Coordination Event Detection and Initiator Identification in Time Series Data

Chainarong Amornbunchornvej
University of Illinois at Chicago
camorn2@uic.edu

Ivan Brugere
University of Illinois at Chicago
ibruge2@uic.edu

Ariana Strandburg-Peshkin
Princeton University
astrandb@princeton.edu

Damien R. Farine
Max Planck Institute for Ornithology
dfarine@orn.mpg.de

Margaret C. Crofoot
University of California, Davis
mccrofoot@ucdavis.edu

Tanya Y. Berger-Wolf
University of Illinois at Chicago
tanyabw@uic.edu

ABSTRACT

Behavior initiation is a form of leadership and is an important aspect of social organization that affects the processes of group formation, dynamics, and decision-making in human societies and other social animal species. In this work, we formalize the Coordination Initiator Inference Problem and propose a simple yet powerful framework for extracting periods of coordinated activity and determining individuals who initiated this coordination, based solely on the activity of individuals within a group during those periods. The proposed approach, given arbitrary individual time series, automatically (1) identifies times of coordinated group activity, (2) determines the identities of initiators of those activities, and (3) classifies the likely mechanism by which the group coordination occurred, all of which are novel computational tasks. We demonstrate our framework on both simulated and real-world data: trajectories tracking of animals as well as stock market data. Our method is competitive with existing global leadership inference methods but provides the first approaches for local leadership and coordination mechanism classification. Our results are consistent with ground-truthed biological data and the framework finds many known events in financial data which are not otherwise reflected in the aggregate NASDAQ index. Our method is easily generalizable to any coordinated time-series data from interacting entities.

CCS CONCEPTS

•**Information systems** → *Spatial-temporal systems; Data mining;* •**Applied computing** → *Sociology;* Economics;

KEYWORDS

leadership, spatio-temporal data, collective behavior, influence, social networks, coordination, initiation

1 INTRODUCTION

Who is the trend-setter whose opinion many follow at the moment? Which zebra initiated the flight from a lion? Whom does the elephant herd follow to water? In all these scenarios, the initiator might not be the one who is speaking the loudest or positioned at the front of the group *after* the group has already agreed to follow [11, 30]. Thus, in order to identify those initiators or trend-setters, we must also determine the moment of the group's decision to follow.

MILETS'17, Halifax, Nova Scotia, Canada

DOI: 10.475/123_4

Coordination Initiator Inference Problem: An agreement of a group to follow a common purpose is manifested by its coalescence into a *coordinated* behavior. The process of initiating this behavior and the period of decision-making by the group members necessarily precedes the coordinated behavior. **Given time series of group members' behavior, the goal is to find these periods of decision-making and identify the initiating individual, if one exists.**

Initiating a group's behavior is a form of leadership [33, 37]. Leadership is an important aspect of the social organization, formation, and decision-making of groups of people in online and offline communities, as well as other social animals. Understanding the dynamics of emerging leadership allows researchers to gain insights into how social species make decisions. Until recently, many works defined leaders by their physical or behavioral characteristics rather than by observing processes of interaction [24].

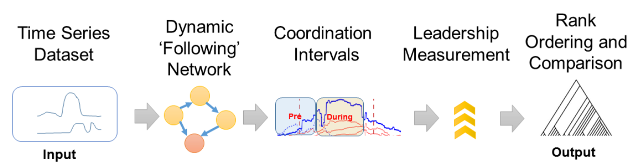


Figure 1: A high-level overview of the proposed framework

The availability of data from physical proximity sensors, GPS, and the web opens up the possibility of measuring leadership as the process of initiation in online activities, face-to-face human interactions, animal populations, and aggregate social processes such as economic activity. This paper presents the new computational problem of inferring leader identity in the context of successful initiation of coordinated activities among groups of individuals or other entities, as well as proposes the first automated method for unsupervised leader identification. The method uses only time series activity data of entities, with no additional information. The proposed approach automatically determines (1) the time interval of group coordination, (2) the time when the (possibly implicit) decision for that coordinated activity was made, (3) the identity of the coordination initiator, and (4) the mechanism by which the group came to follow the initiator.

1.1 Related work

Coordinating patterns of individual activity is a challenge that all social organisms face. Diverse strategies–from democratic to dictatorial–have emerged to allow members of groups to reach consensus [7]. Leadership plays a key role in organizing the collective (i.e. group) behaviors of social organisms ranging from humans [11] to hymenoptera [36]. It potentiates complex patterns of cooperation and conflict (e.g., lions [16], hyaenas [4], meerkat [23], chimpanzees [12], humans [13]), organizes group movements [8, 11], and may prevent free-riding [18].

In the context of group behavior and decision-making in biology and sociology, leaders are individuals who successfully induce a group of others to follow them to a common goal, state, or behavior [8, 25, 33, 37]. Biological studies showed that leaders may be context-specific [8, 29] and the important initiators of particular group activities are not necessarily the individuals found at the top of their group's social dominance hierarchy [5, 31].

Substantial interest currently exists in identifying leaders and determining how they influence the behavior of others in their social environment. Previous work in several domains defined leadership according to physical characteristics (*e.g.,* size, sex [37]), positions in location-based social networks [26], rule-based models [29], physical trajectory and association patterns [2, 22].

Computationally, most previous work uses a global notion of leadership and creates a global, static leadership ranking over the entirety of the input data [3, 14]. Other domain-specific methods infer leadership from implicit pairwise dyadic dominance or leader-follower interactions [2, 21, 26]. Some methods define an explicit network over the dyadic interactions or use a known network topology [35] and use network measures, such as PageRank and HITS, or cascade size to identify leaders [3].

Leadership has also been studied in explicit social network settings. From a social network perspective, leaders can be characterized as influential individuals who have many followers that imitate the leader's actions [14], and, thus, successfully take a group from one behavioral state to another. Much of the computational work has focused on the problem of influence maximization (IM)–i.e. how individuals are able to maximize their impact on the behavior of the group as a whole by iteratively affecting local network neighborhoods [15, 19]. This approach assumes that the network structure is known.

There is a clear gap between the biosociological definitions of leadership in group decision-making and the existing computational approaches. Currently, there are no computational approaches that (1) view leaders as initiators of group behavior change, which can (2) identify the timing of the process of the change initiation and the group's decision-making in (3) arbitrary contexts, under (4) a variety of leadership models.

1.2 Our contributions

In this paper, we focus on the definition of leadership as the initiation of coordinated activities. We aim to close the gap between the biosociological view of the role of leaders in group decision-making, the computational formalism, and methodology.

Therefore, the first part of our contribution is establishing and formalizing this **new computational problem of coordination initiation inference**. We call it the COORDINATION INITIATOR INFERENCE PROBLEM. Our formulation is a **generalization** of many related leadership and initiation inference computational problems. We explicitly relate existing leadership and influence propagation problems as special cases of our formulation. The new formulation uses only the time series of individual behavior as input, with no assumption of additional information such as demography, prior history, dominance hierarchy, or a network structure. The problem formulation aims to identify different local instances of behavior initiation, allows the identity of the initiator to be instance-specific, and makes no assumption on the leadership or behavioral model.

Our additional contribution is in proposing a computational **solution framework** to this new COORDINATION INITIATOR INFERENCE PROBLEM. We propose a general, scientifically grounded, unsupervised, and extendable framework with few assumptions for identifying individuals who lead a group to a state of coordinated activity (or, more generally, an entity that induces group coalescence). Our framework is capable of:

- **Detecting coordinated activity events:** discovering coordination intervals and decision-making periods leading to that coordination;
- **Identifying initiators:** identifying the initiators of this coordinated behavior, that is, the individuals who succeeded in leading the group to coordination, specifically locally to each coordination instance; and
- **Classifying the group coordination model:** characterizing the type of the group's transition behavior to coordination according to interpretable, dynamic models.

We demonstrate the framework's ability to analyze leadership in coordinated activity on synthetic and real datasets over several domains. We compare our framework with state-of-the-art methods for leadership identification for the special cases of our problem where such methods are applicable. For many instances of our new problem, there are no existing methods. We demonstrate that existing solutions fail and do not extend to these instances. We use synthetic simulated data to validate each aspect of the framework. We analyze two biological datasets – GPS tracks of a baboon troop and video-tracking of fish schools, – as well as stock market closing price data of the NASDAQ index. The results are consistent with ground-truthed biological data. Moreover, the framework finds many known events in financial data, which are not otherwise reflected in the aggregate NASDAQ index. Our approach is easily generalizable to any coordinated activity in time series data of interacting entities.

2 PROBLEM FORMALIZATION

Given a collection of time series, we want to find initiators of highly coordinated patterns. To formally state the COORDINATION INITIATOR INFERENCE PROBLEM, we need to formalize notions of "coordination" and "initiation."

First, we define an intuitive notion of a FOLLOWING RELATION, as "two individuals performing the same sequence of actions (or generating time series values) with some fixed delay." Formally:

*Definition 2.1 (*FOLLOWING RELATION). Let $U = (\vec{u}_1, \dots, \vec{u}_t, \dots)$ and $W = (\vec{w}_1, \dots, \vec{w}_t, \dots)$ be m-dimensional, arbitrary-length time series. If for all $t \in \mathbb{N}$, there is a fixed 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 $W \preceq U$. We denote $W \prec U$ if $\Delta t > 0$.

Lemma 2.2. *Let U and W be time series such that $W \preceq U$ and $U \preceq W$, then U and W are equivalent time series denoted $U \equiv W$.*

Proof. There are two cases when both $W \preceq U$ and $U \preceq W$. First, $W = U$ and $U = W$ (that is, $\Delta t = 0$ in both following relations). Clearly, $W \equiv U$. Second, $W \prec U$ with $\Delta t_w > 0$ and $U \prec W$ with $\Delta t_u > 0$. Then, by definition, $\vec{w}_t = \vec{u}_{t+\Delta t_w}$ and $\vec{u}_t = \vec{w}_{t+\Delta t_u}$. Therefore, $\vec{w}_t = \vec{w}_{t+\Delta t_w+\Delta t_u}$. Thus, if $W \preceq U$ and $U \preceq W$ then $W \preceq W$ (and similarly $U \preceq U$). Thus, the two time series are identical periodic with a different starting point and therefore equivalent.

□

Lemma 2.3. *The* following relation *is a partial order over time series [10].*

Proof. Antisymmetry: if $W \preceq U$ and $U \preceq W$, then $W \equiv U$ by Lemma 2.2. The following relation is also trivially reflexive and transitive, which, by definition is a partial order.

□

Next, coordination, or intuitively "all individuals performing the same sequence of actions, at possibly varying delays," is formally defined as:

*Definition 2.4 (*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 $U_i, U_j \in \mathcal{U}$, 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 $t \in [t_1, t_2]$.

Finally, the initiator is intuitively "an individual who first performs a sequence of actions, and all other individuals follow," formally defined as:

*Definition 2.5 (*Initiator). Let $\mathcal{U} = \{U_1, \ldots, U_n\}$ be a coordinated set of 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 $U \in \mathcal{U} \setminus \{L\}$, $L \prec U$.

We are now ready to precisely state the problem of identifying the individual who initiates a coordinated behavior:

Problem 1: Coordination Initiator Inference Problem

Input : Set $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ of time series.

Output: A coordination interval $[t_1, t_2]$ and the initiator time series $L \in \mathcal{U}$ that initiated the coordination.

2.1 Useful observations

Let $\mathcal{U}$ be a coordinated set of time series and $L \in \mathcal{U}$ be the initiator. Since $\mathcal{U}$ is a partial order set and $\forall U_i \in \mathcal{U}$, $L \prec U_i$, then, by definition, L is the minimal element. Moreover, $\mathcal{U}$ is a linear order set since for every pair $U_i, U_j \in \mathcal{U}$, either $U_i \prec U_j$ or $U_j \prec U_i$.

*Definition 2.6 (*Following network). Let $\mathcal{U} = \{U_1, \ldots, U_n\}$ be a set of time series. The *following network* $G = (V, E)$ is defined as a directed graph where the set of nodes V has a one-to-one correspondence to the set of time series $\mathcal{U}$, and each edge in E represents a following relation between two time series: $\forall U_i, U_j \in \mathcal{U}$ the edge $e_{i,j} \in E$ if $U_j \prec U_i$.

Recall that PageRank [6] score, π_i, of a node i in a network G is defined as follows:

$$\pi_i = d \sum_{k \in \mathcal{N}_i} e_{k,i} \pi_k / |\mathcal{N}_i| + (1 - d) \quad (1)$$

Where $\pi_i \in [0, 1]$, $d \in (0, 1]$ is a constant number, $e_{k,i} \in \{0, 1\}$ is one if $e_{k,i} \in E$, and $\mathcal{N}_i$ is a set of neighbor nodes of i such that $k \in \mathcal{N}_i$ if $e_{k,i} \in E$.

Lemma 2.7. *Let $G = (V, E)$ be a following network of time series set $\mathcal{U} = \{U_1, \ldots, U_n\}$. If $U_i \preceq U_j$ then $\pi_i \geq \pi_j$.*

Proof. By transitivity, if U_j follows U_i then the followers of U_j are also the followers of U_i. Thus, since $\forall k \in \mathcal{N}_j$, $U_j \preceq U_k$ and $U_i \preceq U_j$, then $\mathcal{N}_j \subseteq \mathcal{N}_i$. Hence, $\pi_i - \pi_j = d \sum_{k \in \mathcal{N}_i \setminus \mathcal{N}_j} e_{k,i} \pi_k \geq 0$. □

As a corollary of Lemma 2.7, since all the time series follow the initiator L within the coordination interval $[t_1, t_2]$, then L has the highest PageRank score in $\mathcal{U}$ during that coordination period. Moreover, Lemma 2.7 allows us to infer the order of following among the time series within the coordination period, as defined by the PageRank values.

3 METHODS

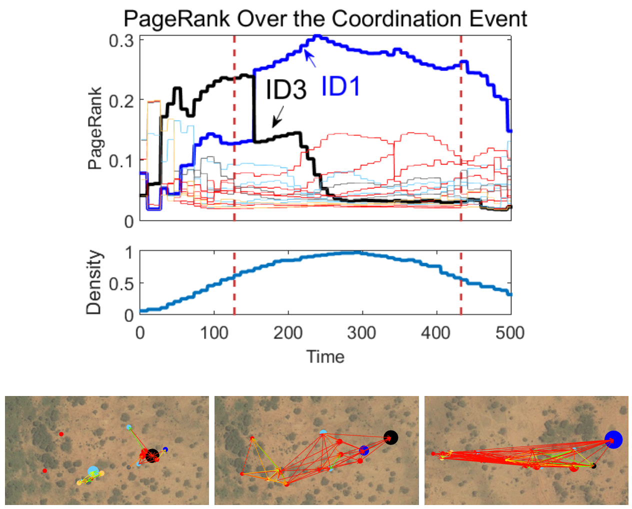
In this section, we present a Framework for Leader Identification in Coordinated Activity (FLICA) as the solution for the Coordination Initiator Inference Problem. On real data, the above formalization is very restrictive, so we relax the *exact* following relation, and *full* coordination to identify 'following' and partial 'coordination' in real applications.[1] Furthermore, multiple coordination events often exist within a set of real time series data. Constructing a single aggregated network would not capture the dynamics of these events. Therefore, FLICA uses a dynamic network approach.

Figure 1 shows the framework overview. At each time step, we infer following relations to construct a sequence of following networks. We then use network density to detect intervals of coordination, and the time series of PageRank values to identify the initiators of these coordination intervals.

3.1 A working example

Figure 2 presents a key example and a brief introduction to our framework, on real GPS trajectory data of olive baboons (*Papio anubis*). Figures 2(b)-2(c) show the leadership of movement of the group by baboon ID3 (Black). Figure 2(d) shows the 'following' network in the coordination interval. Individual ID3 has the largest PageRank in the first two snapshots but the PageRank of individual ID1 (Blue) surpasses ID3 when the network is 'coordinated' (e.g.

[1] However, FLICA using PageRank necessarily provides an exact solution to the Coordination Initiator Inference Problem.



(b) t=50 (c) t=100 (d) t=250

Figure 2: PageRank (top) and density (middle) of the 'following' network over time for an event of baboons' movement which initiates by ID3. (Bottom) The locations of individuals over three different time steps (t = 50, 100, 250), with the 'following' network, and PageRank indicated by node size.

moving together). If we measure the initiator ranking *after* the network has coalesced, then we miss that ID3 initiated coordination and 'built' the network in the pre-coordination interval (to the left of the first dotted red line).

3.2 Following relation inference

For a (U, Q) pair of time series, we use Dynamic Time Warping (DTW) [27] to measure whether U follows Q. DTW is shown to perform better than several other methods in inferring following relation in time series [21] and it is tolerant to noise [28]. Let $P_{U,Q}$ be a sequence of index pairs (i, j) which comprise the DTW optimal warping path of (U, Q). We compute the mean of the signum difference over this sequence of index pairs:

$$\mathrm{s}(P_{U,Q}) = \frac{\sum_{(i,j)\in P_{U,Q}} \mathrm{sign}(j-i)}{|P_{U,Q}|} \quad (2)$$

This function measures the extent of warping between two time series. If time series cannot be shifted one-onto-the-other with a consistent positive or negative sign, $|\mathrm{s}(P_{U,Q})| \approx 0$, then there is no following relation between U and Q. When $\mathrm{s}(P_{U,Q})$ is positive, Q *follows* U, otherwise, U *follows* Q.

3.3 Dynamic following network inference

The set of n m-multidimensional time series $\mathcal{D}$ (*e.g.,* a matrix of size $[n \times m \times t^*]$), a window size parameter ω, and a window shift parameter δ (default is 0.1ω) are the inputs for our framework.

Let the ith time interval be given by: $w(i) = [i \times \delta, i \times \delta + \omega]$. For each $w(i)$, we extract a set of sub time series $\mathcal{Q}_i$ from $\mathcal{D}$. The $\mathcal{Q}_i$ is the $[n \times m \times \omega]$ dimensional matrix of the time series set. Then we construct a following network $G = (V, E)$ as defined in Definition 2.6. The nodes represent the time series from $\mathcal{Q}_i$ and E is a set of edges between time series nodes such that if $U, W \in \mathcal{Q}_i$ and U follows W according to Eq. 2, then $e_{U,W} \in E$ with the edge weight $|\mathrm{s}(P_{U,W})|$. We calculate a following network for each $w(i)$ to construct a dynamic following network $G^* = (V, E^*)$. The pseudo code is given in Procedure 1.

Procedure 1: CreateDyFollowingNetwork

input : A time series set $\mathcal{D}$, ω, and δ
output: A $n \times n \times t^*$ adjacency matrix E^*.

$K \leftarrow (t^* - \omega)/\delta$;
for $i \leftarrow 1$ **to** K **do**
  /* current time interval */
  $w(i) = [(i-1) \times \delta, (i-1) \times \delta + \omega]$;
  /* SubTimeSeries($\mathcal{D}$, $w(i)$) returns all sub time series in $\mathcal{D}$ within the interval $w(i)$ */
  $\mathcal{Q}_i \leftarrow$SubTimeSeries($\mathcal{D}$, $w(i)$);
  $E \leftarrow$CreateFollowingNetwork($\mathcal{Q}_i$) ;
  /* Set all edges within the time interval $[(i-1) \times \delta, i \times \delta]$ to be similar */
  $E^*_{t\in[(i-1)\times\delta,\, i\times\delta]} \leftarrow E$;
end
$\mathcal{Q} \leftarrow$SubTimeSeries($\mathcal{D}$, $[K \times \delta, t^*]$);
$E \leftarrow$CreateFollowingNetwork($\mathcal{Q}$) ;
$E^*_{t\in[K\times\delta,\, t^*]} \leftarrow E$;

3.4 Coordination intervals detection

Note that *network density* of the following network serves as the measure of the extent of coordination over all time series pairs (by Definition 2.4, during the coordination interval *every* pair has a following relation.) We can use this observation to identify times of *approximate coordination*.

Given a time series of network densities, denoted by d, over a dynamic following network G^*, and a density threshold parameter λ, the time interval $[t_i, t_j]$ is a λ-coordination interval if $\mathrm{d}(t) > \lambda$ for *all* $t \in [t_i, t_j]$. The *pre-coordination interval* of coordination $[t_i, t_j]$ is the interval $[t_k, t_i - 1]$, where the discrete derivative $\mathrm{d}(t) - \mathrm{d}(t-1) \geq 0$ for all $t \in [t_k, t_i - 1]$. Together, these intervals are one *coordination event*, represented by the 3-tuple of time indices $I = (t_k, t_i, t_j)$. The collection of coordination events is a set $C = \{I_l\}$. All complete event intervals $[t_k, t_j]$ are mutually disjoint in C, and $|C|$ denotes the total number of 3-tuples. For the remainder of our framework, we measure local leadership only on these events in the set C. To reduce the number of intervals generated near the threshold λ, we apply a greedy merging of nearby coordination intervals (taking the range from the window size ω).

3.5 Ranking comparison

On each coordination event $I = (t_k, t_i, t_j)$, let R_I be some ranking of individuals within the pre-coordination interval $[t_k, t_i - 1]$. We focus on ranking within pre-coordination because this is the interval where coordination is initiated. The *global* rank order, denoted by $\hat{R}$, is the average of all R_I where $I \in C$.

We measure **initiator ranking** according to three different methods: PageRank [6], velocity convex hull (VCH), and position convex hull (PCH). Recall, that by Lemma 2.7, if U follows V, then the PageRank of U is less than that of V. Thus, the initiator is expected to have the highest PageRank. VCH measures how often an individual moves faster than others. It represents a model of leadership for movement. This model can be found in many social species [11, 33]. PCH measures how often an individual moves to an area before others. For example, in a flock model [2], a leader is positioned at the front of the group's trajectory.

3.6 Leadership model features

Let the global rank ordering for PageRank be denoted by $\hat{R}_{pr}$, for VCH by $\hat{R}_v$, and for PCH by $\hat{R}_p$. To measure global leadership in our framework, we order individual nodes i based on initiation support with respect to one of these ranking methods, $R_{\bullet,I}$, over all coordination events $I \in C$. For example, $R_{pr,I}$ is PageRank-rank-ordered list at the event I. If an individual i is at 1st rank at I, then $(i,1) \in R_{pr,I}$; i is an initiator. The initiation support for a node i is the fraction of coordination events at which it was ranked 1 (by a ranking measure $R_\bullet$):

$$\text{sup}_\bullet(i) = \frac{\left|\{\forall I \in C \; (i,1) \in R_{\bullet,I}\}\right|}{|C|} \tag{3}$$

We use the Kendall rank correlation coefficient $\tau()$ [20] to compare event-local and global rank-orders. To compare global and local rank orders, we use the mean Kendall rank correlation over all coordination events against the global: $\text{corr}_\bullet = \frac{\sum_{I \in C} \tau(\hat{R}_\bullet, R_{\bullet,I})}{|C|}$. For example, corr_v compares local and global velocity convex hull rank orders.

Similarly, we compute the mean Kendall correlation between local rankings associated with different measures (e.g. VCH, PCH): $\text{corr}_{\bullet,\bullet} = \frac{\sum_{I \in C} \tau(R_{\bullet,I}, R_{\bullet,I})}{|C|}$.

$\text{corr}_\bullet$ formalizes our intuition that leaders consistently move outside of the spatial extent (corr_p), or the distribution of velocity over the population (corr_v). By comparing the global vs. local correlation in rank ordering, we measure the stability of the global ranking is over time.

$\text{corr}_{\bullet,\bullet}$ measures the relationship between higher-order graph structure and simple time series features. Using this measure, we can gain a better understanding of the high-level aspects of initiating coordination. For example, we see whether changing velocity ($\text{corr}_{v,pr}$), or position ($\text{corr}_{p,pr}$) within the group is correlated with network rank position.

4 EXPERIMENTAL SETUP

We evaluate our framework on eight synthetic movement trajectory models and three real datasets.

4.1 Simulation models

4.1.1 *Dictatorship model (DM).*
In this model, we fix a single initiator who initiates movement from initial positions of the population. At the start of the pre-coordination interval, the initiator moves in a fixed direction and acceleration. Other individuals wait for a randomly sampled lag, before following the initiator at a fixed acceleration (with sampled noise in the heading). After a fixed duration of coordinated movement over the entire population, individuals decelerate at random, until stopping. The Switching Dictatorship model (DM-S) selects two fixed individuals over each trial: a single individual as an initiator during pre-coordination, and another single individual as 'initiator' during coordination.

4.1.2 *Hierarchical model (HM).*
This model is a variation of DM, where we fix a number of individuals (n=4) to follow the previous individual in the sequence, after a sampled lag. The remainder of individuals in the population follow exactly one of these high-ranking individuals, allocated in decreasing proportion per rank. The Switching Hierarchical model (HM-S), similarly to DM-S, selects unique pairs of individuals for each hierarchy level, switching after the pre-coordination interval as in DM-S.

4.1.3 *Event-based model (EM).*
This model is a variation of the Dictatorship model where each coordination event has a different, unique initiator. For example, in one of our applications, a troop of baboons may follow an initiator to food in the morning, and follow a different initiator in the evening to the sleeping site. No existing methods can infer these two situations except our framework.

4.1.4 *Initiator model (INIT-k).*
In this model, we fix k initiators who initiate movement from random initial positions of the population. At the start of the pre-coordination interval, all initiators move on a single target. Non-initiators move in randomly sampled directions with a fix velocity, then follow their initiator after a random time lag. After the pre-coordination period, all individuals move toward a single target, without following their initiator. We run simulations for INIT-1 and INIT-4 initiator models.

4.1.5 *Crowd model (CM).*
This model [29] is a collective movement model where k (=4) informed individuals move toward a target, and the remaining (=16) uninformed individuals move in a linear combination of a direction toward the group's centroid, and the average direction of the group.

4.1.6 *Linear Threshold model (LT).*
This model [19] initiates individual movement by propagation of a linear threshold process on the dynamic network, defined by the k-nearest neighbors at the current time-step. The model is parameterized by ρ, the proportion of these k neighbors required to be infected in order to initiate movement. Once activated, the individual follows a single initiator. The initial probability of activation for each individual is 0.5. We explore the parameter space on combinations of: $k \in \{3,5,10\}$ and $\rho \in \{0.25, 0.50, 0.75\}$.

4.1.7 *Independent Cascade model (IC).*
This model [19] is another propagation process similar to LT. At each time step, each active individual moves toward the initiator and independently attempts to activate its k-nearest neighbors with the probability of ρ. If the individual fails to activate a neighbor, it cannot attempt to activate the same neighbor again. We explore the same sample parameter space as in the LT model.

4.1.8 *Random model.*
In this model, there is no 'following' relations. At the start of the pre-coordination interval, all individuals start moving to a fixed direction, independently of others in the

population. We expect the relative positions of individuals to yield some following relations only by chance.

4.2 Synthetic trajectory simulation

For each of the above models, we generate a trial of synthetic data consisting of 20 individuals, and 20 separate coordination events, for a total of 12,000 total time-steps. Each coordination event has pre-coordination and coordination intervals of 200 time-steps each. Following the coordination interval is another 200 time steps of a post-coordination before repeating. We generate 100 trials for each of models. In total, we have 2,700 simulation datasets.

4.3 Real datasets

4.3.1 Baboon trajectories. High-resolution GPS collars track 26 individuals of a troop of olive baboons (*Papio anubis*) living in the wild in Mpala Research Centre, Kenya [9, 31]. The data consists of latitude-longitude location pairs for each individual at one observation per second. We analyze a subset of 16 individuals whose collars remained functional for a nine day period (419,095 time steps).

4.3.2 Fish schools trajectories. The movement of a fish school of golden shiners (*Notemigonus crysoleucas*) are recorded by video in order to study information propagation over the visual fields of fish [32]. Each population contains 70 fish, with 10 trained, labeled fish who are able to lead the school to feeding sites over 24 separate coordination events. The task is to correctly identify trained fish by initiator ranking.

4.3.3 Stock closing-price time series. We collected daily closing price data for stocks listed in NASDAQ, using Yahoo! Finance.[2] These time series are from January 2000 to January 2016 (4169 time-steps). We remove symbols with a large amount of missing data, leaving a total of 1443 symbols in our dataset. Our analysis focuses on discovering large, known events and crises in an unsupervised way, and to explore initiators and sectors involved in these coordination events.

4.4 Evaluation

For synthetic datasets, we use three evaluation approaches:

- **Global leadership**: For each method, we extract network and/or rank statistics over the entire time series, and report only a single aggregate initiator ranking. We compare the known ground truth ranking (used to generate the data) against the ranking of each method, reporting precision. We measure precision of identifying the true initiator, on DM, LT, IC, and INIT-1 models. For the HM model, we compare the *exact* top-4 ranking against the ground truth (order matters); The evaluation is the same for CM and INIT-4 models, except the *exact* top-4 ranking constraint is relaxed (i.e. we compare top-4 *sets*).
- **Local leadership**: For evaluation data in this case, we use the ground truth ranking for each local coordination event, and the time intervals of each event. We report average precision over each discovered pre-coordination interval. We evaluate the EM model using this approach. We report only the FLICA result, since it is the only method capable of producing local ranking.
- **Initiator leadership**: For each coordination event, we measure the initiator of coordination event before coordination occurs (e.g. in the pre-coordination interval). This individual may not be highly ranked after coordination (see: Figure 2). We report the precision of global leadership considering only the pre-coordination intervals. Since only FLICA identifies pre-coordination intervals, we compare against other methods' global leadership. This evaluation demonstrates that global leadership is distinct from coordination initiation. We evaluate DM-S and HM-S models using this approach.

4.5 Compared leadership methods

Table 1: Leadership inference methods

Method	Input	Time complexity
FLICA	Time series	$O(n^2 \times t \times \omega)$
FLOCK [2]	Trajectory	$O(n^2 \times t)$
LPD [21]	Time series	$O(n^2 \times t \times \omega^3)$
IM [19]	Network	$O(n^2 \times t)$

We demonstrate the performance of our framework by comparing with previous work on influence and leadership [2, 19, 21]. These methods can infer only global initiator ranking, while our proposed framework (FLICA) can detect individual coordination events, handles switching initiator, and performs leadership model classification. Therefore, we use the global leadership identification task to compare FLICA's performance with the prior works. We report the best results under varying parameters for competing models. The time complexity of each method is shown in Table 1.

First, the FLOCK model [2] identifies leaders who influence the norm direction vector of the group. Second, LPD [21] creates an aggregate 'following' network from time-lag features. A node is scored by breadth-first traversal on reversed 'following' edges. Visited neighbors' contribution is inverse-proportional to the geodesic distance. For the purposes of our simulation, we use sliding Euclidean distance alignment (e.g. analogous to cross-correlation) because LPD does not scale to the size of our simulations under DTW (see: Table 1). Finally, for influence maximization (IM), we use the independent cascade model for the 1-seed selection problem [19], on the network derived from [2]. The network describes the probability of any individual A sharing the same direction as B, and in the front of B.

To make leadership comparison possible, we report the global leadership rank ordered list for each method as follows. First, we create rank order lists for FLICA under PageRank. The FLOCK model, however, does not have the explicit ranking score, so we rank individuals based on decreasing time duration of leadership. Third, LPD assigns individuals with higher scores a higher rank. Finally, since IM uses the probabilistic network of influence, we construct the realization of this influence network. A node influences any node to which it has a directed path in the realized network. We rank individuals based on the expectation of nodes influenced by that node over 1000 realized networks.

[2] http://finance.yahoo.com/

Table 2: Precision of leadership identification on simulation models. (* indicates the $std \geq 0.1$).

Models/Methods	FLICA	FLOCK	IM	LPD
DM	**1**	**1**	**1**	**1**
HM (Top-4)	**1**	0.25	**1**	**1**
LT	**0.99**	0.98*	0.99*	0.93*
IC	**1**	**1**	**1**	0.99
CM (Top-4)	**1**	**1**	**1**	0.99
INIT-1	**1**	**1**	**1**	**1**
INIT-4 (Top-4)	**0.74***	0.35*	0.51*	0.21*
DM-S	**1**	0	0.02*	0.25*
HM-S (Top-4)	**1**	0	0.5	0.51
EM	**0.92**	-	-	-
Random	0.01	0	0.01	0.17*

5 RESULTS

5.1 Identifying leaders

In each simulation, we have the label of the true initiator(s). For each of the simulation trials, our method identifies the 'initiator' and 'rank ordered lists' (see Section 4.4). We set a window size ω by the TWIN heuristic [34] on the network density, window shift size $\delta = 0.1\omega$, and the λ threshold at the mean of the network density time series $\mathrm{d}(t)$.

Table 2 reports precision on PageRank rank ordered lists over all synthetic model simulations. We compare against previous work, FLOCK [2], IM [19], and LPD [21] models which produce a single ranking over the entire trial.

The white rows in Table 2 report precision of leadership identification for a fixed initiator across all coordination events (global leadership). Gray rows report precision of initiator leadership where leaders change between pre-coordination and coordination intervals in the event (DM-S, HM-S), or precision of local leadership where the initiator changes per coordination event (EM). The rows labeled 'Top-4' report precision in identifying any of the multiple unordered initiators (CM, INIT-4) or precision *for the correct hierarchical order* (HM, HM-S).

On the white rows, FLICA is robust across all simulations, while FLOCK, IM, and LPD perform well other than on INIT-4 simulations (e.g. with multiple initiators). However, in gray rows ("initiator switching") previous methods fail almost completely since they are unable to detect leadership prior to coordination. When the coordination state is more prevalent than the pre-coordination decision-point, ranking will favor an individual who happens to lead the dynamics in the coordination state (but may not have initiated the state).

The row reporting EM results is a special case of precision. Because we know each coordination event has a unique initiator, ranking individuals across all coordination events will fail. Instead, we report precision in identifying the initiator of *each* coordination event. Since previous work generates only aggregate rankings, precision for these methods are not reported.

5.2 Case study: trained initiators in fish schools

We identify the top-k global initiators of the fish school trajectory dataset (see: Section 4.3.2), where we have the labels of 'trained' individuals expected to lead the school to feeding sites. Table 3 reports precision of identifying trained fish as initiators over 24 trials. The Initiator column is precision of predicting a trained fish as a global initiator. The Top-4 rank column is precision of identifying trained fish as the top-4 ranking individuals. Similar to the simulation models, FLICA performs best overall, again suggesting that dynamic following network representation captures 'following' better than other features.

Table 3: Initiator identification precision in fish (* indicates the $std \geq 0.1$).

Ranking	Initiator	Top-4 rank
FLICA	**0.83***	**0.61***
FLOCK [2]	0.0	0.0
IM [19]	0.0	0.02
LPD [21]	0.17*	0.18*

5.3 Case study: finding "initiators" of stock market events

We apply our leadership framework to stock market closing price data of the NASDAQ index. An 'initiator' in this context measures the extent that a stock increases or decreases in value before a large group of other stocks (e.g. a coordinated group). We apply the framework without any special consideration to the domain, only to qualitatively validate that we can discover known, large events.

Figure 3 shows the network density of the inferred 'following' network over time, where we discover coordination events with λ threshold at the 75th percentile of the network density time series. Pre-coordination and coordination intervals are shown in red and green, respectively. We find significant economic events such as the 2000 tech collapse, and 9/11. More interestingly, we discover known events which are reflected in the network density signal but not the NASDAQ index. For example, we discover a technical econometric event, where the "TED Spread" (a surrogate of national credit risk) begins fluctuating in July 2007, and a small market failure in August 2011. Matching our intuition, the top-ranked companies in the coordination event associated with the year 2000 collapse are primarily in IT and semiconductors, including eBay and SanDisk in the top 10.

5.4 Leadership model classification

Recall, that we proposed several initiator rankings and ranking correlations (Section 3.6). Here, we do leadership model classification on each simulation trial using the proposed features derived from the rank correlations: corr_p, corr_v, $\mathrm{corr}_{p,pr}$, $\mathrm{corr}_{v,pr}$ and sup_{pr}. A classifier takes these features and produces a leadership model label, one per trial of the simulation model in the evaluation hold-out. We use 10-fold cross validation on Random Forests [17] over

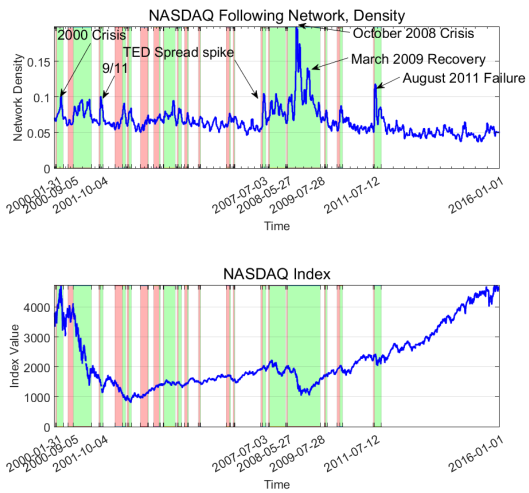


Figure 3: (Top) NASDAQ 'following' network density and (Bottom) NASDAQ index value. Pre-coordination and coordination intervals are shown in red and green, respectively. The framework detects many known events in financial data (labeled above). Many of these events are not reflected in the NASDAQ index.

the 2700 total trials and report mean precision and recall across folds. Table 4 reports the classification results for each simulation model. We combine some models into a shared label because they share similar characteristics when we project them into our feature spaces. For example, DM, and DM-S models always have high corr_v but low corr_p.

Figure 4 visualizes sub-spaces of the full feature-space. Figure 4 (Top) shows the maximum support (sup_{pr}) over all individuals for this trial vs. the corr_v (the rank correlation between global and local VCH ranking) and corr_p. The sup_{pr} axis (x-axis) describes how 'dictatorial' (e.g. consistent) the leadership is across coordination events. DM therefore has high support, while EM (distinct leaders per coordination event) has low sup_{pr}. The corr_v and corr_p axes describe consistency between local and global convex hull rankings. HM has high velocity ranking because leaders accelerate in a consistent sequence, yielding consistent individuals movement outside of the VCH in the previous time step. The random model produces high corr_p because relative positions within the group are somewhat consistent. Therefore, a consistent set of individuals expand the PCH from the previous time step.

Figure 4 (Bottom-Right) reports the mean rank correlation between PageRank rank ordering, against PCH and VCH ranking *in each coordination event.* At the origin (0, 0), ranking from the inferred 'following' network is uncorrelated with time series feature rankings in position or velocity. Following our intuition, the Random simulation has the lowest cross-domain feature correlation, while DM and HM have highest correlation between these domains. As the simplest simulations, DM and HM both dictate that leaders will have regular position (e.g. the front of the group), or velocity (accelerating in sequence before others). Simulations such as CM, LT, IC have indirect relationships between relative position and velocity vs. the following network ranking.

5.4.1 *Baboon leadership model characterization.*

A key aspect of our simulation modeling is that we can characterize real datasets according to how they map into these feature-spaces, compared to synthetic models. We compute each rank correlations over high-confidence baboon events, labeled "Baboon" in Figure 4, thresholded at the 99th percentile of density. We observe that within different sub-spaces, the baboon ranking is similar to Random or Linear Threshold, and has low maximum support for global vs. local rank correlation features (e.g. corr_p). We see this rank correlation between both cross-domain axes (Figure 4 (Bottom-Right)). This suggests that in aggregate, baboon leadership is heterogeneous and context-driven, though overall closer to the Linear Threshold influence model (as biologically expected). This analysis provides a strategy for hypothesis testing and generation on contrasting time-scales and sub-spaces.

Table 4: Random forest classification of synthetic leadership models using proposed features

Model	Precision	Recall	F score
DM, DM-S	0.86	0.80	0.81
HM, HM-S	0.69	0.98	0.80
LT, IC, INIT-k	0.99	0.97	0.98
CM	0.75	0.94	0.80
EM	1	0.54	0.64
Random	0.98	0.95	0.97

6 CONCLUSIONS

We narrow the gap between the biosociological view of leadership in group decision-making and the computational approaches to leadership inference. The work presented in this paper formalizes a **new computational problem**, namely COORDINATION INITIATOR INFERENCE PROBLEM, and proposes the concrete, simple yet powerful, unsupervised general framework as a solution. The framework is capable of (1) identifying events of coordinated group behavior, (2) identifying leaders as initiators of these events, and (3) classifying the type of leadership process at play. We validate the accuracy of our framework in performing all three of these tasks using 2,700 simulated datasets. Since there are no methods for local leadership inference and leadership model classification, we compared our framework with the state-of-art methods for global leadership identification. Our method performance is consistently competitive and its abilities go beyond other approaches in all datasets. We further show that the framework can provide insights on real-world data, including data on collective animal movement and the economy. The methodology presented here is general and applicable to a wide variety of domains where coordination across many individuals or entities is observed. Moreover, our framework is highly flexible, and can easily be extended to incorporate other models of leadership or other features used in model classification, depending on the details of the system being analyzed. For reproducibility, we provide our code and simulation datasets at [1].

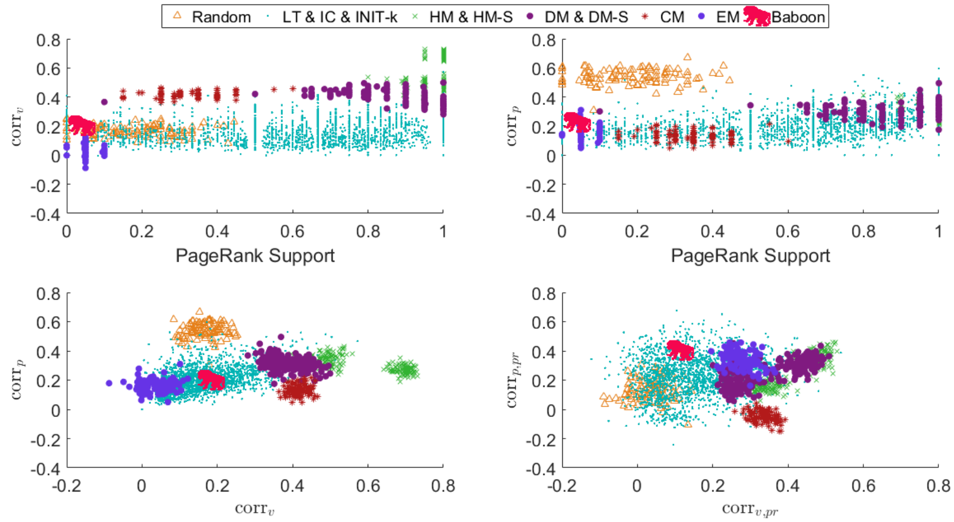


Figure 4: Comparison of feature spaces of leadership model classifications on simulations and real data

REFERENCES

[1] C. Amornbunchornvej and I. Brugere. FLICA: A Framework for Leader Identification in Coordinated Activity website. *UIC Laboratory for Computational Population Biology*, http://compbio.cs.uic.edu/FLICA.html, 2016.

[2] 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.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter. *WSDM*, 65–74. ACM, 2011.

[4] E. E. Boydston, T. L. Morelli, and K. E. Holekamp. Sex differences in territorial behavior exhibited by the spotted hyena (*Hyaenidae, Crocuta crocuta*). *Ethology*, 107(5):369–385, 2001.

[5] L. J. Brent, D. W. Franks, E. A. Foster, K. C. Balcomb, M. A. Cant, and D. P. Croft. Ecological knowledge, leadership, and the evolution of menopause in killer whales. *Current Biol.*, 25(6):746 – 750, 2015.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *WWW*, 107–117, 1998.

[7] Conradt, L. and Roper, T. J. Group decision-making in animals. *Nature*,421(6919):155–158, 2003.

[8] 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.

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

[10] E. W. M. Ben Dushnik. Partially ordered sets. *American Journal of Mathematics*, 63(3):600–610, 1941.

[11] J. Dyer, A. Johansson, D. Helbing, I. Couzin, and J. Krause. Leadership, consensus decision making and collective behaviour in humans. *Philos Trans R Soc Lond B Biol Sci*, 364(1518):781–789, 2009.

[12] I. C. Gilby, Z. P. Machanda, D. C. Mjungu, J. Rosen, M. N. Muller, A. E. Pusey, and R. W. Wrangham. 'impact hunters' catalyse cooperative hunting in two wild chimpanzee communities. *Philos Trans R Soc Lond B Biol Sci*, 370(1683), 2015.

[13] L. Glowacki and C. von Rueden. Leadership solves collective action problems in small-scale societies. *Philos Trans R Soc Lond B Biol Sci*, 370(1683), 2015.

[14] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Discovering leaders from community actions. *CIKM*, 499–508, 2008.

[15] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. *WSDM*, 241–250, 2010.

[16] R. Heinsohn and C. Packer. Complex cooperative strategies in group-territorial african lions. *Science*, 269(5228):1260–1262, 1995.

[17] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, 1998.

[18] P. L. Hooper, H. S. Kaplan, and J. L. Boone. A theory of leadership in human cooperative groups. *J Theor Biol*, 265(4):633–646, 2010.

[19] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. *KDD*, 137–146, 2003.

[20] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81-93, 1938.

[21] 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. *Proc. IEEE PerCom*, 56–64, 2013.

[22] D. Lusseau and L. Conradt. The emergence of unshared consensus decisions in bottlenose dolphins. *Behav Ecol Sociobiol*, 63(7):1067–1077, 2009.

[23] R. Mares, A. J. Young, and T. H. Clutton-Brock. Individual contributions to territory defence in a cooperative breeder: weighing up the benefits and costs. *Philos Trans R Soc Lond B Biol Sci*, 279(1744):3989–3995, 2012.

[24] P.G. Northouse. Leadership: Theory and practice. *Sage*, 2016.

[25] O. Petit and R. Bon. Decision-making processes: The case of collective movements. *Behavioural Processes*,84(3):635–647, 2010.

[26] H. Pham and C. Shahabi. Spatial Influence - Measuring Followship in the Real World. *ICDE*, 2016.

[27] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Sig. Process.*, 26(1):43–49, 1978.

[28] M. Shokoohi-Yekta, J. Wang, and E. Keogh. On the Non-Trivial Generalization of Dynamic Time Warping to the Multi-Dimensional Case. *SDM*, 289–297.

[29] W. Song and S. Quanbin. Computer Simulation of Leadership, Consensus Decision Making and Collective Behaviour in Humans. *PLoS ONE*, 9(1):1–12. Public Library of Science, 2014.

[30] J. C. Stewart and J. P. Scott. Lack of correlation between leadership and dominance relationships in a herd of goats. *J Comp Physiol Psych*, 40(4):255–264, 1947.

[31] 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.

[32] A. Strandburg-Peshkin, C. R. Twomey, N. W. F. Bode, A. B. Kao, Y. Katz, C. C. Ioannou, S. B. Rosenthal, C. J. Torney, H. S. Wu, S. A. Levin, and I. D. Couzin. Visual sensory networks and effective information transfer in animal groups. *Current Biology*, 23(17):R709–R711, 2016.

[33] 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), pp.1995-2004. 2008.

[34] R. Sulo, T. Berger-Wolf, and R. Grossman. Meaningful selection of temporal resolution for dynamic networks. *MLG*, 127–136. ACM, 2010.

[35] J. Sun and J. Tang. A Survey of Models and Algorithms for Social Influence Analysis. *Social network data analytics*, 177–214. Springer, 2011.

[36] D. A. M.P. Weinstein. Leadership behaviour in sawfly larvae *Perga dorsalis* (*Hymenoptera: Pergidae*). *Oikos*, 79(3):450–455, 1997.

[37] E. O. Wilson. *Sociobiology: The New Synthesis.* Cambridge, MA: Belknap of Harvard UP, 2000.