Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 06-006

Identifying Clusters in Marked Spatial Point Processes: A Summary of Results

Sandeep Mane, James Kang, Shashi Shekhar, Jaideep Srivastava,
Carson Murray, and Anne Pusey

March 20, 2006

Identifying Clusters in Marked Spatial Point Processes: A Summary of Results

Sandeep Mane*, James Kang,
Shashi Shekhar, Jaideep Srivastava
Department of Computer Science
University of Minnesota,
Minneapolis, USA

Carson Murray†,
Anne Pusey†
Department of Ecology, Evolution & Behavior
University of Minnesota,
St. Paul, USA

ABSTRACT

Clustering of marked spatial point process is an important problem in many application domains (e.g. Behavioral Ecology). Classical clustering approaches handle homogeneous spatial points and hence cannot cluster marked spatial point process. In this paper, we propose a novel intuitive approach,*Merge Algorithm*, to hierarchically cluster marked spatial point process. This approach treats all spatial point processes in a dendrogram's sub-tree as a single spatial point process while clustering. The resulting dendrogram for marked spatial point process needs be analyzed by a domain expert to identify clusters. To remove the subjective nature of the clusters identified, we propose a novel statistical method, *Cluster Identification Algorithm*, to partition a dendrogram into clusters. This approach identifies (*cuts*) a dendrogram's sub-tree as a cluster if that subtree's intra-subtree similarity is significantly higher than inter-subtree similarity. Experiments with Jane Goodall Institute's chimpanzee ecological dataset from the Gombe National Park, Tanzania which shows that our proposed methods identified clusters which were compatible to those identified by domain experts.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms

Keywords

Marked Spatial Point Process, Besag's L-function, spatial clustering, dendrogram cutting.

*Contact information: smane@cs.umn.edu

†Also affiliated to The Jane Goodall Institute's Center for Primate Studies, University of Minnesota, St. Paul, USA.

1. INTRODUCTION

The problem of clustering of marked spatial point process is a generalization of the problem of clustering spatial points; where instead of a single spatial location for each category, we have multiple spatial locations for each category. Each such category is a spatial point process. Additional information about the probability distribution (e.g. Poisson distribution) of spatial locations for each process may be available. Identifying clusters in marked spatial point process is an important problem found in many application domains. For example, in Behavioral Ecology, ecologists are interested in finding clusters of individuals based on their space usage, which usually consists of several spatial points for each individual. Another example is that of clustering of customers, based on all spatial locations visited by them in a shopping mall.

Classical clustering techniques (see Jain et al. [10]; Han et al. [7]) deal with only homogeneous spatial points, and hence cannot handle clustering of such marked spatial point process (See Apendix A). Thus, for example, it may be difficult to obtain a representative point (e.g. mean) for a cluster of spatial point processes in K-means clustering, whereas hierarchical clustering algorithms use pairwise dissimilarity between spatial point processes and thus lose the relationships between all spatial point processes within the cluster. New techniques are needed for clustering of marked spatial point process which motivates this current research.

In this paper, we formally define the problem of clustering marked spatial point process. We then propose a novel intuitive approach,*Merge Algorithm*, to hierarchically cluster the marked spatial point process and thus produce a dendrogram for marked spatial point process. This approach treats all spatial point processes in a sub-tree as a single spatial point process and thus computes its dissimilarity that other sub-trees (single spatial point process or clustered spatial point processes treated as one) of the dendrogram. Most of the techniques for automatically identifying clusters from dendrograms make use of some heuristic approach and hence are also subjective [14]. Normally, the resulting dendrogram is analyzed by the domain expert to determine the clusters of spatial point process. However, such clusters identified are affected by domain expert's perceptions and hence subjective in nature. Hence, to address the issue of automatically identifying clusters from a dendrogram, we propose a statistical approach,*Cluster Identification Algorithm*, to automatically partition the dendrogram into clusters. This approach

is based on the intuition that for a cluster, the intra-cluster similarity must be significantly higher than the inter-cluster similarity. The results of these two proposed approaches on a Jane Goodall Institute ecological dataset [6] were preferred over other approaches by domain scientists.

To summarize, the main contributions of this paper are:

1. A novel clustering approach,*Merge Algorithm*, is proposed for clustering marked spatial point process.
2. A new approach,*Cluster Identification Algorithm (CIA)* , is proposed, to statistically "cut" (partition) a dendrogram into clusters.
3. Experimental results using these algorithms on Jane Goodall Institute's chimpanzee datasets show that these approaches provide an interesting insight into the clustering of chimpanzees.

The rest of this paper is organized as follows. Section 2 describes the basic concepts and then formulates the problem of clustering marked spatial point process. Section A discusses the limitations of classical clustering approaches in the context of marked spatial point process. Section 3 explains our first proposed approach,*Merge Algorithm* and our second proposed approach,*Cluster Identification Algorithm*, for automatically identifying significantly clustered spatial point processes in a marked spatial point process. Section 4 presents the results of these proposed approaches using a real-world ecological dataset. Section 5 provides a brief survey on related literature. Section 6 concludes this paper and describes future work.

2. BACKGROUND

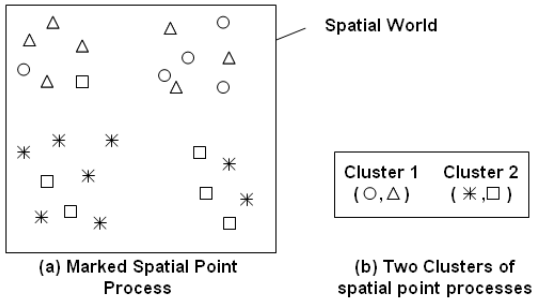
This section first introduces the basic concepts used in the paper and then formally defines the problem of clustering marked spatial point process.

2.1 Basic Concepts

In this sub-section, we explain the basic concepts which are useful for clustering of marked spatial point process.

1. A *spatial point process* [5] is a stochastic process on a given spatial region, a realization of which consists of spatial locations on that region. Spatial point processes can have three different types of interactions, namely, (i) *Complete Spatial Randomness* (CSR), the spatial locations are randomly distributed over the spatial region; (ii) *Clustering*, the spatial locations show a marked clustering in space; or (iii) *Repulsion*, the spatial locations show a marked repulsion from each other. In case of CSR, the spatial point process is proved to have a homogeneous Poisson distribution with intensity λ.
2. A *spatial world* is a well-defined geographic region of interest. All spatial point processes are observed in a common spatial world, i.e. all spatial points of each spatial process are present inside that spatial world. Without loss of generalization, the spatial world is assumed to be represented as two dimensional plane. Also, the spatial world is assumed to be a user specified input.
3. A *marked spatial point process* consists of several spatial point processes on a common spatial region, such that spatial locations for each category are identified by a unique label. Figure 1a illustrates an example of marked spatial point process observed over a given spatial world (rectangular boundary). This marked spatial point process consist of four different spatial point processes, all locations for a spatial point process are identified by a unique category.
4. *Clustering of marked spatial point process* consists of identifying groups of several spatial point processes based on their similarity. This is illustrated in figure 1b. Clustering for marked spatial point process is different from classical clustering. The main difference lies in the fact that in the former case, the items that are to be clustered are not points but set of points. Thus, this can thus be viewed as a generalization of spatial clustering, where in classical clustering is the case when each spatial point process (in marked spatial point process) has only one spatial point (though the assumption of probability distribution for the classical approach must be verified some way). Another important difference is that in clustering marked point processes, spatial locations belonging to a particular mark are usually assumed to have a particular probability distribution (usually homogeneous Poisson distribution); whereas in classical clustering, there is no assumption about the probability distribution for each spatial location.
5. For quantifying the interaction between two spatial point processes, we use the Besag's L-function [2, 3], which is a modified version the the Ripley's K-function [15]. The K-function is a spatial statistical measure to quantify the second-order interaction between spatial point processes. It measures the difference in the covariance (or correlation) between the observed and the expected pairs of points at a certain distance from each other. It is an isotropic (independent of direction) measure and allows the measurement of local as well as global interactions (by varying the distance at which it is measured). Formally,

Figure 1: Clustering of Marked Spatial Point Process.



DEFINITION 1. ***(Ripley's K-function)*** *The K-function $K_{XY}(r)$ between two spatial point processes X and Y at distance r is defined as ,*

$$\widehat{K}_{XY}(r) = An^{-1}m^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{I_r(d_{xy})}{w_{xy}} \qquad (1)$$

where $I_r(d_{xy})$ is an indicator function which is zero if distance between x and y is greater than r or 1 otherwise. n and m are the number of points in X and Y respectively, w_{xy} are the edge correction weights, and A is the area of spatial world.

Since the scale for $\widehat{K}_{xy}(r)$ is not linear in r, Besag's L-function is usually used to provide an easy linear interpretation of the interactions among point patterns. An important observation is that the variance of the function $\sqrt{\frac{K_{xy}(r)}{\pi}}$ is almost constant.

DEFINITION 2. ***(Besag's L-function)*** *The expected value of L-function is thus defined as –*

$$\widehat{L}_{XY}(r) = \sqrt{\frac{\widehat{K}_{XY}(r)}{\pi}} - r \tag{2}$$

A positive L-function represents similarity whereas negative L-function represents repulsion between spatial point processes. Hence, the negation of Besag's L-function is a dissimilarity measure.

2.2 Problem definition

Formally, the problem of clustering marked spatial point process is defined as follows:

PROBLEM DEFINITION 1. *Given a marked spatial point process $\mathbb{M} = \{ \mathbb{S} \mid \mathbb{S}$ is a spatial point process $\}$, a dissimilarity measure D and weights (w_1, w_2) , find clusters of spatial point processes by maximizing the objective function φ given by:*

$$\varphi : \sum_{\substack{\forall\ c_1,c_2 \in \\ set\ of\ clusters}} \mid w_1\ D_{\text{inter-cluster}}(c_1, c_2) - w_2[D_{\text{intra-cluster}}(c_1) + D_{\text{intra-cluster}}(c_2)] \mid \tag{3}$$

subject to the constraints,

(i) Each spatial point process is a stationary process, i.e. there is no time associated with the points in each spatial point process.

(ii) Each spatial point process has a Poisson distribution.

(iii) For the boundary cases, (i.e. clusters c_1 or c_2 or both consist of a single spatial point process), the dissimilarity measure $D_{\text{intra-cluster}}(c_i)$, $(i \in [1, 2])$, is taken as zero.

The weights (w_1, w_2) are used to assign different importance to the inter-cluster dissimilarity and intra-cluster dissimilarity. In general, other objective functions may be defined for simultaneously minimizing intra-cluster dissimilarity and maximizing inter-cluster dissimilarity of clusters.

Consider the example shown in Figure 1(a). Different possible partitions $\mathbb{S} = \{S_i \mid S_i$ is a partition of $\mathbb{M}\}$ of marked spatial point process can be enumerated as,

Algorithm 1 Merge Algorithm

Input:

- A bounding polygon defining the spatial world
- A marked spatial point pattern within the spatial region
- Distance 'r' for clustering
- A dissimilarity measure D between two spatial point processes.

Output:

- A hierarchical clustering of marks.

Pseudo-code:

1. Initialize the set S = $\{m|m$ is a mark which uniquely identifies all locations of spatial point process $\}$
2. Repeat until the size of S = 1
3. For each m_i,$m_j \in$ S,
4. Compute $D_{m_i,m_j}(r)$ each pair of spatial point processes with marks m_i and m_j.
5. End for
6. Find spatial point processes with marks $m_i^{'}$ and $m_j^{'}$ such that $D_{m_i^{'}m_j^{'}} = min\{D_{m_i^{'}m_j^{'}}\}$
7. Cluster spatial point processes $m_i^{'}$ and $m_j^{'}$.
8. Set marks $m_j^{'} = m_i^{'}$
9. Update set S = $\{S \setminus m_j^{'}\}$
10. **End repeat**

$$S_1 = \{(\circ, \triangle), (*, \square)\},$$
$$S_2 = \{(\circ, *), (\triangle, \square)\},$$
$$\vdots \qquad\qquad \vdots$$

The problem of clustering of marked spatial point process is to identify a partition S_j which maximizes the objective function given by equation 3. For example, for the marked spatial point process in the figure 1(a), it is shown that S_1 is the preferred clustering (Figure 1) as it maximizes the objective function.

3. PROPOSED APPROACH

In this section, we propose the Merge algorithm for clustering marked spatial point process. We illustrate this algorithm using an example and then explain how it is different from other hierarchical clustering algorithms. Then we address the problem of identifying clusters from a given dendrogram. An approach is proposed for identifying clusters in a dendrogram using t-test.

3.1 Merge Algorithm

We propose a new hierarchical algorithm called *Merge Algorithm*. The main idea in this algorithm is described in Algorithm 1: For each pair of spatial point process, the dissimilarity measure between them is computed (line 4 in Algorithm 1). Then, the pair of spatial point processes which has minimum dissimilarity (i.e. maximum similarity) are clustered (lines 6-7). In next iteration, the two clustered

Algorithm 2 Cluster Identification Algorithm (CIA)

Input:

- Hierarchical clustering of marks (i.e. dendrogram)
- A dissimilarity measure D between two spatial point processes.
- Confidence threshold α for t-test

Output:

- Marked Spatial Point Process Clusters.

Pseudo-code:

1. For each subtree T of dendrogram,
2. Compute the dissimilarity measure of the subtree D_T.
3. End for.
4. Let $ED_{T,T-1} = D_T - D_{T-1}$ be the edge difference between each sub-tree (T-1) and the subtree T at its parent.
5. Compute one-sided (upper bound) t-test with $(1-\alpha)-\epsilon$ where ϵ is a correction for the boundary case. Let Δ_{upper} be the upper bound.
6. Remove all edges i from the dendrogram s.t. $ED_{T,T-1} > \Delta_{upper}$ to partition the dendrogram into clusters.

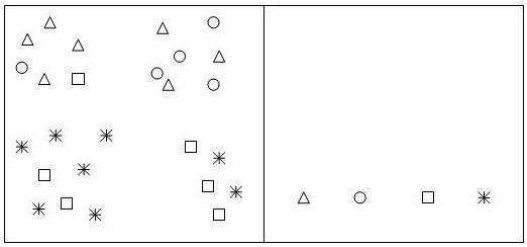
spatial point processes are treated as if it were a single spatial point process, i.e. all the points belonging to both clustered processes are treated as if coming from same spatial point process. Thus, the spatial point processes are repeatedly clustered in a greedy manner to obtain a hierarchical clustering of spatial point processes.

An example of the Merge algorithm is demonstrated in Figure 2. In this example, there are four point processes shown in 2a (left-half) and are the first level of the hierarchical dendrogram (right-half). Then using a distance function such as Ripley's K-function, we can compute the minimum distance between all marked point process pairs such as (□, ∗), (□, △), (□, ◯), (∗, △), (∗, ◯), (∗, □) point processes shown in 2b. Assume that (□, ∗) is selected as the closest pair in this iteration. These point processes are then merged together to form a new marked point process. We then compute the distances between the original two symbols (△ and ◯) with the newly formed point process. As shown in step 3, the △ and ◯ processes are the next minimum and are merged together to form another newly formed marked point process. Finally, in step 4, there are two remaining point processes which are merged together to form the final marked point process and the dendrogram is completed.

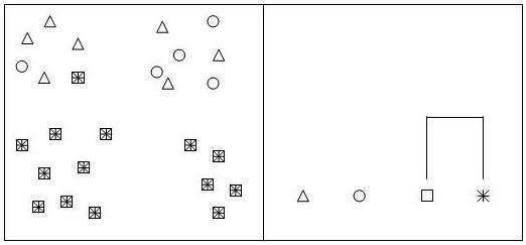
3.2 Cluster Identification Algorithm (CIA)

For clustering using dendrograms, a method is also needed for "cutting" the dendrogram to identify the major clusters in the dendrogram. A widely used approach is to give the dendrogram to a domain expert and ask the expert to identify natural groupings of the data items. This approach has its pitfall, one of which is the subjective nature of identified clusters. To avoid this, we propose a statistical method for

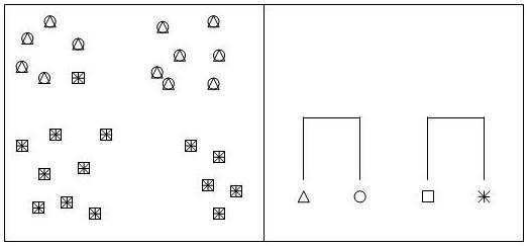
Figure 2: Illustration of Merge Algorithm.



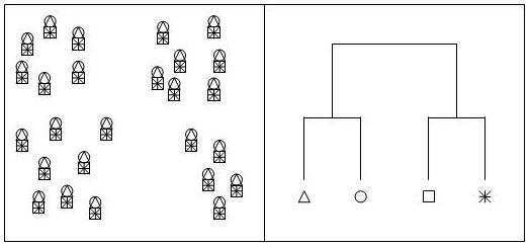
(a) An example of Marked Spatial Point Process (left) and first level of hierarchical dendrogram (right).



(b) Suppose □ and ∗ had the minimum distance between all pairs. Then □ and ∗ are merged together (left) and shown in the dendrogram (right).

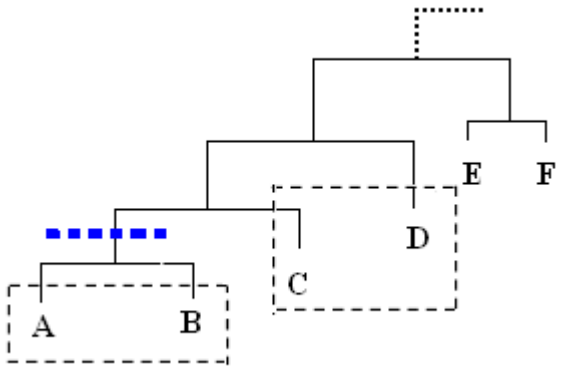


(c) The dissimilarity measure is re-calculated and suppose △ and ◯ are found as the minimum pair and merged (left), then placed in the dendrogram (right).



(d) Finally, the last two clusters are found to be the minimum and are merged (left) and the dendrogram is complete (right).

Figure 3: Iterative use of the threshold $\mathrm{ED}_{\mathrm{upper}}$.



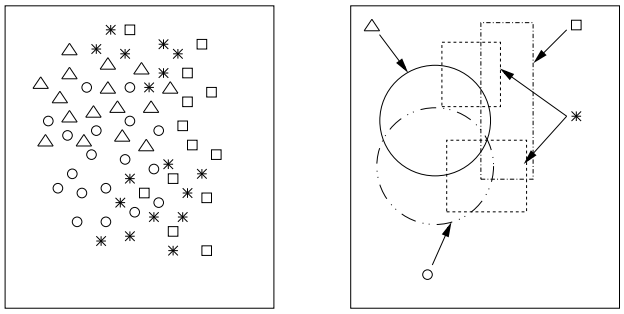
identifying clusters from a dendrogram.

The proposed algorithm, Cluster Identification Algorithm (CIA), is a technique for statistically identifying clusters in a dendrogram. Since each clustered data item has a Poisson distribution, we assume that the difference in dissimilarity metric between adjacent levels of dendrogram has a Gaussian distribution. If the sample size is sufficiently high, then Normality tests can be used to verify whether this assumption holds.

In this approach, the pseudo code of this algorithm is given in Algorithm 2. The main intuition behind this approach is that data items having (statistically) significantly more intra-cluster similarity than inter-cluster similarity will form a distinct cluster. As shown in Algorithm 2, we determine the difference in dissimilarity measure between all adjacent levels of dendrogram (line 4). The method is applicable to all dissimilarity measures as long as the normality assumption holds. Line 5 applies the t-test [4] to all differences of dissimilarity measure between adjacent levels of dendrogram and determine the upper quantile for $(1-\alpha)$. Since a strict upper quantile determined this way may not accommodate the boundary cases well, we instead use a cut threshold $\mathrm{ED}_{\mathrm{upper}}$ based on the quantile for $(1-\alpha)-\epsilon$, where ϵ is thus the correction for the boundary case. The threshold $\mathrm{ED}_{\mathrm{upper}}$ is used to determine which edges of the dendrogram should be cut to obtain clusters. The time complexity is linear since the height values are obtained by a single pass through the dendrogram.

However, a problem occurs for the case where only one of the subtrees at some intermediate level in the dendrogram is cut. This is illustrated in the figure 3, where A, B, C, D, E and F may be data items or clusters. In our approach, when an subtree (A,B) is cut, then both - the left and right subtrees (A,B) and C - are treated as separate clusters. The reason for this is that the two subtrees are already been statistically different (intra-cluster similarity is significantly more in one subtree than the inter-cluster similarity). Now, consider the subtree (B) not having a significant cut. It is combined with higher level cluster/data item - D - (with which it (A,B) and

Figure 4: Clustering Marked Spatial Point Process using MAX vs. Merge Algorithm.

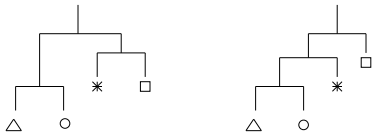


Marked Spatial Point Process

Representative spatial distribution of of spatial point processes

(a) (b)

An Example of Marked Spatial Point Process.



(c) (d)

Two Possible Clusterings of the Marked Spatial Point Process.

the other subtree (B) are clustered along their path to the root of dendrogram), assuming that the maximum distance between any two data items across C and D is not more than $\mathrm{ED}_{\mathrm{upper}}$. However, C may not be combined with the cluster (E,F) as the difference between the intra-cluster distance for C and inter-cluster distance between (((A,B),C),D) and (C,D) is much more than $\mathrm{ED}_{\mathrm{upper}}$. Thus, the idea here is to iteratively use the threshold $\mathrm{ED}_{\mathrm{upper}}$ across multiple levels to determine the cuts, if required.

Thus, an advantage of our proposed overall approach is that it reduces the effort for a domain expert to determine correlated data items and then identify clusters.

4. EVALUATION

4.1 Time complexity of Merge algorithm

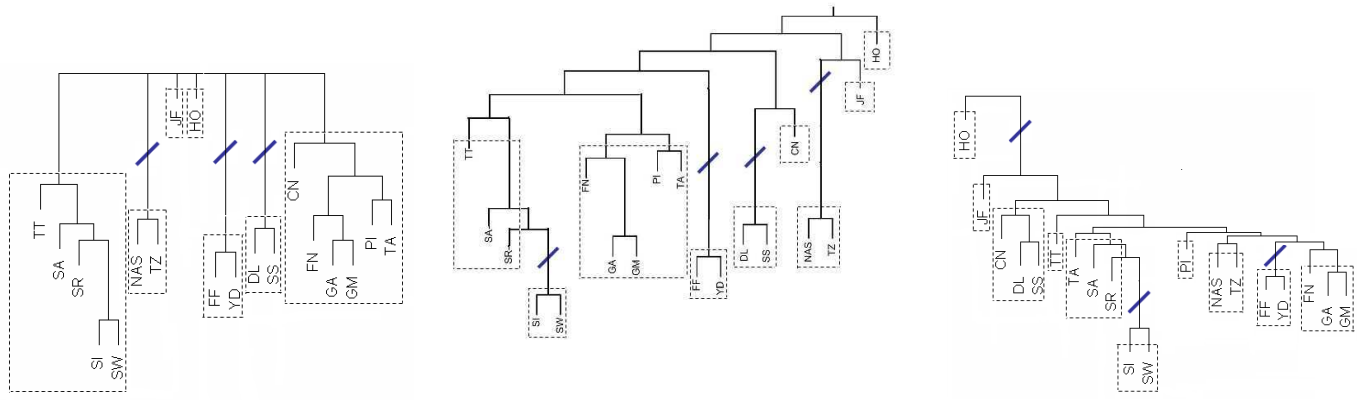
LEMMA 1. *The worst-case computation cost for the Merge algorithm is $O(nN^2)$, where N is the number of total locations in the marked spatial point process and n is the number of spatial point processes in the marked spatial point process.*

Proof:
For the marked spatial point pattern S, let $M = \{m\}$ be the set of all marks and let $|M| = n$. Let $p_1, \ldots, p_n$ be the number of points of each respective mark $m_i \in S$, $1 \leq i \leq n$, and let

$$N = \sum_{i=1}^{n} p_i$$

be the total number of points (irrespective of the marks). Here it is assumed that no two marks have points with same location (i.e. all points are distinct). Note, in the case the

Figure 5: Experimental results of Cluster identification algorithm, using 50%percent kernel chimpanzee locations data for 2001-2002 ($\alpha = 5\%$)



(a) Max algorithm (b) Merge algorithm (c) Single-link algorithm

Table 1: Summary of clusters using Max, Merge and Single-Link clustering algorithms. (Note: The Most common chimpanzees across clusters of the three algorithms are shown in bold.)

Max	Merge	Single-Link
GA,GM,FN,FF,PI,TA,CN	**GA,GM,FN**,FF,PI,TA	**GA,GM,FN**
DL,SS	**DL,SS**	**DL,SS**,CN
YD,FF	**YD,FF**	**YD,FF**
NAS,TZ	**NAS,TZ**	**NAS,TZ**
HO	**HO**	**HO**
JF	**JF**	**JF**
SA,SR,TT,SW,SI	**SA,SR,TT**	**SA,SR**,TA
	SI,SW	**SI,SW**
	CN	TT
		PI

marks are not distinct, the worst case time complexity will reduce further.

Since the Merge algorithm combines two spatial point processes (i.e. two marks) at an iteration, the total number of iterations is $I = n - 1$. At each iteration j, $(1 \leq j \leq I)$, there are $(I - j + 2)$ spatial point processes. At the j^{th} iteration, the worst case cost of computing all pairs of K-function is when all the subsets are of equal size or nearly equal size. For this worst case, the number of points in each spatial point process at the j^{th} iteration is $\left(\frac{N}{I-j+2}\right)$. Thus, the cost of computing the K-function for each pair of marks is $O\left([\frac{N}{I-j+2}]^2\right)$. Since there are $(I-j+2)$ such spatial point processes, the number of K-function values that need to be computed is equal to the number of unique pairs of marks at j^{th} iteration. There are C_2^{I-j+2} such possible unique pairs of marks. Thus, the cost of computation at j^{th} iteration is $O\left((\frac{N}{I-j+2})^2 \times C_2^{I-j+2}\right)$. Summing this over all I iterations, we have the total cost as

$$O\Big(\sum_{j=1}^{I}\Big[\Big(\frac{N}{I-j+2}\Big)^2 \times C_2^{I-j+2}\Big]\Big).$$

This, on simplification, gives,

$$O\Big(N^2 \sum_{j=1}^{I}\Big[\frac{(I-j-1)}{(I-j+2)(2!)}\Big]\Big) \leq O(N^2 I) = O((n-1)N^2).$$

Thus, the worst-case computation cost for the Merge algorithm is $O(nN^2)$.

■

4.2 How is Merge Algorithm different ?

One approach for clustering marked spatial point process techniques is to use an exiting hierarchical clustering technique (e.g. MAX, MIN, Average) using a dissimilarity measure (e.g. L-function) for spatial point process. However, in such an approach, we loose information since only pairwise dissimilarities between spatial point processes are used. Another approach is to use a hyper-graph based clustering approach where interactions between all possible subsets of spatial point processes are used to determine the clusters. However, this is a computationally expensive approach. In Merge algorithm, as spatial point processes are clustered, higher-level interactions between subsets of spatial point processes are used for clustering while requiring less computational costs than the hyper-graph based approach.

An example is illustrated to show that Merge algorithm can identify new, interesting clusters. For comparison, the clustering using MAX algorithm is also shown. Consider the marked spatial point process with four different spatial point processes with marks - square($\Box$), asterix($*$), triangle($\times$) and circle($\bigcirc$) as shown in figure 4(a). Figure 4(b) shows a representative distributions of the points for each spatial point process within the spatial world. We make two assumptions in this example, which can be shown to hold for a marked spatial point process:

(i) The maximum of the dissimilarities between (triangle, asterix) and (circle, asterix) pairs of spatial point processes is more than the dissimilarity between (square, asterix) pair of spatial point processes.

(ii) The dissimilarity between (triangle|circle,asterix) pair of spatial point processes is less than the dissimilarity between (square,asterix) pair of spatial point processes. (The notation "*triangle*|*circle*" means that the two spatial point processes triangle and circle are treated as a single point process.)

For the marked spatial point process, in the first step for both MAX and Merge algorithms, the 'triangle' and 'circle' spatial point patterns will be clustered. For two clusters of spatial point processes, the MAX algorithm assigns the maximum dissimilarity between any two pairs of spatial point processes across clusters as the dissimilarity between the two clusters. Thus in the second step, the MAX algorithm will cluster the square and asterix spatial point processes whereas the Merge algorithm will cluster the asterix spatial point process with the cluster (triangle,circle). The dendrogram using MAX algorithm will be as shown in figure 4(c) while the dendrogram using Merge algorithm will be as shown in figure 4(d). The clustering in figure 4(d) is interesting, as it captures the higher-order interactions between a spatial point process to a cluster of spatial point process. Thus, the Merge algorithm is a very useful approach for finding interesting clusters in marked spatial point process.

4.3 Experimental results

We tested our algorithms with the same dataset investigated in our previous research [12]. Since 1963, Behavioral Ecologists have gathered data about chimpanzee movements in the Gombe National Park, Tanzania [6]. The data used in our analysis is location information for female chimpanzees, divided into different time periods. The location information for each female chimpanzee consists of the set of all locations where a particular female chimpanzee arrived "alone." A female chimpanzee was "alone" if no other adult chimpanzee (except in case of mother-daughter pair) also arrived within five minutes. In addition, we used a 50% kernel to filter the locations of female chimpanzees, thus reducing the chances of having outliers. The spatial region was defined by a minimum convex polygon around all "alone" locations within the time period under consideration.

Based on our previous research, we used the K-function computed at a distance of 250m as a dissimilarity measure for clustering in subsequent analysis. Similar to our previous research [12], we used the MAX (Complete-link) clustering algorithm for constructing the dendrogram. In addition, we also used the Single-link clustering algorithm as well as the proposed "new" Merge Algorithm for clustering. Figure 5 shows the dendrograms obtained for the three algorithms.

Dendrograms are sensitive to minor fluctuations in dissimilarity measure, which makes it difficult to identify the clusters. Hence, CIA was applied to the three dendrograms obtained. Different values of $\alpha \in \{1, 5, 10\}$ were used, but the results of the experiments were similar. Figure 5 shows the result of the statistically identifying the clusters of female chimpanzees for $\alpha = 5$. The diagonal line on an edge represents that that edge has a difference in dissimilarity measure which is more than $\mathrm{ED}_{\mathrm{upper}}$. The dotted boxes represent the cluster of female chimpanzees. Table 1 summarizes the clusters obtained for each clustering approach, with the most common chimpanzees across each row shown in bold type. It is observed that the clusters identified across the three dendrograms are similar, with the similarity being more prominent between the Max and Merge algorithms. The reason that Single algorithm is less similar is be cause it is susceptible to noise, which may be present for peripherally located females. The similarity of results between Max and Merge algorithm is promising and support our hypothesis that a statistical methodology can be used for cutting a dendrogram in case of clustering marked spatial point process.

Due to the small number of samples for t-test, the differences in the inter-cluster and intra-cluster dissimilarity measures could not be evaluated using normality tests. However, we will illustrate that this assumption holds using large synthetic datasets of marked spatial point process in our future work.

5. RELATED WORK

Classical clustering approaches [7] focuses on only the spatial location of the object and ignores the type of the object. Our previous approach [12] was able to use both the object type and the spatial location as part of the clustering technique. We illustrated this problem and two different clustering approaches were proposed. For both approaches, the Besag's L-function [15, 2, 3] was used as a dissimilarity measure. The first approach, which is also of interest in this paper, used the MAX (or Complete Link) clustering algorithm. The identification of clusters in dendrograms thus obtained were difficult for a domain scientist, since the MAX algorithm cannot be easily explained for marked spatial point processes. We also proposed an alternate approach using Ripley's K-function and reverse Cuthill-McKee algorithm for block diagonal matrices. However, the latter approach, being heuristic, may not always give easily visualized clusters. We illustrated these algorithms for studying spatial usage of chimpanzees. The data consisted of spatial locations of different chimpanzees (*Pan troglodytes*) in Gombe National Park, Tanzania, which were shown to form a marked spatial point process.

Andrews [1] shows such an approach used to determine clusters of authors based on the authors correlations to cocitations of medical documents. Other heuristic approaches used for "cutting" the dendrograms include determining the cut based on the distance change between each level for a user specified threshold; finding spatial shapes using random walks [8]; identifying a cut using a metric to identify

correlated genes using micro arrays [11, 13]; obtaining clusters to find classification of words [9]; and finding optimal number of clusters in video sequences [16]. One difficulty with previous cutting methods is that each data item (leaf) in the dendrogram is a single point and it is therefore difficult to use a statistical method to identify the clusters [14]. In case of clustering of marked spatial point patterns, each data item (leaf) is itself composed of several data points having a Poisson distribution. Thus, we can make us of this to statistically identify the clusters in the dendrogram. To the best of our knowledge, no prior work has been done to statistically identify clusters in the dendrogram for marked spatial point patterns.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we discuss the difficulties inherent in of identifying clusters from a dendrogram. We initially explain the problem of clustering marked spatial point processes and then, in continuation with our previous research, propose a new algorithm,*Merge Algorithm*, for clustering them. However, the output of such an algorithm is still a dendrogram, which previously would require a domain expert to identify the clusters. To address this problem, we propose another algorithm,*Cluster Identification Algorithm*, which is used to identify the clusters in the dendrogram. Our experiments using the Jane Goodall Institute's chimpanzee data and the mentioned two algorithm has shown promising results. We were able to statistically identify stable (similar) clusters in all the dendrograms obtained using different clustering approaches. These clusters are observed to be in sync to other chimpanzee behavioral information gathered by behavioral ecologists.

With more complex data being collected and analyzed in terms of clustering, the problem of identifying clusters will become more pronounced. In such cases, we suggest that our new approach can be used, provided that the difference in dissimilarity measure follows Normal distribution. Application of this approach to such domains needs to be explored. In addition, we also plan on more experiments using other data sets, e.g. male chimpanzees' alone location data.

7. ACKNOWLEDGMENTS

Sandeep Mane's and Carson Murray's research were supported by NSF Grant No. IIS-0431141. James Kang was supported by a NSF IGERT fellowship during this research. The authors also thank The Jane Goodall Institute (`http://www.janegoodall.org/`) for the availability of data for this research.

8. REFERENCES

[1] J. Andrews. An author co-citation analysis of medical informatics. *Journal of the Medical Library Association*, 91(1):47–56, 2003.

[2] J. E. Besag. Comments on Ripley's paper. *Journal of the Royal Statistical Society B*, 39(2):193–195, 1977.

[3] N. A. C. Cressie. Statistics for spatial data. New York, 1993. Wiley.

[4] M. H. DeGroot and M. J. Schervish. *Proabability and Statistics, 3rd Edition*. Addison Wesley, 2002.

[5] P. J. Diggle. *Statistical Analysis of Spatial Point Patterns, 2nd Edition*. Oxford University Press: London, 2003.

[6] J. Goodall. The chimpanzees of gombe: Patterns of behavior. Cambridge, MA, 1986. Harvard University Press.

[7] J. Han, M. Kamber, and A. K. H. Tung. Spatial clustering methods in data mining: A survey. In H. Miller and J. Han, editors, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.

[8] D. Harel and Y. Koren. Clustering spatial data using random walks. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 281–286, 2001.

[9] J. Hughes and E. Atwell. The automated evaluation of inferred word classifications. In *European Conference on Artificial Intelligence*, pages 535–539, Amsterdam, The Netherlands, 1994.

[10] A. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), September 1999.

[11] S. Leach and L. Hunter. Comparative study of clustering techniques for gene expression microarray data. In S. Miyano, R. Shamir, and T. Takagi, editors, *Currents in Computational Molecular Biology*, pages 198–199. Universal Academy Press, Inc. Tokyo, 2000.

[12] S. Mane, C. Murray, S. Shekhar, J. Srivastava, and A. Pusey. Spatial clustering of chimpanzee locations for neighborhood identification. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM)*, 2005.

[13] L. McShane, M. Radmacher, B. Freidlin, R. Yu, M.-C. Li, and R. Simon. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. volume 18, pages 1462–1469, 2002.

[14] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. volume 50, pages 159–179, 1985.

[15] B. D. Ripley. The second-order analysis of stationary point processes. volume 13, pages 255–266, 1976.

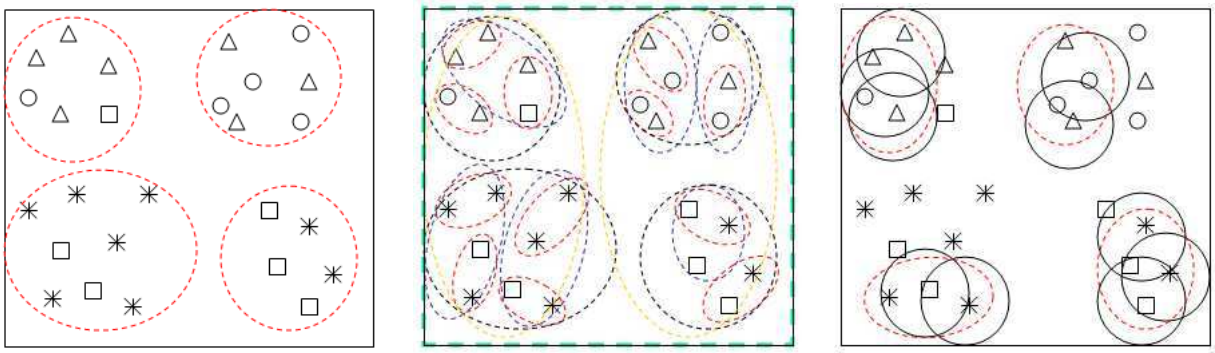
[16] C. Taşkiran, J.-Y. Chen, C. Bouman, and E. Delp. A compressed video database structured for active browsing and search. In *Proceedings of the 1998 International Conference on Image Processing*, Chicago, IL, USA, 1998.

APPENDIX

A. LIMITATION OF CLASSICAL CLUSTERING APPROACHES

Classical clustering approaches focus on the spatial location of each object rather than the type of the object represents. These classical clustering methods tend to fall in three

Figure 6: Classical Clustering Approaches on Marked Point Processes.



(a) Partitioning (b) Hierarchical (c) Density-based

general categories of partitioning, hierarchical, and density-based [7]).

A popular partitioning technique, K-means [7], groups objects together based on their spatial locations shown in Figure 6(a). In this example, the labels of each object are not considered and separates the marked point processes into multiple clusters. This defeats the purpose of clustering marked point processes since we want to group these together where each process is a single entity which cannot be broken up.

Hierarchical clustering is used in various domains and implemented using several methods (i.e. Chameleon [7])(b). In general, hierarchical clustering groups objects together by first creating clusters on the shortest distances between objects until all objects are clustered together. Figure 6 shows an example of this process where the smaller circles represent the first level of clustering, whereas the larger circles represent the next levels of clustering and finally the spatial region is clustered together as one cluster. As in the previous approach, hierarchical has the same limitation as partitioning where the marked point processes are split into multiple clusters.

Another general technique of clustering is density-based where objects are clustered based on their surroundings. Figure 6(c) shows an example of this where the solid circles represent the surrounding regions and the dotted circles are the clusters of the overlapping regions. Two main limitations arise using this approach for marked point processes: (1) as in previous classical methods, the marked point processes are split into multiple clusters, and (2) non-clustered points are considered as noise where every spatial object is important to its own point process.

In general, classical clustering approaches focus on only the spatial location and ignores the type of each object. In our problem, the spatial location and the type of the object is equally important. In the next section, we devise a method to consider both aspects when clustering marked point processes.