Framework for Inferring Leadership Dynamics of Complex Movement from Time Series

Chainarong Amornbunchornvej*   Tanya Berger-Wolf*

Abstract

Leadership plays a key role in social animals', including humans, decision-making and coalescence in coordinated activities such as hunting, migration, sport, diplomatic negotiation etc. In these coordinated activities, leadership is a process which organizes interactions among members to make a group achieve collective goals. Understanding initiation of coordinated activities allows scientists to gain more insight into social species behaviors. However, by using only time series of activities data, inferring leadership as manifested by the initiation of coordinated activities faces many challenging issues. First, coordinated activities are dynamic are changing overtime. Second, several different coordinated activities might occur simultaneously among subgroups. Third, there is no fundamental concept to describe these activities computationally. In this paper, we formalize Faction Initiator Inference Problem and propose a leadership inference framework as a solution of this problem. The framework makes no assumption about the characteristics of a leader or the parameters. The framework performs better than our non-trivial baseline in both simulated and biological datasets (schools of fish). Moreover, we demonstrate the application of our framework as a tool to study group merging and splitting dynamic another biological dataset of trajectories of wild baboons. In addition, our problem formalization and framework enable opportunities for scientists to analyze coordinated activities and generate scientific hypotheses about collective behaviors that can be tested statistically and in the field.

Keywords: leadership, coordination, time series, influence, collective behavior

1 Introduction

Leadership is a process of individuals (leaders) who influence a group to achieve collective goals [8,12]. Leadership plays a key role in solving collective-action problems (e.g. social conflicts, migration, hunting, territorial defense) across social species [8], organizing collective movement [5], as well as collaboration in group's decision making [7,8]. In the context of coordination, which is defined as an emergence of collective actions to achieve the collective goals [18], leadership mainly contributes by fostering collective behaviors in social species ranging from humans [7,8,12] to fish [17].

In nature, leadership can be viewed as a process of initiation of coordinated activity. For example, leadership is a process in which leaders initiate the group's coordinated movement toward a destination [17, 22, 25]. In this process, leaders guide their group's members to follow in the right direction. Understanding how leaders emerge and influence collective behaviors enables scientists to gain insight about synchronization and coordination processes in nature. In this paper, we use the words **'leader'** and **'initiator'** interchangeably.

While many studies on leadership in coordinated activity exist in behavioral research, there are a few computational approaches addressing the leadership of coordination. In social network analysis, Influence Maximization (IM) [10,11,15] is one of the classic problems that focuses on inferring a subset of individuals that maximizes information spreading. However, IM focuses solely on finding potential initiators who initiate the coordination of *information spreading* and, moreover, does not address the question of *when* coordination happens. The method for inferring leaders from online communities actions [9] can be used to identify the *group* being coordinated but it, still, does not provide the information on *when* coordinated activities happen. In movement coordination, [3, 4, 14, 16, 20] propose methods specific to movement activity for finding leaders *during* group's movement intervals but none of them can be used to identify the time of the process of coordination. There also exist many works regarding collective behavior [5,27,28] and implicit leaders. In this model, leaders can influence their group implicitly and leaders' identity might be unknown to the group. Still, none of the works in this category can be used to infer the time of the periods of coordination.

Since leadership is a collective process [12], considering only dyadic interactions is not enough to infer a leadership instance. Therefore, the works in [2, 14, 16] proposed leadership frameworks that are based on a network representation of time series.

In the context of coordination leadership, the method of leadership inference in [2] provides a solution for identifying coordination events, the initiators of these events, as well as proposes an approach for the classification of the types of leadership models acting on a group. However, the framework in [2] cannot be used to infer *multiple* coordinated activities which can occur

*Department of Computer Science, University of Illinois at Chicago. {camorn2,tanyabw}@uic.edu

simultaneously because the notion of multiple factions is not employed by the framework. Our main purpose is to close the gap in the study of coordination leadership.

1.1 Our Contributions First, we introduce the novel computational problem of leadership identification in multiple coordinated activities, namely FACTION INITIATOR INFERENCE PROBLEM. We formalize the problem and analyze its theoretical properties and implications. Second, we propose a high-performance framework for FACTION INITIATOR INFERENCE PROBLEM by combining several existing methods in a principled and novel manner. Our framework is capable of:

- **Detecting intervals of multiple coordination:** inferring intervals when different coordinated activity in multiple groups may appear simultaneously;
- **Identifying leaders:** identifying the initiators of these coordinated activities, the individual who initiates each coordination and the group that follows; and
- **Discovering the events of merging and splitting of coordination:** identifying the time when a coordinated group is separated into smaller sub-groups or merged with another coordinated group.

FACTION INITIATOR INFERENCE PROBLEM: To reach collective goals, group's members must coordinate with each other. Multiple factions within a big group may exist solving their sub-tasks in helping the entire group achieve the collective goals. **Given time series of individual activities, our goal is to identify periods of coordination and the subsequent coordinated activity, find factions of coordination if more than one exist, as well as identify leaders of each faction**

We demonstrate the ability of the framework to infer leadership in multiple coordinated groups on both simulated and biological datasets. Since we propose the new problem and framework and no other approaches exist, we compare our framework against a non-trivial baseline, which is the modification of the closest existing approach in leadership inference. Our approach is flexibly generalizable to any multiple coordinated activities from any time series data.

1.2 Influence Maximization vs. FACTION INITIATOR INFERENCE PROBLEM

- **Coordination Mechanism:** Majority of Influence Maximization work uses Independent Cascade and Linear Threshold models as main coordination mechanisms. Yet, there are other models, such as Hierarchy, Dictatorship or other non-network based models that can be represented as coordination mechanisms. The new problem we formalize in this paper, FACTION INITIATOR INFERENCE PROBLEM, generalizes to all coordination mechanism types. In this paper, the framework that we propose to solve FACTION INITIATOR INFERENCE PROBLEM is able to provide a solution from datasets that generated by any kind of coordination mechanism.
- **Coordination Event:** Influence Maximization focuses mainly on an information spreading event happening in a social network. However, information spreading is considered to be one type of a coordination event, whereas there are other types that are not necessary to be network-based events. For example, a coordinated movement activity of animals is a coordination event that has animals coordinating their trajectories, not necessarily through a wave-like spread of information in a network, to reach a group destination. Our proposed framework can handle all types of coordination events, which is not limited to only information spreading.
- **The dynamics of coordination:** In influence Maximization, majority of papers focus on how to infer a single set of initiators that can maximize influence in a given network. However, in a single dataset, there might be many coordination events and each event can have different initiators. Moreover, these coordination events that have different initiators might happen simultaneously. The framework we propose here aims to address the dynamics of coordination within data. Our framework is capable of inferring when each coordination event happens and who are the initiators of coordination events within data.

2 Problem statement and analysis

2.1 Coordination without noise Given a collection of time series, our goal is to find multiple coordination intervals as well as their initiators. We do not assume that the coordination intervals that belong to different coordinated sets of time series are disjoint and allow overlap. We formalize various concepts of coordination and following similar to [2].

DEFINITION 1. (FOLLOWING RELATION) *Let $U = (\vec{u}_1, \ldots, \vec{u}_t, \ldots)$ and $W = (\vec{w}_1, \ldots, \vec{w}_t, \ldots)$ be arbitrary-length time series. If $\forall t \in \mathbb{N}$, there exists a time delay $\Delta t \in \mathbb{Z}^+ \cup \{0\}$, such that $\vec{w}_t = \vec{u}_{t+\Delta t}$, then U follows W, denoted as $W \preceq U$ for any Δt and $W \prec U$ if $\Delta t > 0$.*

DEFINITION 2. (COORDINATION) *Given a set of m-dimensional time series $\mathcal{U} = \{U_1, \ldots, U_n\}$. The set $\mathcal{U}$ is* coordinated *at time t if for every $\binom{n}{2}$ pairs $U_i, U_j \in \mathcal{U}$, there exists either $U_i \prec U_j$ or $U_j \prec U_i$. The* coordination interval *is the maximal contiguous time interval* $[t_1, t_2]$ *such that* $\mathcal{U}$ *is coordinated for every* $t \in [t_1, t_2]$.

DEFINITION 3. (INITIATOR) *Let* $\mathcal{U} = \{U_1, \ldots, U_n\}$ *be a set of coordinated* m*-dimensional time series within some coordination interval* $[t_1, t_2]$. *Then the time series* $L \in \mathcal{U}$ *is the* initiator *time series for the coordination interval if for each time series* $U \in \mathcal{U} \setminus \{L\}$, $L \prec U$.

DEFINITION 4. (FOLLOWING NETWORK) *Let* $\mathcal{U} = \{U_1, \ldots, U_n\}$ *be a set of time series, a directed graph* $G = (V, E)$ *is defined as a following network, where* V *is a set of nodes that has a one-to-one correspondence to the time series set* $\mathcal{U}$ *and* E *is a set of edges, such that* $e_{i,j} \in E$ *if* $U_j \prec U_i$.

We now extend these concepts to the case of multiple coordinated subgroups.

DEFINITION 5. (FACTION) *Given a set of time series* $\mathcal{U}$, *a subset* $F \subseteq \mathcal{U}$ *at time* t *is maximally coordinated, if* F *is coordinated and there is no other coordinated set* $F' \subseteq \mathcal{U}$ *where* $F \subset F'$. *We call such maximally coordinated* F *a faction at time* t.

DEFINITION 6. (FACTION INTERVAL) *The* coordination interval *of a faction* F *or a faction interval is the maximal consecutive time interval* $[t_1, t_2]$ *such that* F *is coordinated for every* $t \in [t_1, t_2]$.

Faction is a structurally maximal subset and its interval is a temporally maximal subset.

LEMMA 2.1. *A time series* W *is a member of a faction* F *if and only if it has an edge to* F*'s initiator* L.

Proof. Let a time series $W \in F$. Since $\forall U \in F \setminus \{L\}$, $L \prec W$. By definition, there is an edge from W to L.

Let $L \prec W$. If W is not in F, then we can add W to F, which will remain a coordinated set but will now violate the maximality of F. Thus, $W \in F$.

According to Lemma 2.1, a faction F is a set of nodes within G such that all nodes within F have a directed edge to L. Note that L always has the out-degree of zero and in-degree of $|F| - 1$ within a coordination interval.

We are now ready to formally state the FACTION INITIATOR INFERENCE PROBLEM.

2.2 Coordination with noise In the previous section, we stated the definitions and properties of the problem of identifying multiple faction initiators in the ideal setting. In this section, we provide the relaxation and the analysis of the problem with noise.

Problem 1: FACTION INITIATOR INFERENCE PROBLEM

Input : Set $\mathcal{U} = \{U_1, \ldots, U_n\}$ of m-dimensional time series

Output: A set of factions $\mathcal{F} = \{F_1, \ldots, F_k\}$, a set of coordinated intervals $\mathcal{T} = \{[t_1^1, t_2^1], ..., [t_1^k, t_2^k]\}$, and the set of initiator time series $\mathcal{L} = \{L_1, ...L_k\}$ where L_i initiated the coordination interval $[t_1^i, t_2^i]$ of the faction F_i

---

DEFINITION 7. (σ-FOLLOWING RELATION) *Let* $\mathcal{U}$ *be a set of time series and* $\mathrm{sim} : \mathcal{U} \times \mathcal{U} \to [0, 1]$ *be some similarity measure between two time series. For any pair of time series* $U_i, U_j \in \mathcal{U}$, *let* $\Delta t_{max} = \underset{\Delta t \in \mathbb{Z}}{\mathrm{argmax}}\, \mathrm{sim}(U_{i,1}, U_{j,1+\Delta t})$ *where* $U_{i,t} \in \mathcal{U}$ *represents a time series* U_i *starting at time* t, *and let* $\mathrm{sim}_{max}(U_i, U_j) = \mathrm{sim}(U_{i,1}, U_{j,1+\Delta t_{max}})$. *Then, for a threshold* $\sigma \in (0, 1]$, *if* $\mathrm{sim}_{max}(U_i, U_j) \geq \sigma$, *then we have:*

- *if* $\Delta t_{max} > 0$, *then* $U_i \prec_\sigma U_j$,
- *if* $\Delta t_{max} < 0$, *then* $U_j \prec_\sigma U_i$,
- *if either* $\Delta t_{max} = 0$ *or* $U_i \prec_\sigma U_j$ *and* $U_j \prec_\sigma U_i$, *then* $U_i \equiv_\sigma U_j$.

Note that if two time series U and W such that $U \prec_\sigma W$ and $W \prec_\sigma U$, there exists more than one position in time Δt_{max} that make both time series maximize their similarity.

DEFINITION 8. (σ-COORDINATION) *Let* $\mathcal{U}$ *be a set of time series, then* $\mathcal{U}$ *is* σ*-coordinated if for every* $\binom{|\mathcal{U}|}{2}$ *pairs* $U_i, U_j \in \mathcal{U}$, *either* $U_i \prec_\sigma U_j$ *or* $U_j \prec_\sigma U_i$ *exists.*

DEFINITION 9. (σ-FACTION) *Let* $\mathcal{U}$ *be a set of time series. A* σ*-faction* $F \subseteq \mathcal{U}$ *is a maximal set such that* F *is* σ*-coordinated, and there is no other* σ*-coordinated set* $F' \subseteq \mathcal{U}$ *where* $F \subset F'$.

DEFINITION 10. (RELAXED FACTION INTERVAL) *Let* $\mathcal{U}$ *be a set of time series, the time interval* $[t_1, t_2]$ *is a faction interval of initiator* L *if for all* $t \in [t_1, t_2]$, *there exists a faction* F_t *such that* F_t *has* L *as its initiator and* $|F_t| > 1$.

2.3 Coordination measure Given a set of time series $\mathcal{U}$, a set of clusters $\mathcal{C} = \{H_1, \ldots, H_n\}$ such that $\bigcup_k H_k = \mathcal{U}$, we define a cluster membership indicator $\delta_{i,j} = 1$ if time series U_i and U_j belong to the similar cluster, otherwise it is zero. **The average coordination measure** Ψ of a set of clusters $\mathcal{C}$ is defined as follows:

$$\Psi(\mathcal{C}) = \frac{\sum\limits_{U_i,U_j \in \mathcal{U}, U_i \neq U_j} \mathrm{sim}_{max}(U_i,U_j)\delta_{i,j}}{\sum\limits_{U_i,U_j \in \mathcal{U}, U_i \neq U_j} \delta_{i,j}}. \tag{2.1}$$

Note that $\Psi \in [0,1]$. If Ψ is close to one, then all time series within the same cluster are highly similar, with some time delay. This implies there exists a high degree of coordination within each clusters in this case. On the contrary, $\Psi \approx 0$ implies no coordination, on average.

Theorem 2.1. *Given a set of time series $\mathcal{U}$ containing a set of σ-faction $\mathcal{F} = \{F_1, \dots, F_n\}$ where $\bigcup_{F_i \in \mathcal{F}} F_i = \mathcal{U}$, then, for all possible sets of clusters, $\mathcal{F}$ maximizes the average coordination measure Ψ.*

The proof of Theorem 2.1 is at the supplementary.

3 Methods

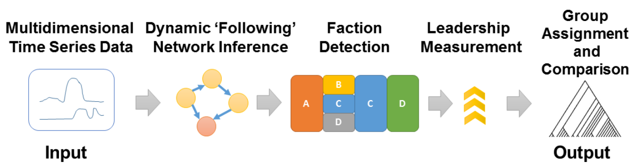


Figure 1: A high-level overview of the proposed framework

The main purpose of our proposed framework is to solve Faction Initiator Inference Problem. We design it to be able to infer a set of factions, relaxing faction intervals, and their initiators from time series. Figure 1 depicts the overview concept of our framework.

3.1 Following network inference Given a set of time series $\mathcal{Q}$ and a similarity threshold σ, for each pair of time series $U, W \in \mathcal{Q}$, our goal is to measure whether either U follows W or no following relation between them. The time series measure we need here should satisfy the properties as follows. First, it can recognize common patterns between two time series if they exist. These common patterns can be noisy, distort, delay, and discontinuous. Second, it can infer time delay between these common patterns.

To infer following relations, we deploy Dynamic time warping (DTW) [21] as a measure of following relation since DTW's warping path can distinguish whether two time series share noisy common patterns, as well as it can also approximately infer time delay of common patterns between time series. Besides, according to the work in [16], DTW performance is superior than other methods in the perspective of following detection among time series.

For any pair of time series $U, W \in \mathcal{Q}$, we deploy the equation from [2], to approximate a following relation as below:

$$\mathrm{s}(P_{U,W}) = \frac{\sum_{(i,j) \in P_{U,W}} \mathrm{sign}(j-i)}{|P_{U,W}|}. \tag{3.2}$$

Where $P_{U,W}$ is the optimal path of DTW. If $(i,j) \in P_{U,W}$, then U at time i is the most similar to W at time j. When $-\sigma < \mathrm{s}(P_{U,W}) < \sigma$, neither U nor W follows each other. We have $U \prec_\sigma W$ if $\mathrm{s}(P_{U,W}) \geq \sigma$. In contrast, $\mathrm{s}(P_{U,W}) \leq -\sigma$ implies $W \prec_\sigma U$. The function is bound by $\mathrm{s}(P_{U,Q}) \in [-1,1]$ and we set $\sigma = 0.5$ for our framework as default.

Then, a following network $G = (V_{\mathcal{Q}}, E)$ is constructed from $\mathcal{Q}$ where $v_k \in V_Q$ represents a vertex of time series $Q_k \in \mathcal{Q}$ and $E_{k,l} = |\mathrm{s}(P_{Q_k,Q_l})|$ if $Q_l \prec_\sigma Q_k$. The pseudo code of following network inference is at the supplementary.

3.2 Dynamic Following network inference As mentioned before, a set of time series $\mathcal{U}$ might consist of multiple overlapping coordination intervals from many factions. Using only summary statistics or an aggregate following network cannot discover these dynamics. Hence, we need to consider each local interval and build a following network to represent the interval. Therefore, we deploy a dynamic following network scheme in our framework, which is a common technique to deal with dynamics of data [13].

The next question is "how long each local interval should it be?" We assume that we have a prior knowledge of time window ω to capture local intervals. Then we will show that we can infer ω from the dataset itself in Section 3.4.

As mentioned previously, we have a set of t^*-length time series $\mathcal{U}$ as the input. We sampling all time series within $\mathcal{U}$ by sliding window intervals and create following networks of these intervals. Let $\omega \in \mathbb{N}$ be a time window and $\delta = 0.1\omega$ (time shift threshold), the i-th sliding window interval is defined by: $w(i) = [(i-1) \times \delta, (i-1) \times \delta + \omega]$. For each interval $w(i)$, we have a set of time series $\mathcal{Q}$. For each time series $U_k \in \mathcal{U}$, there is $Q_k \in \mathcal{Q}$ such that Q_k is a subset of U_k during $w(i)$ time interval. We build the following network for each $w(i)$, then we combine these networks to be a single dynamic network. The pseudo code of the dynamic network creation is at the supplementary.

3.3 Factions detection and coordination intervals For each following network $G = (V, E)$, factions are network components such that all member nodes directly connect to their initiator (Lemma 2.1). We infer factions based on Definition 9 and the coordination intervals of factions are discovered based on Definition 10.

According to Lemma 2.1, initiator nodes have outgoing-degree zero, and all nodes within the similar faction directly connect to their initiator. However, due to the introducing of time window ω, some nodes might not have directed edges to the initiators. Therefore, we relax the constraint of faction membership to make all nodes which have any directed path to an initiator are members of the initiator's faction.

Since a faction is a connected component, which all nodes are reachable from the initiator by inverse paths, we use breath first search algorithm (BFS) to identify all reachable nodes from each initiator node in the following network in order to find members of each faction. Specifically, let $l \in V$ be an initiator node, $\deg_{\text{out}}(l) = 0$, and a faction $F_l \subseteq V$ be a faction which has l as its initiator, then a node $v_k \in V$ is a member of F_l if and only if there is a path from v_k to l in G. The pseudo code of this step is at the supplementary.

Another useful measure about factions that we will use later is a "faction size ratio". Let $G_l = (F_l, E_l)$ be an induced subgraph of G creating from F_l, a faction size ratios of F_l is defined as follows:

$$\text{fs}(F_l) = \frac{|E_l|}{\binom{|V|}{2}} \tag{3.3}$$

3.4 Time window inference In reality, some following relations might not cause by initiators since they either happen by chance or other factors which are not related to influence of leaders. For instance, if a follower is unable to observe a leader's pattern, then the follower cannot be influenced by the leader. Different types of time series have different limitation of 'observation memory', which is the limitation of time delay Δt such that a follower can truly observe and imitate its leader's actions or can get commands from a leader.

Hence, to represent the concept of observation memory limitation, we set a time window ω to limit the length of time delay Δt that we can measure following relations. Moreover, ω help us to prevent the comparison of time series between different coordination events.

Nevertheless, if we set ω too short, we miss inferring some following relations that have $\Delta t > \omega$. On the contrary, long-length ω causes the false positive matching between patterns of different coordination intervals that repeat the same pattern. Therefore, a proper ω^* should be able to infer a higher number of following relations than any ω. Even if some random following relations might appear when we choose ω instead of ω^*, this is not an issue. Since these random following relations appear by chance, they have a little effect to a number of following relations on average.

In our framework, without the knowledge of ω, we use ω such that it maximizes the average coordination measure Ψ (Eq. 2.1). The reason is that a higher coordination measure value implies a higher number of following relations we can infer with ω per faction. Given a dynamic following network based on time window ω, for each time step t, we calculate Ψ_t by designating each faction to be a cluster and creating the last cluster for all time series, which are not in any faction. Then, $\hat{\Psi}_\omega$ is computed from the median of $\{\Psi_1, ..., \Psi_{t^*}\}$. $\hat{\Psi}_\omega$ is used to be a representative coordination measure value of ω. Hence, the optimal ω^* is computed as follows:

$$\omega^* = \underset{\omega}{\text{argmax}}(\hat{\Psi}_\omega). \tag{3.4}$$

3.5 Leadership comparison There are several methods that are widely used for ranking important nodes within the graph. One of the well-known method that consider the high-order relation within a graph is PageRank [19]. In our approach, we deploy PageRank on the following network to rank individuals within each faction and report the rank ordered lists for each time step. Even though PageRank scores are computed from the entire network, we compare individuals' ranking score only within the same faction and create a rank order list for each faction. For each node i within a following network G, the PageRank score is defined below:

$$\pi_i = d \sum_{k \in \mathcal{N}_i^{in}} \frac{E_{k,i}\pi_k}{|\mathcal{N}_k^{out}|} + (1 - d). \tag{3.5}$$

Where $\pi_i \in [0, 1]$ is a rank value of node i, $\mathcal{N}_i^{in}$ is a set of i's followers, $\mathcal{N}_k^{out}$ is a set of individuals k follow, and $E_{k,i} \in [0, 1]$ is an element of adjacency matrix of a following network where k follows i if $E_{k,i} \geq \sigma$.

4 Evaluation Datasets

4.1 Leadership models The evaluation of the framework are conducted based on four models.

4.1.1 Dictatorship Model The Dictatorship Model ('DM') [2] is considered to be the simplest model in the leadership realm. Initially, no movement happen until the leader starts moving to a target, then individuals follow its leader with some time delay until the entire group is coordinated in both direction and velocity. Then, the group gradually stops at the target and starts moving again for the next target.

4.1.2 Hierarchical Model The Hierarchical Model ('HM') [2] is another variation of DM with the hierarchical condition. The hierarchical condition assigns a rank to each individual within a group. A leader has the highest rank. The low-rank individuals follow high-rank individuals with some time delay. In our evaluation model, we assign a linear order hierarchical condition such that ID(1) is a leader and ID(n) is followed by ID($n+1$). The group moves linearly along the line with some noise following its leader.

4.1.3 Independent Cascade Model The Independent Cascade Model ('IC') [15] is one of the influence propagation model in Social network analysis. Initially, everyone has a probability to be activated ρ. Active individuals move toward their leader. For each time step, each active individual simultaneously and independently attempts to activate k-nearest inactive neighbors around it with the probability of success ρ. If success, the inactive individual becomes active at the next time step. Active individuals cannot attempt to activate the same individuals again. Only the leader follows its target and everyone else follows the leader. We explore the parameter space on combinations of: $k \in \{3, 5, 10\}$ and $\rho \in \{0.25, 0.50, 0.75\}$.

4.1.4 Crowd Model In the Crowd Model (CM) [27], there are two types of individuals: informed and uninformed individuals. For each time step, informed individuals move toward the target independently while uninformed individuals keep staying close to both group's position and direction centroids. Therefore, the group direction is implicitly influenced by informed individuals. For each coordination, all informed individuals follow a single target direction vector, while the rest of the group keep staying with the majority.

4.2 Synthetic trajectory simulation We generate time series datasets based on the models described above. For each dataset, it consists of 30 individuals time series of X, Y coordinates. Each time series has a length as 4,000 time steps. The coordination event is a consecutive multiple faction intervals. We have five coordination events for each dataset. For each model above, the coordination event can be divided into two types.

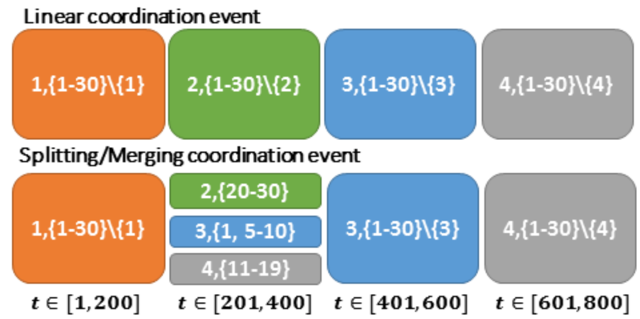


Figure 2: (above) Linear coordination event and (below) Splitting/Merging coordination event . Each block represents a faction such that the first element is the leader ID and the second element is the set of member IDs. The time interval each faction appears is at the last line.

4.2.1 Linear coordination event There are only four factions for each coordination event. The first faction has ID(1) as a leader and others are followers. This faction lasts for 200 time-steps. The next faction is leaded by ID(2) and its coordination interval is $[201, 400]$. The third faction appears within $[401, 600]$ interval and it has ID(3) as a leader. In the last faction, ID(4) leads the group to stop moving and the group completely stop moving around time step $t = 700$. Everyone stops moving within $[700, 800]$, then the group proceeds the next coordination event again.

4.2.2 Splitting/Merging coordination event In this type of coordination event, splitting and merging of factions happen. Within the $[1, 200]$ interval, ID(1) leads a single faction with its direction vector. Then, at $t = 201$, the group is spitted into three factions and they appear within $[201, 400]$ interval. The first faction is leaded by ID(2) with around one third of the previous faction members are followers (Fig 2 below). The ID(2) has its own direction vector. ID(3) leads the second faction with another one third members from the previous faction. ID(3) has a different direction from ID(2). Lastly, ID(4) leads the rest of individuals. ID(4) also has its own direction, which is different from ID(2)'s and ID(3)'s.

At $t = 401$, the factions leading by ID(2) and ID(4) are merged to the faction of ID(3); ID(2) and ID(4) follow the ID(3)'s direction. At the $[401, 600]$ interval, ID(3) completely leads all individuals. Finally, ID(4) leads the faction to stop moving between $t = 601$ and $t = 700$. The group completely stop at the $[701, 800]$ interval. Note that leaders in each faction are informed individuals in the Crowd Model.

In stead of having only one leader for each faction, we have three informed individuals in the Crowd Model.

For each leadership model and its coordination event type, we generated 100 datasets. In total, each model has 200 datasets except IC that we explore all nine possible values of parameters. In total, we have 1,800 datasets for the IC model.

4.3 Real datasets

4.3.1 Baboon trajectories The dataset comes from a GPS collars of olive baboons (*Papio anubis*) troop. The troop lives in the wild in Mpala Research Centre, Kenya [6, 24]. The data consists of time series of latitude-longitude location pairs for each individual per second. 16 individuals whose collars remained functional are analyzed for a case study of the merging coordination event.

4.3.2 Fish schools trajectories The fish dataset is a set of time series of fish positions from the video record. A fish school is golden shiners (*Notemigonus crysoleucas*). The record is used to study information propagation over the visual fields of fish [23]. Each trial contains 70 fish, with 10 fish who trained to leads the group to the feeding sites. The dataset has 24 separate leadership events. The task is to correctly identify trained fish.

5 Evaluation criteria

For each simulation dataset, ground truths of which individual belongs to which faction and who is a leader of each faction are provided. We compared the inference result from each method against these known ground truths to evaluate method performance.

5.1 Individual assignment For all models, for each time step, the accuracy of individual assignment is the number of inferred individuals' factions agree with the ground truth divided by the total number of individuals. Note that, in the Crowd Model, each faction F has a set of informed individuals and individuals belong to F if they follow any inform individual in F.

5.2 Leadership prediction For all models except the Crowd Model, The true positive TP is a number of inferred leaders who agree with the ground truth. The false positive FP is a number of inferred leaders who is not the actual leaders. The false negative FN is a number of actual leaders who are inferred to be non-leaders. In the Crowd Model, TP is a number of inferred leaders who are informed individuals from the right faction. FP is a number of leaders who are uninformed individuals. FN is a number of ground truth factions such that all informed members are non-leaders. We calculated F1-Score to estimate the performance of leadership prediction for each framework.

Table 1: Factions and Leaders identification on simulation models

Dataset	Leadership F1-score		Assignment Acc.	
	mFLICA	FLOCK	mFLICA	FLOCK
DM-L	**0.94**	0.92	**0.89**	0.86
DM-MS	**0.94**	0.91	**0.86**	0.84
HM-L	**0.94**	0.91	**0.94**	0.86
HM-MS	**0.95**	0.90	**0.86**	0.81
IC-L	**0.91**	0.86	**0.86**	0.80
IC-MS	**0.89**	0.85	**0.79**	**0.79**
CM-L	**0.82**	0.64	**0.83**	0.64
CM-MS	**0.75**	0.67	**0.64**	0.55

Table 2: Rank orders median accuracy within factions

Dataset	Top3 Rank Order Accuracy	
	mFLICA	FLOCK
HM-L	0.75	0.78
HM-MS	0.72	0.76

6 Results

6.1 Leadership Identification For each simulation model in Section 4.1, we evaluated results from all datasets using the criteria in Section 5. We set ω time window by a method in Section 3.4, and set time shift $\delta = 0.1\omega$. The results of faction assignment and leaders identification is in Table 1. Each row with the label '-L' is a model with Linear coordination event type (Section 4.2.1) and '-MS' represents a model with Splitting/Merging coordination event type (Section 4.2.2). The 2nd and 3rd columns represent the results of leadership prediction F1-Scores of mFLICA (our proposed framework) and FLOCK framework [3, 26], such that the values in these columns are calculated from the median of all datasets from the similar leadership model. The 4nd and 5rd columns represent individual assignment accuracy results. We took the median of all similar-model datasets to represent each model accuracy. Obviously, mFLICA beated FLOCK in all models. The result implies that the simple framework like FLOCK has a limitation when it needs to deal with complicated noisy models.

In hierarchical models, we reported the result of

Table 3: A school of fish inference median accuracy over 24 trails

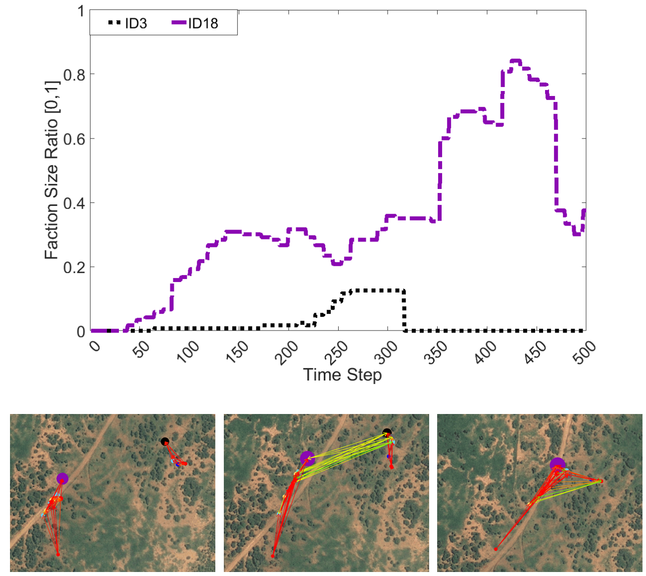
Method	Trained fish factions	Trained fish leaders
mFLICA	0.90	0.88
FLOCK	0.37	0.27

top 3 rank order inference accuracy within each faction in Table 2. The table rows represent leadership model datasets. The columns are accuracy, which determined by the percentage of top-3 individuals from the ground truth appear in the list of top-3 inferred list. Even though mFLICA has a competitive results, the FLOCK framework performs better, which makes sense since the hierarchical model has a linear hierarchy structure and the leader is always in the front of the group's direction, which is exactly matched with the FLOCK assumption.

6.2 Case study: trained leaders in fish schools

We considered that any fish follows trained fish if it is in the faction of trained fish. Among 24 trails of fish movement, the medians of inference accuracy whether any fish follows trained fish are at 2st column in Table 3. We measured the accuracy of inference whether initiators are trained fish in each trial (3nd column in Table 3). According to the result in Table 3, mFLICA performs significantly better than FLOCK in both aspects. This is because the fish datasets are tremendously noisy, and the DTW in mFLICA is more robust to the noise than the simple FLOCK model [16].

6.3 Case study: detecting the group merging event of baboons We used a baboon dataset (see Section 4.3.1) to demonstrate an example application of our framework. The framework can be used to find transitions of coordinated events in real datasets. We deployed the dataset in the period that the merging of two groups happening at Aug 3, 2012, 08:49:01 AM. The length of trajectories are 500 seconds. Figure 3 illustrates the result when merging happens. Before a time $t = 300$, a faction leading by ID(3) (black node) start moving toward the same direction as a faction leading by ID(18) (purple node). This situation can be observed by a faction size ratios of both factions that increase over time. After $t = 300$, ID(3) faction is merging with ID(18)'s faction to be a complete single faction at $t = 400$. After merging, because the faction of ID(18) gains more members, its faction size ratio (Eq. 3.3) increases. Hence, by observing faction size ratios leading by each individuals, we can find merging



(b) $t = 300$ (c) $t = 350$ (d) $t = 400$

Figure 3: The merging coordination event. (top) Faction size ratios (Eq. 3.3) of ID(3) and ID(18) factions. (Bottom) The GPS locations of individuals in the map over three different time steps ($t = 300, 350, 400$), with the 'following' network, and PageRank indicated by node size. ID(3) is black and ID(18) is purple. The red edges have higher edge weights than the light edges.

events (or even spiting events).

7 Discussion

In this paper, we formalized the Faction Initiator Inference Problem and provided an end-to-end general, unsupervised framework as the novel solution, which is competitive and can be used to study any coordinated activities. The framework is competitive against the baseline method in both simulated and real-world datasets. Moreover, we demonstrated that the framework can be used to identify the baboon groups' merging events as well as factions and initiators at each time step. This example implies that our framework opens opportunities for scientists to ask questions about coordinated activities and be able to create scientific hypotheses and test them. Our framework is powerful and almost parameter free (we need only a similarity threshold σ and time shift δ parameter). The scalability bottleneck is the DTW method used to compare time series. The existing DTW lower/upper bound techniques cannot be applied directly in our case since they only compute the distance between time series and not the actual wrapping path needed in our framework. With simpler and faster similarity computation, our framework can become highly computationally scalable. In future, such more scalable approaches should be investigated. Another future work we plan to explore is causality inference, which is closely related to leadership inference in the sense that initiators cause their followers' actions. We are planning to report the Granger causality results for leadership inference in our next paper. The code, datasets, and supplementary file that we used in this paper can be found at [1].

References

[1] mFLICA: code and supplementary. `https://github.com/CompBioUIC/MFLICA`. Accessed: 2017-12-19.

[2] C. Amornbunchornvej, I. Brugere, A. Strandburg-Peshkin, D. Farine, M. C. Crofoot, and T. Y. Berger-Wolf. Flica: A framework for leader identification in coordinated activity. *arXiv preprint arXiv:1603.01570*, 2016.

[3] M. Andersson, J. Gudmundsson, P. Laube, and T. Wolle. Reporting leaders and followers among trajectories of moving point objects. *GeoInformatica*, 12(4):497–528, 2008.

[4] A. Y. Carmi, L. Mihaylova, F. Septier, S. K. Pang, P. Gurfil, and S. J. Godsill. Inferring leadership from group dynamics using markov chain monte carlo methods. In *Modeling, Simulation and Visual Analysis of Crowds*, pages 325–346. Springer, 2013.

[5] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513–516, 2005.

[6] M. C. Crofoot, R. W. Kays, and M. Wikelski. Data from: Shared decision-making drives collective movement in wild baboons, 2015.

[7] J. R. Dyer, A. Johansson, D. Helbing, I. D. Couzin, and J. Krause. Leadership, consensus decision making and collective behaviour in humans. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1518):781–789, 2009.

[8] L. Glowacki and C. von Rueden. Leadership solves collective action problems in small-scale societies. *Phil. Trans. R. Soc. B*, 370(1683):20150010, 2015.

[9] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 499–508. ACM, 2008.

[10] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

[11] X. He and D. Kempe. Robust influence maximization. In *Proceedings of the ninth ACM SIGKDD*, pages 1–10. ACM, 2016.

[12] M. A. Hogg. A social identity theory of leadership. *Personality and social psychology review*, 5(3):184–200, 2001.

[13] P. Holme. *Temporal networks*. Springer, 2014.

[14] D. M. Jacoby, Y. P. Papastamatiou, and R. Freeman. Inferring animal social networks and leadership: applications for passive monitoring arrays. *Journal of The Royal Society Interface*, 13(124):20160676, 2016.

[15] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD*, pages 137–146. ACM, 2003.

[16] M. B. Kjargaard, H. Blunck, M. Wustenberg, K. Gronbask, M. Wirz, D. Roggen, and G. Troster. Time-lag method for detecting following and leadership behavior of pedestrians from mobile sensing data. In *Proceedings of the IEEE PerCom*, pages 56–64. IEEE, 2013.

[17] J. Krause, D. Hoare, S. Krause, C. Hemelrijk, and D. Rubenstein. Leadership in fish shoals. *Fish and Fisheries*, 1(1):82–89, 2000.

[18] T. W. Malone and K. Crowston. The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)*, 26(1):87–119, 1994.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[20] H. Pham and C. Shahabi. Spatial influence - measuring followship in the real world. In *ICDE16*, pages 529–540, May 2016.

[21] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[22] J. E. Smith, J. R. Estrada, H. R. Richards, S. E. Dawes, K. Mitsos, and K. E. Holekamp. Collective movements, leadership and consensus costs at reunions in spotted hyaenas. *Animal Behaviour*, 105:187–200, 2015.

[23] A. Strandburg-Peshkin and et al. Visual sensory networks and effective information transfer in animal groups. *Current Biology*, 23(17):R709–R711, 2013.

[24] A. Strandburg-Peshkin, D. R. Farine, I. D. Couzin, and M. C. Crofoot. Shared decision-making drives collective movement in wild baboons. *Science*, 348(6241):1358–1361, 2015.

[25] S. Stueckle and D. Zinner. To follow or not to follow: decision making and leadership during the morning departure in chacma baboons. *Animal Behaviour*, 75(6):1995–2004, 2008.

[26] T. E. Will. Flock leadership: Understanding and influencing emergent collective behavior. *The Leadership Quarterly*, 27(2):261–279, 2016.

[27] S. Wu and Q. Sun. Computer simulation of leadership, consensus decision making and collective behaviour in humans. *PloS one*, 9(1):e80680, 2014.

[28] C.-H. Yu, J. Werfel, and R. Nagpal. Collective decision-making in multi-agent systems by implicit leadership. In *AAMAS'10*, pages 1189–1196, May 2010.