

Modeling Herds and Their Evolvments from Trajectory Data

Yan Huang¹, Cai Chen², and Pinliang Dong³

¹ Department of Computer Science
University of North Texas, Denton, Texas, U.S.A
[huangyan@unt.edu]

²,³ Department of Geography
University of North Texas, Denton, Texas, U.S.A
[caichen@unt.edu², pdong@unt.edu³]

Abstract. A trajectory is the time-stamped path of a moving entity through space. Given a set of trajectories, this paper proposes new conceptual definitions for a spatio-temporal pattern named *Herd* and four types of herd evolvments: *expand*, *join*, *shrink*, and *leave* based on the definition of a related term *flock*. Herd evolvments are identified through measurements of *Precision*, *Recall*, and *F-score*. A graph-based representation, *Herd Interaction Graph*, or *Herding*, for herd evolvments is described and an algorithm to generate the graph is proposed and implemented in a Geographic Information System (GIS) environment. A data generator to simulate herd movements and their interactions is proposed and implemented as well. The results suggest that herds and their interactions can be effectively modeled through the proposed measurements and the herd interaction graph from trajectory data.

Keywords: Spatio-temporal Data Mining, Spatial Patterns, Spatial Evolvments, Herd Evolvments

1 Introduction

Many moving objects exist in the air, on the ground, or in the ocean. The detection and description of spatio-temporal patterns are essential for better understanding of the behaviors of moving objects (animals, vehicles, and people). For example, models of movements can be used to study the ecology of animal behaviors, habitat preferences, and the dynamics of population densities [4]. Other application examples of the analysis of moving objects include studies in socio-economic geography [8], transport analysis [19], defense, and surveillance [18]. With increased accessibility to data collected by Global Positioning Systems (GPS), radio transmitters, and other location-aware devices, the processing, storage, management, mining and analysis of data, information and knowledge related to moving objects have been a research focus in the last few years.

Trajectory analysis has been the focus of many research efforts recently and is of particular interest in many fields. A trajectory is a sequence of time-stamped point locations describing the path of a moving object with identity e over a period of time. Given a set of trajectories of a set of entities, the problem that this paper deals with is to mine grouping dynamics of moving entities described by their trajectories over time. Assuming the time used by an entity between any two consecutive locations is the same (uniformed sampling), then a trajectory can be represented by $tj(e, \tau) = \langle p_1, p_2, \dots, p_\tau \rangle$, where τ is the length of the trajectory and $tj(e, \tau)[t]$ represents the point locations visited by the entity e at time snapshot $t(1 \leq t \leq \tau)$.

Recently, algorithms to find *flocks* have been proposed to identify groups of entities that travel together for an extended period of time [10]. Formally, given the trajectories of a set of n entities, a time interval I of at least k consecutive snapshots, and a distance r , a flock $f(m, k, r)$ is a set of at least m entities such that for every snapshot t in time interval I , there is a disk of radius r that contains all the m entities [10].

The concept of *flock* is based on several parameters to be provided by a user: the minimum number of entities m , the duration of a time interval k , and the radius r . The focus of *flock* is more on query-based data exploration. However, there are several limitations with a flock-based approach: (1) Entities, e.g. caribous, do not always gather in circular shapes. Finding at least m entities in a radius r , may not give an accurate picture of a *flock*, which most likely roams in arbitrary shapes; (2) Given m , k , and r , the flocks found will have many overlaps. For example, flock F_1 may consist of $m = 100$ entities E traveling within radius r for a time interval I of at least length $k = 1000$. For a query with $m = 50, k = 500$, any subset of E will be qualified as a flock and be returned as a result, which will lead to C_{100}^{50} flocks (choose 50 from 100 entities). Furthermore, for each flock in the C_{100}^{50} flocks, any sub-interval of the I will result in a flock, leading to $C_{100}^{50} \times (1000 - 500 + 1)$ flocks. Although algorithms for discovering longest flocks have been proposed [10], the problem of combinatorial explosion related to entities has not been addressed; (3) Flocks *move* and *evolve* over time. Because of these limitations, there is a need to discover how flocks interact with each other over time throughout the observations. In this paper we focus on modeling *group* traveling patterns and the evolvments and interaction of these groups.

The major contributions of this paper are: (1) A new concept *herd* is proposed for spatio-temporal patterns along with four types of spatial evolvments: *expand*, *join*, *shrink*, and *leave*; (2) Clustering-based methods are used to detect herd snapshots in trajectory data, and mathematical measurements of *Precision*, *Recall*, and *F-score* are proposed to identify herd evolvments; (3) A graph-based representation for herd evolvment is designed and implemented as an extension to a Geographic Information System (GIS). A herd evolvment simulator is implemented.

2 Related Work

Time is an essential dimension for analyzing and interpreting real-world evolution. In paper [6], the authors presented a set of design patterns for modeling spatio-temporal processes expressed in an object-relationship data model. They also claimed that describing geometric transformations of an independent spatial entity implies changes on four orthogonal attributes: shape (form of its boundary), size (area of its interior), orientation (compass direction of its major and minor axes), and location (position of its gravity center measured with geographic or Euclidean coordinates). In order to integrate phenomena that change over space and time in real-world phenomenon, a better understanding of the underlying components of change and how people reason about change are needed. In paper [12], the authors proposed a qualitative representation of changes. It offers a classification of changes based on object identity and the set of operations that either preserve or change identity. These operations can be applied to single or composite objects and combined to express the semantics of sequences of change. The authors also developed a visual language to represent the various types of change, and provided examples to illustrate the application of this language. Later, the authors used a temporal zooming approach to detect and navigate the spatio-temporal changes [11, 13]. These approaches do not deal specifically with trajectory data.

Trajectory data analysis can be divided into two basic categories: single trajectory and multiple trajectory analysis. Since single trajectory data normally depicts one specific moving entity, single trajectory data analysis mainly focuses on looking for individual spatial patterns, and creating predictive models for the moving entity [2, 20]. The predication models are useful for applications such as providing real time traffic information if the next trip stops can be predicted. In some applications, object movements obey periodic patterns or follow similar routes over regular time intervals. Effective data mining algorithms have been proposed to find spatio-temporal periodic patterns [5, 17].

In recent years, there has been increased interest in analyzing spatial-temporal patterns and moving paths of wild animals [1] using multiple-trajectory data analysis. Geographic data mining approaches have been proposed to detect generic spatial-temporal patterns such as flock, leadership, convergence, and encounter in geospatial data [15, 3, 16]. A method to calculate the longest duration of flocks and meetings in spatial-temporal data has been proposed [10]. Definition of moving clusters and efficient algorithms to identify them have been proposed [14]. A recent work tries to find a set of individual trajectories that share the property of visiting the same sequence of places with similar travel times [9], which is related but not the focus of this paper.

Our work is related to and goes beyond finding moving clusters or identifying flocks. We provide a clustering-based definition of *herds* and further model *quantitative* changes and *qualitative* changes of the herds and their interactions.

Our work is also related to spatial clustering. For static datasets, clustering analysis can be either used as a stand-alone tool to get insight into the distribution of a dataset, to identify areas for further analysis, or as a preprocessing

step for other algorithms operating on detected clusters. Clustering algorithms can be classified into partition based, hierarchical clustering, density based, and model based methods. A density based clustering algorithm, e.g. DBSCAN [7], is attractive when one needs to identify arbitrary shaped clusters. In some applications, the location and content of spatial clusters may change over time, which requires a formal definition for moving clusters and algorithms for automatic discoveries [14].

3 Herd Evolutions

In this section, we introduce the ideas of herd and herd evolutions through visual illustrations and formal definitions. Figure 1(a) illustrates six trajectories (O_1, O_2, \dots, O_6) over two time snapshots. From Figure 1(a), suppose we have A, B clusters in snapshot t_i and they will evolve into A, B, C clusters in snapshot t_{i+1} , where C is a new cluster formed in t_{i+1} . Please note that between the two time snapshots, the details of the trajectories are not perceivable due to the sampling granularity. Also, a cluster should include more entities than those illustrated in the figure.

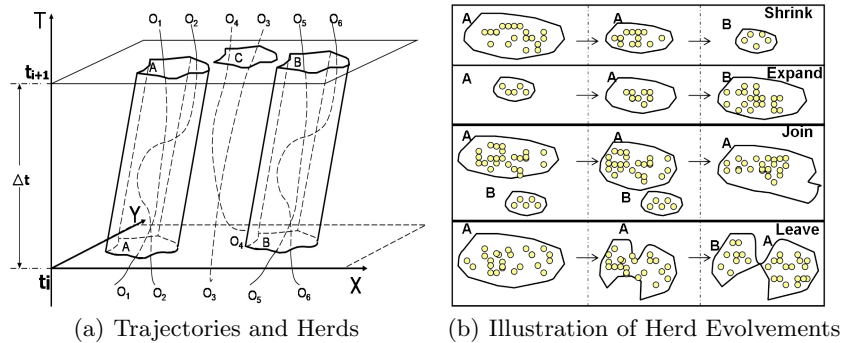


Fig. 1. The Spatial Evolutions

Thus, if we take a snapshot of the locations of n entities at time t , we can spatially identify proximate groups of entities. We call each group consisting of a set of entities at time t a *herd snapshot*. Several herd snapshots over time may be related to each other and represent the *movements* and *evolutions* of herds over time. The mobility of entities causes *herd movement* whereas the membership change of entities causes *herd evolution*.

The groups in each snapshot can be identified through various ways and we propose to use clustering algorithms for this purpose. A spatial cluster is a group of entities gathered in spatial proximity. Various clustering algorithms can be used including partitioning based, hierarchical, model based, and density based.

We argue that a density-based clustering is more suitable in this application due to its following properties: (1) the ability to construct non-spherical clusters of arbitrary shapes; (2) the robustness with respect to noise in the data; (3) the ability to discover an arbitrary number of clusters without the need of specifying the number of clusters to find.

Once the herd snapshot concept is established, we investigate how the membership changes of herd snapshots can result in the dynamics of formation and deformation of new herds. For example, when a herd snapshot $H(t)$ of 100 entities at time t is joined by a few other entities, e.g. 5 entities, in the next snapshot, do we still have the same herd H ? What about 2,000 entities joined $H(t)$? Do we still have the same herd H ? To summarize, the research issue here is how to characterize the herd evolutions using membership changes.

Let $H(t)$ represent a herd snapshot at time t and the herd H itself was formed at sometime t_0 before t . H 's changes in the subsequent time snapshots after t_0 can be classified into: *quantitative* and *qualitative* changes. In a *quantitative change*, members leave H and new members may join H in small quantities. However, the herd can still be “reasonably well” represented by the initial members of H when H was formed, i.e. $H(t_0)$. In a *qualitative change*, members of H change so much that $H(t_0)$ is not a “good” representative of the herd anymore. Of course, the question is how to precisely define “reasonably well” and “good”. We present the intuitive meaning of the four categories of qualitative changes here and propose the formal definitions in section 4:

- Expand: A herd H formed at time $t - i$ *expands* into a new herd H' at time t if $H'(t)$ contains “many” of the members of $H(t - i)$ but also contains “substantial” new members.
- Join: A herd H_1 formed at time snapshot $t - i$ *joins* a herd H_2 formed at time snapshot $t - j$ to form a new herd H at time t , if $H(t)$ is “similar” to the herd H_2 and contains majorities of the members from herd H_1 .
- Shrink: A herd H formed at time $t - i$ *shrinks* into a new herd H' at time t if $H'(t)$ contains “many” of the members of $H(t - i)$ but also contains “substantial” other members not in $H'(t)$.
- Leave: A herd H_1 *leaves* a herd H formed at time snapshot $t - i$ at snapshot t if the majorities of members of $H_1(t)$ are also in $H(t - i)$. Further more, the herd formed by the remaining members, denoted by $H_2(t)$, is “similar” to the herd H .

Figure 1(b) illustrates the four kinds of evolutions in three snapshots. In the *shrink* case, herd A went through a *quantitative change* first and then a *qualitative change*, to shrink into a new herd B ; in the *expand* case, herd A went through a *quantitative change* first and then a *qualitative change*, to expand into a new herd B ; in the *join* case, herd B joined herd A to form a new herd with name A due to the dominance of A in the new herd (reasons for using label A for the new herd will be discussed in section 4); in the *leave* case, herd A went through a *quantitative change* first and then herd B leaves herd A and the remaining herd is still labelled A (reasons for labeling one of the new herds as A will be cleared after section 4).

4 Measurements of Herd Evolutions

The *Precision*(P), *Recall*(R) and *F-Score*(F) measurements have been traditionally used for evaluating the performance of information retrieval systems. For a query Q and the collection of documents retrieved by Q , the measures assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to Q . Intuitively, *recall* is the fraction of the documents that are relevant to the query and are successfully retrieved; and *precision* is the fraction of the documents retrieved that are relevant to what the user is querying for. The formal definitions are presented as follows:

$$Recall = \frac{|relevant\ documents| \cap |retrieved\ documents|}{|retrieved\ documents|}$$

$$Precision = \frac{|relevant\ documents| \cap |retrieved\ documents|}{|relevant\ documents|}$$

With the definitions of precision and recall, we can expound the meaning of F measure: the weighted harmonic mean of precision and recall. The formal definition of the F-score or balanced F-score is:

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

If we treat moving entities in trajectory data as documents in retrieval system, we can adopt the above measurements to the areas of spatio-temporal analysis. Thus, let $H(t)$ be a herd snapshot at time t , i.e. the members of herd H at time t , and let $H'(t+i)$ be another herd snapshot at time $t+i$, we adopt the *precision* (P), *recall* (R), and *F-score* (F) measurements to model the relationship between $H(t)$ and $H'(t+i)$ as follows:

$$R(H(t), H'(t+i)) = \frac{|H(t) \cap H'(t+i)|}{|H(t)|}$$

Intuitively, $R(H(t), H'(t+i))$ measures the percentage of $H(t)$ that continue to exist in $H'(t+i)$. Thus, the more the entities left between t and $t+i$ from $H(t)$, the lower the value of $R(H(t), H'(t+i))$ is.

$$P(H(t), H'(t+i)) = \frac{|H(t) \cap H'(t+i)|}{|H'(t+i)|}$$

Intuitively, $P(H(t), H'(t+i))$ measures the percentage of $H'(t+i)$ that come from $H(t)$. Thus, the more the new entities joined in $H(t)$ between t and $t+i$, the lower the value of $P(H(t), H'(t+i))$ is.

$$F(H(t), H'(t+i)) = \frac{2 \times P(H(t), H'(t+i)) \times R(H(t), H'(t+i))}{P(H(t), H'(t+i)) + R(H(t), H'(t+i))}$$

$F(H(t), H'(t+i))$ represents the combined results of members left and new members joined and it ranges from 0 to 1, where 0 indicates $H(t)$ is completely

different from $H'(t+i)$ and 1 indicates $H(t)$ and $H'(t+i)$ consist of exactly the same members.

4.1 Measuring Generic Herd Evolvments Using R, P, F

In this section, we describe the scenario and criteria to precisely define evolvments of herds.

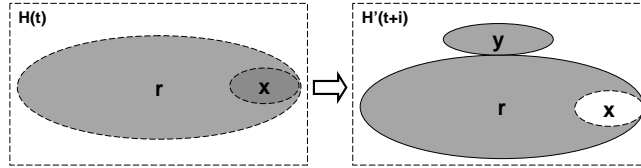


Fig. 2. Generic Herd's Evolvments from t to $t+i$ (all symbols, e.g. r , denote the areas delineated by their closest boundaries)

In Figure 2, let the $|r \cup x|$ represent the members of a herd snapshot $H(t)$ at time t . At time $t+i$, suppose we detect another herd snapshot $H'(t+i)$ consisting of $r \cup y$, where x is the set of entities that escaped from $H(t)$ between t and $t+i$ time snapshots, and y is the set of entities that joined in $H(t)$ between t and $t+i$ time snapshots, then the P, R , and F can be formulated using r, x and y . The *recall* is:

$$R(H(t), H'(t+i)) = \frac{|r|}{|r| + |x|}$$

and the *precision* is:

$$P(H(t), H'(t+i)) = \frac{|r|}{|r| + |y|}$$

And *F-score* will be (please note that r, x, y are disjoint):

$$F(H(t), H'(t+i)) = \frac{2 \times P \times R}{P + R} = \frac{2 \times \frac{|r|}{|r \cup y|} \frac{|r|}{|r \cup x|}}{\frac{|r|}{|r \cup y|} + \frac{|r|}{|r \cup x|}} = \frac{2 \times |r|}{2 \times |r| + |x| + |y|}$$

Thus, we propose the following criteria to define the *quantitative* and *qualitative* evolvments of herd:

1. When the sum of the number of the escaped members (i.e. x) and the number of the newly joined members (i.e. y) is less than the number of the remaining members of $H(t)$ in $H'(t+i)$ (i.e. r), we think that conceptually the herd can still be reasonably represented by the original members of H and claim

VIII

that $H(t)$ underwent a *quantitative change* to $H'(t+i)$ and $H'(t+i)$ is the same herd as $H(t)$. That is, given $|x| + |y| < |r|$ or equivalently when:

$$F(H(t), H'(t+i)) > \frac{2}{3}$$

the change is *quantitative*;

2. When the sum of the number of the escaped members (i.e. x) and the number of the newly joined members (i.e. y) is no less than the number of the remaining members of $H(t)$ in $H'(t+i)$ (i.e. r), conceptually, we think the herd is NO longer the same herd and has undergone a *qualitative change*. We say that $H(t)$ underwent a *qualitative change* to $H'(t+i)$ and $H'(t+i)$ is NOT the same herd as $H(t)$.

That is, given $|x| + |y| \geq |r|$, or equivalently when:

$$F(H(t), H'(t+i)) \leq \frac{2}{3}$$

the change is *qualitative*. This case can be further divided into the following scenarios:

- (a) When $|x| \geq |r|$ and $|y| < |r|$, or equivalently,

$$R(H(t), H'(t+i)) = \frac{|r|}{|r \cup x|} \leq \frac{1}{2}, P(H(t), H'(t+i)) = \frac{|r|}{|r \cup y|} > \frac{1}{2}$$

we say that $H(t)$ shrank or left by others into $H'(t+i)$. In this case, the escaped members (i.e. x) outweigh the remaining members of $H(t)$ in $H'(t+i)$ (i.e. r) and the remaining members of $H(t)$ in $H'(t+i)$ (i.e. r) outweigh the newly joined members (i.e. y).

- (b) When $|y| \geq |r|$ and $|x| < |r|$, or equivalently,

$$R(H(t), H'(t+i)) = \frac{|r|}{|r \cup x|} > \frac{1}{2}, P(H(t), H'(t+i)) = \frac{|r|}{|r \cup y|} \leq \frac{1}{2}$$

we say that $H(t)$ expanded or be joined into $H'(t+i)$. In this case, the newly joined members (i.e. y) outweigh the remaining members of $H(t)$ in $H'(t+i)$ (i.e. r) and the remaining members of $H(t)$ in $H'(t+i)$ (i.e. r) outweigh the escaped members (i.e. x).

- (c) Otherwise, $H(t)$ and $H'(t+i)$ do not have a relationship. If $H(t)$ does not find any other herd snapshot at $t+i$ related to it, H simply disappears at time $t+i$.

Please note that the threshold of $\frac{2}{3}$ for the *F-score* and the threshold of $\frac{1}{2}$ for the precision and recall are decided by the ‘‘majority rules’’ and may be modified based on biological rules to fit specific application domains.

We summarize the conditions of various *qualitative changes* in Table 1.

Table 1. Qualitative Changes Based on P, R When $F \leq \frac{2}{3}$

$F < \frac{2}{3}$	$P > \frac{1}{2}$	$P \leq \frac{1}{2}$
$R \leq \frac{1}{2}$	Shrink or Split	No Relationship
$R > \frac{1}{2}$	No Relationship	Expand or Merge

4.2 An Additional Concern in Labeling New Herds

If two or more herd snapshots are merged into a larger one at time $t + i$, will more than one of the merging herds claim that the newly formed herd is the same herd as themselves (both changes are not *qualitative*)? If this happens, we have an identity problem where we do not know how to call the newly formed herd.

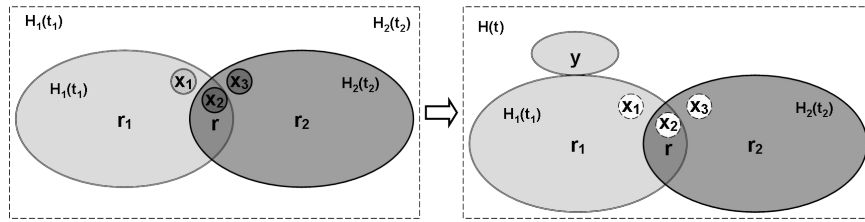


Fig. 3. Herds Formed at t_1 and t_2 Merge at t (all symbols, e.g. r , denote the areas delineated by their closest boundaries)

In Figure 3, assume we have two herds H_1 and H_2 formed at t_1 and t_2 respectively, where $H_1(t_1)$ contains $r_1 \cup r \cup x_1 \cup x_2$ and $H_2(t_2)$ contains $r_2 \cup r \cup x_2 \cup x_3$. The two herds merged in the snapshot t into $H(t)$. At the same time $x_1 \cup x_2 \cup x_3$ escaped and did not participate in $H(t)$. But y joined at the same time. So at the end, the member set of $H(t)$ consists of $r_1 \cup r_2 \cup r \cup y$. To capture the inheritances among herds, we should name *Herd* $H(t)$ either H_1 or H_2 if one of F -scores is larger than $\frac{2}{3}$. If both herds $H_1(t_1)$ and $H_2(t_2)$ went through *quantitative changes* and became $H(t)$, i.e. $H(t)$ is both $H_1(t_1)$ and $H_2(t_2)$, we have problems in labeling the descendant herds (e.g. $H(t)$ in this case). Here, based on conceptual definition of *quantitative change*, if both herds went through *quantitative changes*, we have:

$$\begin{aligned}
 |x_1 \cup x_2| + |r_2 \cup y| &< |r_1 \cup r| \\
 |x_2 \cup x_3| + |r_1 \cup y| &< |r_2 \cup r|
 \end{aligned}$$

In fact, it is possible for the two equations above to hold. For example, when x_1, x_2, x_3, y are empty, $r_1 = r_2$, and r is not empty, it is obvious both equations hold.

In order to preserve herds evolvments, we propose a ranking strategy to establish the inheritance relationship among *Herds* at different snapshots. Specifically, if a herd at the current time snapshot has *F-scores* greater than $\frac{2}{3}$ with several herds formed previously, we rank the herds according to the *F-scores* and chose the one with the highest *F-score* as the label for the current herd.

5 Herd Interaction Graph

Now we are ready to introduce the *Herd Interaction Graph* (or *Herding*) using various herd evolvments defined in the previous sections. For a given starting time t_{start} that we start to observe the herds, we find the *core member set* CM of a herd H . The core member set CM defines and represents the herd H until the actual member set of H deviates qualitatively from CM . Then H disappears and a new herd may emerge with a new CM' . Formally, we define the concept of a *core member set* as follows:

Definition 1 (Core Member Set) Given the trajectories of a set of n entities of length τ and a clustering algorithm to cluster the locations of the entities at each time snapshot: (1) When $t = t_{start}$, a cluster CM defines and initiates a herd H and is called the core member set of the herd H and can be denoted by $H(t)$. H will continue to exist and be represented by $H(t)$ until H 's actual member set $H(t+i)$ at time $t+i$ is *qualitatively* different from $H(t)$ (where *F-score* $\leq \frac{2}{3}$); (2) When $t \neq t_{start}$, a cluster CM defines and initiates a herd H and is called the core member set of the herd H and can be denoted by $H(t)$ if and only if: CM is *qualitatively* different from any existing *core member set* formed in the previous time snapshots (where *F-score* $\leq \frac{2}{3}$); (3) A herd H together with its core member set CM disappears at t when there is no clusters found at t is a quantitative change of CM .

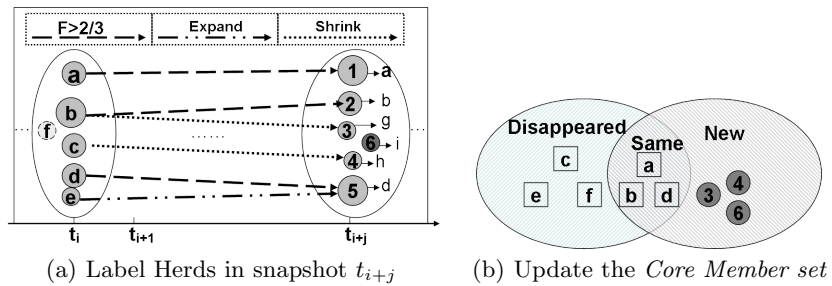


Fig. 4. Core Member Set in Herds

In Figure 4(a), at the beginning time snapshot t_i , we assume there are 6 herds each represented by its core member set. They are labeled as a, b, c, d, e, f .

After j time snapshots at snapshot t_{i+j} , we found 6 clusters labeled as 1 to 6 respectively. We need to see if these clusters are just some quantitative change of the earlier herds represented by their core member set or newly formed herds using R , P and F -score. We calculate their relationships with the core member set of a, b, c, d, e, f using R , P , and F -score. It turned out that 1, 2, 5 are quantitative changes of a, b, d respectively, thus the herds a, b, d continue to exist and are still represented by their core member set at t_i . Furthermore, because cluster 5 is an expansion of cluster e (cluster 5 is also herd d), cluster e joined cluster 5. In a similar vein, cluster 3 (reabeled as g) is a shrinking of cluster b (cluster 2 is also herd b), so herd 3 left herd b . In addition, herd f disappeared, herd 4 (reabeled as h) is a shrinking of herd c , a new herd 6 (reabeled as i) appears. Figure 4(b) shows that in two snapshots, some herds disappear, some are new and the core member set of the common herds, e.g. b, d , will not change (although there are lost members and newly joined members).

5.1 Basic Algorithm to Generate Herd Interaction Graph

We are now ready to introduce our algorithm to generate the *Herd Interaction Graph* by determining the *quantitative changes* and *qualitative changes* in *Herds' Evolvments*.

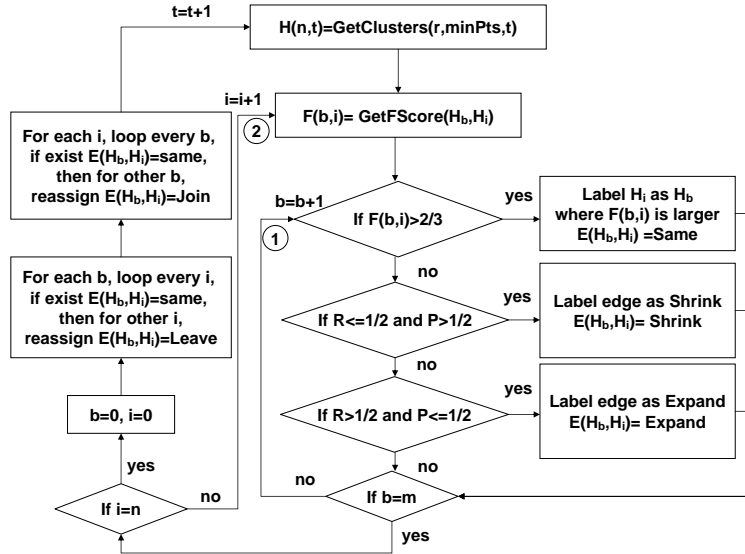


Fig. 5. Flow Chart of the Algorithm to Generate Herd Interaction Graph

In Figure 5, we first use procedure $GetClusters(r, minPts, t)$ to find clusters from the trajectory data at snapshot t via a clustering algorithm, e.g. DBSCAN,

with two parameters: radius (r) and minimum points (minPts), and return the results in the array $H(n, t)$ where n is the number of clusters found. Here we assume we have m herds B formed previously represented by their *Core Member Sets*. To decide if the change from a cluster H_i and a core member set H_b is *quantitative* or *qualitative*, procedure $\text{GetFScore}(H_b, H_i)$ calculates the F -score between them. We use $E(H_b, H_i)$ to represent the relationship, e.g. *expand*, between herd H_b and a cluster H_i .

In case of a *quantitative change* ($F \geq \frac{2}{3}$), we rank the F -scores, label H_i as H_b , and assign $E(H_b, H_i) = \text{same}$, where $F(b, i)$ is the largest for all $b \in B$ if multi-ancestors are found. In case of a *qualitative change*, based upon the criteria we have discussed, we can determine that 1) herd H_b shrinks into herd H_i formed in this snapshot if $P > 1/2$ and $R \leq 1/2$; we then tentatively label the edge between H_b and H_i as *shrink* ($E(H_b, H_i) = \text{shrink}$); 2) herd H_b expands into herd H_i formed in this snapshot if $P \leq 1/2$ and $R > 1/2$; We tentatively label the edge between H_b and H_i as *expand* ($E(H_b, H_i) = \text{expand}$). After both loop 1 and loop 2 are over, we begin starting loop for b, i again and check if we need to reassign edges with *join* or *leave* by using the numbers of incoming and outgoing edges of herds.

Furthermore, for each b , we loop every i , if there is no H_i as the quantitative change of H_b , i.e. $E(H_b, H_i) = \text{same}$, we can determine that $H(b)$ disappeared. Similarly, for each i , we loop every b , if there is no H_b as the quantitative change of H_i , i.e. $E(H_b, H_i) = \text{same}$, we can also determine that $H(i)$ is newly formed.

5.2 Intuitive Visualization of the Herd Interaction Graph

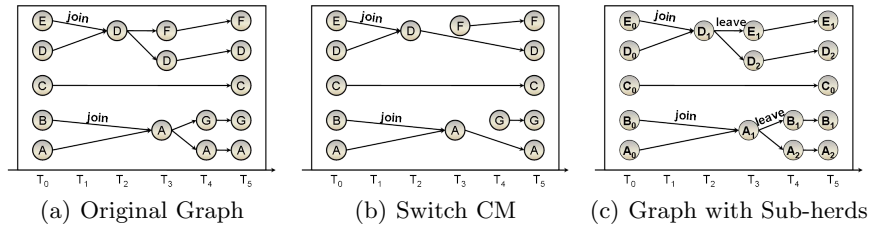


Fig. 6. Herd Interaction Graph

We are now able to produce the interaction graph (see Figure 6(a)) based upon the scenarios and algorithm we have proposed so far. We pay close attention to the intuitive meaning of the herd interaction graph to a user.

In Figure 6(a), for example, with herd E joining D , D will have to appear at snapshot T_2 although it does not have a qualitative change. Because the core member set of a herd does NOT change unless the herd discontinues to exist, new herds, e.g. F , will continue to compare with $D(T_0)$ which is the core member

set of D rather than with $D(T_2)$. However, when F either left or shrank from $D(T_0)$, we need to draw a line from D at T_2 to F at T_3 as shown in Figure 6(b) which can be easily interpret as F left (or shrank) from $D(T_2)$ (D at T_2) instead of $D(T_0)$ (D at time T_0). As a result, there is a discrepancy between the actual meaning of a node in the graph and the intuitive meaning a user may observe.

To represent their interactions, we have to change the mechanism of producing interaction graphs by switching the core member set of a herd to *status-based core member set* of a herd whenever a herd label appears more than one time (see Figure 6(c)). Consequently, a herd will be labeled using the same herd name but with subscriptions. For example, in Figure 6(c), D appears in multiple snapshots, namely T_0, T_2, T_3 and T_5 , and are labeled as D_0, D_1 , and D_2 where the D_2 at T_5 is added to signify the ending of D .

Thus, in order to effectively symbolize and represent the evolvments and interaction of herds through the herd interaction graph, we introduce the concept of *sub-herd* to represent the core member set updates for more intuitive visualization.

Definition 2 (Sub-herd(H)) Given a herd H formed at time t_0 whose core member is represented by $H(t_0)$, the sequential appearance of H on the herd interaction graph even without *qualitative change* are called *sub-herds* of H . Each sub-herd is represented by their members at the time when they appear on the graph (called *status-based core member set*) and is labeled as H_j where j is an increasing integer.

One more point that needs to be emphasized here is that, in Figure 6(b), a new herd G forms in snapshot T_4 calculated by using F, P, R , i.e. $F(G(T_4), A(T_0)) \leq \frac{2}{3}$. However, when we look backward to the herds formed in historical snapshots and determine which herd is the most similar to this new *herd* (the greater the F value is, the more similar the two *herds* are), we can detect that B in snapshot T_0 is actually not very different from G , and G should be labeled as B_1 . In addition, using the core member set of A at T_3 , we conclude that B_1 is a shrink of A_1 . Combined with A_2 , we decide B_1 left A_1 as shown in Figure 6(c). The situation for F in Figure 6(b) is similar. In addition, if *herd* $A(t)$ disappears in next snapshot $t+1$, this herd is forced to appear on graph to highlight its completeness. For example E_1 disappears at time T_5 in Figure 6(c).

Overall in Figure 6(c), there are five herds A, B, C, D and E at the first snapshot. Then herds B joined B at snapshot T_3 and then split into two herds similar to the original herds A and B at T_4 ; herd C stays the same throughout the entire observation; The behaviors of D and E are similar to those of A and B .

6 Algorithm Validations

We describe a herd simulator to facilitate the validation of the proposed concepts and algorithms for detecting herd and their interactions in this section. We also discuss the algebraic cost of the algorithms. Both the simulator and herd graph generation algorithms are implemented as an ArcGIS 9.x extension and source codes and details are available at <http://groucho.csci.unt.edu/herding/>.

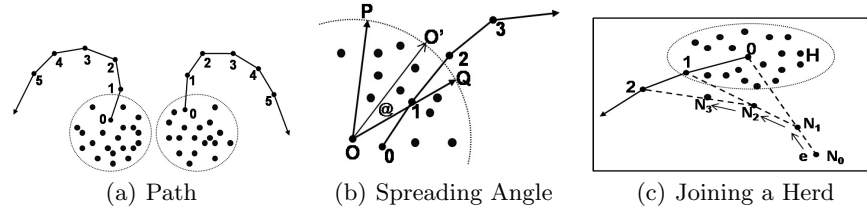


Fig. 7. Simulating Herds

6.1 Simulating Herds and Their Interactions

We have developed a herd simulator using VB.net to allow users to pre-define the moving pattern for each herd through a mouse. A set of clusters will be initially generated at time t_{start} . Then, as shown in Figure 7(a), a user can specify the path for each cluster (herd) where each point represents a snapshot (from 0 to 5 in the graph). Furthermore, a user can specify a spreading angle of a herd so that the core member of a herd will move randomly along the specified path and within the spreading angle. For example, in Figure 7(b), the moving entity O can randomly move to the territory within the area of PQO in the next snapshot. Generally, the greater the spreading angle, the more likelihood for the herd to split and disappear.

By specifying the paths of herds to intersect, merge, and disappear as well as specifying the spreading speed of the herds, users can simulate the interaction of herds, and then apply our proposed algorithm to identify the herd interaction graph to verify their intended herd interactions.

To simulate the common effect of individuals joining an existing herd gradually over time, our simulator allows each entity not yet in a herd to randomly choose a herd and try to catch up with that herd over time by following the path of that herd using maximal speed. In Figure 7(c), the entity e chooses to follow herd H and always heads to the path of H over time, i.e. N_0 to N_1 to N_2 to N_3 .

6.2 Generating the Herd Interaction Graph

The herd interaction graph representing the evolvments and movements of herds is then produced. Figure 6(c) is a screenshot generated by our tool, which is implemented as an ArcGIS 9.x extension. We used the Microsoft Visual Studio 2005 as our project IDE (see Figure 8(a)) and stored the trajectories into a geodatabase (See Figure 8(b)).

Figure 8(c) shows a few more screen shots of our extension to ArcGIS 9.x. In this run, we took a sample dataset with 500 moving entities stored as a geodatabase in ArcGIS and imported into ArcGIS desktop for further data manipulation.

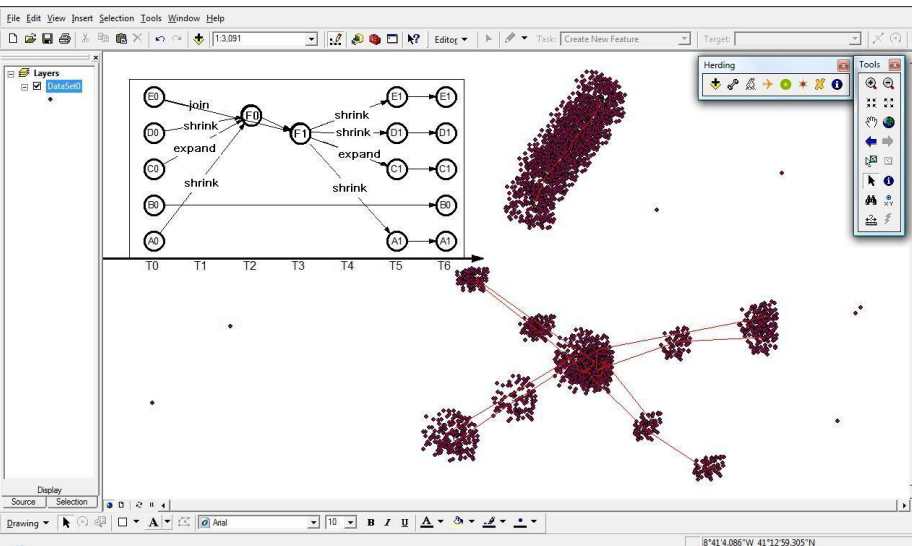
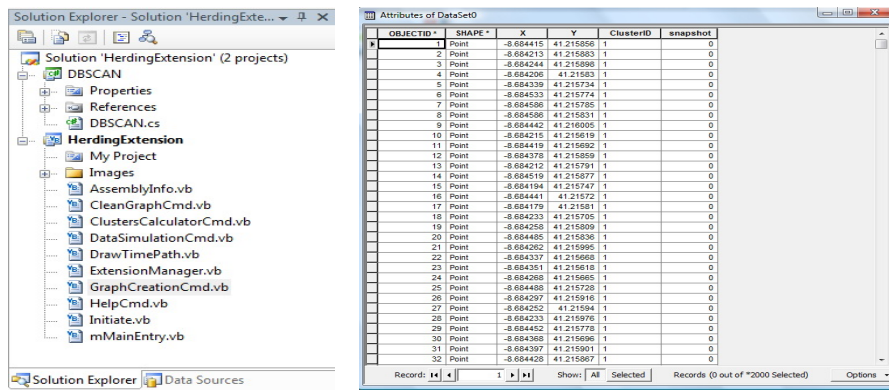


Fig. 8. Screenshots

6.3 Algebraic Cost Model of the Algorithm

We present the algebraic cost model of the proposed algorithm in this section and delay the optimization and performance evaluation to future work due to the space constraint. Suppose we have N entities in our trajectory data, the running time of the clustering algorithm, i.e. *DBSCAN*, will be $O(N \log N)$. If we have T snapshots overall, the cost will be $O(T \times N \log N)$.

Let C represent the average number of clusters and k be the average number of core members for each cluster in each snapshot, then the cost of processing one snapshot will be $O(k^2 \times C^2)$ in order to determine the membership relationships of each pair of clusters using *Precision*, *Recall* and *F-score* measurements. Thus, the overall cost is: $O(T \times N \log N \times k^2 \times C^2)$.

7 Conclusion and Future Work

In this paper, 1) we introduced herds and their *quantitative* and *qualitative* changes during evolvments; 2) we defined *quantitative* and *qualitative* changes by leveraging measurements used in information retrieval, namely *Recall*, *Precision* and *F-score*; 3) we defined four types of *qualitative* evolvments: *expand*, *join*, *shrink* and *leave*; 4) we introduced an effective method to identify these four types of spatial evolvments; 5) we developed a herd simulator to produce trajectories data in a Geographic Information System (GIS) environment; 6) we presented a graph-based representation - *Herd Interaction Graph* to represent herd interactions. The results suggest that herds and their interactions can be effectively modeled through the proposed measurements and the herd interaction graph from trajectory data. We also released the source code of the relevant implementations for public use. We plan to look into further optimization heuristics to improve the performance of the algorithm. We also plan to apply our algorithms to human mobility data to detect social events and their patterns.

References

1. M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, 1999.
2. Ashbrook and D. Starner. Learning significant locations and predicting user movement with GPS. In *Sixth International Symposium on Wearable Computers*, pages 101–108, 2002.
3. M. Benkert, J. Gudmundsson, F. Hübner, and T. Wollé. Reporting flock patterns. In *ESA'06: Proceedings of the 14th conference on Annual European Symposium*, pages 660–671, 2006.
4. D. R. Brillinger, H. K. Preisler, A. A. Ager, and J. G. Kie. An exploratory data analysis (eda) of the paths of moving animals. In *Journal of Statistical Planning and Inference* 122, pages 43–63, 2004.
5. H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 82–89, 2005.

6. C. Claramunt, C. Parent, and M. Theriault. Design Patterns for Spatio-temporal Processes. In *IFIP 2.6 Working Conference on Database Semantics, DS7*, 1997.
7. M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.
8. A. Frank, J. Raper, and J.-P. Cheylan. Life and motion of spatial socio-economic units. In *Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 2001.
9. F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 330–339. ACM, 2007.
10. J. Gudmundsson and M. van Kreveld. Computing longest duration flocks in trajectory data. In *GIS '06: Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 35–42. ACM, 2006.
11. K. Hornsby. Temporal zooming. In *Transactions in GIS*, pages 255–272. Blackwell Publishing, 2001.
12. K. Hornsby and M. J. Egenhofer. Qualitative representation of change. In *COSIT '97: Proceedings of the International Conference on Spatial Information Theory*, pages 15–33, 1997.
13. K. Hornsby and M. J. Egenhofer. Shifts in detail through temporal zooming. In *DEXA Workshop*, pages 487–491, 1999.
14. P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *In International Symposium on Advances in Spatial and Temporal Databases (SSTD)*, pages 364–381, 2005.
15. P. Laube and S. Imfeld. Analyzing relative motion within groups of trackable moving point objects. In *2nd International Conference on Geographic Information Science*, pages 132–144, 2002.
16. P. Laube, M. van Kreveld, and S. Imfeld. Finding remote detecting relative motion patterns in geospatial lifelines. In *Developments in Spatial Data Handling*, pages 201–215. Springer Berlin Heidelberg, 2005.
17. N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W. Cheung. Mining, indexing, and querying historical spatiotemporal data. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 236–245. ACM, 2004.
18. J. Miller and J. Han. Detecting outliers from large datasets. In *Geographic data mining and knowledge discovery*, pages 218–235. Taylor and Francis, 2001.
19. Y. Qu, C. Wang, and X. S. Wang. Supporting fast search in time series for movement patterns in multiple scales. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 251–258, 1998.
20. N. Sumpter and A. J. Bulpitt. Learning spatio-temporal patterns for predicting object behaviour. *Image Vision Comput.*, 18(9):697–704, 2000.