International Journal of Geographical Information Science
Vol. 00, No. 00, Month 200x, 1–31

research

Using Interactions and Dynamics for Mining Groups of Moving Objects from Trajectory Data

Corrado Loglisci
(*Received 00 Month 200x; final version received 00 Month 200x*)

Recent advances in tracking technology enable the gathering of spatio-temporal data in the form of trajectories. Analyzing trajectories can convey knowledge useful for prominent applications and designing computational solutions for mining groups of moving objects may turn out to be a valuable means for a wide class of problems related to mobility. The task of group mining has been investigated by considering mostly the spatial closeness and similarity of the trajectories, while little attention has been paid to the relationships between the trajectories and time-changing nature of the trajectories. The relationships may provide evidence of interactions between the moving objects. The time-changing nature may provide evidence of dynamics of the movements. Therefore, interactions and dynamics can be sources of information that one can consider to discover new forms of groups. In fact, groups of objects may be of interest not only when the objects move together or move close from each other, but also when they come from different places, change direction, join together and then move away from each other. Motivated by this, we introduce the concept of crews and propose a computational solution to discover crews. A crew gathers moving objects with similar interactions and similar dynamics. The proposed computational solution relies on *i)* new movement parameters, which explicitly consider interactions and dynamics, and *ii)* a distance-free clustering algorithm, which groups objects based on the similarity of the movement parameters. We conduct extensive experiments on real-world trajectory data, present a quantitative evaluation of the quality of the crews and perform comparisons with a baseline algorithm and with an algorithm of group pattern mining. The empirical results provide interesting insights on the relevance of some parameters in the construction of the crews.

Keywords: Trajectories, Moving Objects, Crews, Interactions, Dynamics.

* Email: corrado.loglisci@uniba.it *Department of Computer Science - Universita' degli Studi di Bari Aldo Moro, Italy*;

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

1. Introduction

The adoption of position-aware technologies, such as telemetry and GPS devices, has stimulated the development of solutions to collect movement data and has opened up a category of challenges regarding moving objects. Movement data thus constitute a very precious resource for the study of computational models, for the design of data analysis techniques and for the development of advanced solutions for real-world scenarios, such as crowd dynamics monitoring, human mobility understanding, public security and emergency management (Laube 2014, Dodge *et al.* 2008). A recurring problem is finding out collective movements that represent "common" behaviors of different objects and , in most approaches, it is faced with trajectory data analysis techniques aiming at discovering groups of objects that move close to each other for a time duration Dodge *et al.* (2008), Long and Nelson (2013b), Mazimpaka and Timpf (2016).

Benkert *et al.* (2006) and Gudmundsson and van Kreveld (2006) introduced the seminal notion of *flock*. It refers to a collection of objects moving together over a time interval of pre-fixed duration, such that for every time-stamp of the time interval there is a disk of pre-fixed radius that contains the objects. Ong *et al.* (2011) embed the duration into the flock notion to capture traffic jams. Jeung *et al.* (2008) extended the concept of flocks by relaxing the constraints on the shape and size. They formalized the notion of *convoy* and proposed an algorithm to grouping objects that are density-connected to each other within a generic geometric shape. However, the convoy has a practical brittleness due to the requirement on the temporal contiguity (Mazimpaka and Timpf 2016), that is, the objects are put together only when they move close to one another over an uninterrupted time interval.

Motivated by this, Li *et al.* (2010) relax the temporal constraints and propose the notion of *swarm* intended as group of objects that move together on disjointed time instants, while Zheng *et al.* (2014) introduce the notion of *gathering*, which is characterized by core members, which stick to the group, and by objects can enter and leave. Temporal discontinuity has also been studied in Wood (2013) by gathering individuals that move with continuous or intermittent *spatial coherence*. Three definitions of spatial coherence have been proposed: the first two types maintain the classical connotation of closeness (shared location), while the third type works on the reciprocal behavior of the individuals.

In the above-mentioned works, the trajectories are analyzed as sequences, considered independent of each other, composed of positions, considered unrelated to each other. This is a restrictive perspective because *i)* limits the potential we can extract from the movement of individual objects and from the movement of a collection of objects, and *ii)* overlooks two sources of information, that is, the presence of relationships between trajectories and time-changing nature of the movements. Indeed, the presence of relationships between the trajectories may provide evidence of the interactions between objects (Doncaster 1990). In fact, the objects can interact not only when they stay in the same place, but also when they are far apart and move towards a common location to meet. To account for the interactions, we should take the trajectories as correlated processes and consider forms of relationships different from the one due to the sole co-presence. The time-changing nature in intrinsic in movements and expresses the dynamics of physical properties of the motion (Laube *et al.* 2007). The objects can slow down, speed up and change direction. Dynamics characterizes not only individual objects, but also the movement of a collection of objects. For instance, when two objects travel together, the inter-distance can change. To account for dynamics, we should consider the variations of

physical properties, even between two consecutive positions, and analyze the positions as correlated observations of an evolving process.

There is a growing interest in interactions and dynamics in the recent research (Konzack *et al.* 2017, Long and Nelson 2013a, Dodge *et al.* 2016), while, in the past, very few attempts have been done. Andersson *et al.* (2008) define the concept of *leadership* in order to model the interaction between an individual (leader), who is moving ahead, and other individuals, who follow the leader for an uninterrupted time interval. Interactions are studied jointly with dynamics in the pioneer research of Laube *et al.* (2005), in which the notion of interaction is interpreted as relative motion. Relative motion is based on the comparison of motion attributes of individual objects. The motion attributes denote the dynamics of motion and correspond to quantitative movement parameters (e.g., azimuth[1]), which are computed over consecutive time instants relatively to a conventional reference system. Groups of objects that move in spatial proximity and that have the same motion attributes represent the so-called *Relative Motion (Remo) patterns*.

However, the method of extraction of the Remo patterns relies on a pattern matching approach, which requires domain knowledge and expert intervention to define a "template-pattern". This means that only the groups which satisfy the (pre-defined) templates will be extracted, while others will be discarded, regardless of the collective movements they represent. Another drawback of the Remo patterns is that they recognize only the interaction due to spatial proximity and model only the dynamics of individual objects. This way, the analysis is limited to discover groups with objects staying the same location or objects moving in a relatable space (Long and Nelson 2013b). This is quite unrealistic in practical situations because objects can be equally members of a group without necessarily move close or travel similar paths (Konzack *et al.* 2017).

In this paper, we study how discovering groups of moving objects by leveraging upon interactions and dynamics. Our main contributions are the formalization of the concept of *crew* and a computational solution to discover crews from raw trajectory data. A crew gathers moving objects that have similar interactions and similar dynamics. These two characteristics cannot be directly obtained from the raw trajectory data for the reasons discussed above. Thus, we define new movement parameters that represent interactions of pairs of objects and changes of motion of pairs of objects. These parameters correspond to spatio-temporal primitives able to describe *i)* the movement of an object relatively to the movement of another object (interactions), and *ii)* the variations of the physical properties of motion of an object relatively to the movement of another object (dynamics). Using interactions and dynamics to group moving objects allows us to track the dynamics of a collective.

However, while the changes can be captured by analyzing the objects in a conventional reference system, we cannot do the same with the interactions, because they should be modeled as perceived by a moving observer involved in the interaction. The approach we follow is to defining new trajectory primitives able to represent the movement of an object relative to another one (Noyon *et al.* 2007). This idea has been also explored by Andrienko *et al.* (2013), who propose to transform the physical space into an abstract space defined by the position of the group center and direction of the group's movements. In such abstract space, the authors represent the relative positions and movements of the individuals with respect to a reference entity, which, contrarily to the current paper, is the rest of the group.

The computational solution we propose is structured in two main steps. In the first

[1]The azimuth is the horizontal angle of an observer's bearing, measured clockwise from the north direction.

step, we transform the original descriptive space based on raw trajectory data into a feature space based on new movement parameters. This results in a representation based on vectors, where a vector represents the movement of a pair of objects over consecutive time instants. The second step is in charge of discovering valid crews. It performs an ad-hoc clustering algorithm in order to group vectors characterized by similar movement parameters. The members of a crew are pairs of objects whose movements are similar in terms of the new parameters.

Through experiments on real trajectory data (specifically, GPS data) of different categories of moving objects (specifically, pedestrians and wild animals), we investigate the applicability of the computational solution to real-world case studies. Moreover, we perform *i)* a qualitative evaluation by discussing interactions and dynamics expressed by the crews, *ii)* a quantitative assessment of the quality of the discovered groups and *iii)* comparative experiments with the swarms (Li *et al.* 2010), which is the kind of groups closer to the crews. The Remo patterns, seemingly the groups closest to the crews, were not considered for the experimental comparison because they require pre-defined templates, which are not necessary in our work.

The rest of the paper is organized as follows. The next section provides the basic concepts of this work. Section 3 illustrates the proposed computational solution as a framework structured in two main components. Section 4 reports the details of the extensive experiments we performed on two real-world trajectory data. Finally, some conclusions are drawn in Section 5 to close the paper.

2. Basics

Before illustrating the computational solution in detail, in this section we first explain the crews through an example and then provide the fundamental notions and formulation of the trajectory data mining problem studied.

2.1. *Motivating Example*

To give a concrete idea of the crews, we report an example in public security and safety (Mazimpaka and Timpf 2016), where interactions and dynamics can be used to identify places and moving objects (individuals or pedestrians) that have a high potential to cause threats or be targeted by threats. In that scenario, we need to monitor individuals even when they do not stay in the same location or even when they are distant from any target place. In this example, we also discuss the differences between the crews and some existing notions of groups.

In Figure 1, six individuals are tracked over seven time instants. If we rely on spatial closeness only, we may capture, as a flock, a convoy and a swarm, the movement of the individuals $\{o_1, o_2, o_3\}$ observed over the sequence of time instants $\{\tau_1, \tau_2, \tau_3, \tau_4, \tau_5\}$. Those groups may denote the behaviour of suspect individuals that move close to each other uninterruptedly. On the other hand, by combining spatial closeness and temporal discontinuities (Li *et al.* (2010)), we may capture, as a swarm, the movement of the individuals $\{o_4, o_5, o_6\}$ over the sequence of (non-consecutive) time instants $\{\tau_0, \tau_1, \tau_5\}$, while no flock or convoy would be detected (Figure 1a). This group would denote the behavior of suspect individuals that move close to each other with some interruptions. Considering the azimuth as a motion attribute, we discover also Remo patterns (Laube *et al.* 2005). In particular, there are two trend-settings, one composed of the individuals

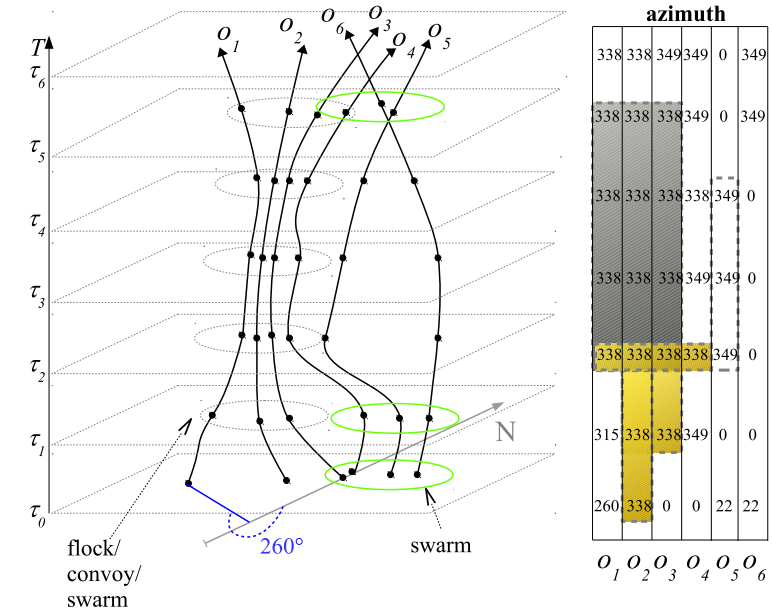


Figure 1. Comparison between flocks, convoys, swarms, Remo patterns and crews. a) Examples of flocks, convoys, swarms (left side) and Remo patterns (right side). b) Examples of crews.

$\{o_1, o_2, o_3\}$, the other composed of the individuals $\{o_2, o_3, o_4, o_5\}$ respectively. The first pattern covers the time-instants $\{\tau_1, \tau_2\}$ and has the individual o_2 as a trend-setter, while the second pattern covers the time-instants $\{\tau_2, \tau_3\}$ and has the individual o_4 as a trend-setter. We can also find a composite pattern defined by *concurrence* and *constancy* with the individuals $\{o_2, o_3, o_4, o_5\}$ over the time-instants $\{\tau_5, \tau_6\}$. This pattern may denote the behavior of suspect individuals that move close to each other and follow the direction

Table 1. Summary of the characteristics of different notions of groups of moving objects.

	no spatial constraint	shape relaxation	temporal discontinuity	interaction	motion changes	no pre-defined templates
flock						X
convoy		X				X
swarm		X	X			X
gathering	X	X				X
remo		X	X	X	X	
leadership				X		X
crew	X	X	X	X	X	X

of 338° north (Figure 1a). It should be noted that without proximity, no Remo pattern is retrieved.

However, by observing the individuals $\{o_1, o_2, o_3\}$ over the entire sequence $\mathcal{T}$, we see they are far apart at the beginning (τ_0), then they come close and follow similar paths ($\{\tau_1, \tau_2, \tau_3, \tau_4\}$) until they leave ($\{\tau_5, \tau_6\}$). This may reveal the movement of suspects coming from distant locations (for instance, airport or train station), they move close towards a target place and finally go towards separate locations (Figure 1b). We should note that this movement models also the above-mentioned groups (Figure 1a), but those groups does not model that movement. In fact, the algorithms for detecting flocks, convoys and swarms cannot discover that movement because they rely on only the spatial closeness. Even the Remo patterns cannot do it because they consider neither the changes within a collective (e.g., the distance between two individuals decreases) nor the interactions (e.g., the individuals which are apart and come close may be related to each other). To capture that movement, we need to *i)* relax the constraint of the proximity, *ii)* account for the interactions and motion changes and *iii)* using the similarity of interactions and motion changes when building the groups. This can be done by considering ad-hoc movement parameters able to characterize the movement of several individuals simultaneously rather than describe their movements independently on each other. Examples of these parameters are *azimuth shortest distance* and *inter-distance ratio* (Figure 1b), which describe pairs of individuals. The former accounts for the shortest distance between the azimuths formed by the two individuals over two time instants, the latter denotes the change of the spatial distance between the positions of the individuals over the same time instants. In this context, we can build a crew with the pairs $\{o_1, o_2\}$ and $\{o_2, o_3\}$, based on the similarity of azimuth shortest distance and inter-distance ratio on consecutive time instants. In particular, there is similarity on the values of azimuth shortest distance and inter-distance ratio respectively, for the pairs $\{o_1, o_2\}$ and $\{o_2, o_3\}$: they fall in [22,45) and [1,2) respectively over $\langle \tau_0, \tau_1 \rangle$. We see that also in the time instants $\langle \tau_1, \tau_2 \rangle$, $\langle \tau_2, \tau_3 \rangle$ and $\langle \tau_3, \tau_4 \rangle$, the values of azimuth shortest distance are similar and fall in [0;22) as well as the values of inter-distance ratio that fall in [0,1). For the same reason, the crew covers also the movements in $\langle \tau_4, \tau_5 \rangle$and $\langle \tau_5, \tau_6 \rangle$. Using the same principle, we can build a crew with the pairs $\{o_4, o_5\}$ and $\{o_5, o_6\}$. It is also models the swarm $\{o_4, o_5, o_6\}$ (Figure 1a). A concise representation of the differences between crews and other groups is reported in Table 1.

2.2. *Problem Formulation*

Here we provide fundamental notions and formulate the problem studied in this work. Frequently used symbols are reported in Table 2. Let $\mathcal{O}$=$\{o_1, o_2, \ldots, o_n\}$ be the set of all moving objects and $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_m\}$ be the time domain. The trajectory of an object o is represented by a poly-line that is given as a finite sequence of positions (fixes), each associated with a time instant of $\mathcal{T}$. The trajectory is denoted with $tr(o) : \langle (p_1, \tau_1), (p_2, \tau_2),$

Table 2. List of symbols frequently used.

Symbol	Explanation
$\mathcal{O}$	the set of moving objects
$\mathcal{T}$	the time domain
$tr(o_u)$	the trajectory of the object o_u
$\mathcal{F}$	the set of the movement parameters
F_l	the l-th movement parameter describing the movement of a pair of objects
$\langle \tau_i, \tau_j \rangle$	a sequence of two time instants of $\mathcal{T}$
$\mathcal{G}$	a pair group
o_r	a reference object of a pair group
o_s	a participant object of a pair group
z_l	the value associated with the pairs of a similarity-based pair group for the movement parameter F_l
$[z_{l_{lower}}, z_{l_{upper}}]$	a range of values of a movement parameter F_l
Π	a timeline (a series of sequences of time instants)
PG	a pair group
PC	a cluster of pairs of objects
SPG	a similarity-based pair group
μ	the minimum number of objects that a valid crew has to contain
γ	the maximum temporal gap in the timeline of a valid crew

$\ldots, (p_m, \tau_m)\rangle$, where $p_i \in \mathbb{R}^2$ is the geo-spatial position sampled at $\tau_i \in \mathcal{T}$. The time instants of a trajectory may not be equally distanced. Two different trajectories may have different time instants and therefore they may have different lengths. We call a sequence of time instants two-by-two consecutive arranged as $\langle \tau_i, \tau_{i+1}, \ldots, \tau_j, \tau_{j+1}, \ldots, \tau_h, \tau_{h+1} \rangle$ $(i < j < h)$ a *timeline*.

In this work, the raw trajectory data are projected into a feature space built with the movement parameters $\mathcal{F}$=$\{F_1, \ldots, F_l, \ldots, F_f\}$. A movement parameter is represented by a function F_l, which maps the positions of two objects o_u and o_v sampled at the time instants $\langle \tau_i, \tau_{i+1} \rangle$ to a numeric value z_l. This is abbreviated as $F_l|_{\langle \tau_i, \tau_{i+1} \rangle}(tr(o_u), tr(o_v)) \rightarrow z_l$. Intuitively, the movement parameters characterize the movement of two objects and specifically account for the interactions of the objects and dynamics of their motion. We determine their values on two consecutive time instants in order to capture the shorter variations of the interactions and motion, that is, the variation from one time instant to the next one. This choice allows us to characterize the trajectories at finer resolutions (Dodge *et al.* 2009).

Now, we formalize notions useful to state the concept of crews and design the algorithm to discover crews. For each notion, we provide a formal statement and an informal explanation.

Definition 2.1: [Pair Group] Let

(i) $\mathcal{O}$ be the set of moving objects
(ii) $\mathcal{G} = \{(\mathrm{o}_r, \mathcal{R}) \mid \mathrm{o}_r \in \mathcal{O}, \mathcal{R} \subseteq \mathcal{O} \setminus \{\mathrm{o}_r\}\}$

then $\mathcal{G}$ is a pair group.

A pair group (PG) consists of $|\mathcal{R}|$ pairs that have one object in common. Its purpose is to model the movements of the pairs of objects formed with the object o_r. We build the pair groups by taking one object at a time as the *reference* object and combining it with all the others (*participants*), so the pair groups are used to model the movements of all the pairs. For simplicity, we will denote the pair (o_r, o_s) by specifying always the reference as the first element. For instance, given the objects $\{o_1, o_2, o_3, o_4\}$, the set $\{(o_2, o_1), (o_2, o_3), (o_2, o_4)\}$ is a pair group.

Definition 2.2: [Pair Cluster] Let

(i) $\mathcal{O}$ be the set of moving objects
(ii) $\mathcal{G} = \{(\mathrm{o}_r, \mathcal{R}) \mid \mathrm{o}_r \in \mathcal{O}, \mathcal{R} \subseteq \mathcal{O} \setminus \{\mathrm{o}_r\}\}$
(iii) $[z_{l_{lower}}, z_{l_{upper}}]$ be a range of values of the movement parameter F_l

(iv) $\mathbb{1}(\cdot)$ be an indicator function that, given the value z_l of the movement parameter F_l, returns true if $z_l \in [z_{l_{lower}}, z_{l_{upper}}]$, otherwise it is false,

then $\mathcal{G}$ is a Pair Cluster iff $\forall(o_r, o_s) \in \mathcal{G}$, $\forall F_l \in \mathcal{F}'(\mathcal{F}' \subseteq \mathcal{F})$, $\exists\ \tau_i \in \mathcal{T}$ s.t. the indicator function $\mathbb{1}(F_l|_{\langle \tau_i, \tau_{i+1} \rangle}(tr(o_r), tr(o_s)))$ returns true.

Intuitively, a pair cluster (PC) comprises pairs whose values of a movement parameter, computed on consecutive time instants, fall in the same range.

Definition 2.3: [Similarity-based Pair Group] Let

(i) $\mathcal{G}$ be a pair group
(ii) $\Pi = \langle \tau_i, \tau_{i+1}, \ldots, \tau_h, \tau_{h+1} \rangle$ be a timeline
(iii) $[z_{l_{lower}}, z_{l_{upper}}]$ be a range of values of the movement parameter F_l
(iv) $\mathbb{1}(\cdot, \cdot)$ be an indicator function that, given two values of a movement parameter F_l, returns true if they fall in $[z_{l_{lower}}, z_{l_{upper}}]$, otherwise it is false,

then $\mathcal{G}$ is a Similarity-based Pair Group iff $\forall(o_r, o_s)$, $(o_r, o_t) \in \mathcal{G}$, the indicator function $\mathbb{1}\left(F_l|_{\langle \tau_j, \tau_{j+1} \rangle}(tr(o_r), tr(o_s)), F_l|_{\langle \tau_j, \tau_{j+1} \rangle}(tr(o_r), tr(o_t))\right)$ returns true, $\forall 1 \leq j \leq m : \tau_j \in \Pi, \tau_{j+1} \in \Pi$.

Intuitively, a similarity-based pair group (SPG) refers to a pair group in which the values of the movement parameters, computed on the timeline Π, fall in the same range.

An example of an SPG built with two movement parameters is illustrated in Figure 1b, where we use the azimuth distance and distance-ratio defined as in Section 2.1. We can compute the values of the movement parameters on the fixes sampled at time instants $\langle \tau_0, \tau_1 \rangle$. For instance, the azimuth distance between the azimuths of o_1 and o_2 (300° and 260°) is 40, the azimuth distance between the azimuths of o_2 and o_3 (260° and 290°) is 30, the azimuth distance between the azimuths of o_2 and o_4 (290° and 50°) is 240. By supposing the ranges [0,21), [22,45), [46,359] for azimuth distance and [0,1), [1,2), [2,10] for distance ratio, we can build an SPG with the pairs (o_2, o_1) and (o_2, o_3) on the timeline $\langle \tau_0, \tau_1 \rangle$. Indeed, the values of the azimuth distance fall in the same range, that is [22,45), and the values of the distance ratio fall in the same range, that is [1,2), respectively. The pair (o_2, o_4) is not a member because the value of the azimuth distance is not included in [22,45) but in [46,359) and the value of the distance ratio is not included in [1,2) but in [0,1), therefore it cannot be considered similar to the others two.

Having defined the concepts of pair group, pair cluster and similarity-based pair group, we can define the crews formally.

Definition 2.4: [Crew] Let $\mathcal{C} = \{\mathcal{S}_1, \ldots, \mathcal{S}_s\}$ be a set of similarity-based pair groups. It is a crew iff $\forall \mathcal{S}_u, \mathcal{S}_v \in \mathcal{C}$

(i) $\mathcal{G}_u = \mathcal{G}_v$,
(ii) $\langle \tau_i, \tau_j \rangle \cap \langle \tau_h, \tau_k \rangle = \oslash$, $\forall 1 \leq i < j \leq h < k : \tau_i, \tau_j$ time instants of the timeline of $\mathcal{S}_u$ and τ_h, τ_k time instants of the timeline of $\mathcal{S}_v$, respectively.

A crew is characterized by *i)* one pair group, *ii)* the timelines of the SPGs of $\mathcal{C}$ *iii)* the ranges of values associated with the movement parameters, for each timeline. The union of the timelines of the SPGs $\{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ results in the timeline of the crew.

Intuitively, Definition 2.4 states that a crew is a moving group composed of one reference object and a set of participant objects, where the members (pairs) have similarities in terms of movement parameters along a sequence of time instants. For instance, in Figure 1b, the pairs (o_2, o_1) and (o_2, o_3) have similar movement parameters along the timeline $\langle \tau_0, \tau_1, \tau_4, \tau_5, \tau_5\ \tau_6 \rangle$ (azimuth distance in [22,45), distance ratio in [1,2)) and along the

timeline $\langle \tau_1, \tau_2, \tau_2, \tau_3, \tau_3, \tau_4 \rangle$ (azimuth distance in [0,22), distance ratio in [0,1)).

To formulate the problem of discovery of crews, we consider some constraints to guide the analysis towards the identification of significant groups. The constraints are user-defined requirements expressed as input thresholds. In particular, we have *i)* γ, the maximum temporal gap (in terms of time instants) between two consecutive similarity-based pair groups; *ii)* μ, the minimum number of objects required to build a crew.

Now, we can formally state the problem:

Assume we are given: the set of moving objects $\mathcal{O}$, set of trajectories $\{tr(o_1), \ldots, tr(o_n)\}$, maximum temporal gap γ and minimum number of objects μ.

Discover the crews compliant to Definition 2.4 and satisfying the input-thresholds γ and μ.

3. The Framework

The framework implements the computational solution for the problem formulated above. It is structured in two main components. The first component (*Raw trajectory data pre-processing* in Figure 2) processes the raw trajectory data and uses a vector-based representation in order to capture the interactions and dynamics of pairs of objects. A vector describes the movement of a pair, in terms of the movement parameters $\mathcal{F}$, on two consecutive time instants τ_i, τ_{i+1}. We distinguish the movement parameters in two different categories. The second component of the framework (*Discovery of crews* in Figure 2) performs a hierarchical clustering algorithm on the vectors and discovers valid crews. The clustering algorithms outputs similarity-based pair groups, which are combining together to discover valid crews.

3.1. *Trajectory Data Preparation*

The trajectory of a moving object in the real world is always a continuous line, but for storage and analysis purposes, it is represented in a discrete form, which corresponds to a sequence of fixes. These may contain noise, outliers and gaps due to the instrumental factors of the tracking devices (e.g., the precision of the device) and to physical factors (e.g., the existence of obstacles), which lead to an irregular timing of the trajectories

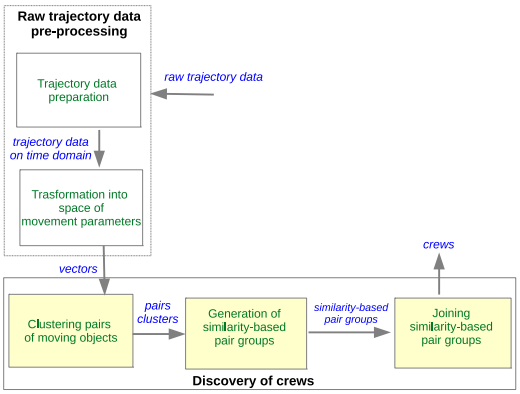


Figure 2. The proposed framework is structured in two main components. The first component performs projects the raw trajectory data into a feature space defined with the parameters $\mathcal{F}$. The second components generates similarity-based pair groups and combines them to discover valid crews.

and to an unfair analysis whether the fixes are processed as recorded. To overcome this problem we propose a data preparation step, which includes techniques typically used to make the trajectories more reliable (Dodge *et al.* 2009). In particular, we filter out outliers by removing the positions whose inter-distance from the previous position is greater than a predefined experimentally set threshold. Then, the missing points (caused by the filtering) are filled by means of a linear interpolation operation, which generates new fixes at the same rate of the sampling of the raw data. Finally, for each object, we generate regular trajectories by transforming the original temporal axis in the time domain $\mathcal{T}$. In particular, this operation maps one equal-width time interval (of the original time axis) to one time instant τ_i of $\mathcal{T}$, to which we associate the weighted moving average of the original fixes. This way, the fixes sampled at the beginning of the time interval will have less weight, while those collocated at the end will have more weight.

We clarify that the construction of the time-domain $\mathcal{T}$ from the original time axis is not finalized to define two temporal granularities, but to prepare trajectory data for one level of temporal granularity. This may overlook crews that can be evident at higher level of analysis, but the cross-granularity discovery is beyond this work.

3.2. *Movement parameters for Interactions and Dynamics*

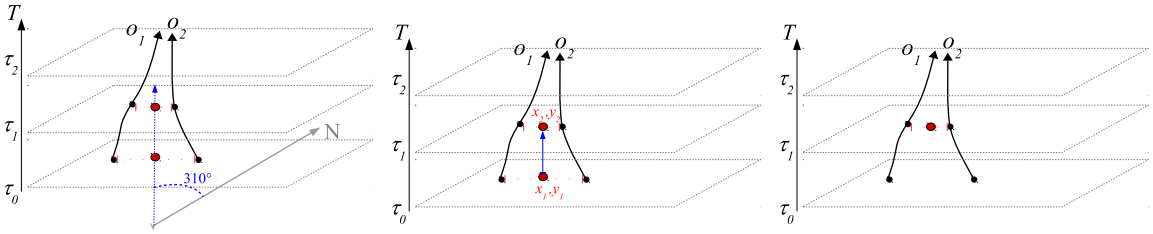
To capture interactions and dynamics we introduce new movement parameters, which, more precisely, are defined to characterize the reciprocal behavior of an object relative to another object and changes of motion of interacting objects. The definition of new parameters is necessary why most existing movement parameters has two main limitations, the representation of the movement of an object regardless of the behavior of the others and the modeling of the motion change of individual objects only (Dodge *et al.* 2008), therefore they cannot be adapted to the current scenario. However, for this step, we use some existing primitive and derivative parameters. In this work, we identify two categories of movement parameters: the first category (*Pairwise Dynamics Parameters, PDP*) is defined to represent the variations of geo-spatial primitives that describe the pair as a sole moving entity, while the second category (*Interactions-wise Dynamics Parameters, IDP*) is defined to represent the variations of physical derivatives and geo-spatial primitives that describe the movement of an object relative to another object. All the movement parameters $\mathcal{F}$ are computed in the time domain $\mathcal{T}$, except two parameters, which are defined on original time axis. In the following we provide a detailed description.

3.2.1. Pairwise Dynamics Parameters

- *Azimuth.* This denotes the direction in which the pair (o_r, o_s) is going over two consecutive time instants $\langle \tau_i, \tau_{i+1} \rangle$. An example is reported in Figure 3a. We compute it as the angle measured clockwise between the line passing through two middle points and north. The two middle points are associated with the two time instants separately and are located at half the distance between the fixes of the objects respectively. The contribution of the parameter Azimuth is to finding pairs of objects that move in similar directions and that have similar changes of direction over two time instants.
- *Displacement.* This denotes the distance between the fixes of the pair (o_r, o_s) recorded in the two time instants $\langle \tau_i, \tau_{i+1} \rangle$. An example is reported in Figure 3b. We compute it as the Euclidean distance between the two middle points, which are determined as in the case of the parameter Azimuth. The contribution of the parameter Displacement is to identifying pairs of objects that keep similar distances over two time instants. Another concept of displacement has been proposed by Long and Nelson (2013a), but,

equally to the ours, it can re-formulated in terms of Euclidean distance.

- *Position.* This is a primitive parameter and denotes the position of the pair (o_r, o_s) at the end of the two time instants $\langle \tau_i, \tau_{i+1} \rangle$. An example is reported in Figure 3c. It has two components, latitude and longitude. We compute it as the middle point of the two fixes at the second time instant. The parameter Position allows us to identify pairs that move around a shared location over two time instants.



(a) (b) (c)

Figure 3. The movement parameters of the category Pairwise Dynamics Parameters account for the variation of the geo-spatial primitives describing the pair of objects as a sole moving entity ((a) Azimuth, (b) Displacement, (c) Position).

3.2.2. Interaction-wise Dynamics Parameters

- *Azimuth Distance.* This denotes the difference between the directions in which the two objects o_r and o_s move respectively. We compute it as the distance between the angles that correspond to the two azimuths. The azimuth of an object is measured clockwise between the line passing through the two fixes of the object and north. For instance, in Figure 4a, the azimuth of o_1 and o_2 is 315° and 290° respectively, whereas the angular distance is 25°. The parameter Azimuth Distance complements the parameter Azimuth, in that it quantifies how far the directions of the two azimuths are. This parameter is used to group pairs whose changes of direction are similar. A typical situation is when the pairs are moving from different directions and proceed in the same direction.
- *Distance Ratio.* This denotes the variation of the inter-distance between the two objects o_r and o_s over two time instants $\langle \tau_i, \tau_{i+1} \rangle$. We compute it as the ratio between two inter-distances, that is, the Euclidean distance between the two fixes at the second time instant τ_{i+1}, divided by the Euclidean distance between the two fixes at the first time instant τ_i. The result is a positive real number, which exceeds 1 when the inter-distance increases by a factor higher than 1. For instance, in Figure 4b, the ratio is 70/100. This parameter is used to group pairs that are coming close (or moving away) by keeping similar distances. Moreover, it allows us to model the cases in which one of the two objects moves away from (or comes close to) the other one at a relatively high speed, which may be a reasonable motivation for longer distances from a time instant to the next one. For instance, when a pedestrian catches a vehicle to move away from another pedestrian. It should be noted that this parameter provides a quantification of the distance between the objects when one changes direction (see Azimuth Distance).
- *Tortuosity Ratio.* This denotes how the tortuosity of the trajectory $tr(o_r)$ of an object changes, compared to the tortuosity of another trajectory $tr(o_s)$. An example is reported in Figure 4c. The tortuosity refers to the degree of windingness or bending of the trajectory within a time interval (Laube *et al.* 2007). To compute the Tortuosity

Ratio, we use the pre-processed raw trajectories as returned by the linear interpolation. In particular, given i) Y_{τ_1,o_1} and Y_{τ_1,o_2} the tortuosity values of two objects o_1 and o_2 computed on the fixes included in the time interval associated with τ_1, and ii) Y_{τ_2,o_1} and Y_{τ_2,o_2}, the tortuosity values of the two objects computed on the fixes included in the time interval associated with τ_2, the Tortuosity Ratio is equal to $\frac{1+|Y_{\tau_2,o_1}-Y_{\tau_2,o_2}|}{1+|Y_{\tau_1,o_1}-Y_{\tau_1,o_2}|}$. A tortuosity value is determined as the ratio between the sum of the distances of consecutive fixes of the trajectory included in the time interval and the length of the path. It has normalized to the range [0,1], as proposed by Dutton (1999). The range of Tortuosity Ratio is [0.5,2], thus values close to 0.5 indicate that the tortuosities of the trajectories become identical in the second time interval, while values close to 2 indicate that the tortuosities are completely different in the second time interval and eventually value 1 indicates that the tortuosities have not been changed. The rationale behind this parameter is that of grouping pairs based on the similar changes in terms of windingness of the trajectories. Clearly, this does not mean that only the pairs with similar tortuosities may be members of a crew, but also pairs with different tortuosities but similar variations relative to two time intervals.

- *Speed Ratio.* This denotes how the speed of an object o_r changes compared to the speed of another object o_s. In particular, we account for the absolute variation of the respective speeds. To compute it, we use the pre-processed raw trajectories as returned by the linear interpolation and determine first the average speeds of the two objects separately and then the final ratio. An example is reported in Figure 4d. In particular, given i) V_{τ_1,o_1} and V_{τ_1,o_2} the speed values of the two objects o_1 and o_2 computed on the fixes included in the time interval associated with τ_1, and ii) V_{τ_2,o_1} and V_{τ_2,o_2} the speed values of the two objects computed on the fixes included in the time interval associated with τ_2, the Speed Ratio is equal to $\frac{1+|V_{\tau_2,o_1}-V_{\tau_2,o_2}|}{1+|V_{\tau_1,o_1}-V_{\tau_1,o_2}|}$. The average speed is calculated on consecutive fixes and it has normalized to the range [0,1]. The range of the Speed Ratio is [0.5,2], thus values close to 0.5 indicate that the objects, on the second time interval, travel with similar speeds, values close to 2 indicate that the speeds differ greatly as time goes by, while when the speeds do not change the value of Speed Ratio is 1.

As a result of the first component of the framework (Figure 2), we have vectors built for all the sequences of consecutive time instants $\langle \tau_i, \tau_{i+1} \rangle$ of the time domain $\mathcal{T}$. A vector is the result of the transformation of the fixes of two objects (o_r, o_s) sampled at two consecutive time instants $\langle \tau_i, \tau_{i+1} \rangle$, it has $|\mathcal{F}|$ dimensions and contains the values $F_l|_{\langle \tau_i, \tau_{i+1} \rangle}(tr(o_r), tr(o_s))$ of each movement parameter F_l.

3.3. *Discovery of Crews*

The second component of the framework is in charge of discovering crews from the vectors previously extracted. It should be noted that the vectors depict the smallest interactions and shortest changes. Indeed, they capture the interactions of the smallest set of objects, that is, a pair of objects, over the shortest time interval, that is, two consecutive time instants. This gives us some hints on how designing the method for the second component. Starting from the consideration that a crew involves more than two objects and may cover more than two time instants, the key idea is to consider the vectors as *building blocks* and build valid crews by combining the vectors properly. To do this, the second component first finds out pair clusters (PCs), then generates similarity-based pair groups (SPGs) and finally builds the crews from SPGs.

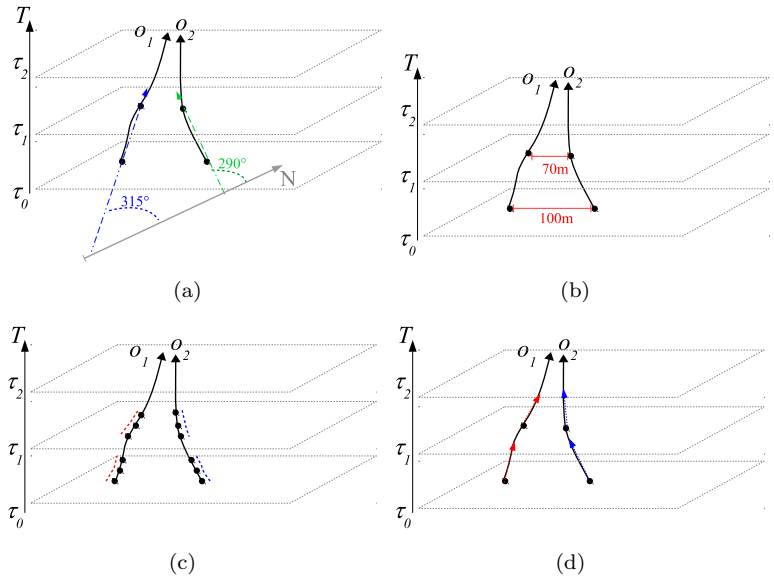


Figure 4. The movement parameters of the category Interactions-wise Dynamics Parameters account for the variation of the physical derivatives and geo-spatial primitives describing the movement of an object relative to another object ((a) Azimuth Distance, (b) Distance Ratio, (c) Tortuosity Ratio, (d) Speed Ratio).

The algorithmic choice of combining together vectors to generate SPGs reminds the multi-scale analysis, whose purpose is recognizing groups of different sizes that have the same collective movements (e.g., Wood and Galton (2009)). However, compared to the current work, there is a substantial difference: accordingly to Definition 2.4, the crews have well-distinct behaviors and cannot share the same collective movement, which instead is permitted in the multi-scale analysis.

3.4. *Clustering Pairs of Moving Objects*

To find out PCs, we propose a hierarchical clustering algorithm that groups vectors based on the similarity of the parameters $\mathcal{F}$. It does not rely on costly functions of distance, as instead many clustering algorithms do, and decides the membership of an element (pair) to a cluster by means of a test performed on the values of the parameters $\mathcal{F}$, which requires less computation.

The algorithm follows the principle of the conventional decision trees, traditionally used for predictive tasks (Frank *et al.* 1998, Loglisci and Malerba 2017), and induces a tree-like structure from the vectors returned by the first component. The nodes of the tree are associated to subsets of vectors: the vectors contained in the root are similar with respect to one movement parameter, while the vectors contained in the leaves are similar with respect to all the parameters. In particular, an internal node of the tree is characterized by *i)* a movement parameter F_l (Section 3.2), *ii)* a pair cluster L (Definition 2.2), *iii)* a threshold value c (whose range is that of the parameter F_l) and it is connected to other nodes by means of two-way branches. A branch goes from a "starting node" to an "ending node" and encodes the clustering function by testing the threshold value c against the values of the parameter F_l of the vectors contained in the starting node. More precisely, a branch guides a subset of the pairs of the PC of the starting node towards the ending node, that is, the set of the vectors that satisfy the test. Thus, the

vectors of an ending node represent the pair of objects (o_r, o_s), for which the indicator function $\mathbb{1}(\cdot,\cdot)$ is true for the parameter F_l. The parameter F_l we assign to a node is selected with a criteria based on the (dis)similarity of the vectors reaching the node. More precisely, we select the parameter F_l that maximizes the reduction between the dissimilarity computed on the vectors of the node and dissimilarity of two subsets. This operation uses the formula

$$DissReduction = \sigma^2(L)_{F_l} - \frac{\sigma^2(left)_{F_l} * |left| + \sigma^2(right)_{F_l} * |right|}{|L|}, \tag{1}$$

where $\sigma^2(L)_{F_l}$ ($\sigma^2(left)_{F_l}, \sigma^2(right)_{F_l}$) is the dissimilarity computed on the values of the parameter F_l for the vectors of L ($left, right$), which is defined as follows:

$$\sigma^2(L) = \sum_{i=1}^{n} p_i * (v_i - \bar{v})^2, \tag{2}$$

where, $v_i (i = 1 \dots n)$ are the values of the parameter F_l in the set L, p_i is the probability of observing v_i in L and $\bar{v}$ is the mean of the values $v_1, \dots, v_n$.

Procedurally, the tree is built by means of an algorithm that recursively splits the initial set of the vectors into subsets of decreasing size. The algorithm starts from the root (which contains the complete set of the vectors) and, moving downwards, creates new (ending) nodes, to which assigns subsets of the vectors of the previously created (starting) node. The ending nodes of an execution of the algorithm become the starting nodes of next execution. More precisely, after having appended new nodes to the tree, the algorithm examines one node at a time. It builds two-by-two subsets of vectors for each movement parameter F_l, but takes only those that maximize the dissimilarity reduction *DissReduction* (Equation 1). The selected subsets are identified by a threshold value, which splits the values of the movement parameter F_l into two ranges. Thus, one subset covers one range and comprises the vectors whose values of the parameter fall in that range (Definition 2.2 and Definition 2.3). The recursive procedure terminates when the initial dissimilarity (computed on the root) has been been reduced for all the parameters by a factor fixed by the user ($required\_reduction \in [0, 1]$) or when all the leaves have a number of pairs of objects smaller than μ^2. In the latter case, the tree might not have all the movement parameters. An illustration of the algorithm of tree induction is reported as Appendix.

3.5. *Building Crews from Similarity-based Pair Groups*

To discover crews with similar interactions and similar dynamics, we should generate SPGs with the highest number of similar movement parameters. The organization of the tree built above suggests to considering the PGs collocated at the leaves. Indeed, by the effect of the dissimilarity reduction (Equation 1), the vectors contained in the leaves have the highest number of similar movement parameters, compared to the vectors of the internal nodes. To generate SPGs, we need two types of information (Definition 2.3), one has a spatial connotation and is related to the ranges of values $[z_{l_{lower}}, z_{l_{upper}}]$, the other one has temporal connotation and is related to the timeline Π. The spatial information can be obtained from the tree; indeed, the path from the root to a leaf directly provides the list of the movement parameters and relative ranges of values. To determine the

timelines, we have to perform a further step of analysis on the pairs contained in the leaves. More precisely, for each leaf, we first gather the vectors by pair group (that is, pairs having an object in common) and then generate the timeline of each pair of objects (o_r, o_s) by using their time instants $\langle \tau_i, \tau_{i+1} \rangle$. Finally, we generate the timeline with the time-instants which are common to all the pairs. To do this, we adapt the technique of computation of intersections among sets proposed in Layer *et al.* (2013), which returns the intervals common to sequences of disjointed time intervals. Our adaptation first matches a primary timeline against remaining timelines and then evaluates the intersection between the time instants of the primary timeline and time instants of the other timelines by means of two binary search operations. These operations work on two sorted lists of time instants, one is composed of the time instants τ_i, the other is composed of the time instants τ_{i+1}. The intersecting time instants are thus sorted by chronological order and combined to form the candidates. Finally, we select the timeline that satisfies a user preference criterion. The preference criterion selects the final timeline and consequently the pair group of the SPG. There are two alternative preference criteria, one criterion, denoted as *maxDuration*, chooses the pair group having the longest timeline, while the second criterion, denoted as *maxObjects*, chooses the pair group with more objects. The criterion choosen holds also for the selection of the primary timeline used when searching intersections. Indeed, the option *maxDuration* picks the longest timeline present in the pair groups, while the option *maxObjects* picks the shortest timeline. Finally, we analyze all the SPGs built from the leaves, but consider only those that satisfy the threshold of the minimum number of objects μ.

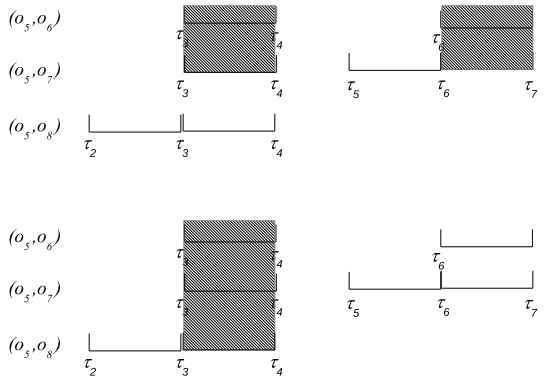


Figure 5. The timelines of candidate similarity-based pair groups are generated from the time instants of the vectors on the leaves. The resulting timeline is selected with an user preference criterion.

An illustration is reported in Figure 5, where there are the timelines of the pairs (o_5, o_6), (o_5, o_7) and (o_5, o_8) of a leaf. We can extract two candidates, one is obtained from the intersection of the time instants of the pairs (o_5, o_6) and (o_5, o_7) and covers the timeline $\langle \tau_3, \tau_4, \tau_6, \tau_7 \rangle$, the other one is obtained from the time instants of the pairs (o_5, o_6), (o_5, o_7) and (o_5, o_8) and covers the timeline $\langle \tau_3, \tau_4 \rangle$. Thus, by applying the criterion *maxDuration* to this example, we obtain the timeline at the top of Figure 5.

The construction of the crews is completed by an algorithm that first selects the SPGs with same reference object o_r and same participant objects o_r and then joins their timelines in chronological order. Procedurally, the algorithm considers one reference object o_r at a time and works on the timelines in which o_r is present. It builds one crew at a time by extending incrementally its timeline. A crew is not built from scratch, but it is obtained by extending the timelines of the crews, which have been previously built, with

other admissible time instants. This strategy guarantees the construction of maximal crews, so there are no crews contained in each other. In the following, we first describe how algorithm operates when builds the crews with same reference object o_r and then provide an explanatory example.

The algorithm starts with the early time instants of the reference object o_r and proceeds by selecting the time instants that have not been considered before. It appends newly selected time instants to the timeline of the current crew. Once the time instants have been used, they are marked, so we will not have duplicated crews. The current crew is extended only if two conditions are satisfied, *i)* the new time instants are distant from those already appended at most γ, and *ii)* the new SPG contains the same participants. When there are no time instants for the current crew, the timeline is completed. Then, the algorithm continues with the time instants that have not been considered before and with those that have been considered before but that have not been marked yet. Thus, the algorithm seeks other crews by evaluating (sub)timelines of previously built crews. More precisely, it removes the time instants in the same order with which they have been inserted and performs new joins with other admissible time instants. The construction of the crews having o_r as the reference stops when all the time instants have been marked and there are no time instants that can be used for new crews. The algorithm terminates when all the reference objects have been scanned. Finally, the crews that have less than μ objects are filtered out.

An illustration is reported in Figure 6 with the timelines of three SPGs ($\langle A, D, I\rangle$, $\langle B, C, E\rangle$, $\langle G, H, F\rangle$ and [1]). Suppose o_r=o_2, γ=3. The algorithm starts with the early time instants, that is, A and evaluates the closer time instants, that is G, which cannot be used because there are not the same participants. Thus, the algorithm evaluates B and joins it with the current crew $\langle A\rangle$ because B has the same participants and is distant from A less than γ. For the same reason, also D and E are joined, so the algorithm builds the crew $\langle A, B, C, D, E\rangle$ and marks A,B,C,D and E. The time instants F are not considered because there are not participants in common, so the algorithm restarts by removing E from the crew $\langle A, B, C, D, E\rangle$ and seeking new joins with other admissible time instants. It picks I, which is not marked, and uses it to build the crew $\langle A, B, C, D, I\rangle$, then marks I. Next, the algorithm removes I and recovers the crew $\langle A, B, C, D\rangle$, which cannot further be extended because there are no admissible time instants after D. The same analysis is done also with $\langle A, B, C\rangle$, $\langle A, B\rangle$ and $\langle A\rangle$. After that, the procedure creates a crew with G, which cannot further be extended because its distance from H exceeds γ.

performs a join with C because A is not yet marked. Then, the algorithm considers the closer time instants, that is G, but cannot be used because there are not the same participants. Next, the algorithm evaluates B and joins it with the current crew $\langle A\rangle$ because B has the same participants and is distant from A less than γ. For the same reason, also C, D and E are joined, so the algorithm builds the crew $\langle A, B, C, D, E\rangle$ and marks A,B,C,D and E. The sequence of time instants F is not considered because there are not participants in common. The algorithm restarts by removing E from the crew $\langle A, B, C, D, E\rangle$ and seeking new joins with other admissible time instants. Thus, it finds I, which is not marked, and uses I to build the crew $\langle A, B, C, D, I\rangle$, then marks I. Next, the algorithm removes I and recovers the crew $\langle A, B, C, D\rangle$, which cannot further be extended because there are no admissible time instants after D. The same analysis is done also with $\langle A, B, C\rangle$, $\langle A, B\rangle$ and $\langle A\rangle$. After that, the procedure creates a crew with

[1]For simplicity, we replace the notation of the time instants $\langle \tau_i, \tau_{i+1}\rangle$ with capital letters.

G, which cannot further be extended because its distance from H exceeds γ.

The algorithm accounts also for the cases in which two SPGs have exactly not the same pairs of objects, but they have pairs in common. For instance, given the sets $\{(o_5, o_6), (o_5, o_7)\}$ and $\{(o_5, o_6), (o_5, o_7), (o_5, o_8)\}$, we may use the pairs $\{(o_5, o_6), (o_5, o_7)\}$, which are present in both sets.

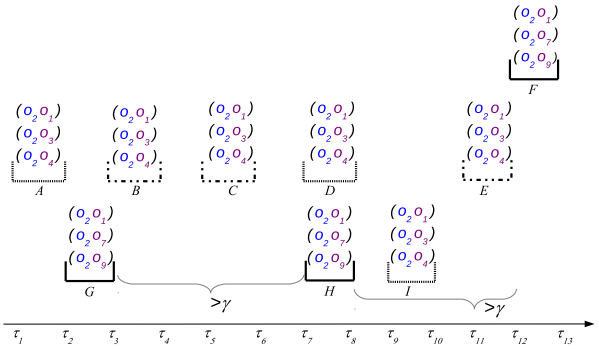


Figure 6. The crews are built by joining similarity-based pair groups under the constraints imposed by the maximum temporal gap γ and equality of the participants.

4. Experiments

The proposed framework has been implemented in Java and applied to trajectory data of real-world moving objects. In particular, we considered the data of two different kinds of moving objects: humans moving in the urban space and animals observed in the natural environment of a forest. The trajectories of these two types of objects have evident differences. Humans move in spaces constrained by physical restrictions due to buildings and road networks, they have different travel modes and may move together or travel/walk alone. The trajectories can refer to daily life routines (such as, going at home, going at work), paths which are periodically traveled (going at soccer stadium on weekends), or exceptional routes (alternative paths in case of emergency situations). On the contrary, wild animals move in a space that is characterized by the morphology of the territory, their movements are often related to changes of natural phenomena and changes of the ecosystem and, therefore, have an intrinsic seasonal component. Animal movements do not raise privacy concerns, which are instead frequent when analyzing human trajectories.

In the scenario of human mobility, there is a category of tasks in which we need to determine social ties between individuals. This type of information can be used, for instance, in social media, to suggest potential friendships on the basis of similar trajectories with other social profiles, or in transportation, to recommend trajectories that friends may know. For the experiments we used the Geolife dataset[1] (Zheng *et al.* 2009), which collects the trajectories of outdoor movements of humans in a period of over three years (from April 2007 to August 2012). These trajectories were recorded by different GPS loggers and GPS phones, and are sampled mostly every 1-5 seconds. The application of

[1] `https://www.microsoft.com/en-us/download/details.aspx?id=52367`

our framework to the Geolife dataset aims at i) discovering social ties between individuals who do not know each other and ii) following social ties when the individuals do not stay in the same location.

In the scenario of animal movements, often ecologists are interested in understanding how individuals or groups behave within their environment. The normal activities, such as grazing, migration and mating, are related to factors dependent of the environment (e.g., habitat preferences, water sources and vegetation availability) and to factors dependent of the relationships within or between the species. These activities could not take place if there are no interactions. For the experiments we used the Starkey dataset[1] (Rapp and Pacific Northwest Research Station 2006), which contains the radio-telemetry locations, tracked in north-eastern Oregon, of three species of animals (elks, deer, cattle) monitored over the period May 1993–August 1996. The sample rate of recording ranges from 20 minutes to 2 hours. In Starkey, our contribution is addressed to two categories of use cases. First, the understanding of the possible factors that may cause unnatural behavior exhibited both by individuals and by groups, for instance, elks move away because of hunting, herds fast graze because of human presence. Second, the characterization of several activities (e.g., walking, grazing) through different configurations of movement parameters.

The characteristics of the datasets are summarized in Table 3.

4.1. *Experimental setup*

The experiments have been performed along to several perspectives. We used several variants of the framework distinguished by user preference criterion (*maxDuration*, *maxObjects*). We tested the influence of the input thresholds (μ and γ) on the resulting crews and on the time consumption. To estimate the quality of the crews with an objective evaluation, we defined two quantitative measures by following the principle of some measures of internal clustering validation (e.g., Modified Hubert statistic (Hubert and Arabie 1985), Davies-Bouldin index (Davies and Bouldin 1979)). These measures give us an indication of how specific and distinct the movement behavior modeled by a crew is, so the higher the measure the better the quality. In particular, the first measure (Q_M) denotes the distance between two crews in terms of the spatio-temporal information associated with the movement parameters, regardless of the pair groups involved, while the second measure Q_P specializes Q_M by considering also the pair groups.

In the following, we report the formulation of the two measures:

$$Q_M = \frac{1}{|\Gamma'| \times |\Gamma''|} \sum_{\tau_i \in \Gamma'} \sum_{\tau_j \in \Gamma''} \frac{\Theta_T(\tau_i, \tau_j) + \Theta_S(\tau_i, \tau_j)}{2}, \tag{3}$$

where, Γ' and Γ'' are the timelines of the two crews, $|\Gamma'|$ and $|\Gamma''|$ refer to the number of the time instants of the two timelines respectively.

[1] `www.fs.fed.us/pnw/starkey/data/tables`

Table 3. Summary of the characteristics of the trajectory datasets.

	$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{T}\rvert$	avg fixes per object	types of objects
Geolife	182	17656921	166633	pedestrians
Starkey	128	166826	1304	wild animals

The term $\Theta_T(\tau_i, \tau_j)$ accounts for the dissimilarity of the timelines and it is computed as the distance between the time instants $\langle \tau_i, \tau_h \rangle$ and the time instants $\langle \tau_j, \tau_k \rangle$ as follows:

$$\Theta_T(\tau_i, \tau_j) = \begin{cases} 0 & \text{if } \langle \tau_i, \tau_h \rangle \cap \langle \tau_j, \tau_k \rangle \neq \oslash \\ \frac{max(\tau_h, \tau_k) - min(\tau_i, \tau_j)}{max(\mathcal{T}) - min(\mathcal{T})} & \text{otherwise.} \end{cases} \tag{4}$$

Intuitively, the closer the time instants, the smaller the dissimilarity, the worse the quality[1].

The term $\Theta_S(\tau_i, \tau_j)$ instead denotes the dissimilarity of the movement behaviors of the two crews, and more precisely, accounts for the distance between the ranges of values $[z_{l_{lower}}, z_{l_{upper}}]$ of the movement parameters. To compute this, we resort to dissimilarity functions for interval-valued data defined in Symbolic Data Analysis (Diday and Esposito 2003):

$$\Theta_S(\tau_i, \tau_j) = \frac{\sum_{F_l \in \mathcal{F}} \delta([z_{l_{lower}}, z_{l_{upper}}]|_{\tau_i}, [z_{l_{lower}}, z_{l_{upper}}]|_{\tau_j})}{|\mathcal{F}|}, \tag{5}$$

where $[z_{l_{lower}}, z_{l_{upper}}]|_{\tau_i}$ and $[z_{l_{lower}}, z_{l_{upper}}]|_{\tau_j}$ are the ranges of the parameter F_l in the time instants $\langle \tau_i, \tau_h \rangle$ and $\langle \tau_j, \tau_k \rangle$. The dissimilarity function $\delta(\cdot, \cdot)$ is in its turn the combination of three measures, $\delta_\pi(\cdot, \cdot)$, $\delta_s(\cdot, \cdot)$, $\delta_c(\cdot, \cdot)$. In particular, δ_π indicates the relative position of the two ranges in their entire interval of values, δ_s indicates the relative sizes of one range with respect to the other without considering the common sub-ranges (intersections) and δ_c is a measure of the non-common sub-ranges. Intuitively, when there is high similarity between the ranges of the parameters, the value of $\Theta_S(\tau_i, \tau_j)$ is small and consequently the quality is small.

The formulation of the second measure (Q_P) is reported in the following:

$$Q_P = \frac{\sum_{p \in \mathcal{G}'} \sum_{q \in \mathcal{G}''} \Theta_G(p, q)}{|\mathcal{G}'| \times |\mathcal{G}''|}, \tag{6}$$

with

$$\Theta_G(p, q) = \begin{cases} 1 \text{ if } p \neq q \\ Q_M \text{ if } p = q \end{cases}. \tag{7}$$

where, $\mathcal{G}'$ and $\mathcal{G}''$ are the pair groups of the two crews. Intuitively, the quality is higher when the pair groups and movements associated with a crew are well separated from those of other crews. Both quality measures, Q_M and Q_P range in [0,1].

4.2. *Comparative experiments*

We performed comparative experiments between the proposed framework and two algorithms, namely *swarm* and *DISCRETIZATION*. The first competitor is the algorithm

[1]The operators max(,) and min(,) work on the time domain $\mathcal{T}$ as on the set of the integers.

proposed in Li *et al.* (2010) and discovers dense groups of objects (swarms) characterized by temporal discontinuity, which is also a characteristic of the crews. The second competitor is a baseline and differs from the proposed framework in the algorithm of construction of the tree. In particular, the choice of the movement parameter to associate with an internal node is not based on the dissimilarity reduction of the vectors of the previously created node, but it is determined a-priori on the basis of the standard deviation of the movement parameters computed on the complete set of vectors (Loglisci *et al.* 2014). More precisely, the movement parameters are arranged from the root downward to the leaves by decreasing standard deviation. The respective nodes are connected by means of n-way branches. A branch is associated with one of n bins obtained by applying an equal-width discretization technique to the movement parameter F_l, associated with the node. We performed several trials and identified the best tradeoff between the size of the internal nodes and minimum number of the objects in the leaves with a discretization to five bins, that is, with 5-way branches. The trees of DISCRETIZATION are balanced (contrarily to those built by our framework) and have a depth equal to the number of the movement parameters plus one (leaves), that is, $|\mathcal{F}|+1$.

The reason behind the use of DISCRETIZATION is to compare our framework with a solution which adopts the same feature space ($\mathcal{F}$), but implements a different notion of similarity of the movement parameters. This affects the construction of the crews and thus the quality.

4.3. *Results and Discussion*

We now present and discuss the results on the number of valid crews (*#crews*), running times (*times*) and quality measures that were obtained by manually tuning one of the two input thresholds (μ and γ) at a time and leaving the other one fixed. For both datasets, we fix the value of *required_reduction* to 0.2 (Section 3.4) and the threshold used for the filtering technique (Section 3.1) to three times the standard deviation of the distances between consecutive fixes, as suggested by Dodge *et al.* (2009). The width of the time interval (Section 3.1) is 30 minutes for Geolife and 45 minutes for Starkey.

We organize the discussion in four perspectives: plots of the statistics, visualization of the distribution of the crews over space, interpretation of some crews of the two datasets and contribution of specific movement parameters in discovering crews.

4.3.1. Influence of the input thresholds

Figure 7 illustrates the number of crews discovered at different values of μ and γ. More precisely, the results are averages computed on the numbers of crews generated by running the framework with *maxObjects* and *maxDuration*. In Figure 7a, we see that when the minimum number of the objects required to form the groups is higher, the number of crews is relatively smaller. This is expected because at high values of μ we should seek a higher number of participants with the same movements, which is more difficult to obtain compared to few participants (e.g., μ=3). Contrary to μ, the threshold γ weakly influences *#crews* (Figure 7b) because it has no immediate impact on the number of SPGs, but it has influence on the length of the timeline. Indeed, when γ is large, we have few crews but with longer duration.

Figure 8 illustrates the time consumption required at different values of μ and γ respectively. More precisely, the results are averages computed on the running times obtained by running the framework with *maxObjects* and *maxDuration*. We see that high values of μ lead to a lower time consumption. The reason behind this is two-fold: *i)* the algorithm

of tree induction requires less running time (it terminates with larger leaves), and *ii)* the algorithm for the discovery of the crews works on a smaller set of SPGs. The low number of SPGs, generated when increasing the threshold γ, explains the decreasing tendency of Figure 8b.

Figure 9 illustrates the quality of the crews at different values of μ and γ. The results are averages computed on the values (ranging in [0,1]) obtained by running the framework with *maxObjects* and *maxDuration*. We see that the quality generally decreases at high values of μ. This can be attributed to the increase of the number of objects that SPGs have in common, which causes the decrease of the dissimilarity of the movement parameters of the crews. Trivially, the highest quality is reached when there is only one crew (Figure 9a, Starkey). Another observation can be drawn on the effect of the threshold μ on the number of crews and on the quality. We see that the greater the set of discovered crews, the higher the quality. Although this may seem contradictory, we should note that at low values of μ the search space is greater, therefore we can nimbly find out crews with very different movements (high quality), while, at high values of μ, the (few) resulting crews appear to be located in a limited subspace of search, where the movements may be very similar (low quality).

On the contrary, when increasing the threshold γ (Figure 9b), the quality remains sub-

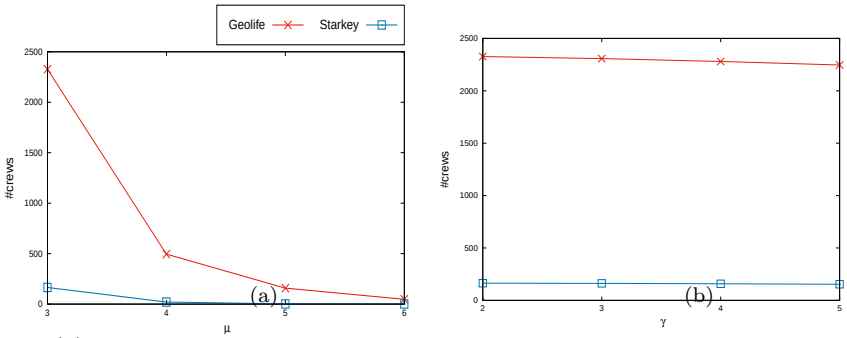


Figure 7. (a) Number of discovered crews at different values of the minimum required number of objects μ, while γ is fixed to 2. (b) Number of discovered crews at different values of the maximum temporal gap γ, while μ is fixed to 3. The results are averages computed on the values obtained by running the framework with *maxObjects* and *maxDuration*.

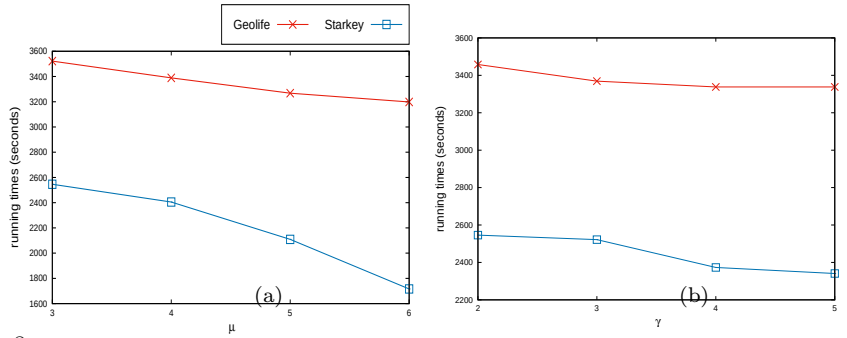


Figure 8. (a) Time consumption at different values of the minimum required number of objects μ, while γ is fixed to 2. (b) Time consumption at different values of the maximum temporal gap γ, while μ is fixed to 3. The results are averages computed on the values by running the framework with *maxObjects* and *maxDuration*

stantially unaltered because the smaller number of the crews counterbalances the higher duration, in the sense that, at high values of γ, there is a small set of crews characterized by relatively high duration. Consequently, the dissimilarity of the movements of the SPGs remains unchanged, also because γ has no immediate impact on the SPGs.

Figure 10 gives a visual and geographic perspective of the crews at two higher values of the thresholds used in the experiments, that is, μ=5,6 and γ=4,5. The map in Figure 10a reports the localization of the areas of Geolife that contain at least 10 crews. The map in Figure 10b reports the localization of the areas of Starkey with at least 5 crews. As to Geolife, there are three main areas for μ (red circles) and five main areas for γ (blue polygons). For Starkey, the distribution of the crews, when we operate on μ, is concentrated in three areas, while there are four areas of crews when tuning γ. As we see, the crews discovered at different values of μ are concentrated in a subset of the areas of the crews discovered at different values of γ. This indicates that the crews with more objects are concentrated in the same areas, while the crews with less objects preserve the similarity for a longer.

4.3.2. Comparisons

We now discuss the results of the number of discovered groups and quality values obtained from the comparison with DISCRETIZATION and *swarm*. For the competitor DISCRETIZATION, as in our framework, there are two variants defined on the criterion of preference *maxDuration* and *maxObjects*. Thus, the results we show here are the averages obtained from the two settings. The experiments of swarm were performed by defining the time-domain with two alternative widths of the time intervals, 60 minutes (swarm 60) and 150 minutes (swarm 150).

Figure 11 reports the counts of the discovered groups. As expected, the increase of the minimum required number of objects (μ) is the strong determinant in the reduction of the number of groups discovered by the three algorithms. Contrarily, the value of γ influences the results of our framework and DISCRETIZATION, but it has no effect on swarm. This is why swarm has no threshold corresponding to γ to filter the groups with respect to the temporal discontinuities.

Figure 12 shows that the quality of the crews determined by our framework is better in general, compared to that of DISCRETIZATION. This is due to the different strategy of tree induction. Indeed, in our framework, the creation of the nodes follows the principle of maximization of the similarity of the properties of the current set of objects, which

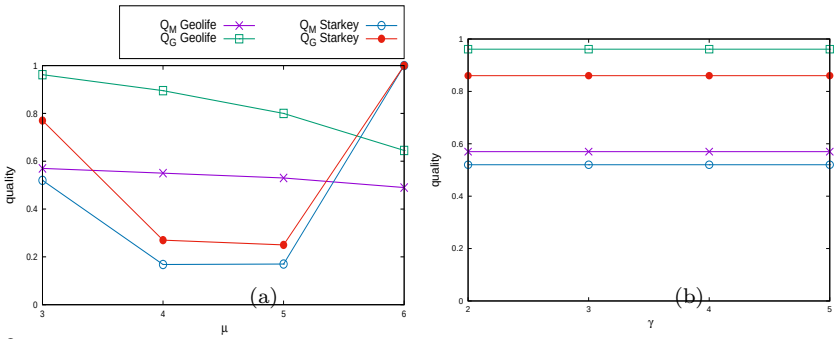


Figure 9. (a) Quality measures Q_M and Q_G computed on the crews discovered at different values of the minimum required number of objects μ, while γ is fixed to 2. (b) Quality measures Q_M and Q_G computed on the crews discovered at different values of the maximum temporal gap γ, while μ is fixed to 3. The results are averages computed on the values by running the framework with *maxObjects* and *maxDuration*

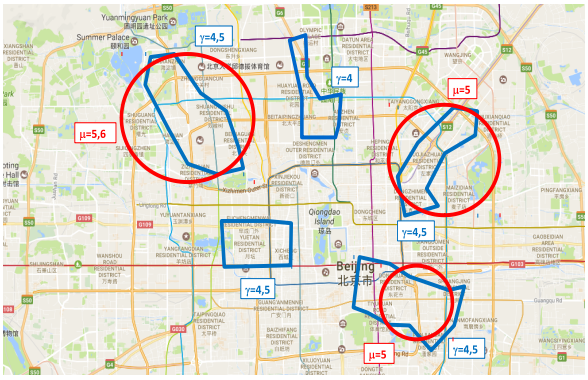
leads to assign smaller sets of vectors to the nodes. This generally makes the "intra-crews" similarity higher and the "inter-crews" similarity lower. Contrarily, the algorithm DISCRETIZATION follows the principle of minimization of the dissimilarity over the complete set of objects and thus does not account for the local similarities, with the result of making the aim of maximization of the similarity arduous.

As to swarm, we cannot use neither Q_M nor Q_G to estimate the quality because they are based on the similarity of the movement parameters and similarity of the timelines, while swarms rely on the spatial proximity, do not use similarity notions and do not consider movement parameters to describe the groups.

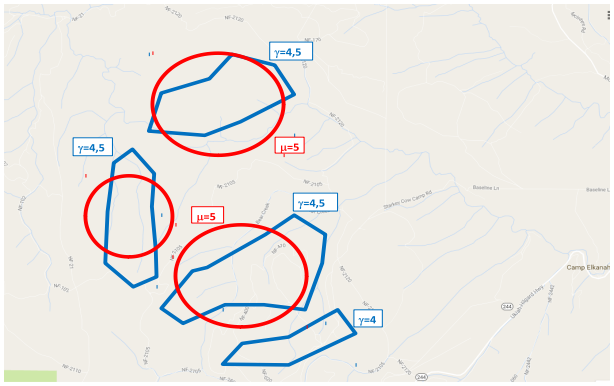
4.3.3. Interpretation of the crews

In this section we report the crews with larger dissimilarity, which is computed as the average of the values of the terms Θ_T and Θ_G between the crews discovered from the two datasets, respectively. We also explain the information the crews depict. For each crew, the movement parameters and timeline are reported.

The following has been discovered from Geolife with μ=3, γ=2, width of time interval =60 mins, *maxDuration*. It involves three objects (two pairs).



(a)



(b)

Figure 10. a) Distribution of the crews of Geolife over the Bejing area (China) b) Distribution of the crews of Starkey over the Starkey Forest area (Oregon-U.S.). The red circles denote the distribution of the crews obtained when tuning threshold μ. The blue polygons denote the distribution of the crews obtained when tuning threshold γ.

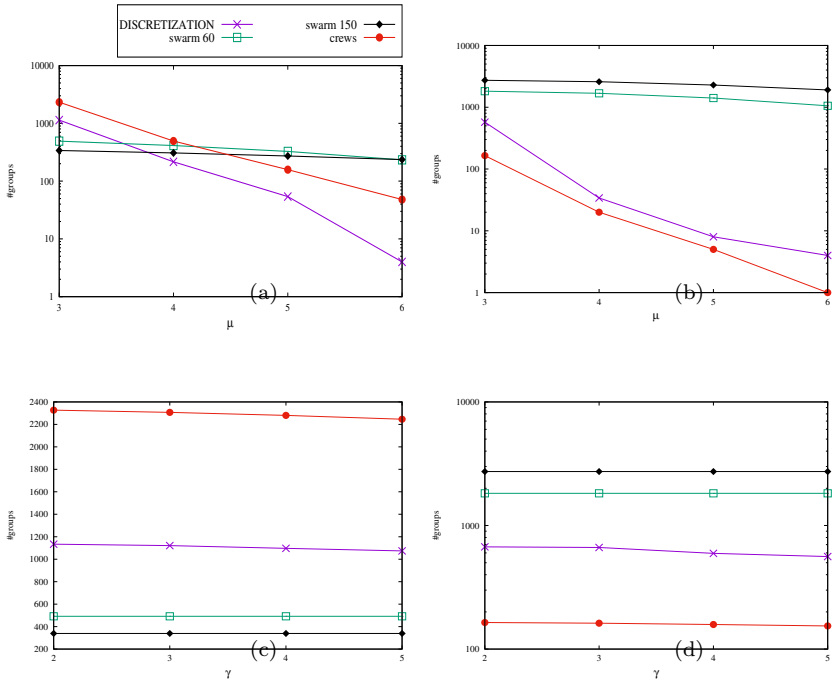


Figure 11. Comparisons between the proposed framework and the competitors DISCRETIZATION and swarm on the number of groups when tuning μ and γ (a) and c) Geolife with μ and γ respectively, b) and d) Starkey with μ and γ respectively).

$\mathcal{C}_1$:
$2009-07-05\_05:00:00, 2009-07-05\_06:00:00$
$\langle azimuth\_distance \in [0°, 47°), azimuth \in [187.5°, 359°], displacement \in [129, +\infty), distance\_ratio \in [0, 0.5)\rangle$

$2009-07-05\_06:00:00, 2009-07-05\_07:00:00$
$\langle azimuth\_distance \in [0, 47°), azimuth \in [187.5°, 359°], distance\_ratio \in [0, 1], displacement \in [0, 85), tortuosity\_ratio \in [0.5, 0.9), azimuth\_distance \in [0, 13°), speed\_ratio \in [1.2, 1.6)\rangle$

$2009-07-05\_07:00:00, 2009-07-05\_08:00:00$
$\langle azimuth\_distance \in [0, 47°), azimuth \in [0, 187.5°), speed\_ratio \in [0.5, 1.9), distance\_ratio \in [1, +\infty)\rangle$

The crew $\mathcal{C}_1$ covers three hours and for each hour (τ_i, τ_{i+1}) the movement parameters change. For instance, for the time instants $\langle 2009-07-05\_05:00:00, 2009-07-05\_06:00:00\rangle$, the two pairs of objects come close ($distance\_ratio \in [0, 0.5)$), their angular distance is less than 47 ($azimuth\_distance \in [0, 47°)$), they travel at least 129 meters ($displacement \in [129, +\infty)$) and they move westwards ($azimuth \in [187.5°, 359°]$).

The following crew has been discovered from Geolife at μ=5, γ=2, width of time interval=60 mins, *maxObjects*. It involves five objects (four pairs).

$\mathcal{C}_2$:
$2009-04-13\_13:00:00, 2009-04-13\_14:00:00$
$\langle azimuth\_distance \in [0°, 80°), azimuth \in [0°, 190°], displacement \in [129, +\infty), distance\_ratio \in [0, 0.5), tortuosity\_ratio \in [0.5, 0.9), speed\_ratio \in [1.1, 2.5]\rangle$

$2009-04-13\_14:00:00, 2009-04-13\_15:00:00$
$\langle azimuth\_distance \in [0, 80°), azimuth \in (190°, 359°], distance\_ratio \in [1, 1.5), displacement \in [0, 134), speed\_ratio \in [0.5, 1.1), tortuosity\_ratio \in [0.5, 0.9)\rangle$

Here, the pairs maintain their angular distance under 80° ($azimuth\_distance \in [0, 80°)$,

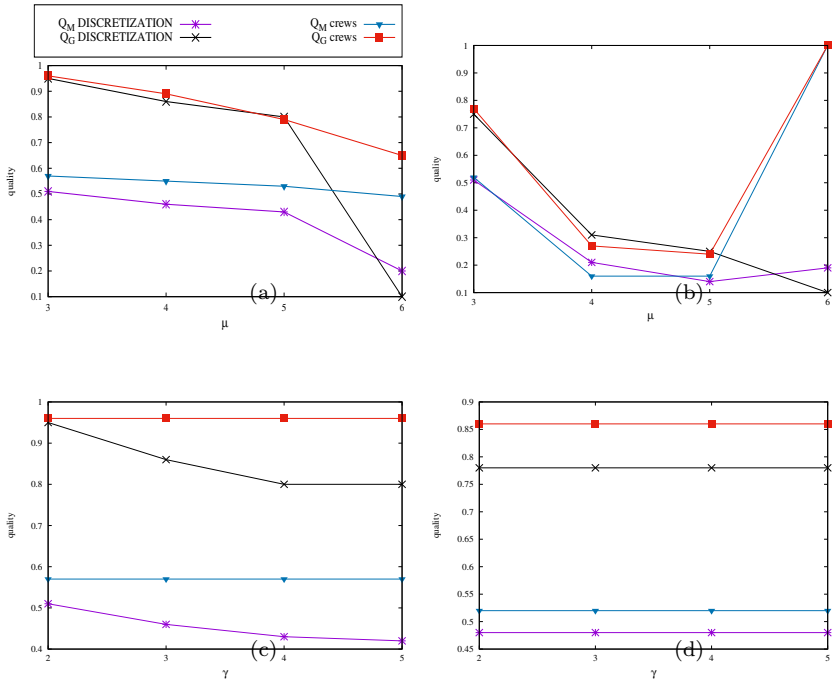


Figure 12. Comparisons between our framework and the competitor DISCRETIZATION on the quality measures when tuning μ and γ (a) and c) Geolife with μ and γ respectively, b) and d) Starkey with μ and γ respectively).

they come close ($distance\_ratio \in [0, 0.5)$), they move eastwards ($azimuth \in [0°, 190°]$), they travel more than 190 meters, with increasing speed, on paths with almost identical tortuosity. The movement is different in the second part of the timeline. They move westwards ($azimuth \in (190°, 359°]$), move away ($distance\_ratio \in [1, 1.5)$), they travel less than 134 meters, at similar speeds ($speed\_ratio \in [0.5, 1.1)$) and with similar tortuosity of the paths ($tortuosity\_ratio \in [0.5, 0.9)$).

The crews we report below have been discovered on Starkey. The following crew has been discovered at μ=3, γ=2, width of time interval =120 mins, *maxInterval*. It involves two pairs of elks.

$\mathcal{C}_3$:
$1993-06-16\_04:00:00, 1993-06-16\_06:00:00$
$\langle azimuth\_distance \in [180°, 359°], azimuth \in [177°, 359°], distance\_ratio \in [0.5, 1), displacement \in [3.5, +\infty)\rangle$

$1993-06-16\_06:00:00, 1993-06-16\_07:00:00$
$\langle azimuth\_distance \in [0°, 180°), azimuth \in [0°, 177°), distance\_ratio \in [0, 0.5), displacement \in [6.1, +\infty)\rangle$

The crew $\mathcal{C}_3$ has a timeline of four hours. In the first part, the pairs are separated by a relatively large distance ($azimuth\_distance \in [180°, 359°]$), but they tend to come close ($distance\_ratio \in [0.5, 1)$), proceed north-westwards ($azimuth \in [177°, 359°]$) and travel at least 3.5 km ($displacement \in [3.5, +\infty)$). In the second part, we see that they are separated by an angular distance of at most 180° ($azimuth\_distance \in [0°, 180°)$), their direction remains unaltered ($azimuth \in [177°, 359°]$), they still come close ($distance\_ratio \in [0, 0.5)$) and travel at least 6.1 km ($displacement \in [6.1, +\infty)$).

The following crew has been discovered on Starkey at μ=3, γ=4, width of time inter-

val= 60 mins, *maxInterval*. It involves four pairs of deer.

$\mathcal{C}_4$:
$1993-05-10\_01:00:00, 1993-05-10\_02:00:00$
$\langle azimuth\_distance \in [180°, 359°]$, $azimuth \in [141°, 359°]$, $displacement \in (-\infty, 3.7)$, $speed\_ratio \in [1.1, 2.5]$, $distance\_ratio \in [0, 0.8)\rangle$

$1993-05-10\_05:00:00, 1993-05-10\_06:00:00$
$\langle azimuth\_distance \in [180°, 359°]$, $azimuth \in [141°, 359°]$, $displacement \in [3.7, +\infty)$, $distance\_ratio \in (0, 0.2)$, $speed\_ratio \in [0.5, 1.4]\rangle$

The crew $\mathcal{C}_4$ covers two hours separated by a temporal gap of three time instants (three hours). We see the pairs travel less than 3.7km in the first part, while in the second part they do more than 3.7km. Their inter-distance decreases (from $distance\_ratio \in [0, 0.8)$ to $distance\_ratio \in [0, 0.2)$). The speeds tend to be similar (from $speed\_ratio \in [1.1, 2.5]$ to $speed\_ratio \in [0.5, 1.4]$).

4.3.4. Insights from crews

We examine the crews visualized in the maps of Figure 10 to draw some qualitative consideration. In particular, our analysis focuses on the identification of the movement parameters that better contribute to the discovery of high quality crews, where the desiderata is large similarity among the pairs of the same crew and little similarity among pairs of different crews. Since the number of ranges produced for a parameter is an indication of the dissimilarity of the values (Section 3.4), the movement parameters relevant for the determination of high quality crews are those with a few ranges. The results of our analysis are illustrated in Table 4, which specifically reports the parameters with the smallest and largest number of ranges used in the crews. We see that the parameters Distance Ratio and Speed Ratio are determinant for the crews discovered from Geolife. This means that the movements have large similarity in terms of inter-distance and relative speed, which is a behavior typical of individuals who know each other. On the contrary, the parameters determinant for the crews discovered from Starkey are generally of the category Pairwise Dynamics. In particular, we see large similarity on Displacement and Azimuth for elks and cattle respectively. This indicates that the elks travel by regular displacements and cattle move at the same direction, which are typical features of migrating animals. We can also note that, for all the three types of animals, the less relevant parameters are of the category Interaction-wise Dynamics and, specifically, are Speed Ratio and Tortuosity Ratio. This denotes a large irregularity of Speed Ratio and Tortuosity Ratio, which is due probably to the presence of threats (e.g., hunting), in the case of cattle, and to the presence of paths within the forest, in the case of elk and deer. The final consideration is deserved to the input thresholds μ and γ. In Starkey, the movement parameters with large similarity remain the same regardless of the threshold we use.

5. Conclusions

In this paper, we have investigated the problem of mining groups of moving objects by accounting for interactions between the objects and dynamics of the movements. Interactions are extracted by capturing relationships between the trajectories. Dynamics is extracted by capturing the time-changing nature of the trajectories. These are two new sources of information that a few works have exploited, although with the limitations that we discussed. Most part of the existing algorithms instead focuses on spatial proximity

Table 4. The movement parameters used in the discovery of crews. The parameters which appear in the crews with the smallest number of ranges (more relevant parameters) are denoted with "+", while those which appear with the highest number of ranges (less relevant parameters) are denoted as "-".

		μ	γ
Geolife	+	Distance Ratio	Speed Ratio
	-	Displacement	Azimuth Distance
Starkey (elks)	+	Displacement	Displacement
	-	Speed Ratio	Tortuosity Ratio
Starkey (deer)	+	Azimuth Distance	Azimuth Distance
	-	Tortuosity Ratio	Tortuosity Ratio
Starkey (cattle)	+	Azimuth	Azimuth
	-	Speed Ratio	Speed Ratio

and path similarity, which lead to mining groups of objects that move together or stay close from each other. The current work opens to a new type of collective movement, which is based on similar interactions and similar dynamics and which does not require spatial constraints on the objects.

The proposed framework addresses some critical points of the problem at the hand. First, we have defined new movement parameters in order extract interactions and dynamics from raw trajectory. The movement parameters model pair-wise interactions and changes of physical properties of motion. Second, we have designed an efficient clustering algorithm, which does not rely on costly measures of distance and avoids re-scanning all the data. The clustering algorithm gathers pairs of objects based on similar interactions and similar dynamics.

To provide arguments on the applicability of the framework to a potentially large class of trajectories, we have performed experiments on different types of moving objects. The experiments have been organized in order to *i)* test the influence of the input thresholds on the discovery process and on the quality of the crews, *ii)* perform a quantitative comparison with alternative solutions, *iii)* argue the usefulness of the information conveyed by the crews in real case studies, and *iv)* distinguish the contribute given by each movement parameter in the construction of high quality crews. Some considerations can be drawn. First, the quality of the crews strongly depends on the trajectory data preparation. Indeed, if we use a coarse sampling to define the time instants, then the movement parameters poorly describe interactions and dynamics. This result is not surprising. Second, the experiments on the temporal maximum gap highlight a characteristic of the framework, that is, the property to discover crews that may be developed over time without temporal continuity. This is demonstrated by the quality of the crews. Third, the meaningfulness of the crews and number of the relative members are always not related to each other, so groups of interest are not necessarily those more numerous. This is evident for both data sets.

The emerging Big Data technologies will stimulate us to upgrade the proposed framework in order to deal with Big mobility data problems. Recent studies focus on solutions of parallel computation for the analysis of large *volumes* (Altomare *et al.* 2017), while we plan to investigate the other two challenges of Big data, that is, *velocity* and *variety*. A future direction for the velocity will be re-designing the proposed framework in order to mine trajectory data streams. Techniques of data stream mining based on time-windows could be considered for this task (Loglisci and Malerba 2014). The final purpose will be that of providing a prompt an real-time response on the formation of crews. A future direction for the variety will be exploiting unstructured data (for instance, geographic documents (Loglisci *et al.* 2012a,b)) and crowd-sourced information (e.g., OpenStreetMap, GeoNames) in order to extend the feature space and movement

parameters. This can be done with techniques of geo-tagging able to annotate punctual trajectory data with contextual information. The final purpose will be that of enriching the information expressed by discovered crews. Several works have already investigated, for instance, the integration of geo-tagging with social media data (Comito *et al.* 2016).

Acknowledgments

This work fulfills the research objectives of the FutureInResearch 8GPS5R0 project "Computer-mediated collaboration in creative projects" funded by Apulia Regional Government for Intervento cofinanziato dal Fondo di Sviluppo e Coesione 2007-2013 APQ Ricerca Regione Puglia "Programma regionale a sostegno della specializzazione intelligente e della sostenibilita' sociale ed ambientale - FutureInResearch".

References

Altomare, A., *et al.*, 2017. Trajectory Pattern Mining for Urban Computing in the Cloud. *IEEE Trans. Parallel Distrib. Syst.*, 28 (2), 586–599.

Andersson, M., *et al.*, 2008. Reporting Leaders and Followers among Trajectories of Moving Point Objects. *GeoInformatica*, 12 (4), 497–528.

Andrienko, N.V., *et al.*, 2013. Space Transformation for Understanding Group Movement. *IEEE Trans. Vis. Comput. Graph.*, 19 (12), 2169–2178.

Benkert, M., *et al.*, 2006. Reporting Flock Patterns. *In*: *Algorithms - ESA 2006, 14th Annual European Symposium, Zurich, Switzerland, September 11-13, 2006, Proceedings*, 660–671.

Comito, C., Falcone, D., and Talia, D., 2016. Mining human mobility patterns from social geo-tagged data. *Pervasive and Mobile Computing*, 33, 91–107.

Davies, D.L. and Bouldin, D.W., 1979. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1 (2), 224–227.

Diday, E. and Esposito, F., 2003. An introduction to symbolic data analysis and the SODAS software. *Intell. Data Anal.*, 7 (6), 583–601.

Dodge, S., *et al.*, 2016. Analysis of movement data. *International Journal of Geographical Information Science*, 30 (5), 825–834.

Dodge, S., Weibel, R., and Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33 (6), 419–434.

Dodge, S., Weibel, R., and Lautenschütz, A., 2008. Towards a taxonomy of movement patterns. *Information Visualization*, 7 (3-4), 240–252.

Doncaster, C.P., 1990. Non-parametric estimates of interaction from radio-tracking data. *Journal of Theoretical Biology*, 143 (4), 431 – 443.

Dutton, G., 1999. Scale, Sinuosity, and Point Selection in Digital Line Generalization. *Cartography and Geographic Information Science*, 26 (1), 33–54.

Frank, E., *et al.*, 1998. Using Model Trees for Classification. *Machine Learning*, 32 (1), 63–76.

Gudmundsson, J. and van Kreveld, M.J., 2006. Computing longest duration flocks in trajectory data. *In*: *14th ACM International Symposium on Geographic Information Systems, ACM-GIS 2006, November 10-11, 2006, Arlington, Virginia, USA, Proceedings*, 35–42.

Hubert, L. and Arabie, P., 1985. Comparing partitions. *Journal of Classification*, 2 (1), 193–218.

Jeung, H., *et al.*, 2008. Discovery of convoys in trajectory databases. *PVLDB*, 1 (1), 1068–1080.

Konzack, M., *et al.*, 2017. Visual analytics of delays and interaction in movement data. *International Journal of Geographical Information Science*, 31 (2), 320–345.

Laube, P., *et al.*, 2007. Movement beyond the snapshot - Dynamic analysis of geospatial lifelines. *Computers, Environment and Urban Systems*, 31 (5), 481–501.

Laube, P., 2014. *Computational Movement Analysis*. Springer Briefs in Computer Science Springer.

Laube, P., Imfeld, S., and Weibel, R., 2005. Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19 (6), 639–668.

Layer, R.M., *et al.*, 2013. Binary Interval Search: a scalable algorithm for counting interval intersections.. *Bioinformatics*, 29 (1), 1–7.

Li, Z., *et al.*, 2010. Swarm: Mining Relaxed Temporal Moving Object Clusters. *PVLDB*, 3 (1), 723–734.

Loglisci, C., *et al.*, 2012a. Toward Geographic Information Harvesting: Extraction of Spatial Relational Facts from Web Documents. *In*: J. Vreeken, C. Ling, M.J. Zaki, A. Siebes, J.X. Yu, B. Goethals, G.I. Webb and X. Wu, eds. *12th IEEE International Conference on Data Mining Workshops, ICDM Workshops, Brussels, Belgium, December 10, 2012* IEEE Computer Society, 789–796.

Loglisci, C., *et al.*, 2012b. An Unsupervised Framework for Topological Relations Extraction from Geographic Documents. *In*: S.W. Liddle, K. Schewe, A.M. Tjoa and X. Zhou, eds. *Database and Expert Systems Applications - 23rd International Conference, DEXA 2012, Vienna, Austria, September 3-6, 2012. Proceedings, Part II*, Vol. 7447 of *Lecture Notes in Computer Science* Springer, 48–55.

Loglisci, C. and Malerba, D., 2014. Mining Dense Regions from Vehicular Mobility in Streaming Setting. *In*: *Foundations of Intelligent Systems - 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings*, 40–49.

Loglisci, C. and Malerba, D., 2017. Leveraging temporal autocorrelation of historical data for improving accuracy in network regression. *Statistical Analysis and Data Mining*, 10 (1), 40–53.

Loglisci, C., Malerba, D., and Papadopoulos, A.N., 2014. Mining Trajectory Data for Discovering Communities of Moving Objects. *In*: K.S. Candan, S. Amer-Yahia, N. Schweikardt, V. Christophides and V. Leroy, eds. *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference (EDBT/ICDT 2014), Athens, Greece, March 28, 2014.*, Vol. 1133 CEUR-WS.org, 301–308.

Long, J.A. and Nelson, T.A., 2013a. Measuring Dynamic Interaction in Movement Data. *Trans. GIS*, 17 (1), 62–77.

Long, J.A. and Nelson, T.A., 2013b. A review of quantitative methods for movement data. *International Journal of Geographical Information Science*, 27 (2), 292–318.

Mazimpaka, J.D. and Timpf, S., 2016. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 2016 (13), 61–99.

Noyon, V., Claramunt, C., and Devogele, T., 2007. A Relative Representation of Trajectories in Geogaphical Spaces. *GeoInformatica*, 11 (4), 479–496.

Ong, R., *et al.*, 2011. Traffic Jams Detection Using Flock Mining. *In*: *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III*, 650–653.

Rapp, V. and Pacific Northwest Research Station, P.O., 2006. *Elk, Deer, and Cattle: The Starkey Project*. Science update U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Wood, Z., 2013. Profiling Spatial Collectives. *In*: *Research and Development in Intelligent Systems XXX, Incorporating Applications and Innovations in Intelligent Systems XXI Proceedings of AI-2013, Cambridge, England, UK, December 10-12, 2013*, 95–108.

Wood, Z. and Galton, A., 2009. A taxonomy of collective phenomena. *Applied Ontology*, 4 (3-4), 267–292.

Zheng, K., *et al.*, 2014. Online Discovery of Gathering Patterns over Trajectories. *IEEE Trans. Knowl. Data Eng.*, 26 (8), 1974–1988.

Zheng, Y., *et al.*, 2009. Mining interesting locations and travel sequences from GPS trajectories. *In*: *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, 791–800.

Appendix

A Tree induction algorithm for clustering pairs of moving objects

Here we explain the algorithm for clustering pairs of moving objects through an example (Figure 13). Consider μ=2 and *required_reduction*=0.2. The newly created nodes are stored in a queue structure and will be processed one by one. At the beginning, the algorithm stores the root in the queue (Figure 13- root.vectors), then it proceeds with the generation of the splitting values for each movement parameter and, for each splitting value, it creates two subsets from the set of vectors of the node. The splitting values are taken from the values existing in the node. Subsequently, the algorithm computes the reduction of the dissimilarity between the vectors of the node and the vectors of the two subsets. By supposing that the parameter that guarantees greater dissimilarity reduction is "Azimuth Distance" and the threshold value is 22 (Figure 13-1), the algorithm inserts two branches and two new nodes into the tree. The node on the left contains the vectors, whose values of "Azimuth Distance" are lower than 22, while the node on the right takes the values greater than or equal to 22 (Figure 13-2). By assuming that the dissimilarity of "Azimuth Distance" has been reduced by a factor of 0.1, it is removed from the list of the parameters. The new nodes are stored in the queue. Hence, the algorithm examines the node on the left and, by assuming that the next parameter with maximum dissimilarity reduction is "Tortuosity Ratio" and the threshold value is 1.2, it expands the tree with two new nodes (Figure 13-3), which will not be stored because the number of the contained pairs does not exceed μ^2. The algorithm continues with the node on the right of "Azimuth Distance". The maximum dissimilarity reduction is obtained with the parameter "Displacement" and with the threshold 150 (Figure 13-4). Finally, the new nodes are added to the tree, but are not stored because the conditions of termination are satisfied.

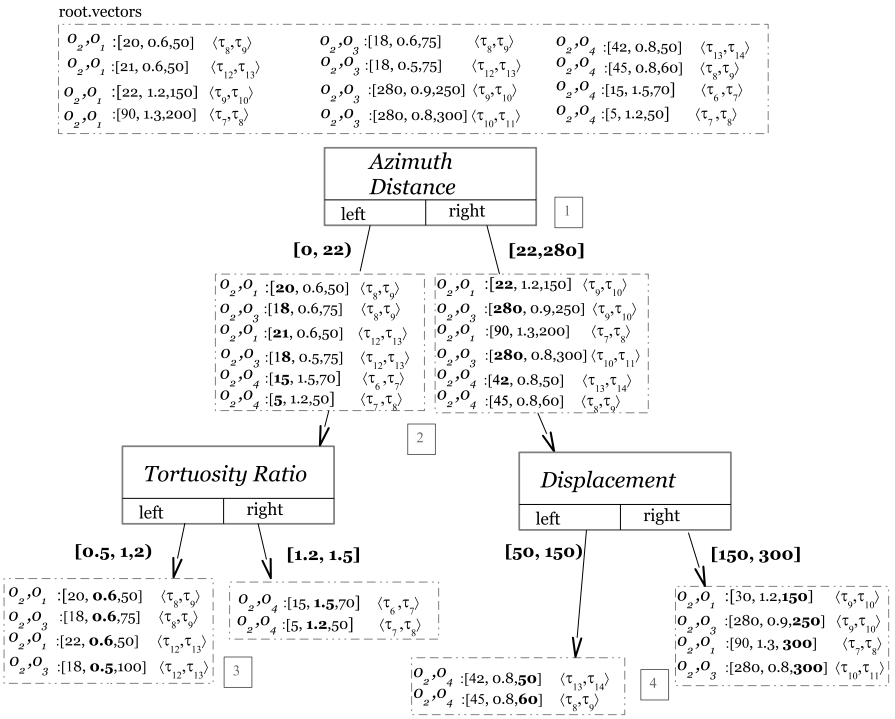


Figure 13. Illustration of the tree induction algorithm.

Time complexity

We conducted the complexity analysis of the three main procedures of the framework, that is, *i)* transformation of the raw trajectory data into vectors, *ii)* clustering pairs of moving objects, *iii)* building crews from similarity-based pair groups. The notation of the input data and thresholds is reported in Table 2. *i)* Let n be the number of objects $\mathcal{O}$ and t be the length of the time-domain $\mathcal{T}$. The vectors are generated for all the pairs of objects n^2 at each sequence of consecutive time-instants $\langle \tau_i, \tau_{i+1} \rangle$, where a number of computations equal to $|\mathcal{F}|$ (movement parameters) is performed for each pair of objects. Thus, the number of operations is $n^2|\mathcal{F}|$ at each sequence, while the time cost is $\mathbf{O}(tn^2)$ for all the non-overlapping sequences of time-instants, considering that $|\mathcal{F}| \ll n, t$. *ii)* The order of the input data of the algorithm of tree induction is tn^2. By assuming that the tree remains "bushy", the depth is $\mathbf{O}(\log\,(tn^2))$. At each node, not all the pairs are considered, but all the pairs are considered at the different depths of the tree, so the amount of work is $\mathbf{O}(tn^2 \log\,(tn^2))$. Considering that at each node all the parameters are evaluated, the time cost is $\mathbf{O}(|\mathcal{F}|tn^2 \log\,(tn^2))$. Also here, $|\mathcal{F}| \ll n, t$. *iii)* Let μ be the minimum required number of objects. The cost of the third procedure is due to the operations of generation and joining of the SPGs. More precisely, the counting of the intersections has time complexity equal to $\mathbf{O}(t\,\log t)$ (Layer *et al.* (2013)) and it is performed for all the leaves, which are $\frac{tn^2}{\mu^2}$ in the worst case, therefore $\mathbf{O}(\frac{t^2n^2}{\mu^2}\,\log t)$. The number of the operations of join is related to the number of pairs μ^2 at each leaf and to the total number of leaves $\frac{tn^2}{\mu^2}$. Therefore, the amount of work is $\mathbf{O}(tn^2 + \frac{t^2n^2}{\mu^2}\,\log t)$.
The total time cost of the framework is:

$$\underbrace{tn^2}_{raw\ data\ transformation} + \underbrace{tn^2 log(tn^2)}_{pair\ clustering} + \underbrace{tn^2 + \frac{t^2n^2}{\mu^2} logt}_{building\ crews}$$

that is, $\mathbf{O}(tn^2(2 + log(tn^2)) + \frac{t^2n^2}{\mu^2} logt)$.