

## A Refined Definition for Groups of Moving Entities and Its Computation\*

Marc van Kreveld<sup>†</sup>, Maarten Löffler<sup>‡</sup> and Frank Staals<sup>§</sup>

*Department of Information and Computing Sciences  
Utrecht University, Utrecht, The Netherlands*

<sup>†</sup>*m.j.vankreveld@uu.nl*

<sup>‡</sup>*m.loffler@uu.nl*

<sup>§</sup>*f.staals@uu.nl*

Lionov Wiratma

*Department of Information and Computing Sciences  
Utrecht University, Utrecht, The Netherlands*

*Department of Informatics, Parahyangan Catholic University  
Bandung, Indonesia*

*l.wiratma@uu.nl*

*lionov@unpar.ac.id*

Received 1 February 2017

Revised 20 December 2017

Published 13 July 2018

Communicated by Seok-Hee Hong, Guest Editor

One of the important tasks in the analysis of spatio-temporal data collected from moving entities is to find a *group*: a set of entities that travel together for a sufficiently long period of time. Buchin *et al.*<sup>2</sup> introduce a formal definition of groups, analyze its mathematical structure, and present efficient algorithms for computing all maximal groups in a given set of trajectories. In this paper, we refine their definition and argue that our proposed definition corresponds better to human intuition in certain cases, particularly in dense environments.

We present algorithms to compute all maximal groups from a set of moving entities according to the new definition. For a set of  $n$  moving entities in  $\mathbb{R}^1$ , specified by linear interpolation in a sequence of  $\tau$  time stamps, we show that all maximal groups can be computed in  $O(\tau^2 n^4)$  time. A similar approach applies if the time stamps of entities are not the same, at the cost of a small extra factor of  $\alpha(n)$  in the running time, where  $\alpha$  denotes the inverse Ackermann function. In higher dimensions, we can compute all maximal groups in  $O(\tau^2 n^5 \log n)$  time (for any constant number of dimensions), regardless of whether the time stamps of entities are the same or not.

We also show that one  $\tau$  factor can be traded for a much higher dependence on  $n$  by giving a  $O(\tau n^4 2^n)$  algorithm for the same problem. Consequently, we give a linear-time algorithm when the number of entities is constant and the input size relates to the

\*A preliminary version of this paper appeared in Proc. 27th International Symposium on Algorithm and Computation (ISAAC 2016).<sup>1</sup>

number of time stamps of each entity. Finally, we provide a construction to show that it might be difficult to develop an algorithm with polynomial dependence on  $n$  and linear dependence on  $\tau$ .

*Keywords:* Moving entities; trajectories; grouping; computational geometry.

## 1. Introduction

Nowadays, inexpensive modern devices with advanced tracking technologies make it easy to track movements of an entity. This has led to the availability of movement data for various types of moving entities (human, animals, vehicles, etc.). Since a tracking device typically returns a single location at each time stamp, each moving entity will be represented by a moving point. Data may consist of just one trajectory tracked over a period of time, or a whole collection of trajectories that are all tracked over the same time period. Note that for the latter case, the locations of each trajectory are not necessarily collected at the same time stamps. It is common to denote the number of trajectories (or moving entities) by  $n$  and the number of time stamps used for each trajectory by  $\tau$ . Hence, the input size is  $\Theta(\tau n)$ . Depending on the application, one of  $n$  or  $\tau$  can be much larger than the other.

To analyze moving object data, a number of methods have been developed in recent times. These methods perform similarity analysis or compute a clustering, outliers, a segmentation, or various patterns that may emerge from the movement of the entities (for surveys see Refs. 3 and 4). These methods are often based on geometric algorithms, because the data is essentially spatial.

One particular type of pattern that has been well-studied is flocking.<sup>5–9</sup> Intuitively, a flock is a subset of the entities moving together (or simply being together) over a period of time. Other names for this and closely related concepts with slightly different definitions are herds,<sup>10</sup> convoys,<sup>11–13</sup> moving clusters,<sup>14,15</sup> swarms,<sup>16,17</sup> mobile groups,<sup>18</sup> traveling companions,<sup>19</sup> and groups.<sup>2,20</sup> Buchin *et al.*<sup>2</sup> introduce a model called the *trajectory grouping structure* which not only defines groups, but also the splitting of a group into subgroups and its opposite, merging. The algorithmic problem of reporting all maximal groups that occur in the trajectories is solved in  $O(\tau n^3 + N)$  time, where  $N \in O(\tau n^4)$  is the output size (the summed size of all groups reported). The algorithm also considers times in between the  $\tau$  time stamps where the locations are recorded as relevant. In between these time stamps, locations are inferred by linear interpolation over time.

In this paper we continue the study of such groups, but we propose a refined definition to the one by Buchin *et al.*<sup>2</sup> We motivate why it captures our intuition better and present algorithms to compute all maximal groups.

### *Previous definition of a group*

The definition of a group by Buchin *et al.*<sup>2</sup> relies on three parameters: one for the distance between entities, one for the duration of a group, and one for the size of a group. We review their definitions next.

For a set of moving entities  $\mathcal{X}$ , two entities  $x$  and  $y$  are *directly  $\varepsilon$ -connected* at time  $t$  if the Euclidean distance between  $x$  and  $y$  is at most  $\varepsilon$  at time  $t$ , for some given  $\varepsilon \geq 0$ . Two entities  $x$  and  $y$  are  *$\varepsilon$ -connected in  $\mathcal{X}$*  at time  $t$  if there is a sequence  $x = x_0, \dots, x_k = y$ , with  $\{x_0, \dots, x_k\} \subseteq \mathcal{X}$  and for all  $i$ ,  $x_i$  and  $x_{i+1}$  are directly  $\varepsilon$ -connected at time  $t$ .

In Ref. 2, a *group* for an entity inter-distance  $\varepsilon$ , a minimum required duration  $\delta$ , and a minimum required size  $m$ , is defined as a subset  $G \subseteq \mathcal{X}$  and corresponding time interval  $I$  for which three conditions hold:

- (i)  $G$  contains at least  $m$  entities.
- (ii)  $I$  has a duration at least  $\delta$ .
- (iii) Every two entities  $x, y \in G$  are  $\varepsilon$ -connected in  $\mathcal{X}$  at all times in  $I$ .

Furthermore, a group  $G$  with time interval  $I$  is *maximal* if there is no time interval  $I' \supset I$  for which  $G$  is also a group, and there is no proper superset  $G' \supset G$  that is also a group during  $I$ .<sup>2</sup>

### Refined definition of a group

One issue with the previous definition is that it does not correspond fully to our intuition. Two entities  $x$  and  $y$  may form a (rather small) maximal group in an interval  $I$  even if they are always far apart, as long as there are always entities of  $\mathcal{X}$  in between them to make  $x$  and  $y$   $\varepsilon$ -connected in  $\mathcal{X}$ . These entities in between are not part of the maximal group, but they do cause  $x$  and  $y$  to be  $\varepsilon$ -connected by the previous definition. This can have counter-intuitive effects especially in dense crowds. To avoid such issues, we refine the definition of a group. In particular, we replace condition (iii) above by:

- (iii') Every two entities  $x, y \in G$  are  $\varepsilon$ -connected in  $G$  during  $I$ .

Note that with this refined definition, a group is like a convoy,<sup>11</sup> with one main difference: a group is defined for both the continuous and the discrete model of a trajectory, whereas a convoy is defined only for the discrete model. We define *maximal* groups analogous to the previous definition by Buchin *et al.*,<sup>2</sup> but now based on the refined definition of groups.

We give two examples that show the difference in these definitions.

First, consider a number of stationary entities  $S$  and two entities  $x$  and  $y$ , see Fig. 1. Entity  $x$  starts (at time  $t_0$ ) to the North of  $S$  and moves around its perimeter to the East. Entity  $y$  starts (at  $t_0$ ) to the South and also moves around the perimeter to the East. After encountering (at  $t_1$ ) each other at the East side, both continue together eastward, away from the stationary entities in  $S$  (ending at  $t_2$ ). By the definition in Ref. 2,  $x$  and  $y$  form a maximal group in the interval  $[t_0, t_2]$ . By our refined definition, they form a maximal group during  $[t_1, t_2]$ , starting when  $x$  and  $y$  are at distance  $\varepsilon$  and actually encounter each other.

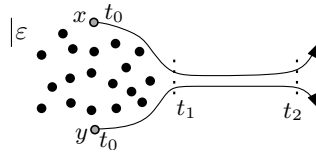


Fig. 1. In the previous definition (see Ref. 2),  $x$  and  $y$  are  $\varepsilon$ -connected during  $[t_0, t_2]$ .

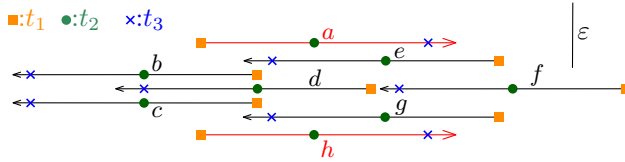


Fig. 2. Entities in  $G = \{a, h\}$  are  $\varepsilon$ -connected using entities not in  $G$ .

Second, the previous definition can even see groups of entities that were never close, see Fig. 2. Here,  $\{a, h\}$  is a maximal group in the interval  $I = [t_1, t_3]$  using the definition in Ref. 2. At each time,  $a$  and  $h$  are  $\varepsilon$ -connected, but through different subsets of entities. By choosing the coordinates carefully, we can ensure that no supergroup of  $\{a, h\}$  is also a group in the same time interval, and hence  $\{a, h\}$  will be maximal. Although  $a$  and  $h$  move in the same direction with the same speed, intuitively they do not form a group because they are too far apart and separated by other entities that move in the opposite direction. With our refined definition, we do not consider  $\{a, h\}$  a group in the interval  $I$ , and hence also not a maximal group.

### Results and organization

We have refined the previous definition for a group of moving entities by Buchin *et al.*<sup>2</sup> and gave two examples and argue why our refined definition can give an intuitively plausible group. From now on, we will use the term “group” to denote a group of entities that comply with our refined definition.

In the following section, we show that for a set  $\mathcal{X}$  of  $n$  moving entities in  $\mathbb{R}^1$  with  $\tau$  time stamps each, the number of maximal groups by the refined definition is  $O(\tau n^3)$ , which is tight in the worst case.

In Sec. 3, we present algorithms to compute all maximal groups in  $\mathbb{R}^1$ . First we consider the case where all trajectories have their vertices at the same time and begin with a basic algorithm for that runs in  $O(\tau^3 n^6)$  time. Subsequent improvements lead to a running time of  $O(\tau^2 n^4)$ . When the time stamps of different trajectories are not the same, our algorithm runs in  $O(\tau^2 n^4 \alpha(n))$  time.

Next, for moving entities in  $\mathbb{R}^d$  ( $d > 1$ ), we model entities and their inter-distance into graphs and show that all maximal groups can be computed in  $O(\tau^2 n^5 \log n)$  time, regardless the uniformity of the time stamps in the trajectories. We show how to achieve this bound in Sec. 4.

In Sec. 5, we consider situations where the value of  $n$  is significantly smaller than  $\tau$ , which is typical in real-life moving entity datasets. We give an  $O(\tau 2^n n^4)$  time algorithm for entities that move in any constant dimension.

Finally, we show an exponential bound on the number of maximal groups that can contain any given time  $t$  in the last section.

## 2. Preliminaries

Let  $\mathcal{X}$  be a set of  $n$  entities moving in  $\mathbb{R}^1$ , given by locations at  $\tau$  time stamps. A trajectory of an entity in  $\mathcal{X}$  can be expressed by a piecewise-linear function which maps time to a point in  $\mathbb{R}^1$ . If  $\mathbb{R}^1$  is associated with the vertical axis and time with the horizontal axis of a 2-dimensional plane, the trajectories of entities in  $\mathcal{X}$  are polylines with  $\tau$  vertices each. We will use the same notation to denote an entity and its trajectory.

Let  $d_{ij}(t)$  be the Euclidean distance between  $i \in \mathcal{X}$  and  $j \in \mathcal{X}$  at time  $t$ . When  $d_{ij}(t) = \varepsilon$ , we say that an  $\varepsilon$ -event occurs. For any  $\varepsilon$ -event  $v$ , we denote by  $t_v$  the time when  $v$  occurs and  $\omega(v)$  the function that returns the two entities that create  $v$ . We assume that no two or more  $\varepsilon$ -events occur at the same time.

Consider an  $\varepsilon$ -event  $v$ ; let  $\omega(v) = \{i, j\}$ . If the distance between  $i$  and  $j$  is more than  $\varepsilon$  immediately before  $t_v$ , then  $v$  is a *start  $\varepsilon$ -event*; if this happens after  $t_v$ , then  $v$  is an *end  $\varepsilon$ -event*. If there is no entity  $k \in \mathcal{X}$  located strictly in between  $i$  and  $j$  at  $t_v$  (so  $d_{ik}(t_v) + d_{jk}(t_v) = \varepsilon$ ), then we say that  $v$  is a *free  $\varepsilon$ -event*.

For clarity, we assume that there are no two parallel edges of trajectories with distance  $\varepsilon$ . This assumption can be lifted without complications; we just need to handle such situations in the appropriate manner.

**Observation 1.** *The number of  $\varepsilon$ -events is  $O(\tau n^2)$ .*

Let  $G$  be a group of entities in time interval  $I$  that is maximal in size. All entities in  $G$  are pairwise  $\varepsilon$ -connected in the interval  $I$ , and hence, there are no free  $\varepsilon$ -events in  $G$  during  $I$ . In the arrangement of trajectories from  $G$ , we define the height of a face as the length of the longest vertical line segment inside the face. Thus, no face has height greater than  $\varepsilon$ .

It is also clear that  $G$  can begin only at a start  $\varepsilon$ -event and end only at an end  $\varepsilon$ -event. Furthermore, we observe that if a start  $\varepsilon$ -event (or end  $\varepsilon$ -event) of  $G$  is not a free  $\varepsilon$ -event with respect to the entities in  $G$ , then before (or after) the interval  $I$ , entities in  $G$  are still pairwise  $\varepsilon$ -connected and we can extend the interval of  $G$ . Therefore,  $G$  can be a maximal group only if both the start  $\varepsilon$ -event and end  $\varepsilon$ -event are free  $\varepsilon$ -events (but this is not a sufficient condition).

**Observation 2.** *There can be at most one maximal group that starts and ends at a particular pair of start  $\varepsilon$ -event and end  $\varepsilon$ -event.*

**Theorem 3.** *For a set  $\mathcal{X}$  of  $n$  entities, each entity moving along a piecewise-linear trajectory of  $\tau$  edges, the maximum number of maximal groups is  $\Theta(\tau n^3)$ .*

**Proof.** Any group  $G$  that starts at a start  $\varepsilon$ -event contains at most  $n$  entities. When a free end  $\varepsilon$ -event involving  $G$  occurs, only group  $G$  ends but a subgroup of  $G$  with fewer entities may continue longer. This can happen at most  $n - 1$  times. Therefore, the maximum number of maximal groups is  $O(\tau n^3)$ .

For the lower bound on the number of maximal groups, we use the same construction given by van Goethem *et al.*<sup>21</sup> Although they use the previous definition by Buchin *et al.*<sup>2</sup> to define a group, the construction itself also works for our refined definition of a group. They start by building a construction in which all  $n$  entities move along lines and show that there can be  $\Omega(n^3)$  maximal groups. Then, repeating the construction  $\Omega(\tau)$  times gives  $\Omega(\tau n^3)$  as a lower bound on the number of maximal groups.  $\square$

The approach to compute all maximal groups is to work on the arrangement  $\mathcal{A}$  of line segments that are the trajectories. For a subset  $G \subseteq \mathcal{X}$  and interval  $I$ , we can remove entities from  $G$  that are separated at a face with height larger than  $\varepsilon$  in  $I$  (corresponding to a free  $\varepsilon$ -event). Only if there are no such faces, the remaining entities in  $G$  can be a group. Note that removing entities in  $G$  involves removing the corresponding trajectories from the arrangement  $\mathcal{A}$ , which can cause new faces that are free  $\varepsilon$ -events.

### 3. Algorithms for Entities in $\mathbb{R}^1$

In this section, first we consider the case where the trajectories have the same time stamps. We present a basic algorithm that computes all maximal groups in  $O(\tau^3 n^6)$  time for entities moving in  $\mathbb{R}^1$ . Then we present a more efficient algorithm that runs in  $O(\tau^2 n^4)$  time. Furthermore, we present an  $O(\tau^2 n^4 \alpha(n))$  time algorithm if the vertices of the trajectories have different time stamps.

#### 3.1. Basic algorithm

We describe a simple algorithm to compute all maximal groups. Let  $\mathcal{V}_s$  and  $\mathcal{V}_e$  be the sets of all start  $\varepsilon$ -events and all end  $\varepsilon$ -events, respectively. Fix one event of each type:  $\beta \in \mathcal{V}_s$  and  $\gamma \in \mathcal{V}_e$ . By Observation 2, there is only one maximal group  $G$  that starts at  $\beta$  and ends at  $\gamma$ . Furthermore, observe that  $G$  necessarily contains the entities  $\omega(\beta) = \{a, b\}$  and  $\omega(\gamma) = \{c, d\}$ , and that if  $G$  is a maximal group on  $I = [t_\beta, t_\gamma]$ , then all entities in  $G$  are on the same side at time  $t_\lambda \in (t_\beta, t_\gamma)$  when a free  $\varepsilon$ -event  $\lambda$  occurs. We then use the following approach to find  $G$  (if it exists):

- (1) Initialize a set  $G$  containing all entities in  $\mathcal{X}$ .
- (2) Build an arrangement  $\mathcal{A}$  induced by the trajectories of the entities in  $G$  on  $I$ .
- (3) A face  $f$  in  $\mathcal{A}$  contains a free  $\varepsilon$ -event  $\lambda$  if (and only if) the height of  $f$  is more than  $\varepsilon$ . If  $f$  has height larger than  $\varepsilon$ , test if (the trajectories of)  $a, b, c$ , and  $d$ , all lie on the same side of  $f$ . If not, there is no maximal group  $G$  that starts

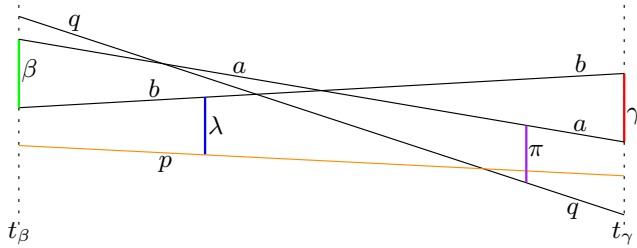


Fig. 3. Removing trajectory  $p$  (due to the free  $\varepsilon$ -event  $\lambda$ ) causes the  $\varepsilon$ -event  $\pi$  to become a free  $\varepsilon$ -event.

at  $\beta$  and ends at  $\gamma$ . If they do pass on the same side, let  $S$  denote the set of entities whose trajectories lie on the other side of  $f$ . Remove these entities of  $S$  from  $G$ , and remove their trajectories from  $\mathcal{A}$ . Observe that new free  $\varepsilon$ -events may appear because removal of a trajectory from  $\mathcal{A}$  merges two faces of  $\mathcal{A}$  into a larger one. See Fig. 3. Repeat this step until there is no more free  $\varepsilon$ -event  $\lambda$  with  $t_\lambda \in (t_\beta, t_\gamma)$ .

- (4) Check that  $\beta$  and  $\gamma$  are now free. If so,  $G$  is a maximal group on  $I$ , and hence we can report it. If not,  $G$  is actually a group during a time interval  $I' \supset I$ . Hence,  $G$  may be maximal in size, but not in duration. We do not report  $G$  in this case.

**Theorem 4.** *Given a set  $\mathcal{X}$  of  $n$  entities in which each entity moves in  $\mathbb{R}^1$  along a trajectory of  $\tau$  edges, all maximal groups can be computed in  $O(\tau^3 n^6)$  time using the Basic Algorithm.*

**Proof.** The number of combination of a pair of start and end  $\varepsilon$ -events is  $O(\tau^2 n^4)$ . Building an arrangement from trajectories of entities takes  $O(\tau n^2)$  time. Removing a trajectory  $e$  and checking new faces in  $\mathcal{A}$  takes time proportional to the zone complexity of  $e$ :  $O(\tau n)$ . Since there are at most  $n$  trajectories to be removed, the whole process to remove entities for each interval  $I$  takes  $O(\tau n^2)$  time. Therefore, the running time of the algorithm is  $O(\tau^3 n^6)$  time.  $\square$

### 3.2. Improved algorithm

The previous algorithm checks every pair of possible start and end  $\varepsilon$ -events  $\beta$  and  $\gamma$  to potentially find one maximal group. To improve the running time, we fix a start  $\varepsilon$ -event  $\beta$  and consider the  $O(\tau n^2)$  end  $\varepsilon$ -events  $\gamma$  in increasing order. We show that we can check for a maximal group on  $[t_\beta, t_\gamma]$  in amortized  $O(1)$  time.

We build the arrangement  $\mathcal{A}$  for all trajectories, starting from time  $t_\beta$ , and sort the end  $\varepsilon$ -events  $\gamma$ , with  $t_\gamma > t_\beta$  on increasing time. We then consider the end  $\varepsilon$ -events  $\gamma$  in this order, while maintaining a maximal set  $G$  that is  $\varepsilon$ -connected in  $G$  throughout the time interval  $[t_\beta, t_\gamma]$ .

Let  $\omega(\beta) = \{a, b\}$  be the entities defining the start  $\varepsilon$ -event  $\beta$ , and let  $G \supseteq \{a, b\}$  be the largest  $\varepsilon$ -connected set on  $[t_\beta, t_\gamma]$ . We compute the largest  $\varepsilon$ -connected set on  $[t_\beta, t_{\gamma'}]$  for the next ending event  $\gamma'$  as follows. Note that this set will be a subset of  $G$ .

Let  $S$  be the set of entities that separate from  $a$  and  $b$  at  $\gamma$ . We remove all trajectories from the entities in  $S$  from  $\mathcal{A}$ . As before, this may introduce faces of height larger than  $\varepsilon$ . For every such face  $f$ , we check if  $a$  and  $b$  still pass  $f$  on the same side. If not, there can be no maximal groups that contain  $a$  and  $b$ , start at  $t_\beta$ , and end after  $t_\gamma$ . If  $a$  and  $b$  lie on the same side of  $f$ , we add all entities that lie on the other side of  $f$  to  $S$  and remove their trajectories from  $\mathcal{A}$ . We repeat this until all faces in  $\mathcal{A}$  that have non-empty intersection with the vertical strip defined by  $[t_\beta, t_{\gamma'}]$  have height at most  $\varepsilon$  (or until we have found a face that splits  $a$  and  $b$ ). It follows that the set  $G' = G \setminus S$  is the largest set containing  $a$  and  $b$  that is  $\varepsilon$ -connected throughout  $[t_\beta, t_{\gamma'}]$ . If  $\beta$  and  $\gamma'$  are free with respect to  $G'$  then we report  $G'$  as a maximal group.

Building the arrangement  $\mathcal{A}$  takes  $O(\tau n^2)$  time, and sorting the ending-events takes  $O(\tau n^2 \log(\tau n))$  time. By the Zone Theorem, we can remove each trajectory in  $O(\tau n)$  time. Checking the height of the new faces can be done in the same time bound. It follows that the total running time is  $O(\tau n^2(\tau n^2 + \tau n^2 \log(\tau n) + R))$  where  $R$  is the total time for removing trajectories from the arrangement. Clearly,  $R$  is bounded by the complexity of the arrangement:  $O(\tau n^2)$ . So, the total running time is  $O(\tau^2 n^4 \log(\tau n))$ .

### **Further improvement**

We can avoid repeated sorting of end  $\varepsilon$ -events by pre-sorting them in a list, and for each start  $\varepsilon$ -event, use this list. The list will contain events that do not concern the entities involved in the start  $\varepsilon$ -event, but this can be tested easily in constant time. Thus, we conclude:

**Theorem 5.** *Given a set  $\mathcal{X}$  of  $n$  entities in which each entity moves in  $\mathbb{R}^1$  along a trajectory of  $\tau$  edges, all maximal groups can be computed in  $O(\tau^2 n^4)$  time.*

Next, we consider finding all maximal groups when the vertices of different trajectories do not have the same time stamps. We use the same idea as in the above algorithm: take one start  $\varepsilon$ -event  $\beta$  at a time and remove trajectories to find all maximal groups containing  $\omega(\beta)$ .

We use a similar strategy to split trajectories vertically into  $\tau$  cells as in Ref. 22, where each cell now contains  $O(n)$  segments of trajectories. It follows that the complexity of each cell is bounded by the number of possible intersections between segments:  $O(n^2)$ . Thus, building the arrangement  $A$  still takes  $O(\tau n^2)$  time. However, by the Zone Theorem for an arrangement of line segments, removing a trajectory in each cell now takes  $O(n\alpha(n))$  time,<sup>23</sup> where  $\alpha(n)$  is the inverse Ackermann Function. Therefore, the total time to remove trajectories in  $A$  is  $O(\tau n^2 \alpha(n))$  time and we obtain:



**Theorem 6.** *Given a set  $\mathcal{X}$  of  $n$  entities in which each entity moves in  $\mathbb{R}^1$  along a trajectory of  $\tau$  edges under the condition that their vertices have different time stamps, all maximal groups can be computed in  $O(\tau^2 n^4 \alpha(n))$  time.*

#### 4. Algorithms for Entities in $\mathbb{R}^d$

In  $\mathbb{R}^d$  ( $d > 1$ ), it is harder to test whether an  $\varepsilon$ -event really connects or disconnects because the two entities may be  $\varepsilon$ -connected through other entities in the group, see Fig. 4. This observation immediately gives the condition for an  $\varepsilon$ -event to be *free*.

We model our moving entities as a graph where vertices represent entities and an edge exists if two entities are directly  $\varepsilon$ -connected. As shown by Parsa,<sup>24</sup> we can maintain the graph under edge updates, while allowing same component queries, in  $O(\log n)$  time per operation.

To compute maximal groups, we start at a start  $\varepsilon$ -event  $\beta$ , where  $\omega(\beta) = \{a, b\}$ , and maintain the connected component  $\mathcal{C}$  throughout the sequence of sorted  $\varepsilon$ -events. At each  $\varepsilon$ -event  $\gamma$ , we remove any vertices that are disconnected from  $\mathcal{C}$  and start again from  $\beta$  in case we remove anything. We stop if  $a$  and  $b$  are disconnected. If  $\beta$  is a free  $\varepsilon$ -event when we reach  $\gamma$  again, we report  $\mathcal{C}$  as a maximal group and continue. See Fig. 5 for an example of how the algorithm works.

We start at  $O(\tau n^2)$   $\varepsilon$ -events and for each, we process  $O(\tau n^2)$   $\varepsilon$ -events. We may need to restart this process up to  $n - 1$  times. In  $\mathbb{R}^d$ , our approach only examine

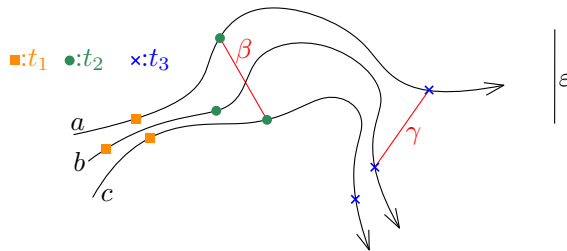


Fig. 4. At  $t_2$ ,  $\beta$  is not a free  $\varepsilon$ -event because  $a$  and  $c$  are still  $\varepsilon$ -connected through  $b$ . At  $t_3$ ,  $a$  and  $b$  create the free  $\varepsilon$ -event  $\gamma$ .

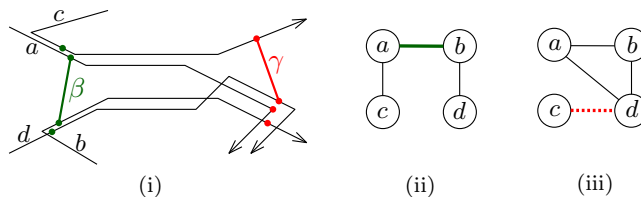


Fig. 5. Two  $\varepsilon$ -events  $\beta$  and  $\gamma$ . (i) The maximal group  $\{a, b, c, d\}$  has a time interval  $[t_\beta, t_\gamma]$ . (ii) Connectivity of the graph after  $t_\beta$ . A new edge appeared connecting  $a$  and  $b$ . (iii) Connectivity of the graph after  $t_\gamma$ . The edge between  $c$  and  $d$  disappeared.

the  $\varepsilon$ -events of entities and does not affected by whether the vertices of trajectories have the same time or not, therefore we obtain the same result for both cases:

**Theorem 7.** *Given a set  $\mathcal{X}$  of  $n$  entities moving in  $\mathbb{R}^d$  along a trajectory of  $\tau$  edges, all maximal groups can be computed in  $O(\tau^2 n^5 \log n)$  time.*

## 5. Algorithms with Linear Dependence on $\tau$

In many real-life situations, GPS is used for collecting trajectory data from human, vehicle or animal movement. Since it may record the position of an entity frequently and over a period of time, the number of vertices in each trajectory is often much larger than the number of moving entities.<sup>25</sup> Therefore, the dependence of the algorithm on  $\tau$  is more important than the dependence on  $n$ . Next, we present a simple algorithm that is linear in  $\tau$ , at the cost of an exponential dependence on  $n$ . In particular, our algorithm will compute all maximal groups in  $O(\tau n^4 2^n)$  time.

We consider all  $2^n$  subsets of  $\mathcal{X}$  in order of decreasing size, while maintaining the set of maximal groups found so far (ordered by increasing starting time). For each subset  $G$  we determine the maximal time intervals during which  $G$  is  $\varepsilon$ -connected, and for each such interval  $I$  we check if  $G$  is dominated by a maximal group  $H \supset G$  on  $I$ . If such a set does not exist, then  $G$  is a maximal group on  $I$ . Notice that we only need to know when the start  $\varepsilon$ -event and end  $\varepsilon$ -event of a particular group occurs. Therefore, this algorithm can be applied for both cases where the vertices of trajectories have the same or different time stamps.

For each subset  $G$ , we consider the  $\varepsilon$ -events generated by the entities in  $G$ . We can compute all these  $O(\tau n^2)$   $\varepsilon$ -events in  $O(\tau n^2 \log n)$  time, by sorting the batches of  $O(n^2)$   $\varepsilon$ -events between two consecutive time stamps separately, and concatenating the resulting  $\tau$  lists. We then go through the  $\varepsilon$ -events in order, and check if  $G$  is  $\varepsilon$ -connected at every  $\varepsilon$ -event. We can easily handle every event in  $O(n^2)$  time, by naively checking if the entities in  $G$  are  $\varepsilon$  connected (we can easily improve on this, but the total running time will be dominated by the number of sets anyway). It follows that we can compute the sequence  $\mathcal{S}_G$  of maximal time intervals on which  $G$  is  $\varepsilon$ -connected in  $O(\tau n^4)$  time. Note that  $\mathcal{S}_G$  contains at most  $O(\tau)$  such time intervals.

For each interval  $I$  in  $\mathcal{S}_G$  we now have to check if  $G$  is a maximal group during  $I$ . The set  $G$  is a maximal group on  $I$  if and only if there is no maximal group  $H \supset G$  on a time interval that contains  $I$ . Since we maintain the maximal groups larger than  $G$  (and the time interval on which they are a maximal group), ordered by increasing starting time, we can iterate through them once, and extract the maximal groups that are a superset of  $G$ . Figure 6 shows the sequence of maximal time intervals found by the algorithm when processing the entity set  $G = \{a, b\}$  and its supersets.

Since by Theorem 3 there are at most  $O(\tau n^3)$  maximal groups, this takes at most  $O(\tau n^4)$  time. Let  $\mathcal{I}$  denote the set of time-intervals corresponding to those groups, ordered by increasing starting time. We now simply scan through  $\mathcal{S}_G$  and  $\mathcal{I}$

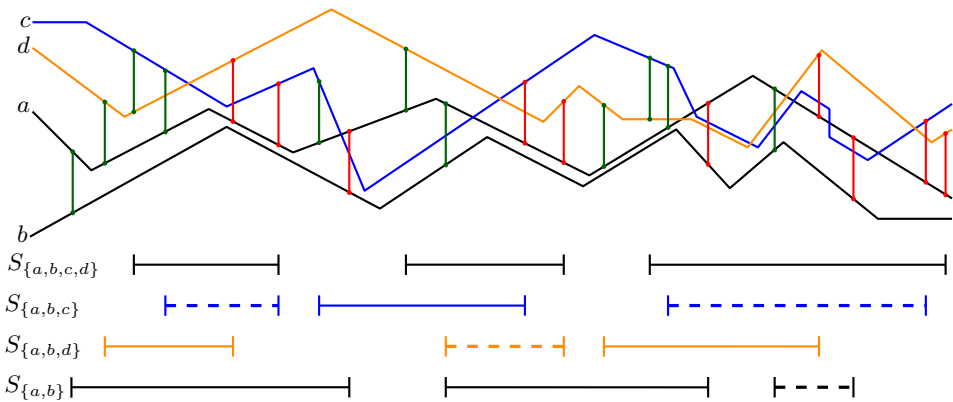


Fig. 6. The sequences of maximal time intervals for  $G = \{a, b\}$  and all of its supersets. The three time intervals at the bottom are found when processing  $G$ ; the other ones were found earlier. The dashed time intervals do not give rise to a maximal group.

simultaneously, while maintaining the time interval in  $\mathcal{I}$  that started earliest and has not ended yet. For every interval  $I$  in  $\mathcal{S}_G$  we can then check if  $G$  is a maximal group on  $I$  in constant time. In total this takes  $O(\tau n^3)$  time. Using a similar simultaneous scan we can add the intervals on which  $G$  is maximal to our set of maximal groups found so far.

It follows that we can compute all time intervals on which  $G$  is maximal in  $O(\tau n^4)$  time. Since we do this for all subsets  $G \subseteq \mathcal{X}$  we obtain the following result.

**Theorem 8.** *Given a set  $\mathcal{X}$  of  $n$  entities in which each entity moves in  $\mathbb{R}^d$  along a trajectory of  $\tau$  edges, we can compute all maximal groups in  $O(\tau n^4 2^n)$  time, using  $O(\tau n^3)$  space.*

## 6. A Lower Bound on the Maximum Number of Maximal Groups at Some Time $t$

The result in the previous section shows that, when  $\tau$  is large but  $n$  is small, we can improve the dependence on  $\tau$  from quadratic to linear. However, we pay for this by having an exponential dependence on  $n$ . This naturally raises the question whether an algorithm with linear dependence on  $\tau$ , but polynomial dependence on  $n$ , is possible. While we do not know the answer to this question, we present a construction which may indicate that such a result is hard to obtain, if possible at all.

We show that the number of maximal groups that contain a given time  $t$  can be exponential in  $n$ , provided that  $\tau$  is sufficiently large. Without the requirement that the maximal groups must span a single moment in time, it is easy to make a construction of trajectories that has a number of maximal groups that is linear in  $\tau$ , even with just two entities, so it is unbounded in  $n$ . Similarly, we can easily construct trajectories that give rise to  $2^n - n - 1$  maximal groups (with a group size of at least 2) that are different in composition using roughly  $2^n$  time stamps by making

these groups consecutive. The construction that we present, where many different maximal groups occur simultaneously, is more involved, and shows that there may be  $\Omega(\sqrt{2}^n)$  maximal groups simultaneously when there are  $\Omega(\sqrt{2}^n)$  time stamps. While the result does not imply any lower bound for the problem of computing all maximal groups, it suggests that it may be difficult to obtain an algorithm that is linear in  $\tau$  and polynomial in  $n$ . Several natural approaches to the problem (based on, for instance, divide-and-conquer) appear not to work due to this construction and the result on simultaneous maximal groups.

**Theorem 9.** *There exists a set  $\mathcal{X}$  of  $n$  entities in  $\mathbb{R}^1$  whose trajectories are defined by  $\Theta(\sqrt{2}^n)$  time stamps, for which the number of maximal groups at some time  $t$  is  $\Omega(\sqrt{2}^n)$ .*

**Proof.** We use a set of  $n = 2k + 2$  entities, denoted  $p_1, \dots, p_k, p'_1, \dots, p'_k$ , and  $q$  and  $q'$ . We are interested in counting the groups that contain  $q, q'$ , and for each  $i$ , exactly one of  $p_i$  and  $p'_i$ . We call any such group  $k$ -max and will show that they are all maximal. A  $k$ -max group  $G$  is encoded by a length- $k$  bitstring where the  $i$ -th bit is 1 if  $p_i \in G$  and it is 0 if  $p'_i \in G$ .

We make a construction with the following properties; the half after  $t_{mid}$  is illustrated for  $k = 3$  in Fig. 7:

- (1) The trajectory of  $p_i$  is the reverse of  $p'_i$ , with respect to  $t_{mid}$  (that is, mirrored in  $t_{mid}$ ), and vice versa.
- (2) A  $k$ -max group starts and ends at free  $\varepsilon_q$ -events of  $q$  and  $q'$ .
- (3) A  $k$ -max group encoded by bitstring  $B$  starts a fraction after time  $1 + B$  and ends a fraction before time  $t_{mid} + 1 + B$ , where  $B$  is interpreted as a binary number.
- (4) There are only  $O(1)$  trajectory vertices of each trajectory within one time unit.
- (5) Each  $k$ -max group is maximal.

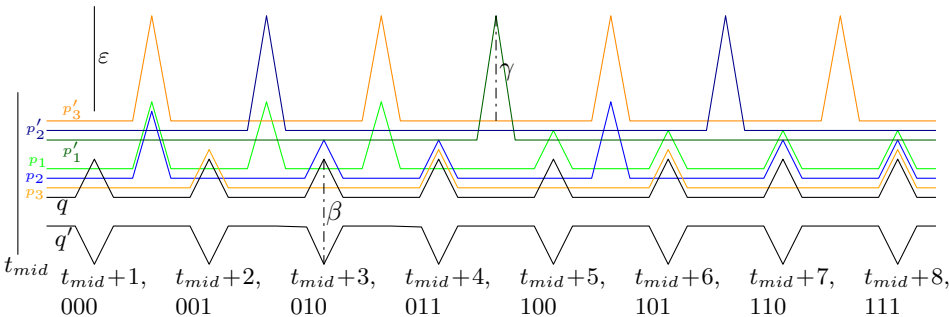


Fig. 7. Right half of the lower bound construction for  $k = 3$ . The times very near the start and end  $\varepsilon$ -events of  $q$  and  $q'$  are shown together with the bitstring of the  $k$ -max group that ends there. At  $t_{mid}$ , all trajectories are in a single point (the trajectories are not shown near  $t_{mid}$ ).

At  $t_{mid}$ , all trajectories pass through a single point to ensure they are continuous when mirroring, and they are pairwise directly  $\varepsilon$ -connected. It is the moment in time for which  $\Omega(\sqrt{2}^n)$  maximal groups exist, as we will show. After  $t_{mid}$ , the entities  $q$  and  $q'$  will have  $2^k$  pairs of  $\varepsilon$ -events: an end  $\varepsilon$ -event directly followed by a start  $\varepsilon$ -event. We call these events  $\varepsilon_q$ -events. Whether these  $\varepsilon_q$ -events are free for a  $k$ -max group  $G$  depends on the time and the bitstring, or equivalently, which entities from  $p_1, \dots, p_k$  are in  $G$ .

The  $\varepsilon_q$ -event at a time  $t$  is free for a  $k$ -max group  $G$  if and only if the bitstring corresponding to time  $t$  is the same as the bitstring of  $G$ . Hence, (assuming that no earlier  $\varepsilon$ -event ends  $G$ )  $G$  will end at the time of its bitstring, so a fraction before  $t_{mid} + 1 + B$ . By symmetry of  $p_i$  and  $p'_i$ ,  $G$  will start a fraction after time  $1 + B$ . In Fig. 7, for example, at time  $t_{mid} + 4$ , the  $k$ -max group  $\{p'_1, p_2, p_3, q, q'\}$  ends.

The other  $\varepsilon$ -events of the trajectories are between two consecutive  $\varepsilon_q$ -events. These  $\varepsilon$ -events involve the entities of  $p'_1, \dots, p'_k$  and the trajectories need three vertices between  $\varepsilon_q$ -events. Their presence ensures that only one of  $p_i$  or  $p'_i$  is in a particular  $k$ -max group. Notice that these  $\varepsilon$ -events will also be the start or end  $\varepsilon$ -events of maximal groups that are supersets of a  $k$ -max group.

Suppose  $p'_i$  is an entity that creates a free  $\varepsilon$ -event  $\beta$  just before a  $k$ -max group  $G$  containing  $p_i$  ends at  $t_{mid} + B$ . Obviously,  $p'_i$  only needs to create such a free  $\varepsilon$ -event once and it follows this is only necessary if the previous  $k$ -max group  $G'$  that ends at  $t_{mid} + B - 1$  is not containing  $p_i$ . However, other  $k$ -max groups that will end after  $G$  might contain  $p'_i$ . Therefore, to prevent this  $\varepsilon$ -event becomes free in the duration of those  $k$ -max groups, we make entities of  $p_1, \dots, p_k$  that are not in  $G$  to keep  $p'_i$   $\varepsilon$ -connected to all other entities. Still, not all of them are needed to prevent  $\beta$  to become free, but only for each entity  $p'_h$  where  $h < i$ , because by the ordering of the bitstrings,  $k$ -max groups contain  $p'_i$  and those entities might end after  $\beta$ . See Fig. 7: before  $\varepsilon$ -event  $\beta$ , only  $p_1$  prevents  $p'_2$  from creating a free  $\varepsilon$ -event (but not  $p_3$ ). Two  $k$ -max groups contain  $p_1, p'_2$  and one of  $p_3$  or  $p'_3$  end after  $\beta$  while  $k$ -max group of  $\{p'_1, p'_2, p_3\}$  ends before  $\beta$ .

**Claim 1.** *If a maximal group containing time  $t_{mid}$  contains at least  $p_i$  or  $p'_i$  for all indices  $i$ , and both  $p_i$  and  $p'_i$  for at least one index  $i$ , then its time interval cannot contain both time  $h$  and  $t_{mid} + h$  for any integer  $h$ .*

**Proof.** Suppose for contradiction  $G$  is a maximal group which contains both  $p_i$  and  $p'_i$ , and its time interval fully contains an interval  $[h, t_{mid} + h]$  for some integer  $h$ ; suppose further that  $i$  is the smallest index for which this is the case. Let  $B$  be the bitstring that encodes the entities with indices  $1 \dots i - 1$ ; let  $B^- = B0111 \dots 1$  be obtained from  $B$  by appending a single 0 and  $k - i$  1s and let  $B^+ = B1000 \dots 0$  be obtained from  $B$  by appending a single 1 and  $k - i$  0s. Then  $G$  starts not earlier than some time between  $B^-$  and  $B^+$ , and ends not later than some time between  $t_{mid} + B^-$  and  $t_{mid} + B^+$ . Refer to Fig. 7. Hence, there is no integer  $h$  such that both  $h$  and  $t_{mid} + h$  are contained in the time interval of  $G$ .  $\square$

The claim directly implies that all  $k$ -max groups are maximal, because by Property 3 they start and end at some time  $h$  and  $t_{mid} + h$ , but adding any other trajectories will cause both  $p_i$  and  $p'_i$  to be in the group for some  $i$ .

Moreover, the  $\varepsilon$ -events created by entities of  $p'_1, \dots, p'_k$  are also the end  $\varepsilon$ -events of the  $2^k - 1$  groups that have more entities than a  $k$ -max group. Let the maximal group contains all entities end at free  $\varepsilon$ -event  $\gamma$  at time  $t_\gamma = t_{mid} + 2^{k-1} + T$  ( $0 < T < 1$ ) created by  $p'_i$ . By the symmetry of the construction and the ordering of the bitstrings, two groups of  $n - 1$  entities not containing either  $p'_i$  or  $p_i$  will end at time  $t_\gamma - 2^{k-2}$  and  $t_\gamma + 2^{k-2}$ , respectively. Then, continuing the same process with the two groups recursively will results on other maximal groups with different entities. Since the start and end  $\varepsilon$ -events of these groups are always start later or end earlier than  $k$ -max groups, then these groups are maximal because their interval will not contain interval of other maximal groups. Clearly, the number of these maximal groups is fewer than  $k$ -max groups because their  $\varepsilon$ -events only occur between two consecutive  $\varepsilon_q$  events. In Fig. 7,  $p'_1$  defines  $\gamma$ , the end  $\varepsilon$ -event for a maximal group containing all entities. Then, maximal group that are not contain  $p_1$  or  $p'_1$  will end before or after  $\gamma$ , respectively.

To build the construction, all trajectories must have a constant times  $2^k$  vertices for the  $\varepsilon$ -events of  $q$  and  $q'$  and a constant number of vertices in between those  $\varepsilon$ -events. Each trajectory in the construction has  $\Theta(\sqrt{2}^n)$  vertices. We conclude that the number of maximal groups that contain time  $t_{mid}$  in this construction is at least  $2^k = 2^{n/2-1} = \Omega(\sqrt{2}^n)$ .  $\square$

## 7. Conclusions and Future Work

In this paper we introduced a variation on the grouping structure definition<sup>2</sup> and argued that it corresponds better to human intuition. The number of maximal groups that can arise in a set of  $n$  moving entities is  $\Theta(\tau n^3)$  in the worst case. We have given an algorithm for trajectories moving in  $\mathbb{R}^1$  that computes all maximal groups and runs in  $O(\tau^2 n^4)$  time. In  $\mathbb{R}^d$ , our algorithm runs in  $O(\tau^2 n^5 \log n)$  time. For the more general case where the input trajectories do not have time-aligned vertices, the algorithm for trajectories in  $\mathbb{R}^1$  can be extended at the cost of an extra factor of  $\alpha(n)$ , while the same result still holds for trajectories in  $\mathbb{R}^d$ .

Furthermore, we presented an algorithm that has only linear dependence in  $\tau$ , at the expense of exponential dependence in  $n$ . Since collections of trajectories are often very large in the number of time stamps and not necessarily in the number of trajectories, this algorithm or a practical variation on it may still be useful. This algorithm is not affected by whether or not the vertices of the trajectories are aligned in time.

The trade-off in the dependence on  $n$  and  $\tau$  gives rise to interesting open problems. Most importantly, is it possible to develop an algorithm whose running time is linear in  $\tau$  and polynomial in  $n$ ? Similarly, can we realize subquadratic

dependence on  $\tau$  without having exponential dependence on  $n$ ? In general, what trade-offs are possible?

Future work includes implementing our algorithms and experimentally showing the differences between the definition of groups by Buchin *et al.*<sup>2</sup> and our refined definition, both qualitatively and quantitatively. It would also be interesting to develop an output-sensitive algorithm that uses considerably less time if the output is small, or under realistic input assumptions. Finally, it would be interesting to investigate whether one can develop algorithms that take geodesic distance into account to define direct  $\varepsilon$ -connectedness instead of the straight-line distance, as was done for the previous definition of a group.<sup>22</sup>

## Acknowledgments

Maarten Löffler is partially supported by the Netherlands Organisation for Scientific Research (NWO), under projects 639.021.123 and 614.001.504. Frank Staals is partially supported by the Danish National Research Foundation (grant nr. DNR84). Lionov Wiratma is supported by the Ministry of Research, Technology and Higher Education of Indonesia (138.41/E4.4/2015).

## References

1. M. van Kreveld, M. Löffler, F. Staals and L. Wiratma, A refined definition for groups of moving entities and its computation, *Proc. 27th Int. Symp. Algorithms and Computation (ISAAC '16)*, Sydney, Australia (2016), pp. 48:1–48:12.
2. K. Buchin, M. Buchin, M. van Kreveld, B. Speckmann and F. Staals, Trajectory grouping structure, *J. Comput. Geom.* **6** (2015) 75.
3. J. Gudmundsson, P. Laube and T. Wolle, Computational movement analysis, *Springer Handbook of Geographic Information*, eds. W. Kresse and D. N. Danko (Springer, Berlin, 2012), pp. 725–741.
4. Y. Zheng and X. Zhou, *Computing with Spatial Trajectories* (Springer, New York, 2011).
5. M. Benkert, J. Gudmundsson, F. Hübner and T. Wolle, Reporting flock patterns, *Comput. Geom.* **41** (2008) 111.
6. J. Gudmundsson and M. van Kreveld, Computing longest duration flocks in trajectory data, *Proc. 14th ACM Int. Symp. Advances in Geographic Information Systems (ACM-GIS '06)*, Arlington, USA (2006), pp. 35–42.
7. J. Gudmundsson, M. van Kreveld and B. Speckmann, Efficient detection of patterns in 2D trajectories of moving points, *GeoInformatica* **11** (2007) 195.
8. M. Vieira, P. Bakalov and V. Tsotras, On-line discovery of flock patterns in spatio-temporal data, *Proc. 17th ACM Int. Conf. Advances in Geographic Information Systems (ACM-GIS '09)* (ACM, Seattle, USA, 2009), pp. 286–295.
9. U. Turdukulov, A. O. Calderon Romero, O. Huisman and V. Retsios, Visual mining of moving flock patterns in large spatio-temporal data sets using a frequent pattern approach, *Int. J. Geographical Inform. Sci.* **28** (2014) 2013.
10. Y. Huang, C. Chen and P. Dong, *Modeling Herds and Their Evolvments from Trajectory Data*, Lecture Notes in Computer Science, Vol. 5266 (Springer, Berlin, 2008), pp. 90–105.

11. H. Jeung, M. Yiu, X. Zhou, C. Jensen and H. Shen, Discovery of convoys in trajectory databases, *PVLDB* **1** (2008) 1068.
12. J. Yeoman and M. Duckham, Decentralized detection and monitoring of convoy patterns, *Int. J. Geographical Inform. Sci.* **30** (2016) 993.
13. H. H. Aung and K. Tan, Discovery of evolving convoys, *Proc. Scientific and Statistical Database Management, 22nd Int. Conf. (SSDBM 2010)*, Heidelberg, Germany (2010), pp. 196–213.
14. P. Kalnis, N. Mamoulis and S. Bakiras, *On Discovering Moving Clusters in Spatio-temporal Data*, Lecture Notes in Computer Science, Vol. 3633 (Springer, Berlin, 2005), pp. 364–381.
15. C. S. Jensen, D. Lin and B. C. Ooi, Continuous clustering of moving objects, *IEEE Trans. Knowledge and Data Eng.* **19** (2007) 1161.
16. Z. Li, B. Ding, J. Han and R. Kays, Swarm: Mining relaxed temporal moving object clusters, *PVLDB* **3** (2010) 723.
17. Z. Li, J. Han, M. Ji, L.-A. Tang, Y. Yu, B. Ding, J.-G. Lee and R. Kays, Movemine: Mining moving object data for discovery of animal movement patterns, *ACM Trans. Intell. Syst. Technol.* **2** (2011) 37:1.
18. S. Hwang, Y. Liu, J. Chiu and E. Lim, *Mining Mobile Group Patterns: A Trajectory-Based Approach*, Lecture Notes in Computer Science, Vol. 3518 (Springer, Berlin, 2005), pp. 145–146.
19. L.-A. Tang, Y. Zheng, J. Yuan, J. Han, A. Leung, C.-C. Hung and W.-C. Peng, On discovery of traveling companions from streaming trajectories, *Proc. 2012 IEEE 28th Int. Conf. Data Engineering (ICDE '12)* Washington, DC, USA (2012), pp. 186–197.
20. Y. Wang, E. Lim and S. Hwang, Efficient mining of group patterns from user movement data, *Data & Knowledge Eng.* **57** (2006) 240.
21. A. van Goethem, M. van Kreveld, M. Löffler, B. Speckmann and F. Staals, Grouping time-varying data for interactive exploration, *Proc. 32th Annual Symp. Computational Geometry (SoCG '16)*, Boston, USA (2016), pp. 61:1–61:16.
22. I. Kostitsyna, M. van Kreveld, M. Löffler, B. Speckmann and F. Staals, Trajectory grouping structure under geodesic distance, *Proc. 31st Int. Symp. Computational Geometry (SoCG '15)*, Eindhoven, The Netherlands (2015), pp. 674–688.
23. M. Sharir and P. K. Agarwal, *Davenport–Schinzel Sequences and Their Geometric Applications* (Cambridge University Press, 1995).
24. S. Parsa, A deterministic  $O(m \log m)$  time algorithm for the Reeb graph, *Discr. Comput. Geom.* **49** (2013) 864.
25. G. Andrienko, N. Andrienko, P. Bak, D. Keim and S. Wrobel, *Visual Analytics of Movement* (Springer-Verlag, Berlin, Heidelberg, 2013).