules without A Support Threshold"*, in proc of the 18th ACM Symposium on Applied Computing (ACM SAC), Melbourne, Florida

5. K. Koperski and J. Han, *"Discovery of Spatial Association Rules in Geographic Information Databases"*, in Proc. 4th Int' l Symp. on Large Spatial Databases (SSD95), Maine, Aug. 1995, pp. 47-66.

6. G. Piatetsky-Shapiro, *"Discovery, Analysis, and Presentation of Strong Rule"*, in Knowledge Discovery in Databases 1991, pp. 229-248.

7. S. Shekhar and S. Chawla, *"Spatial Databases: A Tour"*, Prentice Hall, 2003 (ISBN 013-017480-7), published on June 15th, 2002.

8. S. Shekhar and Y. Huang, *"Discovering Spatial Co-location Patterns: A Summary of Results"*, in proc. of 7th International Symposium on Spatial and Temporal Databases(SSTD01), L.A., CA, July 2001.

9. X. Wu, C. Zhang, and Zhang, S., 2002. *"Mining Both Positive and Negative Association Rules"*, in proc of the 19th International Conference on Machine Learning (ICML-2002)