Understanding Collective Behavior with the Concept of Core Clusters

Daniel Seebacher
University of Konstanz
Germany
Daniel.Seebacher@uni.kn

Daniel A. Keim
University of Konstanz
Germany
Daniel.Keim@uni.kn

Manuel Stein
University of Konstanz
Germany
Manuel.Stein@uni.kn

***Abstract*—Experts in collective behavior have been investigating how animals move together in groups for decades. Technical progress in recent years introducing smaller and more affordable GPS sensors has made it possible to study wider ranges of animals more closely than ever before. However, this results in more and more data, establishing the need for novel visual analytics solutions. State-of-the-art methods allow the recognition of clusters or often visited places of tracked animals but lack ways to recognize which animals are often together in the same groups. To close this gap, we propose an algorithm for the identification of so called core clusters as well as show how core clusters can be enriched with contextual information. Our proposed solution advances the understanding of experts analyzing collective behavior, especially, when inspecting animals with social relationships such as baboons. We demonstrate the usefulness of our approach as well as discuss our findings in a use case with real baboon data.**

***Index Terms*—Collective Behavior, Visual Analytics**

I. Introduction

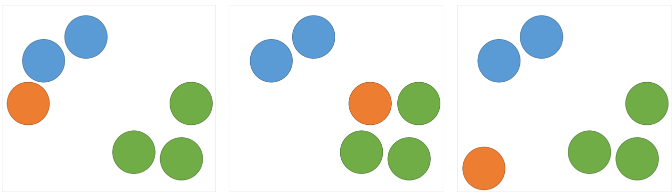
Collective behavior refers to relatively spontaneous and unstructured behavior by large numbers of individuals acting with or being influenced by other individuals [1]. The best known examples of collective behavior are flocks of birds or schools of fish, which protect themselves from predators or improve their search for food through this behavior. Collective behavior can also be observed in other animal groups or even in human everyday life. A question that has not yet been fully answered, however, is how these groups are formed. Baboons, for example, decide whether to join a group based on the size of existing groups or their own and other animals' direction of movement. Social factors can also influence this decision [2]. Still, many questions remain unanswered, such as whether there are groups that remain constant over time, so-called *core clusters*, or whether differences in the group composition depend on the current activity.

To answer such questions, analysts use data collected with the help of GPS trackers delivering motion data in an extremely high spatial and temporal resolution. This data enables a very precise analysis of animal movement behavior and can be used for event or activity detection. Existing approaches to analyze the movement of individuals usually focus on the analysis of individual trajectories or clusters of trajectories. An overview of such existing works is given by Andrienko et al. [3]. An example that can be emphasized is the REMO concept [4]. Here trajectories are first converted into a feature matrix, for example based on the motion azimuth and then a matching of formalized patterns on the matrix is performed. This approach can be used to detect swarm formation behavior or alpha animals. However, animal movement analysis, especially baboons, has so far often been limited to predicting their positions and usually only uses the nearest neighbors for this, although the behavior can be quite complex and also influenced by other factors. One approach to not only predict the position of animals, but characterize the movement of a group and the actions of individuals is presented by Andrienko et al. [5]. Additionally, Cluster analysis is also often used to identify important groups or commonly visited places. However, up to now, more in-depth examinations of these clusters has not been carried out and according to domain experts, this is an important and necessary step to understand the social factors of the collective behavior of baboons and other animals.

In this paper, we contribute an algorithms for the hierarchical creation of core clusters as well as the contextual enrichment of detected core clusters by activity detection. In addition, we provide suitable visualizations for the analysis of identified core clusters and their activities enabling the recognition of patterns not recognizable with purely mathematical models. We demonstrate the usefulness of our approach in a use case including an expert interview.

II. Core Clusters

Many animals have a tendency to form groups as this can help in the search for food or provide protection from predators. The size and composition of these groups is dynamic as groups of animals may form larger groups or may split into smaller groups. However, it can be observed that some animals often meet in the same groups. In order to make this preference recognizable, we propose our concept of core clusters. A core cluster is defined as a set of individuals who often reside together in a common cluster over a certain period of time. The individuals do not have to be permanently in the same cluster, which is why we distinguish between two states: The core cluster is *present* when all associated individuals are in the same cluster and the core cluster is additionally *differentiated* when the core cluster is *present* and no other individual from another core cluster is in the same cluster. Still, not every individual automatically belongs to a core cluster as soon as it is part of the same cluster, since this affiliation can also be a coincidence as illustrated in Figure 1. Here, we identified two core clusters, the blue core cluster with two individuals and the green core cluster with three individuals. The orange individual seems to have no preference, as it first moves with the individuals of the blue core cluster, then moves to the individuals of the green core cluster and then leaves both core clusters. Therefore, it is not assigned to any of these core clusters.



(a) Time A: The orange individual is part of the blue core cluster.
(b) Time B: The orange individual is part of the green core cluster.
(c) Time C: The orange individual is part of no core cluster.

Fig. 1: The orange individual is part of no core cluster since it has no clear preference to join the blue or the green core cluster.

In the rest of this section we show how core clusters can be determined on the concrete example of baboons. We use the baboon data provided by Strandburg et al. [2], [6]. The data include the movements of 26 baboons over a period of 14 days. However, the sensor of one animal failed after only 26 minutes, which is why it was removed from further analysis. Missing values in the remaining data were cleaned up by linear interpolation. The movements of the animals were measured in a time resolution of one second between 6am and 6pm. 95% of the position data showed an error of less than 0.26m. In total, the data set contains more than 10 million data points.

As a core cluster contains a set of individuals often occurring in the same clusters, clusters must be identified at first. We choose the DBSCAN enabling us to recognize a variable number of clusters while individuals do not necessarily have to be part of a cluster. We set *minPts* to 2 so that we can find even the smallest possible clusters. ϵ is the distance threshold value and was chosen in consultation with collective behavior experts so that animals were clearly divided into clusters at sleeping places, which corresponds to a distance of 10m. This clustering was performed for each time point and subsequently we calculated how often animals occurred in the same clusters over the entire measurement period. An excerpt of the result of this calculation is shown in Table I.

Animal ID	Animal ID	Relative Frequency
1	2	0.81
3	4	0.77
3	5	0.75
1	3	0.72

TABLE I: Pairs of animals and the relative frequency how often they occur in the same clusters.

After clustering and calculating the relative frequency of animals appearing in the same cluster, we start to determine the core clusters. Core clusters must consist of at least two individuals and there must be at least two core clusters every day so that the behavior of the core clusters among themselves can be investigated. The procedure is divided into two parts. At the beginning, for every day and every possible size of core clusters, the system searches for frequently occurring sets of elements, so-called core cluster candidates. In the second part, the core clusters are extracted from the candidate sets. For the efficient search of frequently occurring sets of individuals, we employ existing algorithms such as the Apriori algorithm or PrePost+ [7] both being suitable for searching candidates with a *minSupp* $> x$. We start with a *minSupp* = *1* and search for all frequently occurring amounts of individuals. If no core clusters are found, the *minSupp* is reduced iteratively, for example by 10 % per iteration and the search is repeated.

After searching for core clusters, PrePost+ retrieves the following set of potential core clusters $S_1 = \{\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{1,5\},\{3,4\},\{1,2,3\}\}$. In this case, we are, for example, especially interested in core clusters of size two, thus we first remove all sets with less or more than two individuals, leaving $S_2 = \{\{1,2\},\{1,5\},\{3,4\}\}$. Next, all possible sets are formed from S_2, which themselves contain at least two sets:

$$S_{3,1} = \{\{1,2\},\{3,4\}\}$$
$$S_{3,2} = \{\{1,2\},\{1,5\}\}$$
$$S_{3,3} = \{\{1,5\},\{3,4\}\}$$
$$S_{3,4} = \{\{1,2\},\{1,5\},\{3,4\}\}$$

We remove every set $S_{3,x}$ in which one individual occurs in multiple subsets, leaving us with the potential core clusters $S_{3,1}$ and $S_{3,3}$. These sets are merged giving us the final core cluster candidate set for one day $C_1 = S_{3,1} \cup S_{3,3} = \{\{\{1,2\},\{3,4\}\},\{\{1,5\},\{3,4\}\}\}$.

From this set of core cluster candidates, the final core clusters can be extracted. We explain the procedure using the following example where we have candidates for two days, our candidate set from the previous example C_1 and a candidate set for a second day $C_2 = \{\{\{1,5\},\{6,7\}\}\}$. Since there is only one candidate with two core clusters in C_2, these are stored in our set of final core clusters. Subsequently, for all other core cluster candidates, their frequencies are measured on the different days. We get $\{1,2\} = 1$ and $\{3,4\} = 2$. Now we have to determine for the candidates $C_{1,1}$ and $C_{1,2}$ which of them we use as the core cluster. With the help of the frequencies, we cannot make a decision, since both have the same. $C_{1,2}$ already has an existing core cluster, so it is preferred. Thus we get the following core clusters for the days as the final result:

$$\text{Day } 1 : \{\{1,5\},\{3,4\}\}$$
$$\text{Day } 2 : \{\{1,5\},\{6,7\}\}$$

In a final user-driven step, the user is enabled to further refine these core clusters with the help of hierarchical aggregation. Suppose we found the following core clusters:

Day 1:$\{\{1, 2\}, \{4, 5, 6\}\}$ and Day 2:$\{\{1, 2, 3\}, \{4, 5\}\}$. The core clusters are observably different, but contain certain similarities among themselves. We calculate this similarity using an edit distance, using a cost of 1 for replacing an individual and a cost of 0.5 for adding or deleting. This allows us to calculate a hierarchical clustering on the core clusters where the user can combine individual core clusters into larger ones. An end result for this step could look like this:

$$\text{Day } 1 : \{\{\{1, 2, 3\}, \{4, 5, 6\}\}$$

$$\text{Day } 2 : \{\{\{1, 2, 3\}, \{4, 5, 6\}\}$$

The occurrence of the different core clusters on one or more days is shown with an adapted EventRiver [8] visualization. The x-axis represents the time and the y-axis visualizes how present the core cluster is at a certain point in time using a gray bar and how differentiated the core cluster is at this point in time using a colored bar. An example can be seen in Figure 2.

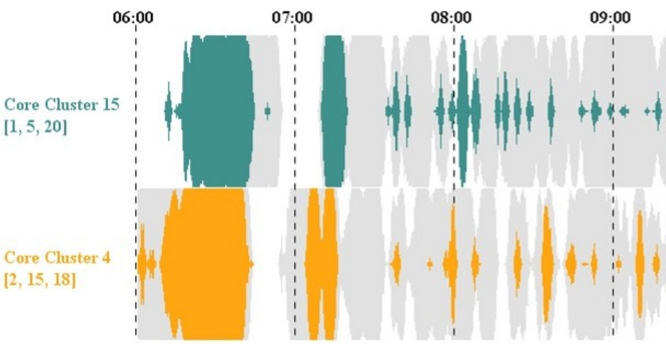


Fig. 2: Visualization of the occurence of core clusters over the course of one day. The time is represented on the x-axis. Gray bars show how present a core cluster is and colored bars show how differentiated a core cluster is, at a given point in time.

In this visualization, however, it is difficult to see which animals often form a core cluster with other animals, which is why we also offer other visualizations such as a matrix visualization showing the frequency of the common occurrence, as shown in Figure 3.

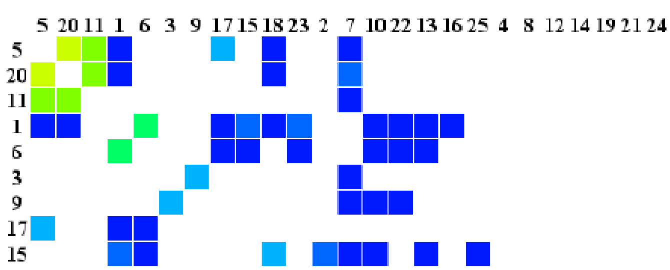


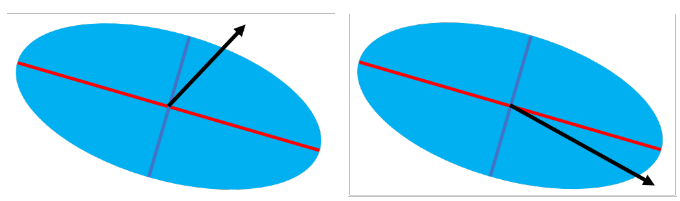
Fig. 3: Detail of the matrix visualization, which shows which animals often occur in the same core clusters. The animals {5, 20, 11} and {1, 6} often appear in the same core clusters.

III. Context Enrichment by Activity Recognition

The recognition of core cluster is a first step to classify collective animal movement behavior. For our baboon example, we now investigate which animals are often in a group together as well as if and how these groups move. In order to make the shown movement behavior understandable, the movement as well as its purpose has to be detected. Therefore, baboon movement patterns must be assigned to different activities. In discussions with collective behavior experts, we identified four activities that would be interesting to identify automatically: Sleeping, waiting, foraging and moving. In the following, we explain how we can enrich the movement of the baboons with context information by presenting methods for the four activities and explaining the design rationales behind them.

The recognition of sleep activity is straightforward. We know that the animals move to a place to sleep in the evening and usually move away from it in the course of the next morning. Therefore, we define the sleep activity as the period in which the animals are within a given r radius of these sleeping places.

If the animals move from their sleeping places, this usually happens to forage. However, not every movement serves the search for food, but can also serve to get to places with more food or to a sleeping place. To decide between foraging and transitioning, we examine how animals move in relation to each other. According to our collective behavior experts, animals such as baboons tend to move forward side by side if they are looking for food allowing them to cover a wider area and, thus, ensuring that all animals find food. However, baboons moving in a line one after the other indicates transitioning between two places. To identify this behavior, we form the major axis of a group of baboons, which corresponds to the line between the two animals furthest apart in the group, and the orthogonal of this major axis. In addition, we calculate the direction of movement of the group by averaging the direction of movement of all animals. If the angle between the direction of movement and the orthogonal is smaller than the angle to the main axis, we see that the animals move rather next to each other and are therefore in search of food. Otherwise, if the angle between the direction of movement and the main axis is smaller, the baboons are transitioning as illustrated in Figure 4.



(a) Foraging Behavior　(b) Movement Behavior

Fig. 4: The main axis of a group of baboons is defined as the line between the two furthest apart animals, here highlighted in red and its orthogonal in blue. If the movement of the group, highlighted by the black arrow, is closer to the orthogonal, then the baboons are moving next to each other, indicating foraging behavior. If the movement of the group is closer to the main axis, the animals are walking in a line, indicating transitioning behavior.

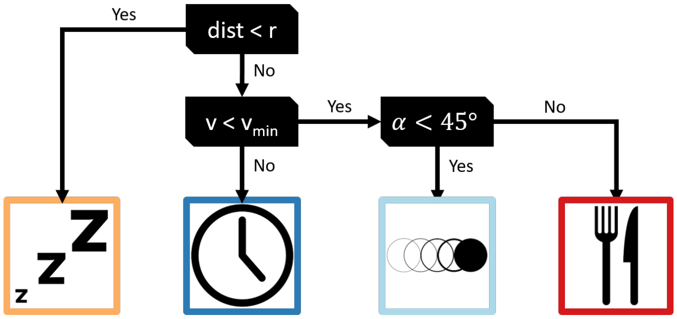


Fig. 5: Flow Chart for the activity detection of baboons.

The last activity is waiting. We define waiting as the activity if the animals do not move but are not at their sleeping place. The complete flow chart for the activity detection of baboons is shown in Figure 5.

IV. Use Case

After detecting core clusters and enriching them with recognized activities, further analyses can be carried out. For example, we can now analyze whether core clusters are present or differentiated during different activities. We can combine different visualizations, for example, the visualization of the appearance of the core clusters with a time visualization of the detected activities. This can be seen, for example, in Figure 6. Here it can be seen that during the main movement time between 08:00 and 12:00 am the core clusters are mainly present, but not differentiated. This means that only a few, large clusters are formed, i.e., the animals search for food together. However, as soon as they arrive at a suitable feeding place, such as between 1 p.m. and 5 p.m., the core clusters are hardly present anymore, i.e., the animals search for food either individually or in small groups.

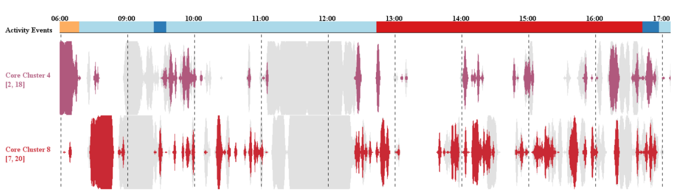


Fig. 6: Comparison of the presence and differentiation of the identified core clusters and the identified activities. The colors of the activities can be seen in Figure 5. Differences can be seen in the presence and differentiation of the core clusters during movement and foraging.

V. Discussion

In this paper, we presented and defined the concept of *core clusters*. In addition, a corresponding algorithm for the recognition of core clusters is presented. This works, very roughly summarized, in two steps. First, clustering the animals at any time using DBSCAN and second, identifying individuals that often occur in the same clusters using the Apriori or PrePost+ algorithm. A possible improvement is the usage of the HDBSCAN [9] (Hierarchical Density-based Spatial Clustering of Applications with Noise) algorithm, since it works without a specific ϵ. Afterwards we presented how the found core clusters can be enriched with context information. For this we use information such as the proximity to the sleeping area, the speed of movement or the direction of movement depending on the cluster spread. These two concepts can help experts to investigate collective behavior. Our use case, for example, shows that differences in the collective behavior of the animals can be observed in their search for food and in their movement.

In order to find out whether our definition of core clusters and the activities makes sense and whether a system that makes it possible to analyze them is useful for experts, we conducted an interview with three domain experts at the University of Konstanz. These experts are the ones responsible for the collection of baboon data and have been working for several years on the analysis of baboons' behavior. The invited experts consider our definition of core clusters to be correct. However, it was pointed out that core clusters could be found because animals sleep together but are no longer in the same groups for the rest of the day. The activities we used were proposed by one of the experts, but the other two experts agreed in the interview that these four activities are sufficient for the time being. For these reasons, the experts definitely see a benefit in this work. Improvement suggestions of the experts are, for example, to include the fields of vision of the baboons, since the animals could form groups further apart if the visibility conditions are better.

Our proposed solutions serves as a first step towards a deeper understanding of collective behavior and there is ample opportunity for further research. We plan to apply our proposed solution to other kinds of animals expressing collective behavior such as fish, birds or dolphins. An extension of the core cluster and activity detection algorithm is also conceivable. At present, some aspects, such as the movement trajectories of the animals, are not yet fully included in these analyses. Finally, we see possibilities to improve and extend the visualizations. In this way, core clusters and activities can also be visualized with the help of spatio-temporal visualizations, such as spatially ordered treemap [10] to allow for the analyzing presence and density of baboons. Here, for example, correlations between core clusters, activities and locations could be identified.

Acknowledgment

This paper has been partially supported by the German Research Foundation DFG within Priority Research Program 1894 (DFG SPP VGI)

The authors wish to thank Daniel Hafner for his implementations and the domain experts for their valuable feedback and discussions.

References

[1] S. E. Barkan, *Sociology: Understanding and changing the social world.* Flat World Knowledge, Incorporated, 2011.

[2] A. Strandburg-Peshkin, D. R. Farine, I. D. Couzin, and M. C. Crofoot, "Shared decision-making drives collective movement in wild baboons," *Science*, vol. 348, no. 6241, pp. 1358–1361, 2015.

[3] N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures," *Information Visualization*, vol. 12, no. 1, pp. 3–24, 2013.
[4] P. Laube, M. van Kreveld, and S. Imfeld, "Finding remo—detecting relative motion patterns in geospatial lifelines," in *Developments in spatial data handling*. Springer, 2005, pp. 201–215.
[5] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi, "Space transformation for understanding group movement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2169–2178, 2013.
[6] M. Crofoot, R. Kays, and M. Wikelski, "Shared decision-making drives collective movement in wild baboons," *Movebank Data Repository*, 2015.
[7] Z.-H. Deng and S.-L. Lv, "Prepost+: An efficient n-lists-based algorithm for mining frequent itemsets via children–parent equivalence pruning," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5424–5432, 2015.
[8] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. Keim, "Eventriver: Visually exploring text collections with temporal references," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 1, pp. 93–105, 2012.
[9] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.
[10] J. Wood and J. Dykes, "Spatially ordered treemaps," *IEEE transactions on visualization and computer graphics*, vol. 14, no. 6, 2008.