AMERICAN UNIVERSITY OF BEIRUT

DATA-DRIVEN RECOVERY OF TIME-EVOLVING CAUSAL INTERACTION NETWORKS AND STOCHASTIC DYNAMICS IN ZEBRAFISH GROUPS

by
ANDRE VARTAN PANOSSIAN

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Physics
of The Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
August 2023

AMERICAN UNIVERSITY OF BEIRUT

Data-Driven Recovery of Time-Evolving Causal Interaction Networks and Stochastic Dynamics In Zebrafish Groups

by

ANDRE VARTAN PANOSSIAN

Approved by:

Dr. Sara Najem, Assistant Professor Physics	Advisor
Dr. Leonid Klushin, Professor Physics	Member of Committee
Dr. Jihad Touma, Professor Physics	Member of Committee

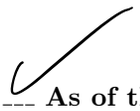
Date of thesis defense: August 28, 2023

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Panossian (Last) Andre (First) Vartan (Middle)

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes



___ **As of the date of submission of my thesis**

___ **After 1 year from the date of submission of my thesis .**

___ **After 2 years from the date of submission of my thesis .**

___ **After 3 years from the date of submission of my thesis .**

Signature

07/09/2023

Date

Acknowledgements

The completion of this thesis is not the result of a lone effort, but rather a tribute to everyone who has offered their assistance, guidance, and encouragement throughout this incredible journey.

I would like to express my deepest gratitude to my advisor, Dr. Sara Najem. Your relentless dedication to knowledge and learning is an inspiration, and our weekend meetings will forever remain as some of my most valuable academic experiences. Your guidance extended far beyond academia, mentoring me in life, and for that, I am eternally grateful. The counsel you have given me will not only shape my professional journey but will resonate well beyond. Thank you for believing in my potential, challenging my ideas, and nurturing my intellectual growth.

I extend as well my profound gratitude to the members of my thesis committee, Dr. Jihad Touma and Dr. Leonid Klushin. Thank you for the insightful discussions, constructive critiques, and continuous encouragement that greatly enriched my work and instilled in me the confidence to challenge the boundaries of my research. Your expertise, patience, and dedication have significantly shaped my academic journey, and the wisdom I gained from our interactions will continue to guide my future endeavors, whether in academic or corporate settings.

I also wish to express my appreciation to the AUB Physics Department as a whole. The environment fostered here - one of intellectual curiosity, mutual respect, and collaborative learning - has not only provided me with a robust foundation in my field of study but has also inspired a profound love for science and a passion for continuous learning. The breadth and depth of knowledge, resources, and support offered by the department were invaluable in the successful completion of this thesis. To every professor, administrative staff member, and fellow student who contributes to making this department a place of growth and exploration, I extend my deepest thanks.

Next, I want to extend my sincere gratitude to my parents, whose steadfast support has been the foundation of this project. Despite our modest upbringing and the complexity of my studies, which perhaps may have confused them, their faith in me never faltered. My quest for knowledge has been motivated by their conviction that education is a goal in and of itself rather than just a means to an end. I dedicate this accomplishment to you, Mom and Dad, because your unwavering love, support,

and sacrifices have been my source of inspiration throughout my path.

I also wish to extend my sincerest thanks to my circle of friends who have formed an integral part of my support system. Arwa, Ali, Layal, Melissa, Dima, thank you for standing by me, sharing in my triumphs and setbacks, and always being there to lift my spirits. Your friendship has made this journey more bearable and enjoyable.

To my friends at work, Omar, Charbel, Hussein, Alexander, and Joseph, your motivation during those moments of self-doubt and falter was invaluable. Your success stories, advice, and belief in my capabilities kept me on track when the road seemed too steep. The friendship we share has been, and will continue to be, a source of strength for me.

This thesis stands as a testament not only to my efforts but also to the collective endeavors of everyone who has been a part of this journey. Each of you has been a piece of the puzzle that is this project, and for that, I am always grateful.

Abstract of the Thesis of

Andre Vartan Panossian for Master of Science
Major: Physics

Title: Data-Driven Recovery of Time-Evolving Causal Interaction Networks and Stochastic Dynamics In Zebrafish Groups

This thesis explores the collective behavior and dynamics of juvenile zebrafish (Danio rerio) shoals of varying group sizes, employing a multifaceted approach grounded in physics, information theory, graph theory, and stochastic analysis. We look at how groups of 4, 10, 60, 80, and 100 zebrafish interact.

We examine two key behavioral metrics: rotation and polarization order parameters, and observe that the decay times derived from the autocorrelation functions of these metrics' time series increase considerably as group size grows. This signals a heightened level of coordination that arises with increased density, with decay rates of the rotation order parameter in the largest group exhibiting a ten-fold difference compared to the smallest group.

To learn more about how coordination and density work together, we used the Optimal Causation Entropy principle (oCSE) to build dynamic, time-evolving, causally-weighted networks that show how zebrafish shoals of sizes 4, 10, and 60 interact with each other. By leveraging these networks, and exploring them using graph theory, we relate the increase in coordination within denser systems to a more consistent, and less volatile causal structure. Within the context of network science, we measure the average number of interacting neighbors, then look at the emergence of leadership, provide a way to quantify it, and compare that across the three different groups.

In the concluding part of this study, we use the Kramers-Moyal equation to combine Kramers-Moyal coefficients with Sparse Regression techniques, also known as equation learning, to derive interpretable, analytical expressions of stochastic differential equations describing the evolution of the rotation and polarization order

parameters for the different group sizes, as well as the coupled differential equation that describes the concurrent evolution of these order parameters.

Collectively, our findings cast light on the intricate relationships between collective behavior, emergent sustained coordination, information sharing, and stochastic dynamics in animal groups, providing a holistic framework for studying such systems.

Table of Contents

Illustrations

Tables

Chapter 1

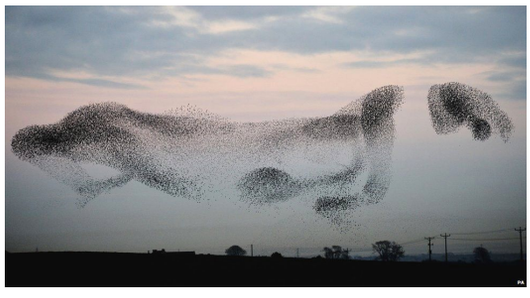
Introduction

1.1 Choreography Of Active Matter Systems

In the rapidly evolving landscape of scientific research, active matter science has emerged as an interdisciplinary field that integrates concepts from physics, chemistry, biology, and materials science to investigate the collective behavior of self-propelled particles. These active systems, which convert energy into motion and exhibit intricate interactions with their environment, represent a striking departure from traditional equilibrium systems, and thus present a unique opportunity for expanding our understanding of the microscopic world [1].

Spanning a wide range of natural and synthetic systems, from the coordinated motion of bird flocks [2] and fish schools [3] to the orchestrated dynamics of cellular and bacterial assemblies [4], active matter science has captured the attention of researchers worldwide. The endeavor to unveil the underlying principles that govern the behavior of these self-driven systems carries immense potential for advancing our knowledge and inspiring the development of innovative technologies.

(a) School of Fish



(b) Flock of Bird

Figure 1.1: Active Matter Systems: Air And Sea

(a) Bacillus Subtilis Clusters

(b) Wildebeest Migration

Figure 1.2: Active Matter Systems: Micro to Macro

1.2 Modeling Collective Dynamics

The study of animal interactions, particularly among groups of individuals exhibiting nonlinear behavior, is an essential and challenging area of research. In such "nonlinear" systems, agents can influence each other in unexpected ways, making the collective behavior of the group difficult to predict or analyze using traditional linear methods that assume a direct proportionality between cause and effect. Understanding these interactions can help researchers gain insight into the underlying mechanisms that govern collective behavior in various species. In [5] Lord draws an analogy to statistical physics, where macroscopic observables are determined by microscopic interactions, coordinated group behaviors appear to emerge from interactions between individuals. It is also worth mentioning that correlational tools [6], are not entirely suited for analyzing the interaction among agents especially when the relationship between cause and effect is not linear, in addition to the fact that correlational tools most often lack the directionality needed when dealing with causal interaction.

Diving deeper into this analogy with statistical physics to better understand collective group motion, researchers pursue three parallel approaches. The first approach is a macroscopic analysis that focuses on observed group or large scale-level behaviors such as group morphology or material-like properties [7]–[9]. The second approach involves a microscopic analysis that determines the nature of the interactions between individuals [6], [10], [11]. Lastly, the final approach deals with the macroscopic consequence of the agent-agent interactions, and specifically how these small-scale interactions give rise to emergent behavior or macroscopic properties [12].

The most scientific attention normally goes to the third approach simply because it is easy to simulate simple models of collective behavior on computer within this agent-based modeling context, notably the classic Vicsek model [12], the Reynolds model [13], and Couzin et al. models [14]. While these studies have been very helpful in expanding our understanding of the macroscopic patterns and ordering that can be found in active collective systems, to claim that these straightforward models accurately depict genuine animal behavior, one must implicitly believe that the interactions between individuals are effectively represented, i.e that these simple models actually reflect reality. We cannot say for sure if that is the case, but we do know that these and most other models frequently make the assumption that animals only interact with other animals that are physically close to them, which gets rid of the complexity of having to deal with the whole network and instead replace it with the much smaller proximity network [15]. In light of that, knowing if spatial local neighbors dominate interactions is crucial to understanding collective motion, while knowing that the introduction of even a small number of long range interaction could qualitatively impact macroscopic behavior in a lattice as was shown by Strogatz et al.[16], in addition to that these models make an assumption about the number of interacting neighbors or the size of the interaction neighborhood.

In this thesis, we employ a novel approach to studying animal interactions by applying the optimal causation entropy (oCSE) principle developed by Sun, Bollt et al. [17], to time series data obtained from optical tracking of the zebrafish, D. rerio, of the wild-type TU strain [18]. The oCSE principle allows us to identify causal relationships and quantify information flow among the fish, forming a network of information flow within the swarm. [19], [20], [21].

This thesis represents a comprehensive exploration into the collective behavior and dynamics of shoals of the juvenile zebrafish, across varying group sizes. Through a multifaceted approach, we have connected the dots between physics, information theory, graph theory, and machine learning, to dissect and understand the interaction patterns within these shoals. Applying the Optimal Causation Entropy principle, we constructed dynamic, causally-weighted networks reflecting interactions within zebrafish shoals. Through these networks, we unraveled the phenomenon of emergent coordination within increasingly dense systems.

In the final phase, we employed the Kramers-Moyal equation [22] and sparse regression techniques [23] to derive analytical expressions of the stochastic differential equations governing the evolution of the rotation and polarization order parameters and their coupling.

Our work highlights the importance of utilizing advanced non-paramteric methods such as the oCSE principle and Kramer-Moyal expansions for studying complex, nonlinear systems. The findings shed light on the interplay between collective behavior, emergent coordination, and information transfer in animal groups. They form a foundational framework for future research into such complex systems.

Chapter 2

Network Science

The motivation behind including a chapter on network science in this thesis stems from the recognition that our research and results frequently employ network-related concepts and terminology. To ensure that readers have a solid understanding of the foundational notions in network science, this chapter will provide a comprehensive overview of the basic principles, concepts, and methodologies within the field. By familiarizing readers with these essential aspects, we aim to facilitate better comprehension and interpretation of our research findings and discussions throughout the thesis.

2.1 Introduction

In today's interconnected world, we are continually influenced by and engage in a multiplicity of dynamic networks that shape various areas of our life. Our evolution as a species has been closely related to our interactions within thousands of years of ecological, biological, social, and other networks. As a result, complex socio-technical ecosystems such as cities, water and electricity systems, and transportation networks have emerged. The Internet and the World Wide Web have further transformed how we access, distribute, and generate knowledge, underlining the growing need of understanding the linkages, trends, and patterns within these large networks in order to successfully address global concerns.

Network science is an interdisciplinary field that is quickly expanding. Its goal is to develop theoretical and practical ways to better understand the structure and function of natural and man-made networks. With foundations in disciplines such as graph theory [24][25], sociology [26] [27], communication research [28], scientometrics [29][30], biology [31], and physics [32]–[35] [36], researchers with varied work styles, techniques, and research interests have been drawn to network science.

Despite the field's diversity, problems exist in the form of parallel, unconnected research strands and inconsistencies in nomenclature and techniques. To address these issues, a truly multidisciplinary approach is required, in which techniques and datasets from one domain can be utilized to further our understanding of networks

in other domains. This interdisciplinary approach has already resulted in the identification of unexpected commonalities between seemingly unrelated systems, such as social networks and the Internet [37]–[39], implying the existence of universal principles and growth mechanisms underpinning multiple networks.

Network science is a relatively new field with many unanswered problems. Researchers must contend with difficulties such as system-dependent limits on node interconnectivity, changing node characteristics over time, and network embedding within natural environments. Furthermore, networks are rarely isolated; rather, they are frequently interconnected and influenced by the broader systems in which they exist.

This chapter will provide a comprehensive introduction to network science, encompassing its fundamental concepts, methodologies, and applications. By presenting a clear and coherent overview of network science, we aim to equip readers with the foundational knowledge necessary to comprehend and interpret the research findings and discussions that rely on network-related terminology and principles throughout the thesis.

2.2 Graphs and Sub-graphs

In this section, we aim to provide essential concepts and terminologies required for understanding networks. It is important to note that different fields within network science have their own unique vocabulary. The most suitable foundation for a precise mathematical portrayal of networks can be found in graph theory, which we will utilize here. In fact, graph theory has its origins in the groundbreaking efforts of Euler to resolve the Königsberg bridges conundrum (Euler, 1736) [40]. Following the introduction of the random graph model by Erdős and Rényi (1959)[41], graph theory has matured into a discipline that provides an abundance of rigorous mathematical and practical findings for network analysis.

Networks, also known as graphs, have a specific structure (or topology) and can contain quantitative information. The structure may or may not be weighted and may or may not be directed. There may exist quantitative information regarding the types, weights, or other attributes of nodes and edges. This section introduces various network types, their definitions, and their representations. We will begin by describing graph structure.

2.2.1 *Undirected Graphs*

An undirected graph $G = (V, E)$ is defined as a countable set V of nodes or vertices, and a set E of edges or links between unordered pairs of different nodes. The nodes are identified by their order i in the set V, and an edge (i, j) connects two adjacent, connected, or neighboring nodes i and j. The graph's size is denoted as N, which corresponds to the cardinality of the set V. The set E has a cardinality of M, which

represents the total number of edges in the graph. When $M = N(N-1)/2$, the graph is called a complete N-graph, because it connects all possible node pairs with edges. Undirected graphs can be graphically represented by a set of dots representing the nodes, connected by lines between the corresponding edges, as seen in Figure 2.1(a-d).

2.2.2 *Directed Graphs*

A directed graph, also known as a digraph, is made up of a set of non-empty countable nodes V and a set of directed edges ED, which are represented as ordered pairs of different nodes. The directed character of the edges is generally expressed in graphic depictions by an arrow denoting the direction of the edge. Refer to Figure 2.1e and 2.1f for an example. It is important to keep in mind that having an edge from node i to node j (i.e., $i \rightarrow j$) in a directed graph does not necessarily imply the presence of the reverse edge $i \leftarrow j$.

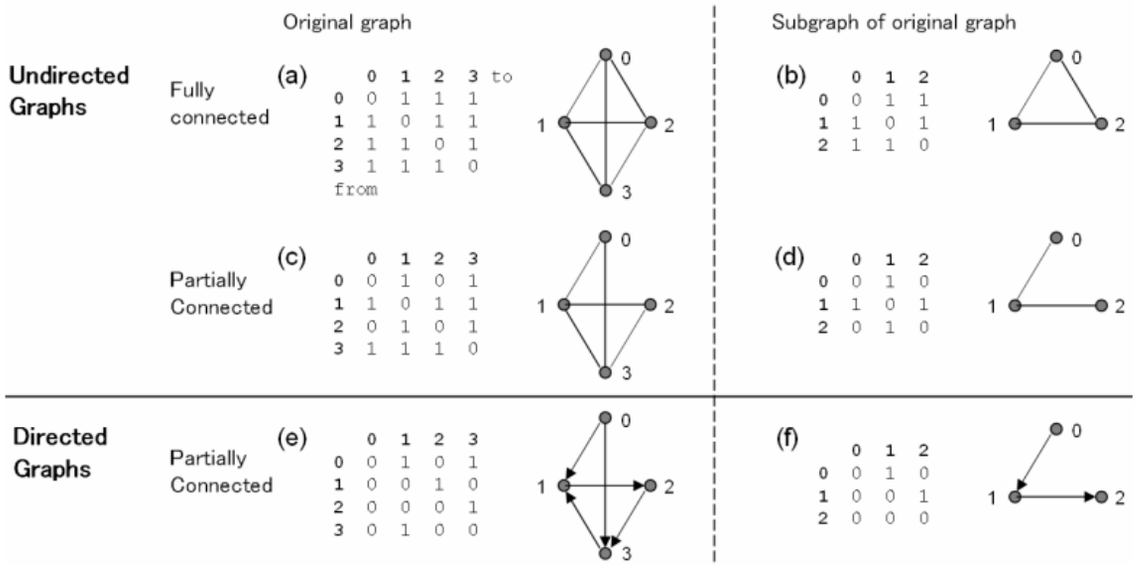


Figure 2.1: Adjacency matrix and graph representations of different undirected and directed graphs

2.2.3 *Trees*

A tree graph is a hierarchical graph with a unique parent node for each edge, also known as a child. A rooted tree is a tree graph that grows from a single parent node. Tree graphs are distinguished by the fact that the number of nodes in a tree always matches the number of edges plus one, which may be represented as $N = E + 1$. Another important feature of trees is that if any edge is removed, the tree becomes unconnected.

2.2.4 *Graph Representation*

A convenient mathematical approach to defining a graph involves using an adjacency matrix $x = x_{ij}$ with dimension $N \times N$. In this representation, $x_{ij} = 1$ if the edge $(i, j) \in E$ and $x_{ij} = 0$ if $(i, j) \notin E$. It is worth noting that for undirected graphs, the adjacency matrix is symmetric, that is $x_{ij} = x_{ji}$, and thus contains redundant information. However, for directed graphs, the adjacency matrix is not necessarily symmetric. Figure 2.1 illustrates the adjacency matrices and corresponding graphical representations for four undirected graphs (a-d) and two directed graphs (e and f). It is interesting to note that the adjacency matrix is also referred to as a sociomatrix in the social network literature.

Using an adjacency matrix to depict the relationships between nodes in a network is a succinct and informative approach to do so. The adjacency matrix can also be used to compute different graph features such as the degree distribution, clustering coefficient, and centrality metrics. This representation can also be utilized in algorithms that include matrix operations, such as finding the shortest path between nodes or computing the adjacency matrix's eigenvalues and eigenvectors. Despite its value, the adjacency matrix representation can be computationally expensive for big graphs, limiting its usefulness.

2.2.5 *Subgraphs*

In graph theory, a graph $G' = (V', E')$ is considered a subgraph of another graph $G = (V, E)$ if all the nodes in V' are contained in V and all the edges in E' are included in E. More formally, $E' \subseteq E$ and $V' \subseteq V$. Figure 2.1 b, d, and f illustrate subgraphs of the graphs shown in Figure 2.1 a, c, and e, respectively. A clique is a complete subgraph of size $n < N$. For instance, the graph shown in Figure 2.1 b is a 3-subgraph of the complete N-graph depicted in Figure 2.1 a. So far, the definitions have been qualitative, describing the structure of a graph. Nevertheless, quantitative information, such as edge weights, can also be assigned to a graph.

2.2.6 *Weighted Graphs*

Real networks often exhibit significant heterogeneity in the capacity and intensity values of their edges [35], [42]–[44]. In social systems, for example, the strength and frequency of contacts are important in characterizing corresponding networks, whereas the volume of traffic between internet routers and the number of passengers using different airlines are important in comprehending the structure of these systems. As a result, it is preferable to build a weighted network that goes beyond simply topological representation and reflects the strength or worth of the links. This can be achieved by associating each edge (i, j) with a weight w_{ij}.

Similar to the adjacency matrix $x = x_{ij}$, we can define a weighted adjacency matrix $W = w_{ij}$. The weighted adjacency matrix can be used to represent undirected weighted graphs where $w_{ij} = w_{ji}$ and directed weighted graphs where $w_{ij} \neq w_{ji}$,

although this may not always be the case. Because it combines both topology and quantitative information, the weighted graph representation provides a deeper description than the unweighted model.

2.3 Graph Connectivity

2.3.1 *Node Degree*

In undirected graphs, the degree k of a node refers to the number of edges connected to it. In contrast, in directed graphs, the degree of a node is defined as the sum of its in-degree and its out-degree, i.e., $k_i = k_{\text{in},i} + k_{\text{out},i}$. The in-degree $k_{\text{in},i}$ of node i is the number of edges pointing towards i, while its out-degree $k_{\text{out},i}$ is the number of edges departing from i. Using the adjacency matrix, we can express the degree of a node as the sum of its in- and out-edges, i.e.,

$$k_{\text{in},i} = \sum_j A_{ji} \tag{2.1}$$

$$k_{\text{out},i} = \sum_j A_{ij} \tag{2.2}$$

Where A_{ji} is the (j,i)-th element of the adjacency matrix, representing an edge pointing towards node i, and A_{ij} is the (i,j)-th element representing an edge departing from node i. Therefore, the degree of a node i can be expressed as the sum of the in- and out-degrees:

$$k_i = k_{\text{in},i} + k_{\text{out},i} = \sum_j A_{ji} + \sum_j A_{ij}. \tag{2.3}$$

For undirected graphs, the adjacency matrix is symmetric, and thus $k_{\text{in},i} = k_{\text{out},i} \equiv k_i$. In this case, the degree of a node is simply the sum of its edges, i.e., $k_i = \sum_j A_{ji} = \sum_j A_{ij}$. For instance, in Figure 2.1 a, node 1 has a degree of three, while in Figure 2.1 e, node 1 has an in-degree of two and an out-degree of one.

2.3.2 *Nearest Neighbors*

The nearest neighbors of a node i are the nodes that are directly connected to it by an edge. In other words, the number of nearest neighbors of a node is equal to its degree. For example, node 1 in Figure 2.1a has nodes 0, 2, and 3 as its nearest neighbors, since these nodes are directly connected to node 1 by edges.

2.3.3 *Clustering Coefficient*

The following explanation follows the Fagiolo Method for computing the Clustering Coefficient, see [45]

Undirected Networks The tendency of a network to form closely connected neighborhoods can be measured by the clustering coefficient (CC). Considering a binary, undirected network (BUN), described by a graph $G = (N, A)$ with N as the number of nodes and A as the adjacency matrix. The element a_{ij} is 1 if nodes i and j are neighbors, and 0 otherwise. For a given node i, the degree k_i represents the number of its neighbors. The extent of clustering for i's neighborhood can be evaluated by the ratio between the number of triangles in G with i as a vertex (t_i) and the total possible triangles that i could have formed. Thus, the CC for node i is given by

$$C_i^A = \frac{1}{2}\sum_{j\neq i}\sum_{h\neq(i,j)} a_{ij}a_{ih}a_{jh} \div \frac{1}{2}k_i(k_i-1) = \frac{A_{ii}^3}{k_i(k_i-1)},$$

where A_{ii}^3 is the ith element of the main diagonal of A^3. The network-wide CC for graph G is then obtained by averaging C_i over the N nodes.

Weighted Undirected Networks Weighted undirected networks (WUN) incorporate the heterogeneity in the capacity and intensity of their connections. Each edge ij in G (where $a_{ij} = 1$) is assigned a value w_{ij} proportional to the link's weight in the network. The concept of node degree in BUNs is replaced with node strength $s_i = \sum_j w_{ij}$. The extension of the CC of node i to WUNs is given by

$$\tilde{C}_i^W = \frac{1}{2}\sum_{j\neq i}\sum_{h\neq(i,j)} w_{ij}^{1/3}w_{ih}^{1/3}w_{jh}^{1/3} \div \frac{1}{2}k_i(k_i-1) = \frac{(W^{[1/3]})_{ii}^3}{k_i(k_i-1)},$$

where $W_{ii}^{1/3}$ is the ith element of the main diagonal of the matrix $W^{1/3}$.

Directed Networks Directed networks involve non-mutual relationships, represented by non-symmetric adjacency or weight matrices. The CC for binary directed networks (BDNs) can be defined in a similar manner to BUNs, but takes into account all possible directed triangles formed by each node, regardless of the edge direction. The CC for node i in BDNs is given by

$$C_i^{DA} = \frac{t_i^D}{T_i^D} = \frac{(A+A^T)_{ii}^3}{2(k_i^{tot}(k_i^{tot}-1) - 2k_i^{\leftrightarrow})},$$

where t_i^D is the number of directed triangles formed by i and T_i^D is the total possible triangles that i could form. The overall CC for BDNs is defined as $C^D = \frac{1}{N}\sum_{i=1}^N C_i^D$. And note that when there are no self-interactions, the number of bilateral edges between i and its neighbors (the number of nodes for which an edge $i \to j$ and $j \to i$ exist is: $k_i^{\leftrightarrow} = \sum_{j\neq i} a_{ij}a_{ji}$).

Weighted Directed Networks The CC for BDNs can be extended to weighted directed networks (WDNs) by substituting A with $W^{1/3}$. Therefore, the CC for node i in WDNs is given by

$$\tilde{C}_i^{DW} = \frac{\tilde{t}_i^D}{T_i^D} = \frac{(W^{1/3} + (W^T)^{1/3})_{ii}^3}{2(k_i^{tot}(k_i^{tot} - 1) - 2k_i^{\leftrightarrow})},$$

where $\tilde{t}_i^D$ is the number of weighted directed triangles formed by i.

2.3.4 *Global Efficiency*

Global efficiency is a measure of the efficiency of information exchange on a network [46], and it is particularly useful for examining indirect paths between nodes. It is the average of the inverse shortest path length and is computed as:

$$E_{glob} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{2.4}$$

where d_{ij} is the shortest path length between nodes i and j, and N is the total number of nodes. High global efficiency implies a high speed of information exchange in the network due to the presence of short paths among nodes.

The efficiency of a network can provide insights into the network's overall structure and organization. High network efficiency is often associated with optimized processing and robustness to perturbations. For instance, a highly efficient social network may facilitate rapid communication and information spread among individuals. Meanwhile, an efficient transportation network enables quick and resource-light routes, contributing to optimal traffic management.

Understanding these measures can thus help to draw inferences about network performance and to devise strategies for network improvement, resilience, and control. These metrics can offer powerful tools for analyzing the structure and function of various complex networks, including social networks, biological networks, and technological networks.

2.4 Modeling Complex Dynamics on Networks

The study of dynamic processes in large-scale complex networks has gained popularity, resulting in ground-breaking breakthroughs in a variety of domains such as social sciences, engineering, medicine, and natural sciences [47], [48]. Researchers have focused on understanding the role of network structures in shaping the dynamics of various systems [49]–[53]. This understanding has contributed to the development of strategies for managing network dynamics [54], [55], improving network resilience [56], and optimizing network performance in different applications [57], [58].

One of the difficulties that researchers encounter is identifying network structure without dramatically changing the underlying system. To get around this problem, researchers used time series data to discover the network structure responsible for the system's behavior.

It is critical to consider directed links that demonstrate "cause and effect" linkages while analyzing and understanding the relationships inside complex networks. In contrast to non-directed relationships, such as correlations, which might occasionally provide a superficial knowledge of the connections, directed interactions often contain valuable information about the underlying mechanics of the system [59].

Various techniques have been employed to infer network structures from time series data, including Granger causality [60], transfer entropy [61], and dynamic Bayesian networks [62]. These methods each have their own set of assumptions and constraints for determining directed relationships within the network. Comparing and testing these methods in real-world applications is critical for improving network inference accuracy and dependability.

Furthermore, network inference techniques can be applied to various fields. For instance, they can be utilized to understand the spread of information or influence in social networks [63], identify the structure of ecological networks and species interactions [5], and analyze transportation networks to optimize traffic flow [64].

In conclusion, the investigation of dynamic processes in large-scale complex networks has led to significant advancements in a wide array of disciplines. Focusing on the role of network structure and utilizing time series data to infer these structures has proven to be a valuable approach for understanding and controlling the dynamic properties of various systems. Accounting for directed "cause and effect" relationships in network inference techniques is crucial for gaining deeper insights into the underlying mechanisms of these complex systems.

Chapter 3

Causal Discovery From Time-Series Data

In this chapter, we will discuss causal discovery methods for time series data, which are essential for our research. The motivation behind this focus lies in the time series data obtained from tracking zebrafish, as it holds valuable information about their behavior. Analyzing this data is particularly important because it allows us to uncover non-local interactions, understand the propagation of information, and understand the emergence of coordination, providing a deeper understanding of the complex dynamics governing zebrafish movement.

3.1 Introduction to Causality

Causality is a pivotal topic in science, with a long and rich history and causal inference now being at the forefront of machine learning breakthroughs in policy evaluation, social science, as well as marketing [65]–[67]. The following introduction is inspired by the elaborate review of [68] on the topic of causality and its history.

The natural and social sciences have long sought cause-effect relationships between variables, events, and objects. Although some mathematicians, such as Russel [69], tried to disprove "causality" in mathematics and physics arguing that causal relationships and physical equations are incompatible, the concept is still used in the language of various sciences, including mathematics and physics. The reason why causality is difficult to define is because math and physics go beyond equations and in a sense causality can be mathematically analyzed as a "flow" between processes. In general, causation is the relationship between events, objects, variables, or states, with the cause typically presumed to precede the effect in time.

Causal relationships are frequently studied in situations influenced by uncertainty, and when we think of uncertainty we think of probability theory. It is in fact the latter that appears to be the most prevalent "mathematical language" employed by many scientific disciplines for causal modeling. Keeping in mind that in numerous disciplines the objective goes beyond merely identifying causal relationships and ex-

tends to measuring or quantifying the relative strengths of these relationships. And while the literature on the topic of causal modeling is vast, touching on mathematical logic, Markov models, Bayesian probability, etc. [68], [70], we concentrate here solely on information-theoretic approaches that perceive causality as a phenomenon that can be "measured" and quantified.

Since the first goal of this thesis is to make use of information-theoretic tools to uncover causality, we center our attention on defining causality within that specific context, which means measuring causal influence from multivariate time series by estimating entropy, mutual information, and discussing the relevant non-parametric ways to estimate these quantities.

One of the early definitions of causality came from Suppes (1970) who proposed that an event X is a cause of event Y if X occurs before Y, the likelihood of X is non-zero, and the likelihood of Y given X is higher than the likelihood of Y occurring alone. Then in 1956, Weiner proposed the first computationally measurable and generic definition of causality [71], arguing that "For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one."

3.1.1 *Granger Causality*

We now attribute the concept of causality in experimental frameworks, that is to say using time-ordered data, to Clive W. J. Granger, the 2003 Nobel laureate in economics. In his Nobel lecture [72], Granger drew inspiration from Wiener's work and outlined two key aspects of causality:

1. The cause precedes the effect
2. The cause provides unique information about the effect, which cannot be found in any other variable.

What Granger found these to imply, is that the more data one has from the causative variable the better the forecast is for the effect variable. This type of causality, known as Granger causality (GC), defines the extent to which one process X_t influences another process Y_t and is based on the concept of incremental predictability. A process X_t is said to Granger-cause another process Y_t if future values of Y_t can be better predicted using both X_t and Y_t past values rather than solely Y_t past values. The standard GC test, developed by Granger [60], is based on a linear regression model:

$$Y_t = a_0 + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \sum_{k=1}^{L} b_{2k} X_{t-k} + \epsilon_t, \tag{3.1}$$

Where ϵ_t are uncorrelated random variables with zero mean, and variance σ^2, L is the specified number of time lags, and $t = L+1, \ldots, N$. The null hypothesis that X_t does not Granger-cause Y_t is supported when $b_{2k} = 0$ for $k = 1, \ldots, L$, reducing the equation to:

$$Y_t = a_0 + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \tilde{\epsilon}_t. \tag{3.2}$$

This linear approach to measuring and testing causality has seen extensive application not just in economics and finance [73], but also across various natural science domains such as climatology [74] and neurophysiology. However, since this only works for linear relationship and most real life systems do not abide by this simplistic model, there was a need for more general methods.

3.1.2 *Transfer Entropy*

While many non-linear extensions to the Granger Causality were introduced, making use of non-parametric regression [75] and local linear predictors [76] to name a few, we turn our attention there to a specific measure, the non-parametric transfer entropy developed by Schreiber [77] for measuring causal information transfer between systems.

Schreiber proposed a non-parametric approach to quantify causal information transfer between systems, known as transfer entropy. This measure is essentially an information-theoretic functional of probability distribution functions. It will be demonstrated that Schreiber's transfer entropy [77] is equivalent to conditional mutual information [78] when conditioned appropriately.

This information-theoretic method is used in climatology [79], physiology [80] and neurophysiology [81]. In the following we will go into the details transfer entropy, its generalization, its importance in the study of dynamical systems and subsequent application. But first we need to work our way up to it.

3.2 Information Theoretic Approaches for Causality Detection

3.2.1 *Definitions and Basics*

Going back to the necessary conditions mentioned in 3.1.1 for the establishing a causal association, we highlight that the first condition is relatively simple when we have access to time series data, which is more and more available for many systems where it didn't exist before. However, the second condition is challenging since it requires evaluating all available information from every variable's time series data; which is why many approximations are made such as neglecting time delay, or limiting ourselves to only few variables (small-size networks with few nodes)

[82], [83], or by partially neglecting the second condition, thus reducing the causal network inference accuracy [84]. Given all of that, it is clear that inferring large scale causal network using available time-ordered data is not only challenging, but had remained unsolved until quite recently, since it not only required theoretical advances but also algorithmic and computational ones. We will later show in that chapter the recently proposed method to challenge that, which we picked and exploited to fit the objective of this thesis[17], [85], [86].

In summary, the classical Granger causality test was designed for linear regression models, but several nonlinear extensions have been proposed. Information-based causality inference addresses the model-dependent limitation in linear Granger causality tests and Schreiber's transfer entropy was one introduced method to measure information flow or effective coupling between two processes regardless of their functional relationship.

Transfer entropy from process Y to process X quantifies the uncertainty reduction of X's future states based on Y's past, given that X's past is already known; it is essentially the mutual information between X's future and Y's history, conditioned on X's history. Although logically sound, this method falls short when applied to multi-variable settings, and fails to detect the network structure properly, that is because transfer entropy was created to identify information flow between two processes [83], [87]. Since we are specifically interested, given a node i, in identifying the other nodes directly impacting i, i.e its direct "causal parents" while avoiding the inference of indirect or spurious causal links as illustrated in Figure 3.1(c), we need to re-think the conditioning of the transfer-entropy based methods that are likely to pick up on indirect influence and dominance of neighboring nodes [83].

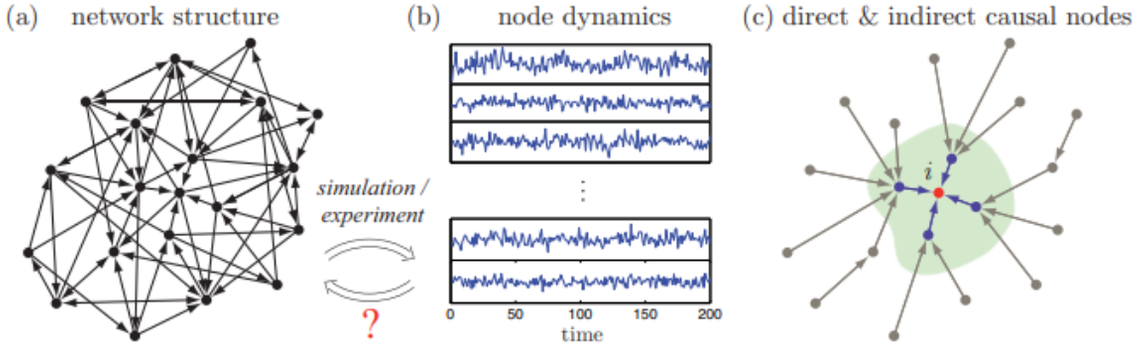


Figure 3.1: The inference of causal network structures underlying observed dynamics (a→b) is possible using high-dimensional time series data from simulations, experiments, and data mining. The goal is to identify the direct causal links of each node i while eliminating non-influential nodes (c) [17]

Knowing how to condition is essential as it can help split causal interactions into direct and indirect ones, and proper conditioning is used widely in network inference methods.[82], [83], [87]–[90]. However, even within this general theme, network

inference requires a theoretically grounded approach that is also algorithmically reliable and efficient. Consequently, as Sun and Bollt put it, there are two crucial steps in causal network inference, the first one in adopting a clear statistic for the causal relationship inference; optimizing this step means focusing on accuracy and generality of the statistic, while the second step is the more challenging one, and that is devising an algorithm that can iteratively apply the first step to learn the causal network. In the second step the focus is now on computational costs and accuracy of the numerical methods that compute the statistics, adding onto that finite-sized data, and it becomes very challenging as the number of nodes in our system grows. Since one of the goals of the thesis is inferring large causal network structure as they contain the most interesting interplay between coordination and information flow, it is important for us to consider an efficient algorithm.

The first and most intuitive thing to consider is testing a candidate causal association by conditioning on all other variables. In other words a directed link $j \to i$ is flagged as truly causal if the value of the statistic remains non-negligible when we condition on all other variables in our system, this is in fact one of the available methods in literature [82]. The issue is obviously clear, the method requires evaluating the statistic in a space with the dimensionality being as high as the whole system which will be computationally too expensive for large networks. Before we go on to the method we have decided to use, it's important to mention one other algorithm that improved upon the previous, namely, The PC algorithm [90]. This algorithm addresses this issue by repeatedly testing candidate causal links conditioned on subsets of the remaining variables [89]. More specifically, a link $j \to i$ is considered to be non-causal if it is negligible when conditioned on some subset of the nodes. That would clearly reduce the dimensionality of the search space and make it as large as the conditioning set that could be much smaller than the system size. The problem however is that if there is no upper bound to the size of the conditioning set, in other words, if the maximum degree or number of causal neighbors of a node is not known beforehand, the PC algorithm will need to search conditioning sets that can be as large as the whole network, which brings us back to the problem of inferring causal network for large network structures. Given all that, a compromise must be made between an algorithm's computational cost and its data efficiency.

3.2.2 *Framework, Terminology, and Assumptions*

Inferring causal networks from high-dimensional time series begins with a theoretical framework. The framework works for linear and nonlinear systems, following the method and terminology in [17] we will work our way up to the oCSE algorithm.

Consider a network (graph) $G = (V, E)$, and $V = \{1, 2, ..., n\}$ is the set of nodes and $E \subseteq V \times V \times \mathbb{R}$ is the set of weighted links. The adjacency matrix $A = [A_{ij}]_n^n$ is defined in the following way:

$$A_{ij} = \begin{cases} \text{weight of the link } j \to i \text{ if } j \to i \text{ in the network,} \\ 0 \text{ otherwise.} \end{cases} \tag{3.3}$$

Let's call the corresponding unweighted adjacency matrix $\chi_0(A)$ defined by its entries by $\chi_0(A)_{ij} = 1$ iff $A_{ij} \neq 0$ and $\chi_0(A)_{ij} = 0$ iff $A_{ij} = 0$. We also define the set of causal parents of i as

$$N_i = \{j | A_{ij} \neq 0\} = \{j | \chi_0(A)_{ij} = 1\}. \tag{3.4}$$

If we have a subset of nodes $I \subset V$, we define its set of causal parents in the following way,

$$N_I = \bigcup_{i \in I} N_i. \tag{3.5}$$

We can define for each node i the stochastic dynamics in the network as such:

$$X(i)_t = f_i \left(A_{i1} X(1)_{t-1}, A_{i2} X(2)_{t-1}, ..., A_{ij} X(j)_{t-1}, ..., A_{in} X(n)_{t-1}, \xi(i)_t \right), \tag{3.6}$$

where $X(i)_t \in \mathbb{R}^d$ is a random variable representing the state of node i at time t, this is practically one of the time series data available or computed, such as speed, acceleration, turn-rate, etc. $\xi(i)_t \in \mathbb{R}^d$ is simply the fluctuation or the noise on node i at time t, and $f_i : \mathbb{R}^{d \times (n+1)} \to \mathbb{R}^d$ models the functional dependence of the state of node i on the past states of nodes j with $A_{ij} \neq 0$. Note that other than the noise term $\xi(i)_t$, the state $X(i)_t$ only depends on the past states of its causal parents, $X(j)_{t-1} (j \in N_i)$.

Once we have quantitative observations on the behavior of each node's dynamic states, practically the time series data, determining the causal system's dynamics is in fact equivalent to finding out three things:

1. The causal network topology, $\chi_0(A)$
2. The connection weights, A_{ij}
3. The specific functional dependencies between nodes, f_i.

These problems are intertwined and each has its own challenges, but let's first consider the problem of finding out the topology $\chi_0(A)$. Mathematically this is expressed as such:

$$\begin{cases} \text{Given:} & \text{Samples of the node states } x^{(i)}t \ (i = 1, 2, \ldots, n; t = 1, 2, \ldots, T). \\ \text{Goal:} & \text{Determine the underlying causal network structure,} \\ & \text{i.e., find } \arg\min_{\hat{A}} ||\chi_0(A) - \hat{A}||_0 \end{cases} \tag{3.7}$$

Sun and Bollt [17], suggest that a practical causation inference method should meet these three criteria:

1. Model-free: no presumptions should be made about the shape or parameters of the underlying process model in order for the method to work.

2. Computationally efficient: the method should be computationally efficient.

3. Data efficient: the method should be able to produce reliable results from a small sample size.

One such method that meets these requirements is the optimal causation entropy principle as proven by Sun and Bollt [17], and it is this method that we aim to use to infer the causal network.

3.2.3 *Markov Assumptions*

We examine the system within a probabilistic framework, assuming stationarity and the existence of a continuous distribution. We also make the following assumptions concerning the conditional distributions $p(\cdot|\cdot)$ that arise from the stationary process defined by (3.6). For every node $i \in V$ and time indices t, t',

$$\begin{cases} (1) \quad \text{Temporally Markov:} \\ \quad p(X_t|X_{t-1}, X_{t-2}, \dots) = p(X_t|X_{t-1}) = p(X'_t|X_{t'-1}). \\ (2) \quad \text{Spatially Markov:} \\ \quad p(X_t^{(i)}|X_{t-1}) = p(X_t^{(i)}|X_{t-1}^{(N_i)}). \\ (3) \quad \text{Faithfully Markov:} \\ \quad p(X_t^{(i)}|X_{t-1}^{(K)}) \neq p(X_t^{(i)}|X_{t-1}^{(L)}) \text{ whenever } (K \cap N_i) \neq (L \cap N_i). \end{cases} \tag{3.8}$$

The first condition implies that the underlying dynamics is a time-invariant Markov process. The second condition ensures that when determining a node's future state, information about the past of any other node becomes irrelevant if knowledge of the past states of all its causal parents N_i (defined in (3.4)) is provided. The third condition guarantees two things, first that the set of causal parents is unique and second that each causal parent has an observable effect independently from the information given by any other causal parent.

3.2.4 *Optimal Causation Entropy*

We quickly review a few information theory basics before getting to the heart of the matter: causation entropy, a model-free information-theoretic statistic used to infer direct causal relationships. [83].

Since we have defined the flow of information from an agent X to another agent Y as the sharing of information between the future states of Y and the past states of X that cannot be explained by any other variable in the system, we need to be able to precisely define the information associated with a variable, as well as the concept of shared information, to that end we begin by introducing Shannon Entropy.

Shannon entropy is used to quantify the information content of a random variable. If $p(x)$ denotes the probability that a measurement of a variable X takes a specific

value x, the uncertainty associated with that variable is defined as its Shannon entropy,

$$H(X) \equiv -\sum_x p(x) \log p(x) \tag{3.9}$$

Now let's add another random variable Y, where the joint distribution with X is given by $p(x, y)$. The joint entropy of these variables as it is called is defined analogous to Eq. (3.9) such that

$$H(X; Y) = -\sum_y \sum_x p(x; y) \log p(x; y) \tag{3.10}$$

Similarly, define the conditional distribution of Y given X as $p(y|x)$, which represents the probability that $Y = y$ given that $X = x$. Conditional entropy is given by

$$H(Y|X) = -\sum_y \sum_x p(x, y) \log p(y|x) \tag{3.11}$$

$$= \sum_y \sum_x p(x, y) \log \frac{p(x)}{p(x, y)} \tag{3.12}$$

The Eqs. (3.9)-(3.12) apply to random variables that take on discrete values, however similar definitions are present for continuous variables by simply replacing the summations by the correct integrals; imagine X is a continuous random variable, we can interpret $p(x)$ as the probability density function of X and use differential entropy in the following way:

$$h(X) \equiv -\int_{-\infty}^{\infty} p(x) \log p(x) dx \tag{3.13}$$

We can similarly define the joint entropy and the conditional entropy between two variables X and Y in the following way (also see Figure 3.2(a))

$$\begin{cases} \text{Joint entropy: } H(X, Y) \equiv H(Y, X) \equiv -\iint p(x, y) \log p(x, y) \, \mathrm{d}x\mathrm{d}y. \\ \\ \text{Conditional entropies:} \\ H(X|Y) \equiv -\iint p(x, y) \log p(x|y) \, \mathrm{d}x\mathrm{d}y, \\ \\ H(Y|X) \equiv -\iint p(x, y) \log p(y|x) \, \mathrm{d}x\mathrm{d}y. \end{cases} \tag{3.14}$$

For an in-depth review of information theory in the continuum limit a full review can be found in [91].

Since the Shannon entropy defined in Eq. (3.9) holds the information content of a random variable, then, the entropy of a pair of random variables or the joint entropy

can also be defined as the sum of the entropy of one variable plus the conditional entropy of the other. That is the famous entropy chain rule [91]

$$H(X;Y) = H(X) + H(Y|X) \tag{3.15}$$

Now introduce another random variable Z, a corollary of Eq. (3.15) states that:

$$H(X;Y|Z) = H(X|Z) + H(Y|X,Z) \tag{3.16}$$

One other important measure is Mutual information, an information-theoretic concept that measures the reduction in uncertainty about X given all of the available information about Y. Mutual information is defined as

$$I(X;Y) = H(X) - H(X|Y) \tag{3.17}$$
$$= H(Y) - H(Y|X) \tag{3.18}$$

When two random variables, X and Y, are considered, the information in X can be divided into information that belongs only to X and information that is shared with Y. The mutual information, $I(X;Y)$, describes the shared information between X and Y, and is defined as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.19}$$

where $H(X,Y)$ represents the entropy of the joint random variable (X,Y).

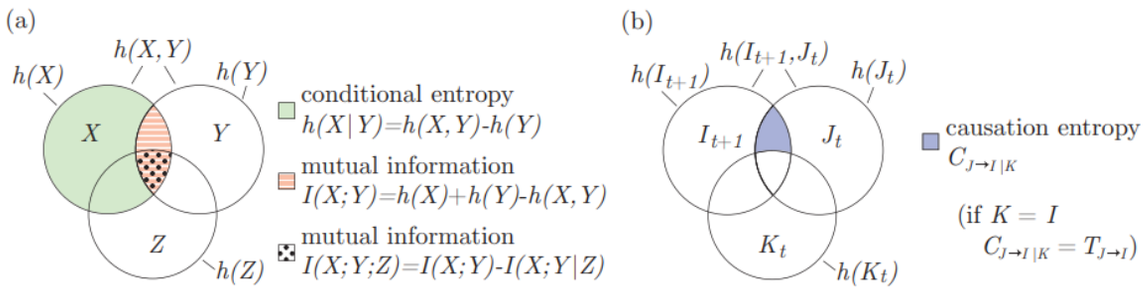


Figure 3.2: Visual Representation of Conditional, Mutual, and Causation Entropy [17]

A visual way of understanding the above is the following (refer to Figure 3.2); conditioning is analogous to removing part of a circle from a Venn diagram, leaving the remaining part. The conditional entropy of Y given X, $H(Y|X) = H(X,Y) - H(X)$, quantifies the uncertainty associated with Y given knowledge about X. Conditional entropy is crucial for understanding swarm behavior since it allows for finding the mutual information between two variables that is not present in a third variable. if we now add another random and discrete variable Z, the so-called conditional mutual information now defined, of X and Y given Z, is:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{3.20}$$

Let me take the time to prove (3.20), starting with the expression for conditional mutual information

$$I(X;Y|Z) = \sum_x \sum_y \sum_z p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

Apply the properties of logarithms to separate the terms inside the logarithm

$$I(X;Y|Z) = \sum_x \sum_y \sum_z p(x,y,z) \left[-\log p(x|z) + \log \frac{p(x,y|z)}{p(y|z)} \right]$$

Using the Bayes' Rule on the second term in the brackets, we get that

$$\frac{p(x,y|z)}{p(y|z)} = \frac{p(x,y,z)}{p(y,z)} = p(x|y,z)$$

Distribute the probability term $p(x,y,z)$ among the three logarithms

$$I(X;Y|Z) = -\sum_x \sum_z p(x,z) \log p(x|z) + \sum_x \sum_y \sum_z p(x,y,z) \log p(x|y,z)$$

Recognize the first term as the conditional entropy $H(X|Z)$

$$-\sum_x \sum_z p(x,z) \log p(x|z) = H(X|Z)$$

Recognize the second term as the negative of conditional entropy $H(X|Y,Z)$

$$\sum_x \sum_y \sum_z p(x,y,z) \log p(x,y|z) = -H(X|Y,Z)$$

Combine the terms to obtain

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

And it becomes evidently clear as well that:

$$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$$

Given all of that we can now translate Transfer Entropy into a type of conditional mutual information. If information in X "flows" to Y, then two conditions must be met

1. There must be information contained in Y at a future time, say $t+\tau$, that is not explained by the state of Y at time t, and

2. This information would be shared by Y at time $t+\tau$ and X at time t. For information to "flow" from X to Y over time τ, it is necessary that $I(X(t); Y(t+\tau)|Y(t)) > 0$. The measure on the LHS is considered as the Transfer Entropy $T_{X\to Y}$

Transfer Entropy is thus:

$$T_{X\to Y} = I(X(t); Y(t+\tau)|Y(t)) \tag{3.21}$$
$$= H(Y(t+\tau)|Y(t)) - H(Y(t+\tau)|X(t), Y(t)) \tag{3.22}$$

Since $H(Y(t+\tau)|Y(t))$ measures the uncertainty of $Y(t+\tau)$ given information about $Y(t)$ and $H(Y(t+\tau)|X(t), Y(t))$ measures the uncertainty of $Y(t+\tau)$ given information about both $X(t)$ and $Y(t)$ then the transfer entropy $T_{X\to Y}$ can be understood as the decrease in uncertainty regarding the future states of Y when the present state of X is given alongside that of Y.

However, real networks have more than two nodes. Transfer entropy cannot distinguish direct and indirect causality in networks without proper conditioning, as mentioned earlier and as is in fact shown by Sun and Bollt [17], [83] that looked at dynamic systems involving a number of nodes larger than two, and since transfer entropy disregards additional sources of information it incorrectly identified couplings between nodes that in fact didn't exist, at least not directly. In other words, indirect and spurious influences will be interpreted as direct. It's obvious from Equation (3.22) that transfer entropy is a pairwise measure, i.e., the contributions to $Y(t+\tau)$ by $Y(t)$ and $X(t)$ are captured, but any additional contribution by a possible third state $Z(t)$ is not accounted for: introducing the generalization of transfer entropy, called causation entropy developed in [83] and [17], which will be the defined below and will be used as the foundational method to uncover causal associations in this thesis. The connections among entropy, transfer entropy, and causation entropy are depicted in Figure 3.2(b).

Causation Entropy (CSE) is a quantity designed to detect direct information channels. Which means that if X, Y, and Z are variables related to three agents, then the Causation Entropy of X to Y given Z is

$$C_{X\to Y|Z} = I(X(t); Y(t+\tau)|Z(t)) \tag{3.23}$$
$$= H(Y(t+\tau)|Z(t)) - H(Y(t+\tau)|X(t), Z(t)) \tag{3.24}$$

In other words, $C_{X\to Y|Z}$ is the information shared between $X(t)$ and $Y(t+\tau)$ that is not already contained in $Z(t)$. Where Z is a third stationary stochastic process or collection of states. To obtain a more general definition of causation entropy, let a set of processes be given by Z. Then, The above value $C_{X\to Y|Z}$ can be read as the causation entropy from X to Y given Z and is always non-negative [17], [83].

It is clear that by setting $Z = Y$, Eq. (3.23) becomes the transfer entropy from X to Y, and thus, the causation entropy can be viewed as a generalization of transfer entropy, since the conditioning set does not necessarily contain information about the past of Y. Finally, note that Z can be selected as the empty set resulting in

$$\begin{aligned} C_{X\to Y|\emptyset} &= C_{X\to Y} && (3.25)\\ &= H(Y(t+\tau)) - H(Y(t+\tau)|X(t)) && (3.26)\\ &= I(X(t); Y(t+\tau)) && (3.27) \end{aligned}$$

Which is just the mutual information between $Y(t+\tau)$ and $X(t)$.

In this thesis the goal is three fold, we first want to apply causation entropy to identify causal relations and their strength among a number of swimming agents, secondly we want to see how this method scales when the number of agents (nodes) grows and what could be then inferred about causal relationships in self-propelled agents, are causal relations only local or can we identify long range interactions? Is the number of causal parents for an agent related to the group size? Is coordination in larger groups linked to a graph theoretic measure of our recovered causal networks, and do we see an ease in information flow in larger groups? How can we quantify leadership, does it arise, if so how long does it persist, and how does all of that changes as the number of agents grow?

As a result of defining causation entropy (CSE) by using conditioning sets, the challenge now lies in selecting the appropriate conditioning sets in order to discover the direct flow of information within a network. In light of this, we will talk about optimal causation entropy, also known as oCSE, which is an algorithmic method that successfully learns the fundamental network interaction structure.

The initial conditioning set is used as a foundation for the oCSE, which then adds as many variables as are required. This stage is known as the discovery phase, and it is immediately followed by the removal phase, during which redundant elements from the set are eliminated. Let $X = X_1, X_2, \ldots, X_m$. Initially, let Z be an empty set, unless there are prior knowledge on what Z or some of the elements it must contain. On each iteration, the variable X_i is added to Z if $C_{X_i\to Y|Z} = \max_{X_j\notin Z} C_{X_j\to Y|Z} > 0$. The discovery phase ends when no such variable can be found from the remaining set of variables.

The final set Z might contain redundancies since it's a set of nodes that communicate with Y because the value of $C_{X_i\to Y|Z}$ can be positive due to indirect information flow from X_i to Y, unless Z contains all other true causal components, but we can't be sure, therefore, the removal phase that follows eliminates elements from Z if they are redundant given other elements in Z. On each iteration, a new member of Z, Z_i, is chosen and removed if and only if $C_{Z_i\to Y|Z\setminus Z_i} = 0$ which means that if the information provided by any of the other variables in the set Z is enough to explain Y, then Z_i is redundant.

Algorithm 1 Aggregative discovery of causal nodes.

Input: Set of nodes $I \subset V$
Output: K (which will include N_I as its subset)
Initialize: $K \leftarrow \emptyset$, $x \leftarrow \infty$, $p \leftarrow \emptyset$.
while $x > 0$ **do**
  $K \leftarrow K \cup \{p\}$
  for every $j \in (V - K)$ **do**
    $x_j \leftarrow C_{j \to I|K}$
  end for
  $x \leftarrow \max_{j \in (V-K)} x_j$, $p \leftarrow \arg\max_{j \in (V-K)} x_j$
end while

Algorithm 2 Progressive removal of noncausal nodes.

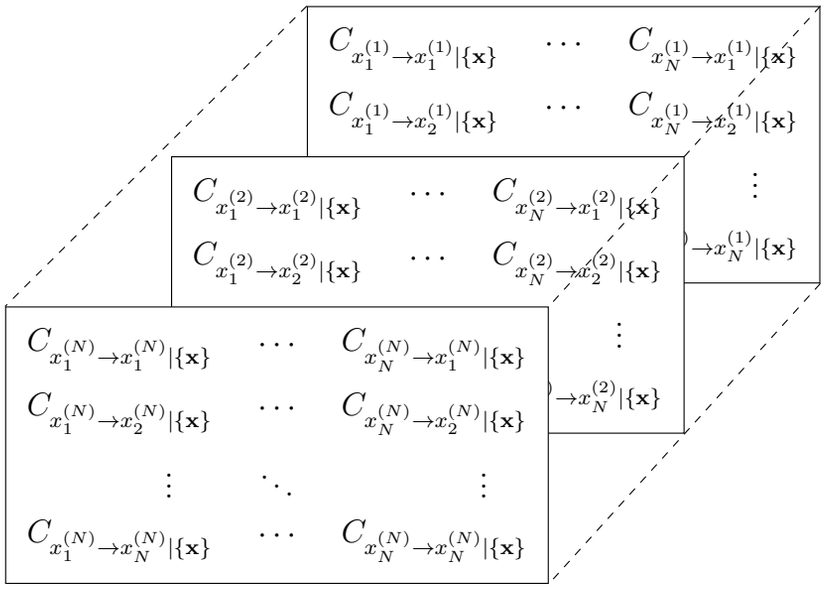
Input: Sets of nodes $I \subset V$ and $K \subset V$
Output: $\widehat{N_I}$ (inferred set of causal parents of I)
for every $j \in K$ **do**
  if $C_{j \to I|(K-\{j\})} = 0$ **then**
    $K \leftarrow K - \{j\}$
  end if
end for
$\widehat{N_I} \leftarrow K$

Once we have treated all the variables of Z, and removed the redundancies, the remaining variables in Z are the direct causal parents of Y, which means that information flows directly from the elements of Z to Y. This assertion was proven in [17] by Sun and Bollt. The causal parent relationship is written as $X_i \to Y$. The set of relationship or direct causal associations $X_i \to X_j$ forms a directed graph in which the variables are the nodes and links or edges represent the direction of causality, and the weights are the values of the causation entropy for the remaining variables after passing the redundancy test.

Having discussed the concept of optimal causation entropy, we can now proceed to explain how this concept can be employed to derive a network from the given time series data. In this network, the interactions among the agents are represented by the causation entropy values. As a result, we can construct an adjacency matrix for a weighted directed graph, which may not necessarily be symmetric.
This adjacency matrix is composed of the causation entropy values, where each entry in the matrix represents the causation entropy for the interaction between two agents, consider the matrix below:

$$\begin{pmatrix} C_{x_1 \to x_1|\{\mathbf{x}\}} & C_{x_1 \to x_2|\{\mathbf{x}\}} & \cdots & C_{x_1 \to x_N|\{\mathbf{x}\}} \\ C_{x_2 \to x_1|\{\mathbf{x}\}} & C_{x_2 \to x_2|\{\mathbf{x}\}} & \cdots & C_{x_2 \to x_N|\{\mathbf{x}\}} \\ \vdots & \vdots & \ddots & \vdots \\ C_{x_N \to x_1|\{\mathbf{x}\}} & C_{x_N \to x_2|\{\mathbf{x}\}} & \cdots & C_{x_N \to x_N|\{\mathbf{x}\}} \end{pmatrix} \tag{3.28}$$

The matrix displays the causation entropy values for all possible directed interactions between the agents. Since we will apply method on a sliding window of data, at the end we will have an evolving adjacency matrix for a weighted directed graph. The result will look something like the matrix below:



This will enable us to better understand the underlying dynamics of the system and identify influential agents or important causal relationships, as well as the number of interacting "neighbors" on average, check whether causal neighbors are or are not necessarily spatial neighbors, and how all that changes with group size.

3.2.5 *Practical Computational Considerations*

Estimating CSE values based on collected data is required for use in practical applications. Because CSE is expressed as a conditional mutual information, it fundamentally necessitates the utilization of a reliable and "good" estimator for the data being collected. The development of such estimators is an important computational and statistical issue that is relevant to a significant portion of the body of research that has been done. Binning techniques, such as histograms, which estimate the probability density p(x) by measuring the frequency of data points within a fixed-size area surrounding the point, are a straightforward way to calculate entropy and related values. These techniques can be used to calculate entropy and related values. Although these methods are simple to understand and put into practice, it has been discovered that they converge slowly. This is especially true for multivariate data sets, which suffer from poor scaling with the embedding dimension. Nonparametric estimators that make use of k-nearest neighbor (knn) statistics are capable of achieving quicker convergence than their parametric counterparts. The core idea entails arriving at an estimate of the density at a particular location by taking into account the distance to the k neighbors who are located closest to that location as

opposed to the neighbors who reside in an area of constant size. A non-parametric estimator based on that idea is the Kraskov-Strögbauer-Grassberger (KSG) estimator [92] for mutual information is data-efficient (resolving structures at the smallest possible scales with k=1), adaptive (providing higher resolution where data is more abundant), and minimally biased (bias primarily stems from nonuniform density at the smallest resolved scale, leading to typical systematic errors scaling as functions of k/N for N points). Here, we make use of an extension of the Kraskov-Strögbauer-Grassberger (KSG) estimator for mutual information [93]. The details of this estimator can be found in the Appendix A.

In algorithmic inference, like oCSE, it is important to figure out if the estimated CSE value $C_{X\to Y|Z}$ should be considered strictly positive, since it is unlikely that we will get 0 exactly when computing. We use a shuffle test for the null hypothesis $C_{X\to Y|Z} = 0$ to figure out how to solve this problem [17]. So following the methodology outlined in [5] and hypothesized by Sun and Bollt[17], given time series samples (x_t, y_t, z_t) of a stochastic process (X_t, Y_t, Z_t), the estimated value of $C_{X\to Y|Z}$ should typically be larger than $C_{X'\to Y|Z}$, where X' represents dummy data obtained by replacing x_t with x'_t, where x'_t is a random permutation of the set x_t. We consider $C_{X\to Y|Z}$ significant if it is greater than a fraction $(1-\alpha)$ of the values of $C_{X'\to Y|Z}$ after running numerous permutations, and α is a pre-chosen significance level.

Each iteration of oCSE necessitates computing $O(n)$ CSE values, assuming a constant sample size and CSE estimator of choice. The computational complexity of oCSE for determining the parents of a single node or variable in a network is $O(Kn)$, where $K \leq n$ is the total number of iterations. Computational complexity for inferring a sparse network is $O(m)$, where $m < n^2$ is the number of links in the network, because the value of K typically corresponds to the target node's degree [17]. Contrast this with the computational complexity of $O(n^2(n-1)k^{-1}/(k-1)!)$ for inferring the entire network using a classical combinatorial-search-based algorithm such as the PC algorithm [90], where k is the maximum degree. Because of its incremental, non-combinatorial nature, oCSE relies more on the density of links than the size of the network n to achieve a given inference accuracy level (for a directed pair).

In summary, the number of data samples required to achieve a given level of inference accuracy for a specific directed pair is less dependent on the network size n when using the oCSE algorithm due to its incremental and non-combinatorial nature. The algorithm relies more on the number of links in the network than on any other single factor. This is helpful because the algorithm's complexity scales with the number of links m rather than the total number of connections n^2, allowing for more effective processing of sparse networks with a large number of nodes, in addition to that oCSE's ability to adapt to varying link densities, and use a nonparametric estimator for conditional mutual information makes it a valuable tool for studying complex systems.

Chapter 4

Learning Stochastic Equations From Data

4.1 Introduction

Differential equations are an essential tool for modeling complex dynamical systems, which are common in a variety of fields, including ecology [94]. The significance of space and stochasticity in the system being investigated will determine which type of differential equations - ordinary, stochastic, or partial - will be employed in the analysis. Even if the rules or interactions on a smaller scale that make up a complex system are straightforward, it may still be possible to translate that complexity into differential equations. To give an example, in the context of an ecological population, these local regulations may have their origins in processes such as the birth and death of organisms [95], the movement and interaction of organisms [96], [97], or the interaction of organisms with their surrounding environment [98], [99]. These micro-level interactions eventually add up to have an effect on the dynamics of macro-level entities such as groups, populations, or even entire ecosystems [99], [100]. The incorporation of empirical data into these models, despite the power and insights offered by such dynamical systems, continues to present a significant challenge.

Now, in the age of big data, we can track the evolution of biological systems through time-stamped, high-resolution records[101], [102]. The data capture behaviors at all levels of biological organization, from cells to animals[102], from groups to entire populations [97], [103], [104] , and from population sizes to population fitness[105], [106]. These data sets are paving the way for a more effective integration of models and data.

In order to capture the dynamics of real-world systems accurately, it is necessary to treat state variables as stochastic, taking into account both the mean properties and the inherent randomness cite [98]. The randomness inherent in biological systems causes unexpected outcomes that do not occur in deterministic systems [100], [101]. A suitable framework for studying these stochastic dynamics can be found in

stochastic differential equations (SDEs). The primary objective here is how can we go from time series data to stochastic differential equations.

Indeed there are many ways, and with the advances in machine learning models on one hand and numerical methods on the other, it is now possible to do so in ways that weren't possible before. Stochastic differential equations can be inferred from time series data using conventional stochastic calculus techniques, which involve the estimation of jump moments [107]–[109] while more recent techniques employing neural networks have also been employed for SDE discovery [110]. On the other hand, in deterministic models, recent advancements in equation learning have enabled the discovery of simple, interpretable differential equation models from time series data [23], [111], which permits and motivates a combination of the jump moments and equation learning. Combining equation learning with jump moments methods is highly effective as is outline by [22], as it facilitates the extraction of straightforward and analytical and understable SDE models directly from data [112]–[114].

To accomplish this, we make use of the Python library Pydaddy (Python library for Data Driven Dynamics) [22], which unifies these various approaches into a single, cohesive framework. This allows us to conduct our research in a more streamlined and effective manner.

4.2 SDE Discovery

A one-dimensional stochastic process can be viewed as a mapping that transports a real time variable t into an appropriate state space that elucidates the random variable $x(t)$ dynamics, which are impacted by random perturbations. A prototypical stochastic process is represented by the stationary Langevin equation, an illustrative example of a stochastic differential equation expressed as:

$$dx(t) = a(x)dt + b(x)dB(t), \tag{4.1}$$

where $a(x)$ denotes the drift term, $b(x)$ symbolizes the diffusion term, and $B(t)$ is an independent Brownian motion. In cases where the dynamics attributes, i.e., $a(x)$ and $b(x)$ are independent of time, these processes are referred to as stationary. Stochastic processes can also be characterized by discontinuities, despite the Langevin equation being continuous in time. A straightforward approach to encompassing these discontinuities is by integrating an elementary Lévy process $L(t)$ into the equation, manipulated by an amplitude $h(x)$ (Applebaum, 2011):

$$dx(t) = a(x)dt + b(x)dB(t) + h(x)dL(t). \tag{4.2}$$

In this extended equation, the interpretation of $a(x)$ and $b(x)$ as drift and diffusion still holds. It should be noted that Langevin processes are merely a specific type of Lévy processes and all such processes are Markovian.

Stochastic processes can be depicted in terms of evolving random variables, which adhere to a stochastic differential equation, or through the progression of their conditional probability density function $p(x, t|x', t')$, which follows a partial differential equation. For instance, if a single particle's movement obeys the Langevin equation, its probability density function $p(x, t|x', t')$ evolves in accordance with the Fokker–Planck equation, articulated as:

$$\frac{\partial p(x,t|x',t')}{\partial t} = -\frac{\partial D1(x)p(x,t|x',t')}{\partial x} + \frac{\partial^2 D2(x)p(x,t|x',t')}{\partial x^2}, \tag{4.3}$$

considered in the stationary scenario, i.e., no explicit time dependence of the coefficients $D_1(x)$ and $D_2(x)$, which are directly linked to the drift and diffusion terms in the first equation:

$$D1(x) = a(x), \tag{4.4}$$

$$D2(x) = \frac{1}{2}b^2(x). \tag{4.5}$$

For discontinuous processes, the Fokker–Planck equation proves inadequate (Risken and Frank, 1996; Stemler et al., 2007; Gardiner, 2009; Tabar, 2019). Discarding the continuity condition, the temporal evolution of the conditional probability density follows the Kramers–Moyal equation:

$$\frac{\partial p(x,t|x',t')}{\partial t} = \sum_{m=1}^{\infty}(-1)^m \frac{\partial^m Dm(x)p(x,t|x',t')}{\partial x^m}, \tag{4.6}$$

where $D_m(x)$ symbolizes the mth Kramers–Moyal (KM) coefficient, defined from the corresponding conditional moments $Mm(x, \tau)$ of the variable x and a time-lag τ:

$$D_m(x) = \frac{1}{m!}\lim_{\tau\to 0}\frac{M_m(x,\tau)}{\tau} = \frac{1}{m!}\lim_{\tau\to 0}\frac{1}{\tau}\langle(x(t+\tau)-x(t))^m|x(t)=x\rangle, \tag{4.7}$$

where $\langle\cdot\rangle$ denotes the expected value. If a stochastic process is 'sufficiently' continuous, the third and all higher KM coefficients become null according to Pawula's theorem (Pawula, 1967a, b) [115], and the Kramers–Moyal equation simplifies to the Fokker–Planck equation. Moreover, the Kramers–Moyal equation allows single particle's motion to assume different functional forms that represent different (discontinuous) stochastic processes. Regardless of these variants, the KM coefficients can be related to the stochastic process properties in a similar manner as equation (4.4-4.5)

In practice, an essential aspect of using a description like the Kramers–Moyal equation is the ability to estimate the coefficients $D_m(x)$ directly from data. To obtain the KM coefficients $D_m(x)$ from a single stochastic process realization, i.e., a singular time series, we appraise the transition probability densities in the limit of a

disappearing time step $\tau \to 0$, which numerically corresponds to considering the minimal increment Δt in the data ($\tau \to \Delta t$):

$$D_m(x) \approx \frac{1}{m!}\frac{1}{\Delta t}\langle (x(t+\Delta t) - x(t))^m | x(t) = x \rangle, \tag{4.8}$$

through which we estimate the various KM coefficients directly from the data.

In the below i will follow through the explanation of the Pydaddy paper [22].
The objective is to use the time series data to determine the drift (f) and diffusion functions (g^2). In particular, we hope to find simple, easily interpretable analytical expressions that characterize f and g^2 quantitatively in addition to qualitatively. The two-step process, which is laid out in greater depth later in the chapter, is as follows:

- Using the so-called jump moments or the Kramer-Moyal coefficients, we first extract the drift and diffusion components from the supplied time series data [109].
- Then, we extract the drift and diffusion functions and use an approach based on sparse regression, also known as equation learning, to find interpretable analytical expressions for these functions [23], [112].

First, we will go over how to find meaningful diffusion and drift functions. Let's pretend x is a d-dimensional state variable and Δt is the sampling interval. We can define the instantaneous drift and diffusion as functions of t and the instantaneous value of the state variable x if the time series is generated by the underlying SDE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x}) \cdot \boldsymbol{\eta}$:

$$\tilde{F}(t; \mathbf{x}) = \frac{\mathbf{x}(t+\Delta t) - \mathbf{x}(t)}{\Delta t} \tag{4.9}$$

$$\tilde{G}(t; \mathbf{x}) = \frac{(\mathbf{x}(t+\Delta t) - \mathbf{x}(t))(\mathbf{x}(t+\Delta t) - \mathbf{x}(t))^T}{\Delta t} \tag{4.10}$$

With these instantaneous drift and diffusion functions $\tilde{F}$ and $\tilde{G}$, we can seek comprehensible mathematical expressions for F and G, using the approach we used, as presented by Brunton et al. [23]

Now, imagine a collection of potential functions $\{F_1, F_2, ...F_k\}$. The F_i might stand in for monomials up to some degree, or it could also be another appropriate basis like a Fourier or Chebyshev basis, and if we have some prior knowledge it could also be problem-specific fundamental functions.

The goal is to express F, the drift function, as a linear combination of a small subset of these potential expressions. To be more precise, we are looking for coefficients ξ_i such that $F(x) = \sum \xi_i F_i(x)$, where a small subset of the ξ_i's are nonzero. As will be seen below, sparse regression is a useful tool for accomplishing this.

To differentiate between the estimated drift and diffusion and the real $\mathbf{f}$ and $\mathbf{g}$, we represent them with capital letters F and G. The estimated F is $\mathbf{f}$ and the estimated G is $\mathbf{g}\mathbf{g}^T$ and finding analytical expressions for drift and diffusion is formulated as a sparse regression problem. The sparse regression procedure for the drift function in the scalar case is described in detail. For the diffusion function, a similar procedure is used.

If we consider the column vector $\mathbf{x}_{T\times 1}$ containing the state variable x sampled at each point of time, i.e., $\mathbf{x}_i = x(i\Delta t)$. T being the total number of observations. Given the expression for the drift we have in eq. (4.9) we can compute the instantaneous drift values, so consider $\boldsymbol{\phi}_{T\times 1}$ as a column vector containing the instantaneous values of the drift, i.e., $\boldsymbol{\phi}_i = \tilde{F}(i\Delta t)$. Assuming we have selected a library of functions $\{F_1, F_2, ...F_k\}$ for the drift function, we define a dictionary matrix $\Theta_{T\times k}$ with the ith column given by $\Theta_i = F_i(\mathbf{x})$. The notation $F_i(\mathbf{x})$ is used as shorthand for evaluating F_i on each entry of x. The sparse regression problem in terms of $\boldsymbol{\phi}$ and Θ corresponds to finding the sparse vector $\boldsymbol{\xi}$ that solves the equation

$$\boldsymbol{\phi} = \Theta\boldsymbol{\xi} \tag{4.11}$$

Finding a sparse $\boldsymbol{\xi}$, requires a procedure called sequentially thresholded least squares (STLSQ). The execution of this algorithm takes place in several stages: first, a solution for $\boldsymbol{\xi}$ is found using ordinary least-squares. After that, and given a predefined sparsity threshold, all entries of $\boldsymbol{\xi}$ smaller than that threshold are set to zero, which will remove the corresponding columns from Θ. The process is then repeated with the remaining terms until no more terms can be removed.

In the STLSQ algorithm, the choice of the sparsity threshold is extremely important, and it must be made in an appropriate manner in order to guarantee that the correct models are recovered. In most cases, the selection of the threshold is determined by an information criterion such as the Akaike Information Criterion (AIC), or by the accuracy as determined by cross-validation. In this instance, the second method is utilized, and k-fold cross-validation is utilized for the purpose of model selection.

The idea behind cross-validation is to train a model with only a portion of the available data (referred to as the training set), and then evaluate the performance of the model using another set of data (referred to as the validation set). A model that performs well on the validation set is presumed to not have over-fit on the noise in the training data and may therefore be more generalizable. This is because good performance on the validation set measures how well the model predicts unseen data. Practically, the dataset gets split in k equal chunks. Each model is fit using $k-1$ chunks and the remaining chunk as the validation set. The procedure is repeated for a given model, each time taking one of the chunks of as the validation set, the average validation error is then calculated for each model. The best model is normally the one with lowest cross-validation error. However, we also require model parsimony, and since the model with the lowest cross validation error may not always be the best, we want to avoid a non-sparse model being chosen. To do that, we sort the

models by complexity and choose the model with the greatest cross-validation error reduction.

In conventional approaches the drift and diffusion are both computed as functions of x using the respective conditional moments. However, in contrast to those, the precision of the estimated functions F and G does not depend on the sampling interval Δt when using the Pydaddy approach. Both the requirement to subsample the time series and the elimination of the bin-width parameter from the estimation procedure are brought to an end as a result of this. This method, which combines sparse regression with the instantaneous drift and diffusion, enables us to do away with two arbitrary parameter choices that were previously involved in the estimation procedure. These choices were the bin width and the subsampling time scale. In contrast to this, the traditional methods necessitate the selection of the value of Δt that is most suitable.

Chapter 5

Results

5.1 Analysis of Fish Data

We begin with time series data of groups of fish, for each fish in any group, we have its $x(t)$ and its $y(t)$. Our data consists of 5 groups of fish containing 4, 10, 60, 80 and 100 individuals, respectively. The first thing we will do is take a look at basic trajectory analysis and collective behavior analysis to get a sense of what our systems look like and how they behave, especially since we are not interested in comparing the behavior of different group sizes.

5.1.1 *Trajectories*

Our fish are swimming freely in a shallow tank with a radius of approximately 25 fish body lengths. The details of the zebrafish rearing, handling, and experimental set-up are found in [18], The figure below shows the trajectory covered by the group of four fish at different time steps.

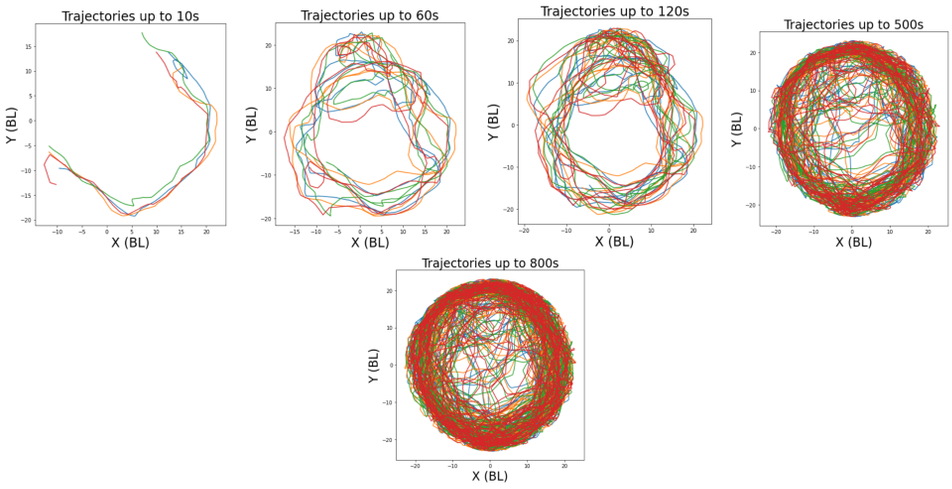


Figure 5.1: Covered trajectories for the group of four fish at different time instances

It is easy to infer from graph 5.1 that the fish tend to swim in circles at the edge of the tank, which prompts us to take a look at the spatial distribution of the fish

within the tank. This is achieved by using the method of Gaussian Kernel Density Estimation (KDE), refer to Appendix B for more details.

Each fish's positions are used to compute a 2-dimensional Gaussian KDE, which effectively gives us a smoothed, continuous approximation of the distribution of the fish's positions. This allows us to identify areas of the tank where the fish tends to spend most of its time (high density) and areas where it rarely goes (low density).

The resulting KDEs are then visualized as heat maps, with color indicating the estimated density of positions. All heat maps share the same color scale, which facilitates the comparison of spatial distributions between different fish. Refer to the below for the results:

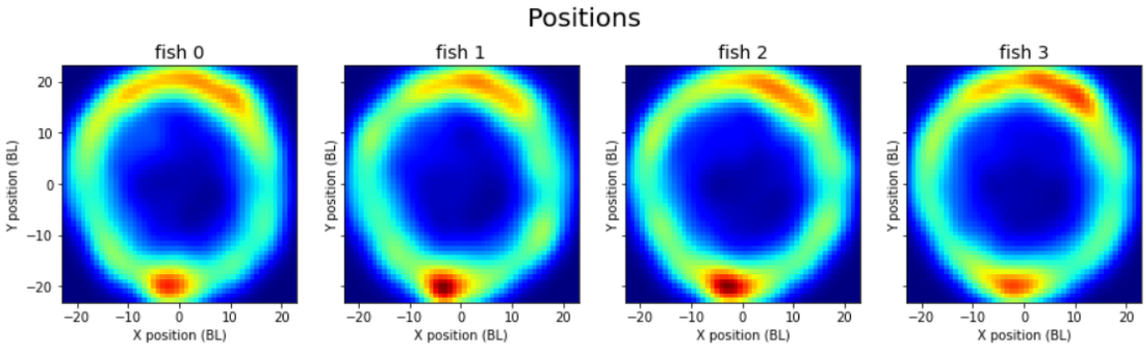


Figure 5.2: Positions heatmap with Gaussian KDE for four fish system

We clearly see how the fish spend most of their time on the periphery of the tank, in addition to that we see no clear distinction between the behavior of any one of the fish, in other words any fish could be substituted for another over long periods of time, this will be rendered clearer when we now look at the distributions of the speeds and acceleration.

5.1.2 *Speeds and Accelerations*

Following the spatial analysis, we now turn our attention to the study of the dynamic behavior of the fish, specifically their speeds and accelerations. Speeds and accelerations are fundamental quantities in the study of motion and provide critical insights into the overall dynamics of the system.

We are particularly interested in examining the distributions of these quantities. Distributions, essentially, tell us about the statistical properties of our measurements: where the bulk of the values lie (central tendency), how spread out they are (variability), and what range of values they take on (extremes). In the case of our fish, examining the distributions of speed and acceleration could reveal typical or average behavior, variations, and possible extremes.

Importantly, we compare these distributions across different fish. This is an es-

sential step as it allows us to investigate whether individual fish behave similarly or distinctly when it comes to their speed and acceleration. Identifying common patterns across multiple individuals can suggest shared behavioral characteristics or responses to the environment.

Moreover, we consider the aggregated distribution of speeds and accelerations from all fish combined. This provides a holistic view of the dynamics of the entire group, going beyond individual behaviors. Comparing this total distribution with those of individual fish allows us to assess whether the group behavior is merely an average of individual behaviors or if there are emergent properties at the group level.

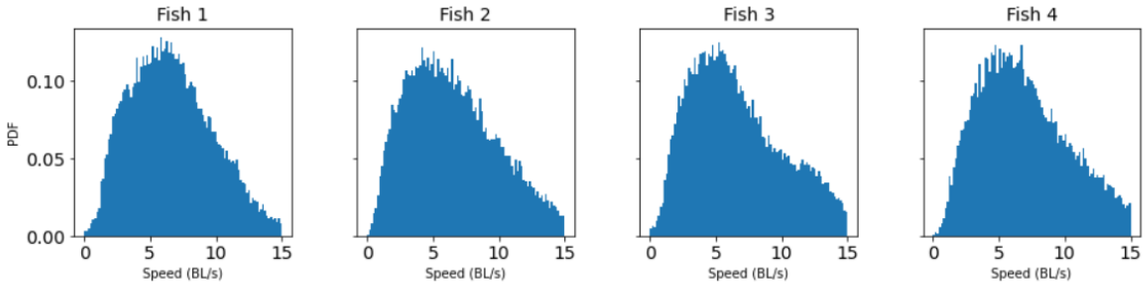


Figure 5.3: Speed PDF for the four fish system

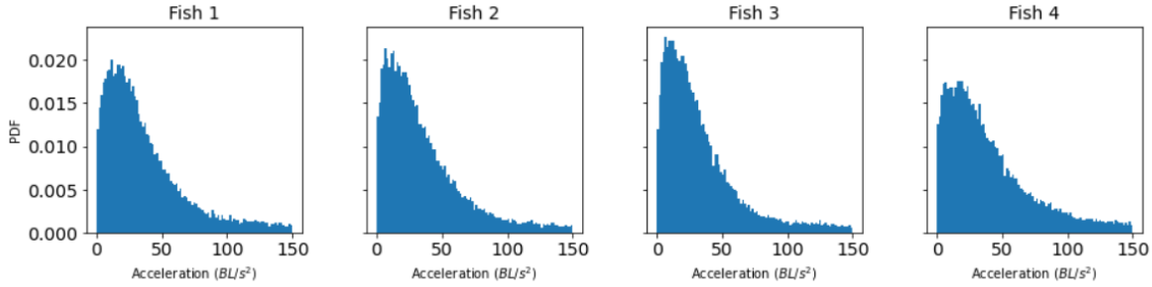


Figure 5.4: Acceleration PDF for the four fish system

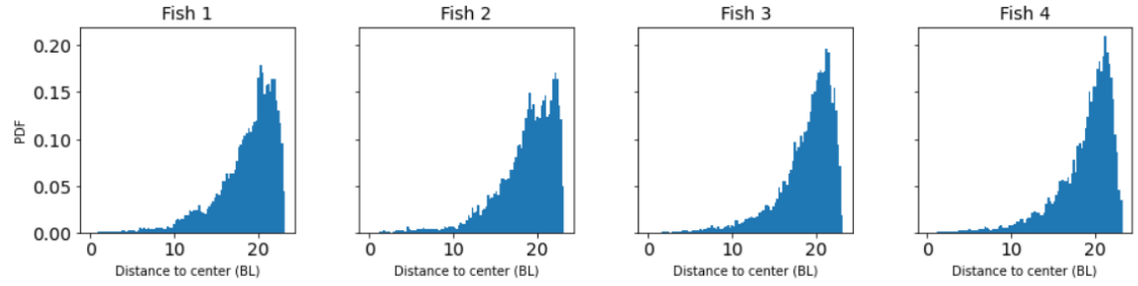


Figure 5.5: Distance to Center PDF for the four fish system

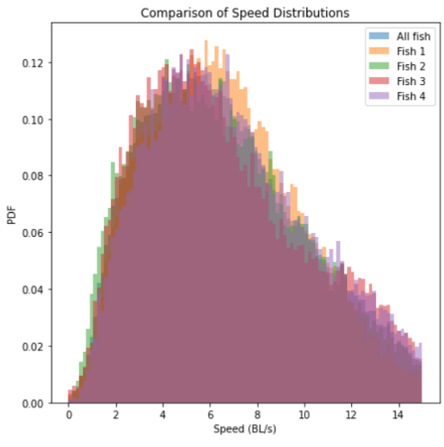
A significant observation emerges: there is no discernible distinction between the distributions of speeds and accelerations among individual fish, nor is there a dis-

tinction between an individual's distribution and the aggregate distribution of all fish combined as will be seen below.

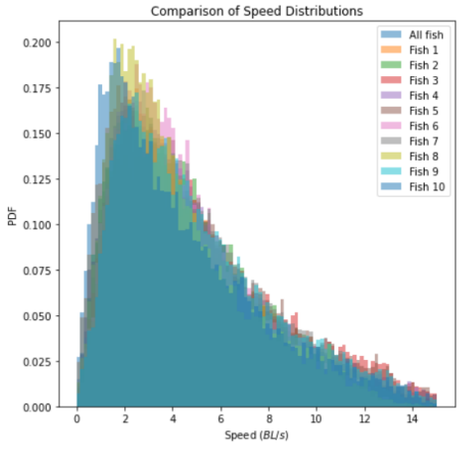
This result is particularly noteworthy as it implies a high level of homogeneity in the dynamics of individual fish within the group, as well as a strong correspondence between individual and group behaviors. Essentially, this suggests that each fish is moving with similar speed and acceleration patterns as its peers, and that the overall group movement is well-represented by the movement of any single individual.

5.1.3 *Comparison Between The Five Groups*

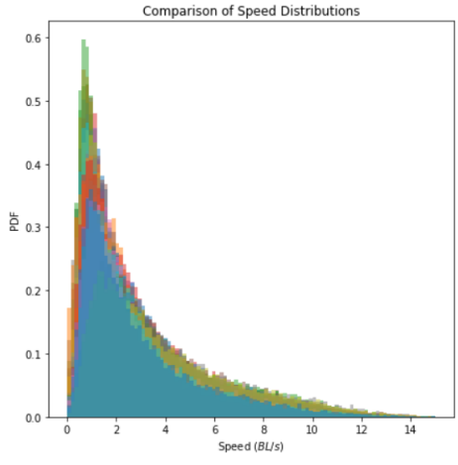
We've looked at the system of four fish so far, it's now worthwhile to look at the distribution from the the different group sizes for the speeds and acceleration, we show below the mentioned distributions for all the fish for each of the group sizes.



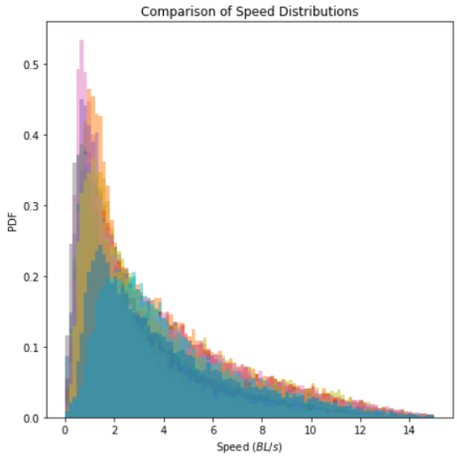
(a) Speed PDF: 4 fish



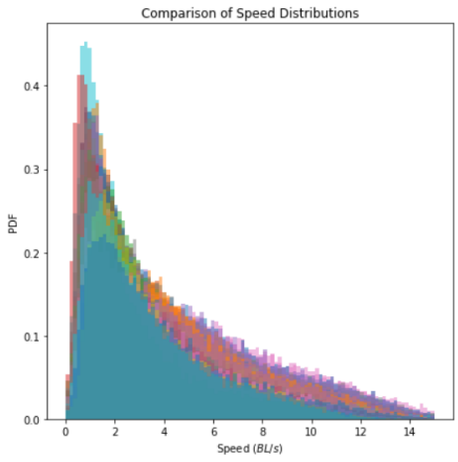
(b) Speed PDF: 10 fish



(c) Speed PDF: 60 fish



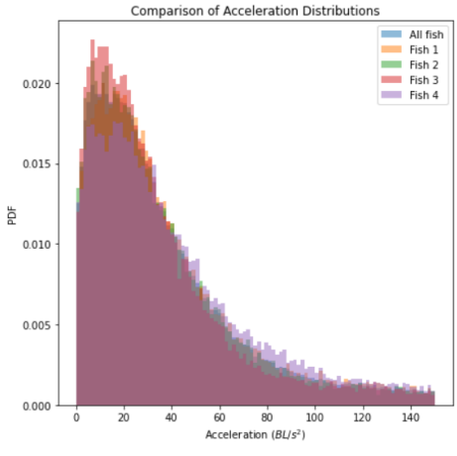
(d) Speed PDF: 80 fish



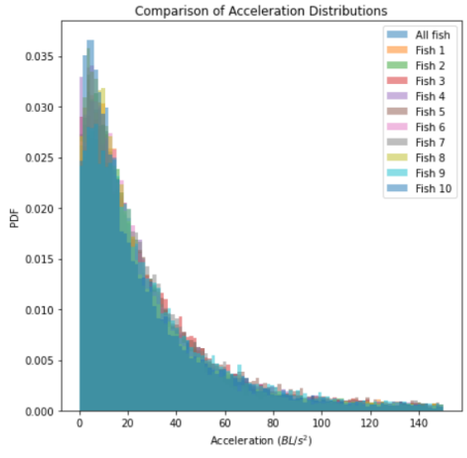
(a) Speed PDF: 100 fish

Figure 5.7: Distribution for the Speeds of the fish in the different systems

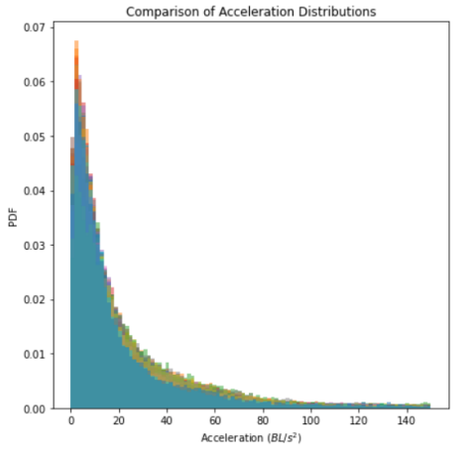
And the same thing for the accelerations:



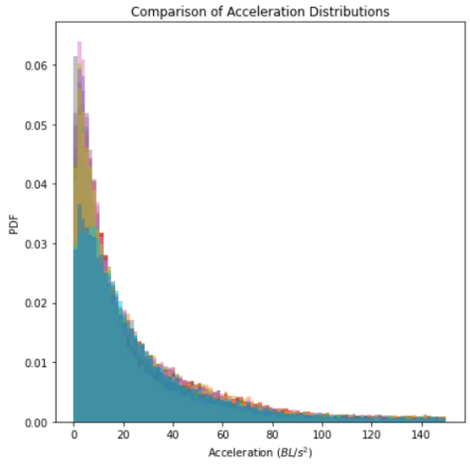
(a) Acc. PDF: 4 fish



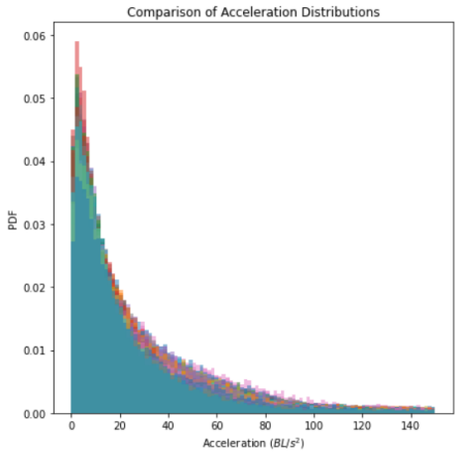
(b) Acc. PDF: 10 fish



(c) Acc. PDF: 60 fish



(d) Acc. PDF: 80 fish



(a) Acc. PDF: 100 fish

Figure 5.9: Distribution for the Acceleration of the fish in the different systems

From the above plots, we see that the distributions of the speeds and accelerations is very much the same for all individuals as well as the aggregate speed and aggregate acceleration for all the agents, regardless of group size. Of course we see that as the group size is large there exists some differences in the distribution since they don't overlap the same way they do for smaller group size, however there's no reason to believe that this is due to any actual significant difference between the individual and purely due to the limited length of the time series being studied, especially that the tank will tend to get crowded as the group gets larger, which means that during the period when the system is studied some individuals will not be able to explore all possible speeds and acceleration the same way that the 4 fish freely swimming did. It is however clear that there's a tendency for the average speed, or the most frequent speed, as well as the acceleration to move to the left as the group size grows, this is to be expected since density and speed are inversely proportional, to that end we look at the aggregate distribution of the speeds and velocity as a function of group size:

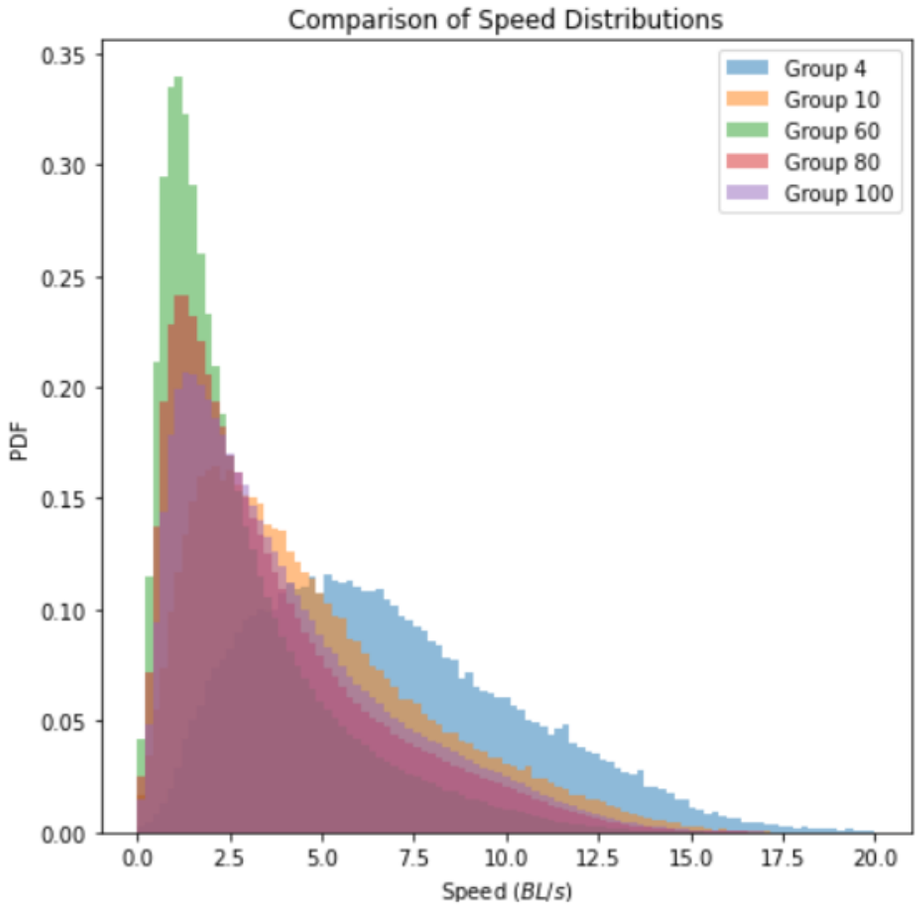


Figure 5.10: Comparison of Speed PDF for all groups

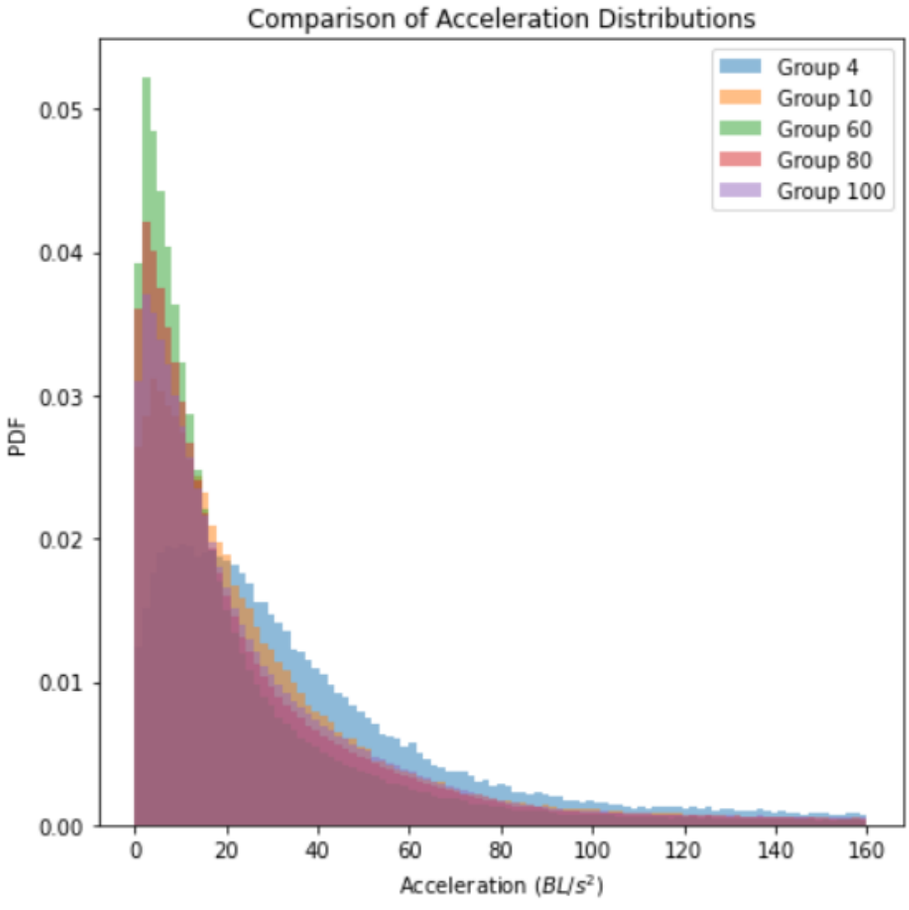
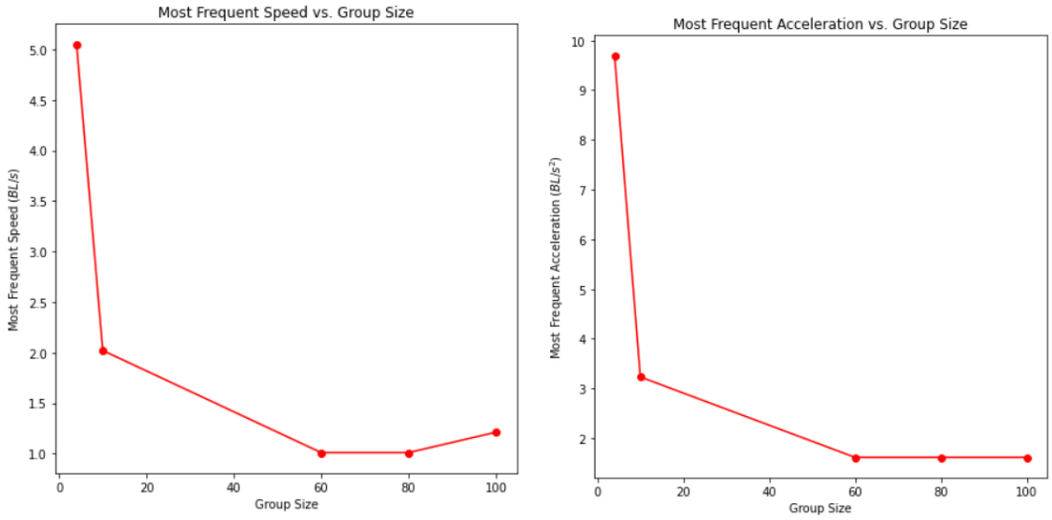


Figure 5.11: Comparison of Acceleration PDF for all groups

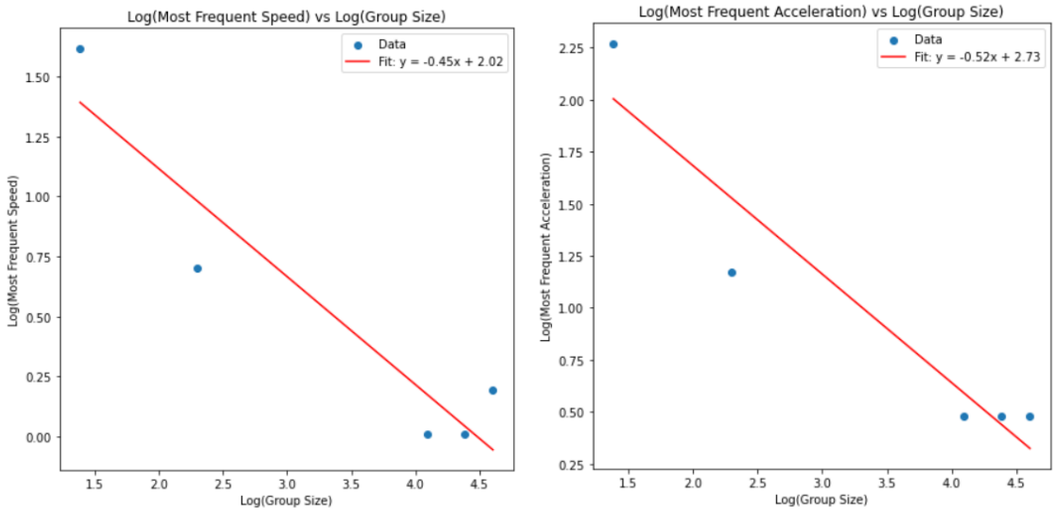
The tendency of the peak to move to the left as the group size increases is now clearer, for both speed and acceleration, we track that peak as a function of group size in the below:



(a) Peak speed vs group size (b) Peak acceleration vs group size

Figure 5.12: Values of the most frequent speeds and accelerations as a function of group size

And below are their corresponding fits on a log-log plot, since we suppose the tendency to be somewhat of a decreasing exponential for both:



(a) Best fit: Peak speed vs group size (b) Best fit: Peak Acc. vs group size

Figure 5.13: Best fits for the most frequent speeds and accelerations vs group size

We infer the relationships between peak velocity (v_{peak}), peak acceleration (a_{peak}) and group size (N) from our data, albeit cautiously given the limited number of data points.

Assuming a power-law relationship between peak velocity (v_{peak}), peak acceleration (a_{peak}) and group size (N), we propose initial relationships of the forms:

$$v_{\text{peak}} = aN^{-\nu} \tag{5.1}$$

$$a_{\text{peak}} = bN^{-\mu} \tag{5.2}$$

where a and b are proportionality constants and ν and μ are the exponents of the power law.
To make the estimation of these parameters more manageable, we transform our data and the relationships by taking the natural logarithm of both sides. This yields the linear equations:

$$\log(v_{\text{peak}}) = -\nu \log(N) + \log(a) \tag{5.3}$$

$$\log(a_{\text{peak}}) = -\mu \log(N) + \log(b) \tag{5.4}$$

In these linear forms, we proceed with a least-squares fitting approach. Upon obtaining the best fit lines, the slopes provide the values of ν and μ, while the y-intercepts correspond to $\log(a)$ and $\log(b)$.
From our data analysis, we derived the following relationships:

$$\log(v_{\text{peak}}) = -0.45 \log(N) + \log(7.5) \tag{5.5}$$

$$\log(a_{\text{peak}}) = -0.5 \log(N) + \log(15.3) \tag{5.6}$$

Converting these back to their original forms gives:

$$v_{\text{peak}} = 7.5N^{-0.45} \tag{5.7}$$

$$a_{\text{peak}} = 15.3N^{-0.5} \tag{5.8}$$

These final expressions represent the power-law relationships between peak velocity, peak acceleration, and group size observed from our data.

5.1.4 *Calculation of Order Parameters*

The order parameters of interest in this study were the rotation order parameter and the polarization order parameter. These parameters were calculated using the time-dependent positions $x_i(t)$ and $y_i(t)$ for each fish i. The methods of calculating these order parameters are detailed below.

Rotation Order Parameter

The rotation order parameter quantifies the coordinated rotation of the fish about a central point. The process of calculating the rotation order parameter involves the following steps:

1. **Calculation of Displacement:** for each fish i, the displacement $\vec{d_i}(t)$ from the center of mass is calculated. If the center is a fixed point (x_c, y_c), this involves subtracting the center coordinates from the fish's position, i.e. $\vec{d_i}(t) = (x_i(t) - x_c, y_i(t) - y_c)$. If the center's position varies with time, the corresponding center position at time t is subtracted, i.e., $\vec{d_i}(t) = (x_i(t) - x_c(t), y_i(t) - y_c(t))$. In our case the center of mass position varies with time.

2. **Calculation of Angular Momentum:** The 2D cross product of the velocity $\vec{v_i}(t)$ of each fish and its displacement vector $\vec{d_i}(t)$ is calculated. This gives the scalar angular momentum $L_i(t)$ of each fish relative to the center, i.e., $L_i(t) = \vec{v_i}(t) \times \vec{d_i}(t)$.

3. **Summation of Angular Momenta:** The angular momenta of all fish are summed to give the total angular momentum $L_{\text{total}}(t)$ at each time point t, i.e., $L_{\text{total}}(t) = \sum_i L_i(t)$.

4. **Average Angular Momentum:** The average angular momentum per fish, also known as the rotation order parameter, is obtained by dividing the total angular momentum by the number of fish N, i.e., $OP_{\text{rotation}}(t) = \frac{1}{N} L_{\text{total}}(t)$.

Polarization Order Parameter

The polarization order parameter characterizes the degree to which the fish are moving in the same direction. It is calculated as follows:

1. **Velocity Calculation:** The velocities $\vec{v_i}(t)$ of the fish are computed by taking the difference between their positions at consecutive time points, i.e., $\vec{v_i}(t) = (x_i(t+1) - x_i(t), y_i(t+1) - y_i(t))$.

2. **Velocity Normalization:** Each velocity vector is normalized to a unit vector, i.e., $\hat{v}_i(t) = \frac{\vec{v_i}(t)}{|\vec{v_i}(t)|}$, where $|\vec{v_i}(t)|$ is the magnitude of $\vec{v_i}(t)$.

3. **Summation of Normalized Velocities:** The normalized velocities are summed to obtain the resultant vector $\vec{R}(t)$ for each time point t, i.e., $\vec{R}(t) = \sum_i \hat{v}_i(t)$.

4. **Polarization Order Parameter:** The polarization order parameter is computed by dividing the magnitude of the resultant vector by the number of fish N, i.e., $OP_{\text{polarization}}(t) = \frac{|\vec{R}(t)|}{N}$.

Here are the results for the different group sizes:

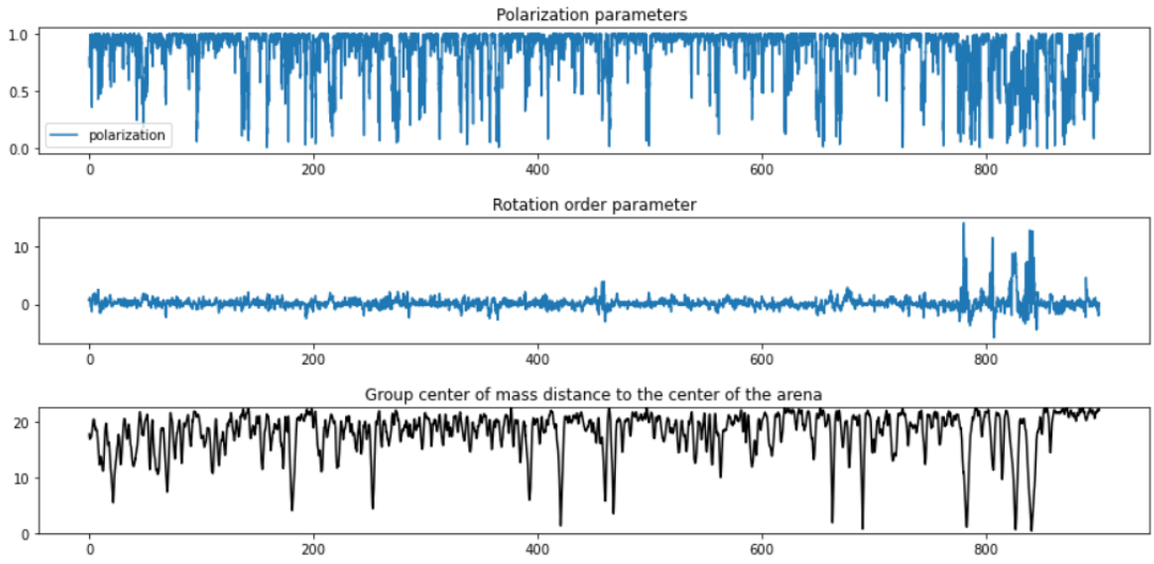


Figure 5.14: Order Parameters for the 4 fish group, time in seconds

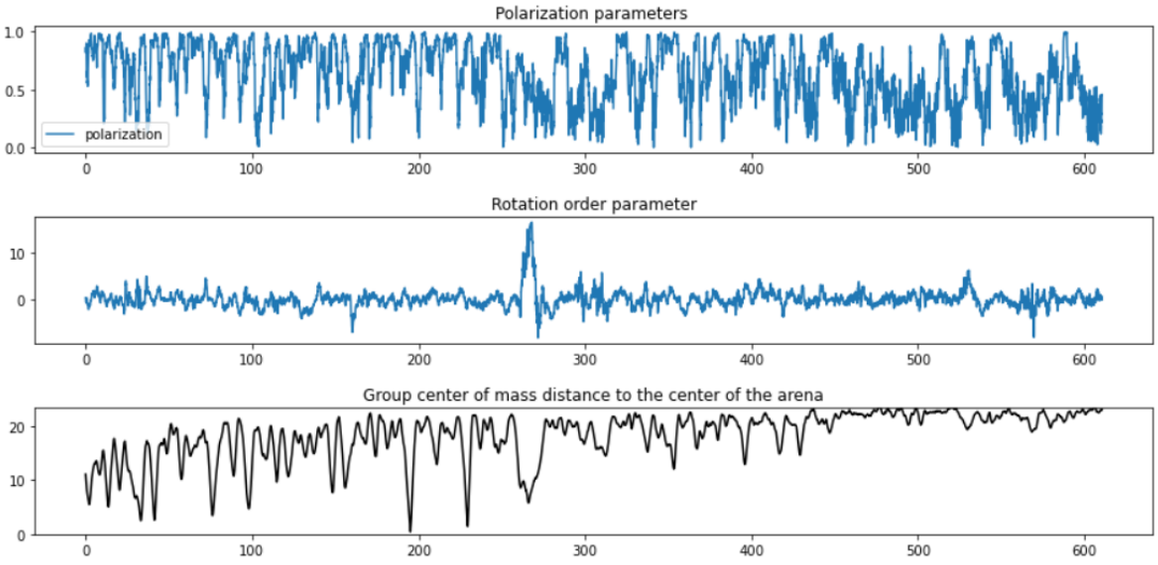


Figure 5.15: Order Parameters for the 10 fish group, time in seconds

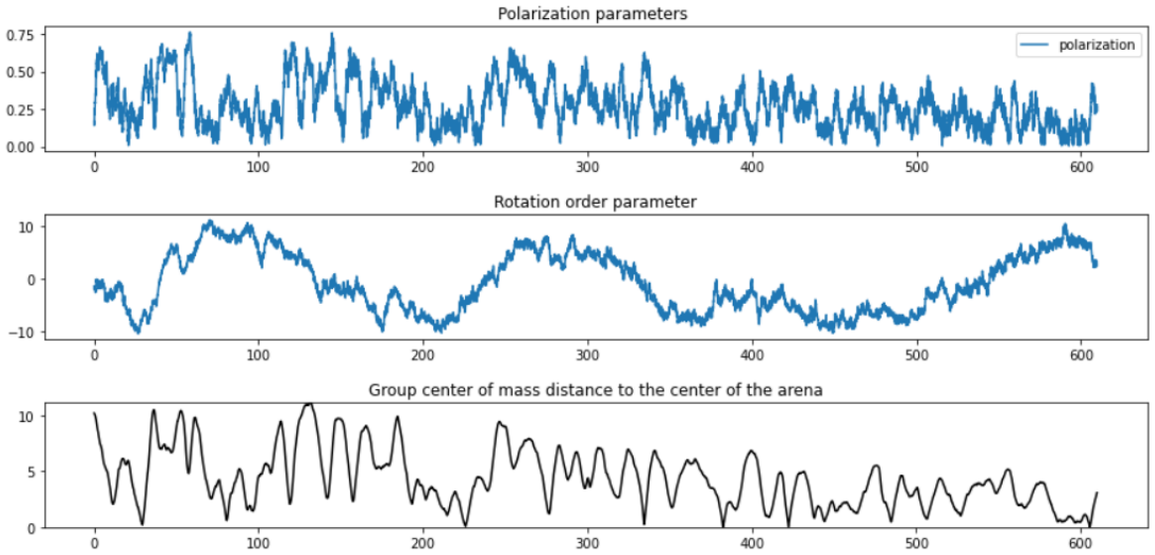


Figure 5.16: Order Parameters for the 60 fish group, time in seconds

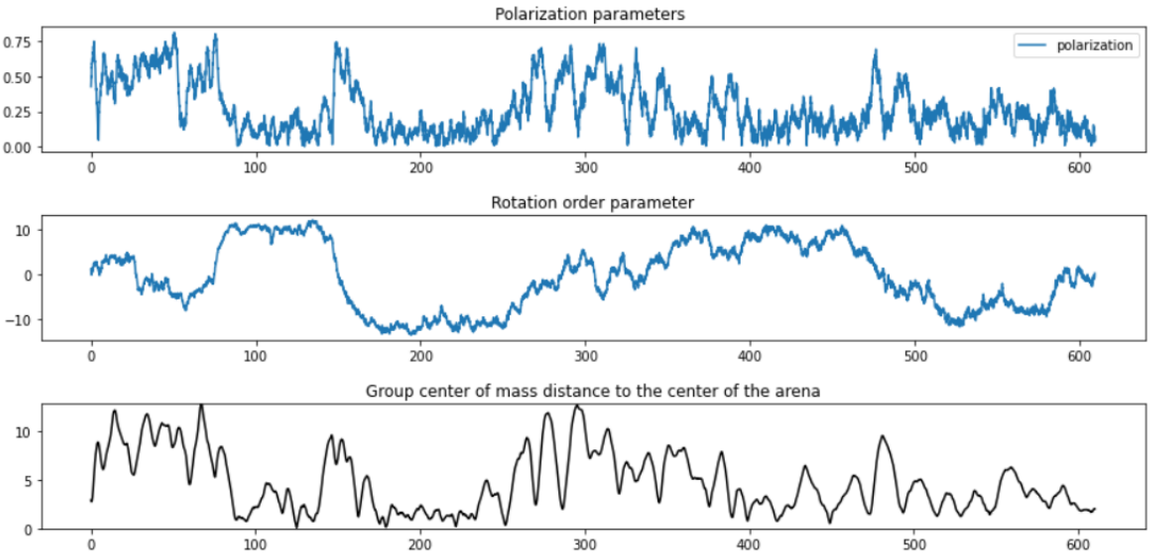


Figure 5.17: Order Parameters for the 80 fish group, time in seconds

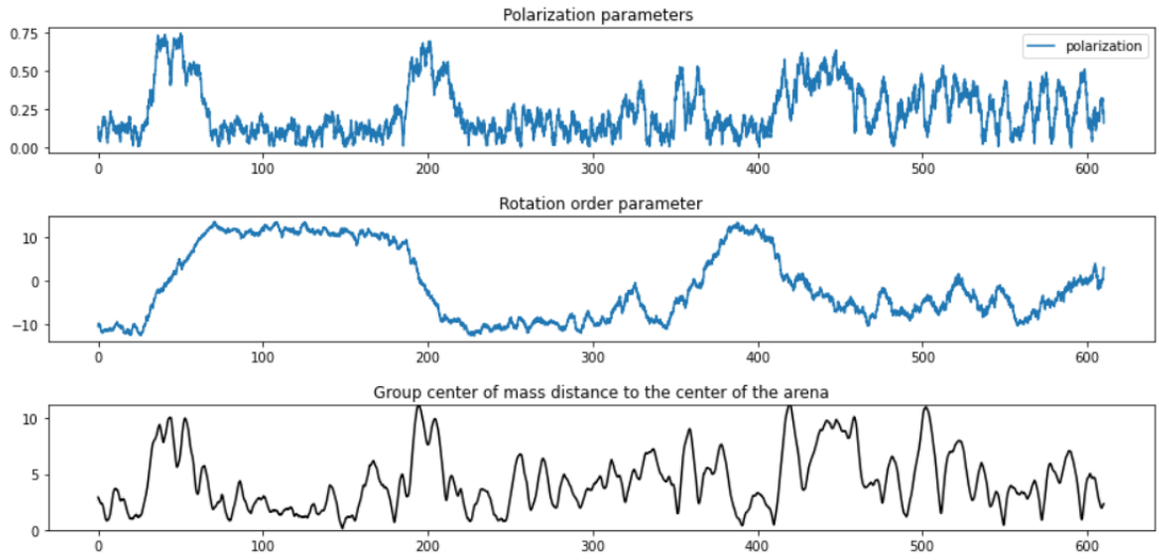


Figure 5.18: Order Parameters for the 100 fish group, time in seconds

Through our examination of the Polarization and Rotation order parameters across different fish group sizes, an intriguing interplay becomes evident as we transition from smaller groups of 4 fish to larger groups of 100 fish.
The Polarization order parameter measures the alignment of the fish group in their direction of movement. A high value indicates that the group members are moving coherently in the same direction, whereas a low value suggests random, uncoordinated movement. On the other hand, the Rotation order parameter provides a measure of the coordinated rotational movement around a group center. A positive value indicates a counterclockwise rotational motion, a negative value implies a clockwise motion, and a value close to zero represents a lack of coordinated rotational movement.

In smaller groups, such as those of 4 and 10 fish, we observed that the Rotation order parameter fluctuates around zero. This suggests that in these smaller groups, rotational motion around a common center is less prevalent, and any rotational behavior that does occur is more likely to be a random occurrence rather than a coordinated group movement. This observation aligns with the theory that smaller groups may lack the necessary structure to form a consistent vortex-like pattern, a behavior that would contribute to a non-zero Rotation order parameter.

As we move to larger group sizes, the Rotation order parameter shows a tendency to stabilize at either -10 or 10 (BL^2/s), indicating a consistent clockwise or counterclockwise rotation of the fish around the group's center of mass. This pattern can be attributed to the increased complexity of the larger groups' dynamics. More specifically, in larger groups, fish have more potential neighbors influencing their movement, leading to emergent rotational behaviors as they strive to align themselves with the nearest neighbors. This coherent rotational motion is a form of collective behavior often seen in larger animal groups, including fish schools and

bird flocks.

The Polarization order parameter, in contrast, does not show a clear trend with increasing group size. This is likely due to the interplay between alignment and rotational behaviors. In larger groups, despite the potential for increased alignment (higher Polarization), the emergent rotational behavior (indicated by the Rotation order parameter) can cause group members to have differing instantaneous directions even though they are following the same overall circular path. This can decrease the Polarization value, removing the expected positive correlation between Polarization and group size.

To quantify things further, and since there seems to be a tendency to persist in a certain state as groups get larger, we thought it would be interesting to look at the decay time of the auto-correlation functions of both the rotation and polarization order parameter, to do so we obtain the auto-correlation function and fit to it a decaying exponential of the for $ae^{-bx} + c$ and considered the decay time to be the time it takes for this exponential to drop by $1/e$ of it's maximum/initial value, below we summarize our findings and compare the different decay times for the different group sizes:

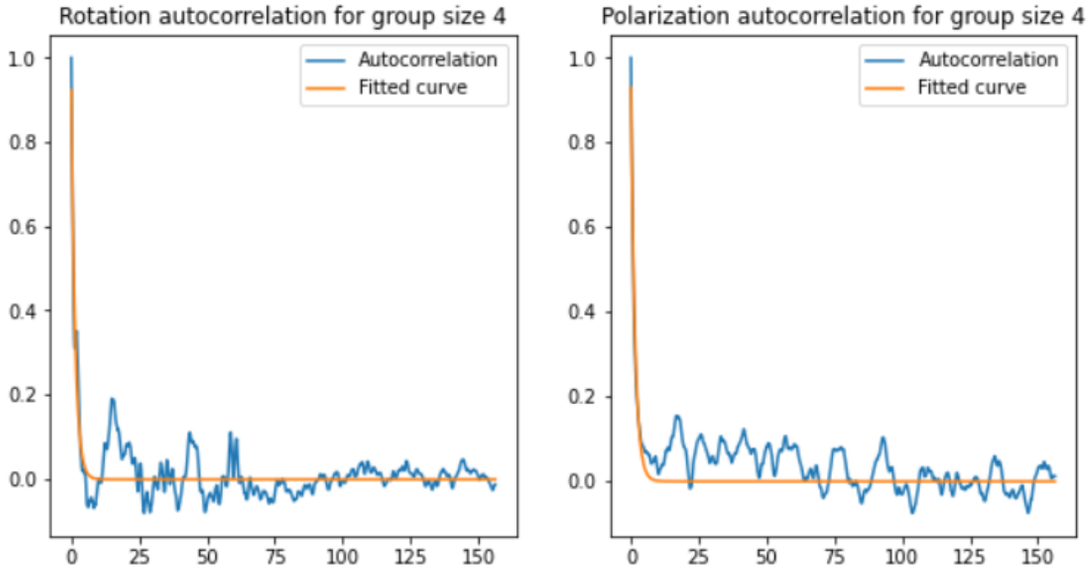


Figure 5.19: Auto-correlation function and exponential fit of order parameters for the group of 4 fish

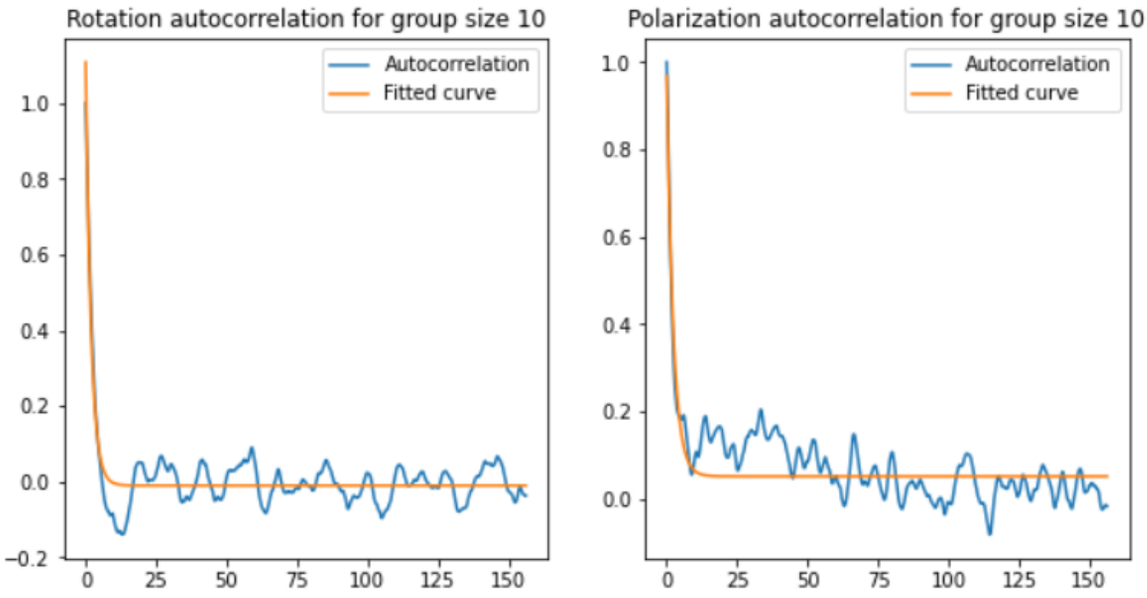


Figure 5.20: Auto-correlation function and exponential fit of order parameters for the group of 10 fish

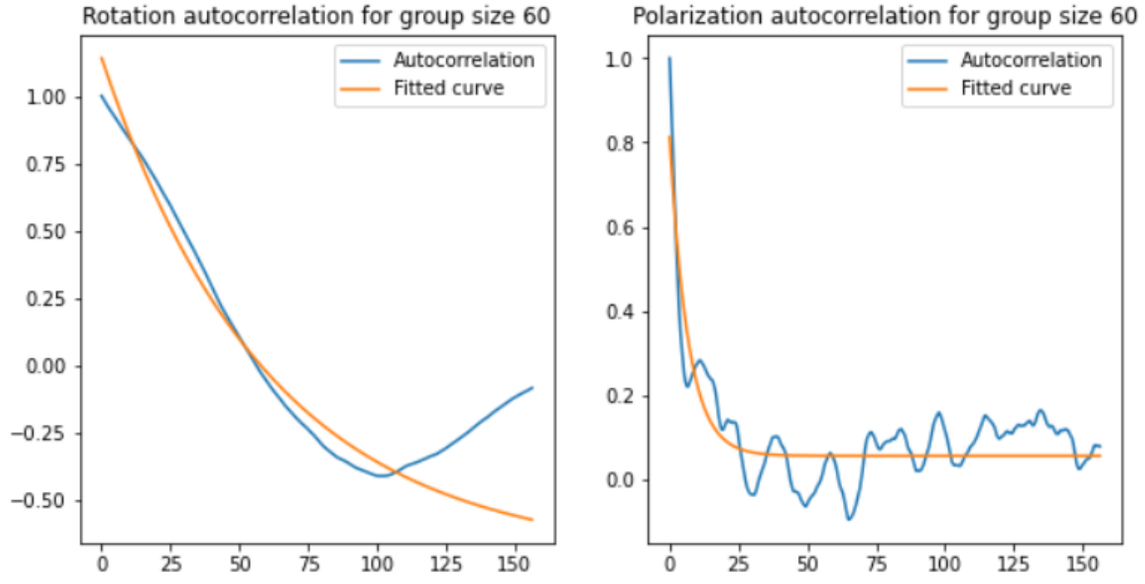


Figure 5.21: Auto-correlation function and exponential fit of order parameters for the group of 60 fish

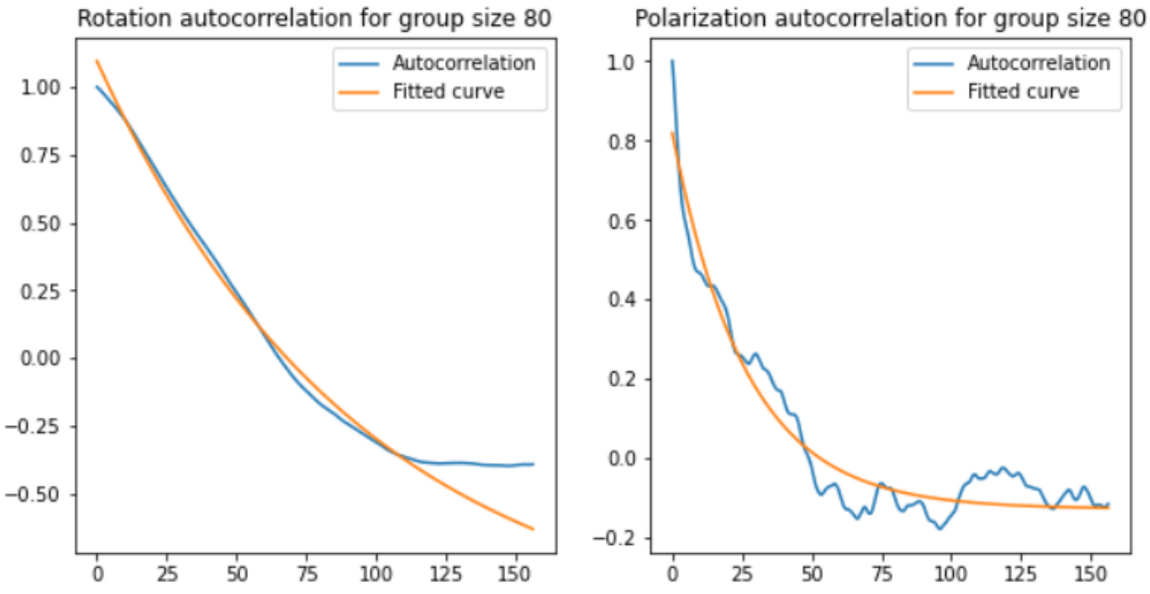


Figure 5.22: Auto-correlation function and exponential fit of order parameters for the group of 80 fish

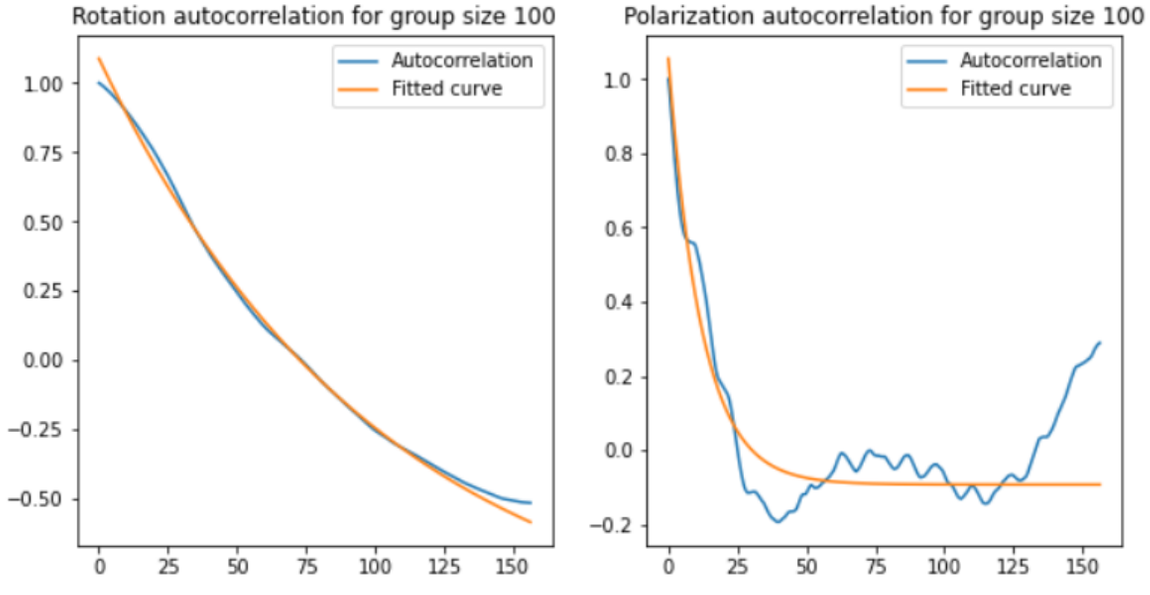


Figure 5.23: Auto-correlation function and exponential fit of order parameters for the group of 100 fish

And here's a summary of the decay rates, we notice clearly a tendency to persist in rotational motion much more than polarization, by the nature of the circular tank and the behavior of these types of active systems:

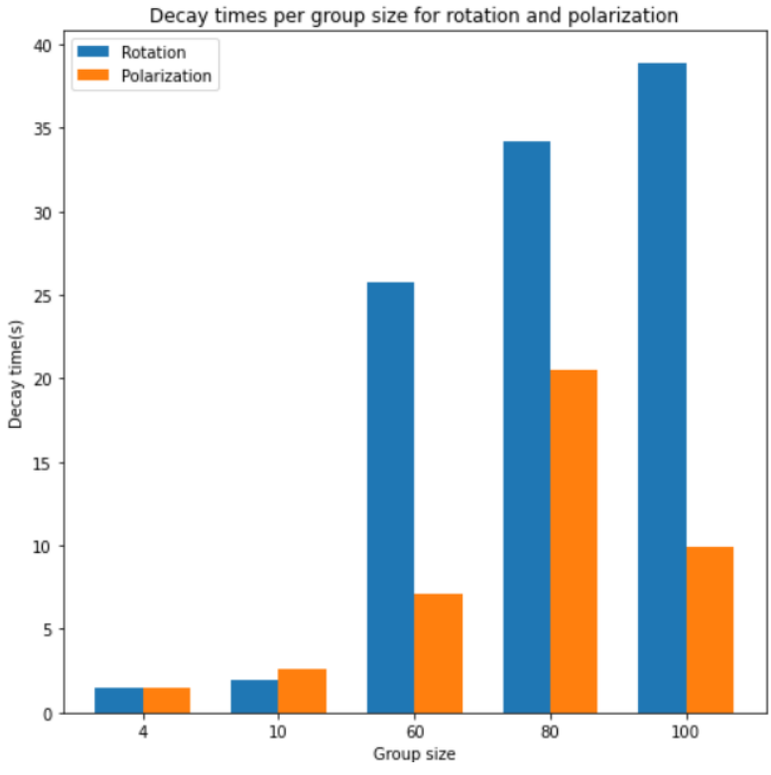


Figure 5.24: Comparison in the decay rates of the order parameters for all group sizes

In summary, our observations highlight the intricate balance and interplay between different forms of collective behavior, specifically alignment and rotation, in fish schools of varying sizes. Understanding these dynamics is key to understanding the complexities of group behavior and has potential implications in various fields, including the study of biological systems, robotics, and crowd control.

5.2 Causation Entropy and Network Recovery

While over long periods it seems that no fish can be distinguished from another, we want to explore whether that holds true for shorter time scales, and we are especially interested in the information flow among the agents during the short time periods.

To that end, we aim to use the Optimal Causation Entropy Principle (oCSE) to recover temporally evolving directed causal network structures for three groups of fish, size 4, 10 and 60 due to computational limitations while performing the recovery on the larger groups of 80 and 100. We believe the group of 60 fish is fairly representative of of the behavior of dense fish systems given all previous analysis of the behavior of the order parameters in the large groups (60, 80 and 100).

Employing the oCSE algorithm explained in its respective section give us three series of directed, temporally evolving, networks where the weight of the links represents

the value of the Causation Entropy (i.e the Causality Strength for our purposes), in the following, we use the acceleration time series for two reasons: first because it represents in active matter what's called "social force", and because it shows the least autocorrelation, we then set the time steps $\tau = 5$, i.e 5 time steps since we find that the average cross-correlation maximum lag to be at 5 time steps (0.16s), we also run 500 trials per hypothesis test, we chose hypothesis tests with significance level $\alpha = 0.1$ in both the aggregation and removal phases of the oCSE algorithm, and K=4 as the KSG estimator parameter, a very small value picks up too much noise, and a very high value will smooth out a lot of important information.

The goal of that analysis is two-fold:

1. First, we're interested in understanding the information transfers at the fish level, in other words understanding if there are prominent/persistent leader fish in any group size, how long does a leader's term last, etc. And knowing how many individuals does each fish influence on average and how many individuals does a fish keep track of on average.

2. Second, we're interested in making use of graph theory and network analysis tools in order to understand information flow on a macro level, hoping that this will give us insight into why larger group sizes seem better coordinated and have longer "memory" times as compared with smaller groups for their rotation and polarization order parameters.

5.2.1 *Causal Parents Evolution*

Now that we have the networks, we begin by looking at the evolution of the causal parents for random nodes (fish) in our system, to show how fish in larger group sizes tend to longer bonds than those in the smaller systems:

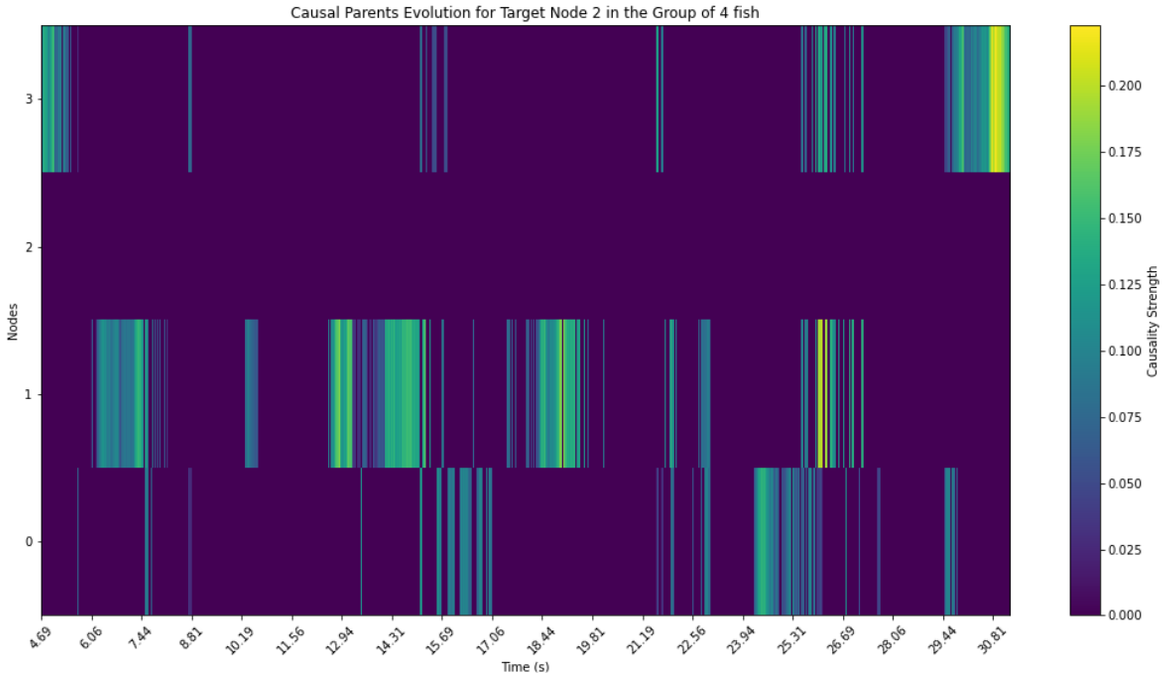


Figure 5.25: Evolution of the Causal Parents for a fish in group 4

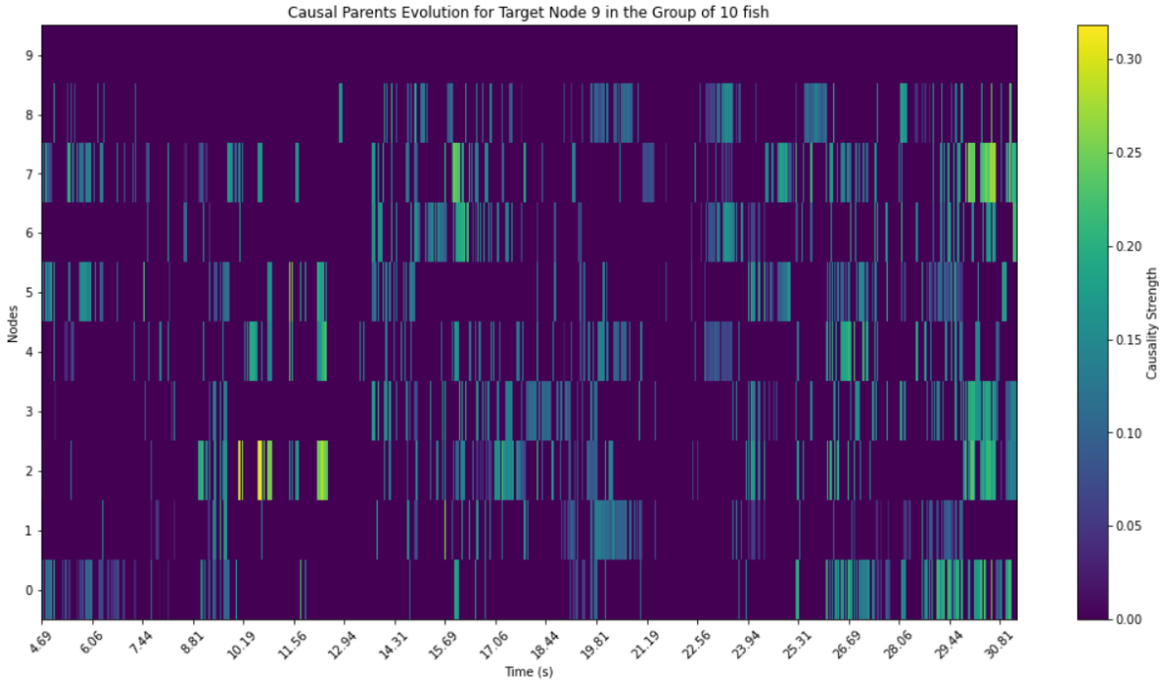


Figure 5.26: Evolution of the Causal Parents for a fish in group 10

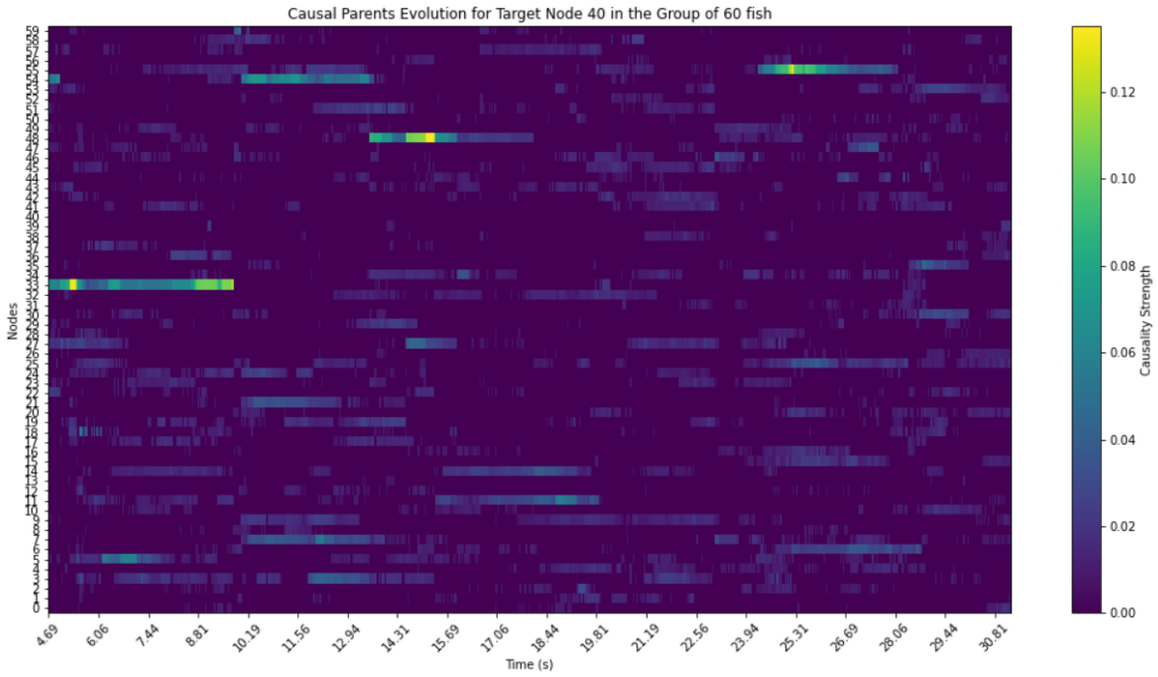


Figure 5.27: Evolution of the Causal Parents for a fish in group 60

It seems from the graphs above that the group of 60 might exhibit formations of causal bonds for longer times than the other two groups, with the group of 10 fish showing the least amount of long term bonds, one reason could be that in the group of 4 fish some causal pairs might appear for some time and tend to persist until another fish comes and knocks out the causal pair, while in the system of 10 it's much more likely for the whizzing fish to keep knocking out causal pairs. Until the density is high enough for some causal stability to arise.

Now we look at the causal networks build over the real space, the way we do that is by considering the average position of each fish during the time-series window upon which we built the causal network, which is 150 frames. And since we know that the average might not be enough information of how that fish moved and could be a misleading representation we show circles centered at the average position of each fish with a radius equal to the standard deviation of the position vector.

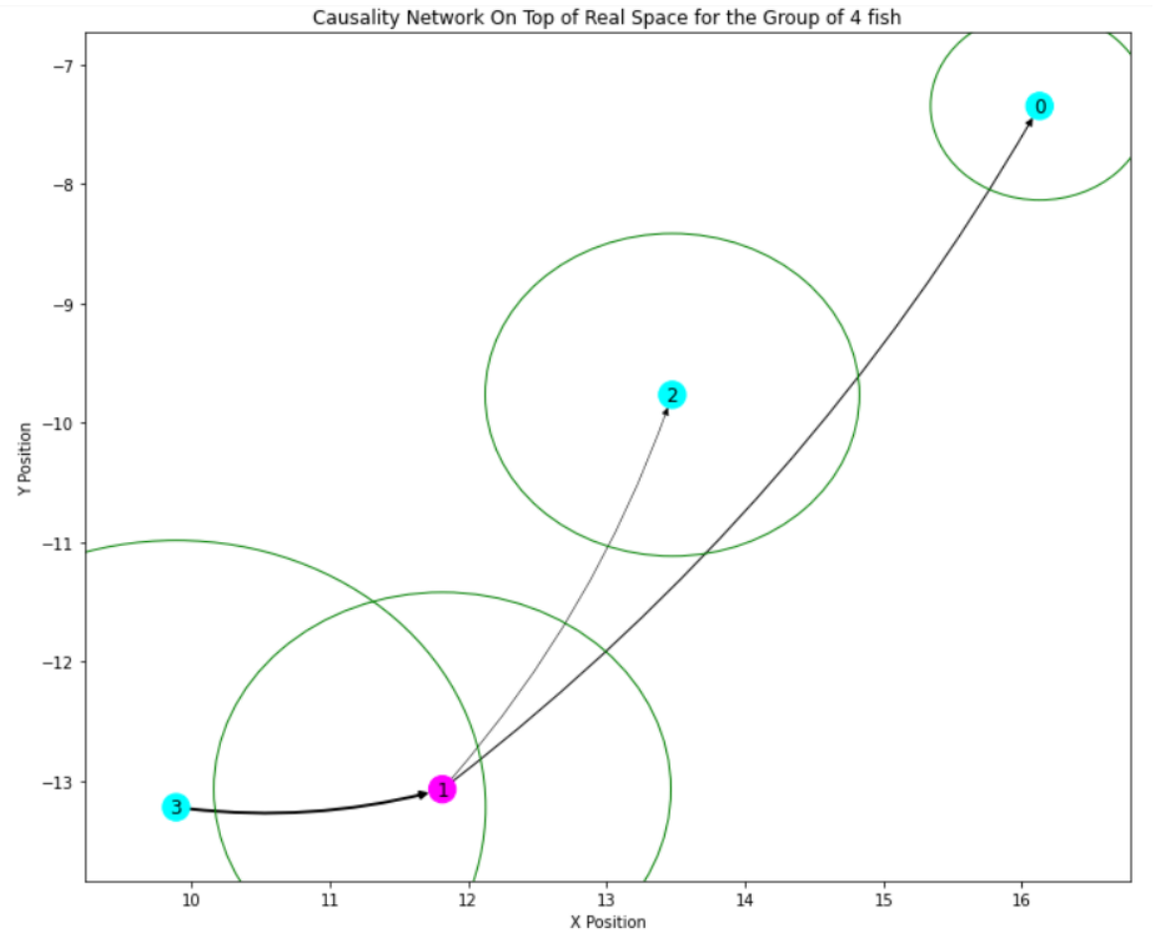


Figure 5.28: Causality network for group 4, where the arrows show the directed causal links, and the red circles have a radius equal to one standard deviation of the position vector of the time window used to build the network

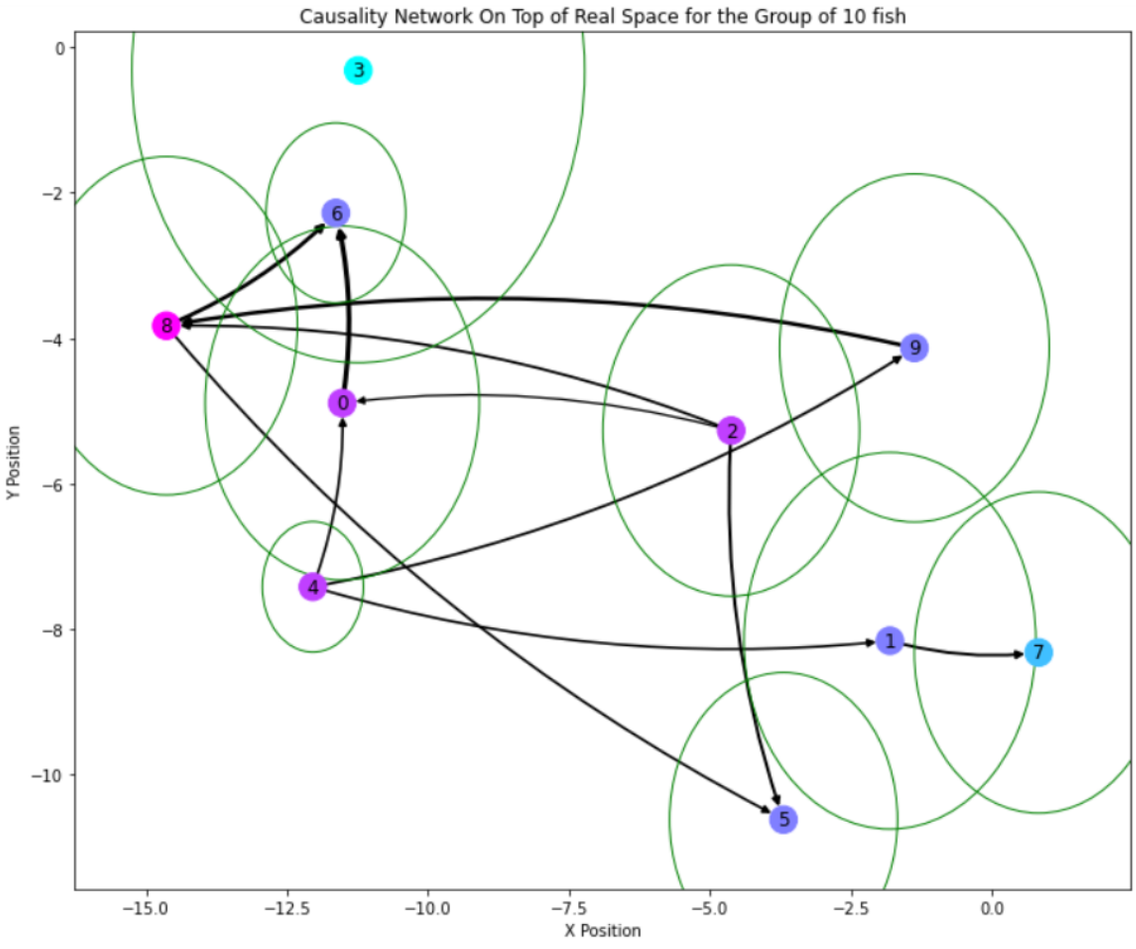


Figure 5.29: Causality network for group 10, where the arrows show the directed causal links, and the red circles have a radius equal to one standard deviation of the position vector of the time window used to build the network

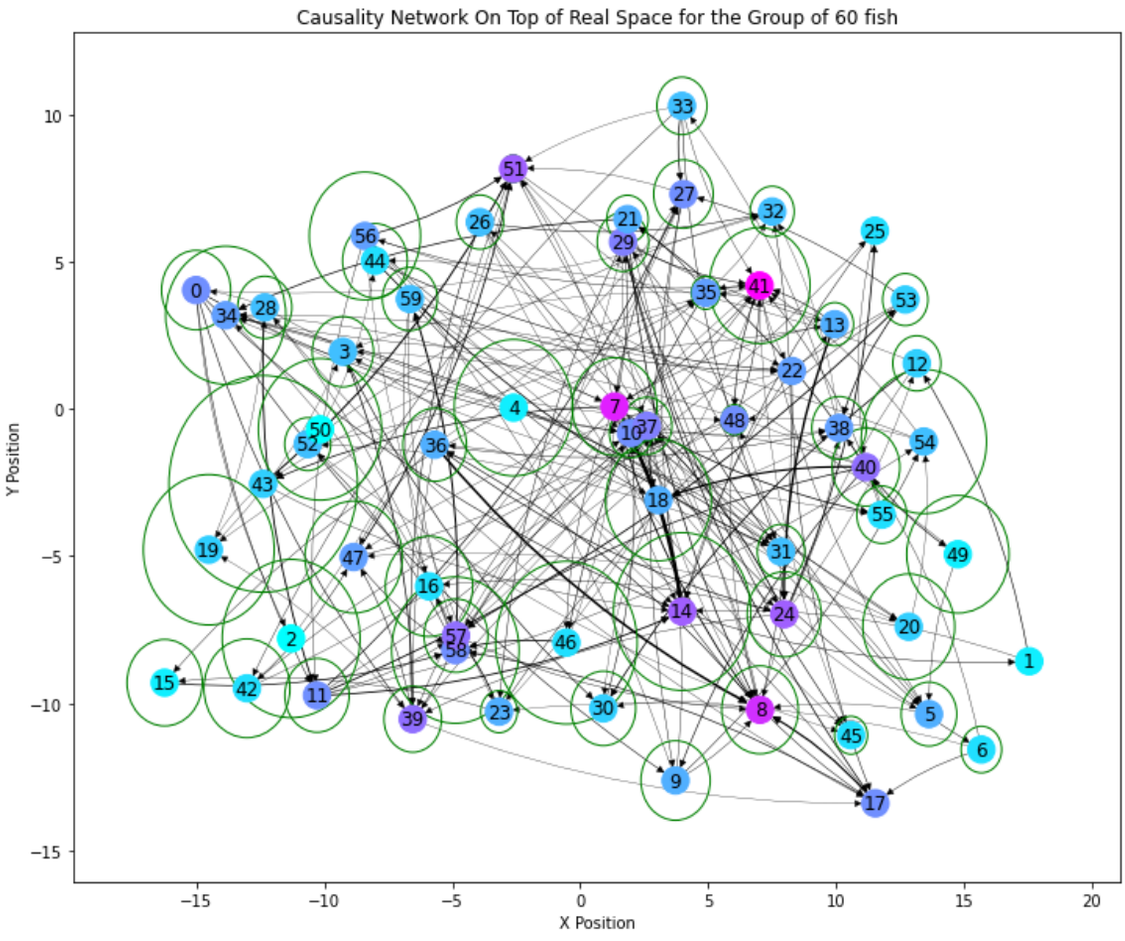


Figure 5.30: Causality network for group 60, where the arrows show the directed causal links, and the red circles have a radius equal to one standard deviation of the position vector of the time window used to build the network

5.2.2 *Leadership and Influence*

Next, since we have temporally evolving networks that tell us how each fish is interacting with it's neighbors as a function of time, we go back to our original questions: does leadership arise? how to quantify it? and how does it change as a function of group size?

To that end we define a "leadership" time series, which is simply the sum of the causal outflow from a fish at any time instant, it's an intuitive measure of leadership. In network science lingo that is called the "Weighted Out-Degree", the sum of the weights of the links leaving a node. We define as well the "followership" time series, simply as the the sum of the causal inflow onto a node. Put simply, the node with the highest out-degree is the leader.

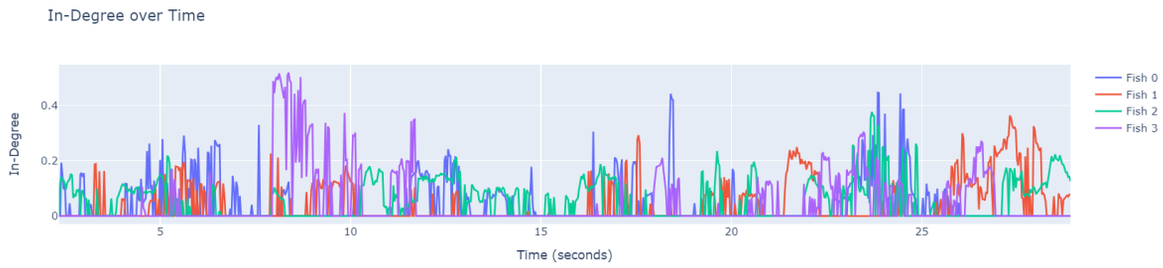


Figure 5.31: In degree time-series for each fish in the group 4, i.e followership time-series

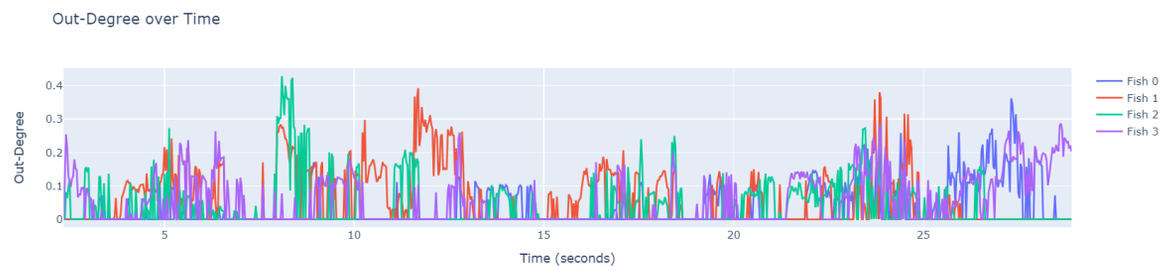


Figure 5.32: Out degree time-series for each fish in the group 4, i.e leadership time-series

Now we look at leadership from another perspective, while out definition of leadership as the total causal outflow from a link is quite intuitive, it's worth making use of a network science concept called betweenness centrality to look at leadership from another perspective. In the study of network science, betweenness centrality is a measure of a node's centrality, or importance, in a network. It quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

Formally, the betweenness centrality $C_B(v)$ of a node v is given by the expression:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{5.9}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v.

In the context of our zebrafish group, nodes represent individual fish and edges represent interactions between them. Importantly, in our network, 'distance' is not spatial but is instead defined in terms of inverse causal strength. Hence, a 'shorter' path in this network represents a path of stronger causal interaction. Therefore, a fish with high betweenness centrality may not directly influence many other fish but plays a critical role in the network due to its position along paths of strong causal interaction.

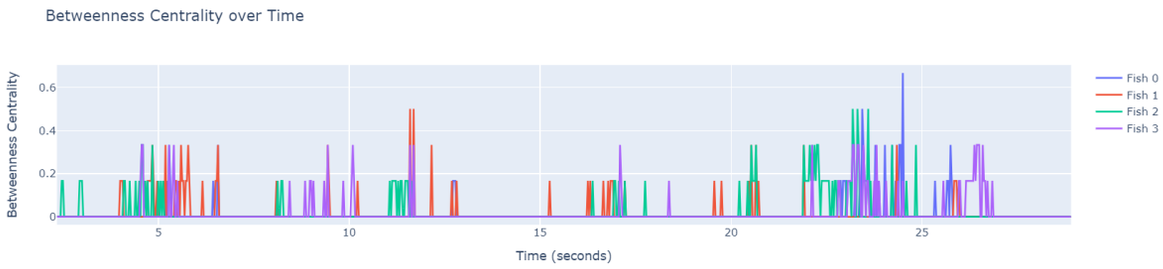


Figure 5.33: Betweenness Centrality time-series for each fish in the group 4

Below are the results for the group 10 and group 60:

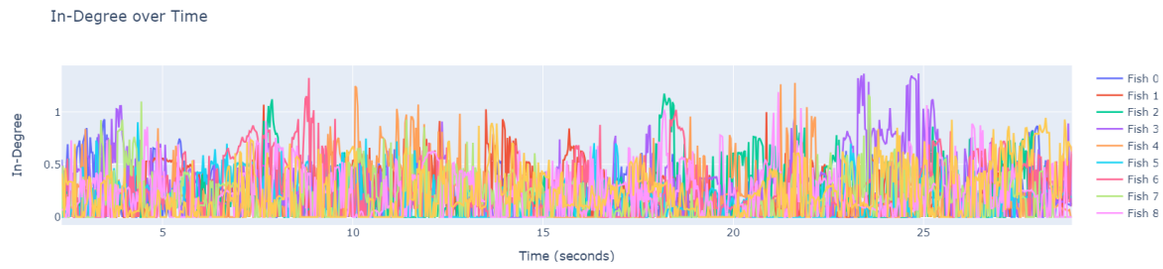


Figure 5.34: In degree time-series for each fish in the group 10, i.e followership time-series

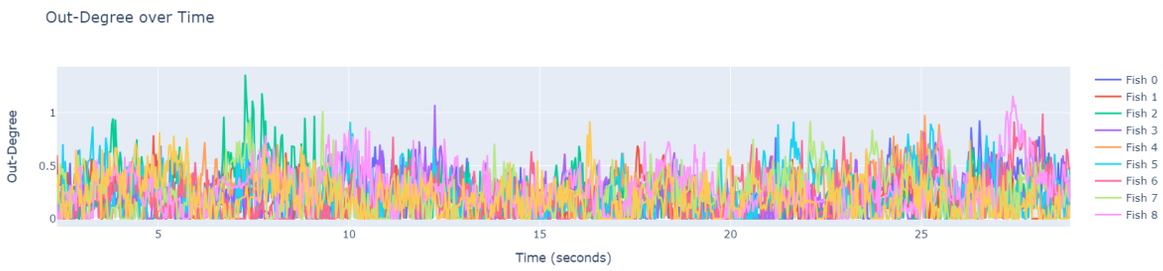


Figure 5.35: Out degree time-series for each fish in the group 10, i.e leadership time-series

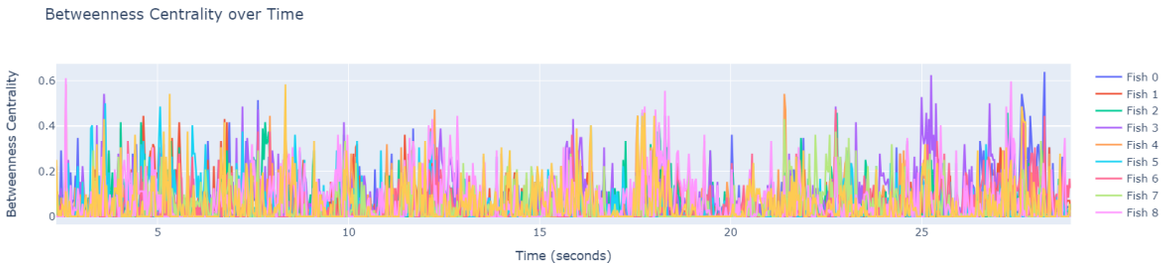


Figure 5.36: Betweenness Centrality time-series for each fish in the group 10

Clearly the noise hides most of the details, so for visual purposes we show the same results but with a moving average of window 15 steps. And we show directly the results of the group 60 with a moving average.

Group 10:

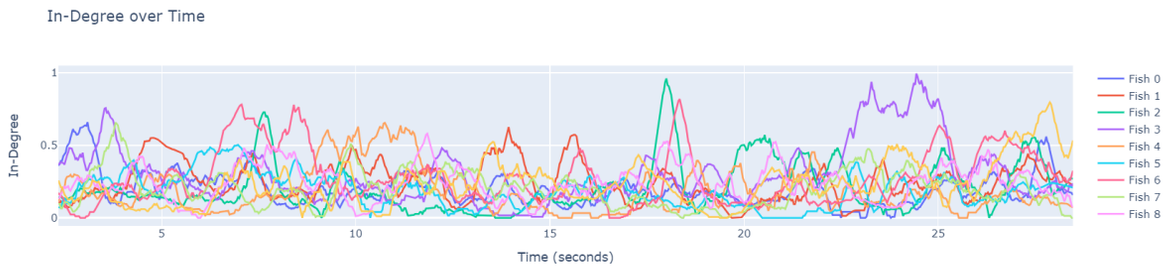


Figure 5.37: Moving Average: In degree time-series for each fish in the group 10, i.e followership time-series

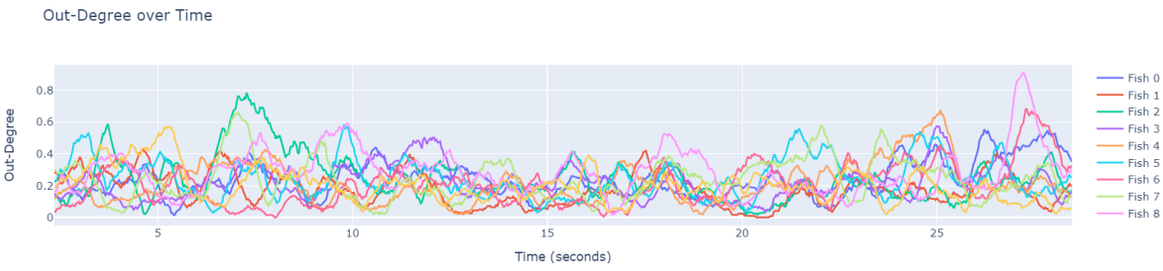


Figure 5.38: Moving Average: Out degree time-series for each fish in the group 10, i.e leadership time-series

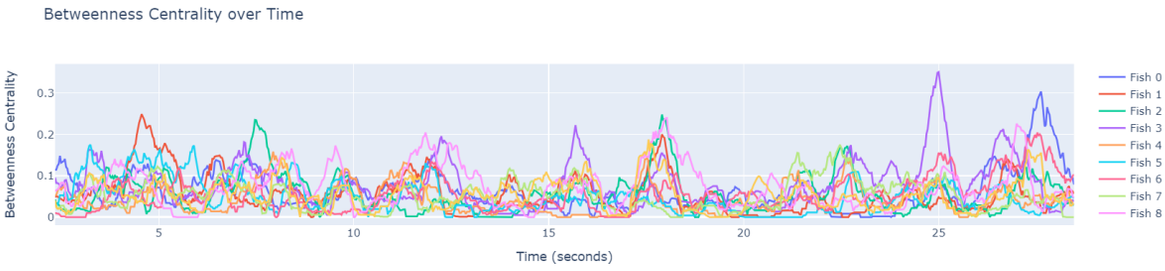


Figure 5.39: Moving Average: Betweenness Centrality time-series for each fish in the group 10

Group 60:

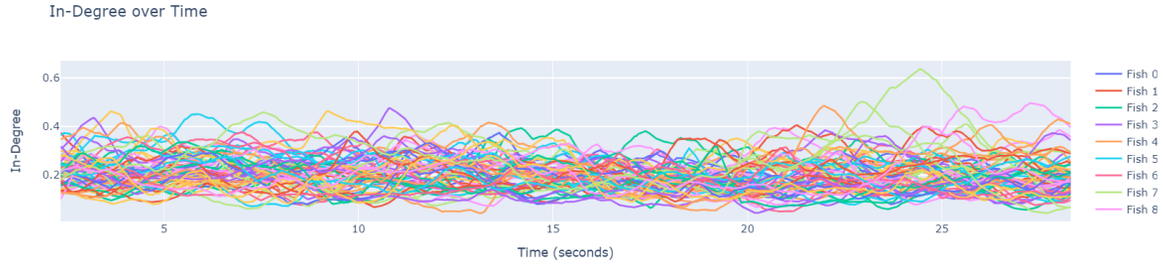


Figure 5.40: Moving Average: In degree time-series for each fish in the group 60, i.e followership time-series

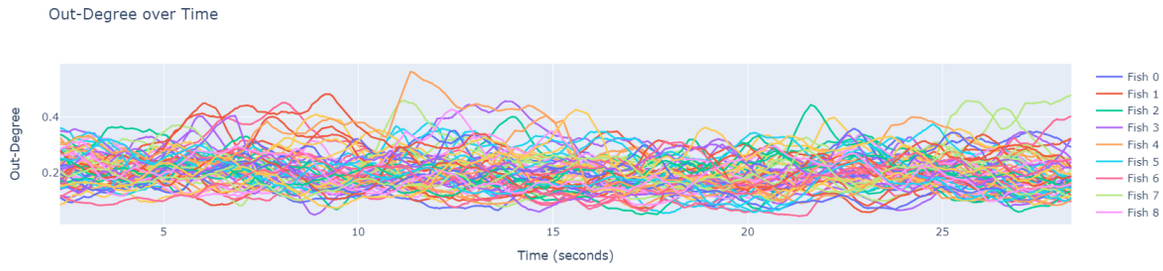


Figure 5.41: Moving Average: Out degree time-series for each fish in the group 60, i.e leadership time-series

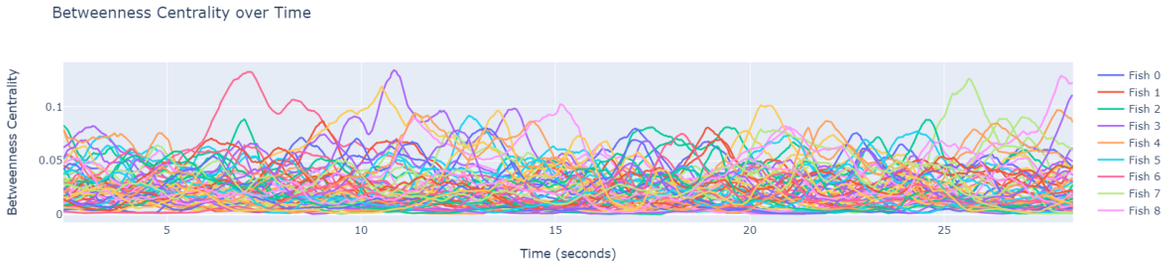


Figure 5.42: Moving Average: Betweenness Centrality time-series for each fish in the group 60

5.2.3 *Coups and Regime Stability*

In the following we will make use of these time-series in two ways, first we will look at evolution of the leader with time and see how that compares for different group size, we will also see how many leadership changes (Coups) happen to know how volatile these system are, and what is the average term that a leader serves before another takes his place. The leader was simply defined at the node with the highest causal outflow for each network.

The below graphs show a white bar marking the leader at each time step:

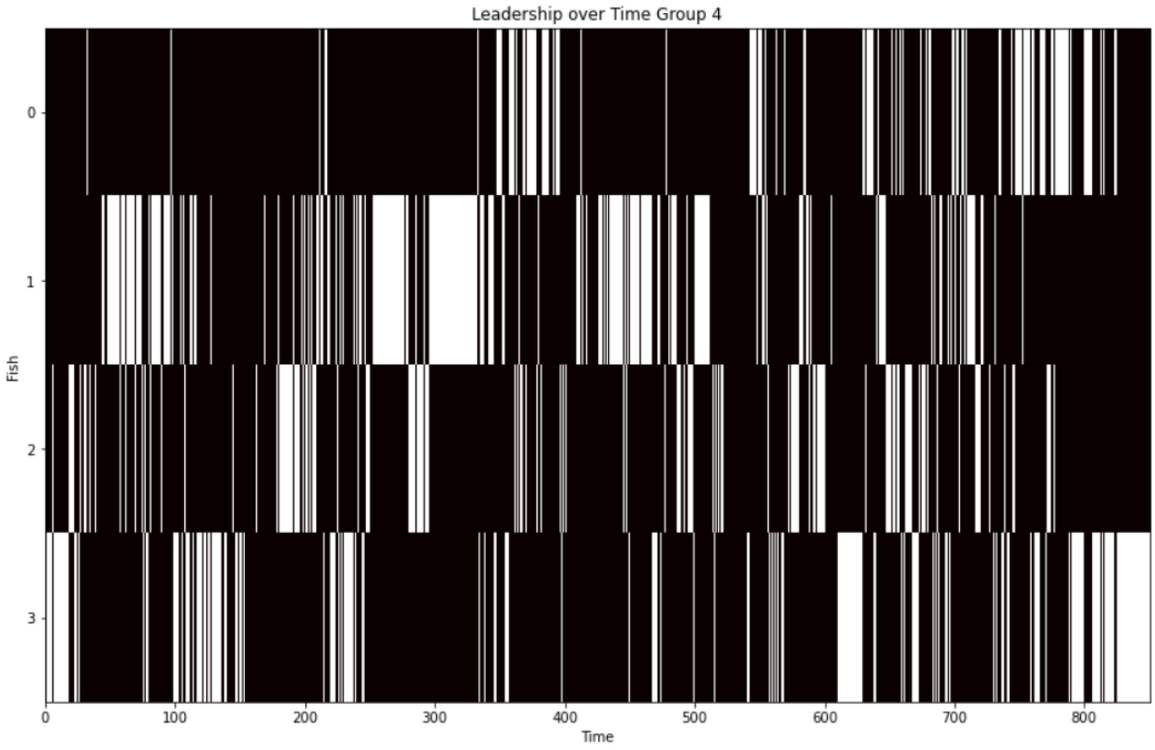


Figure 5.43: Evolution of the leader in group 4

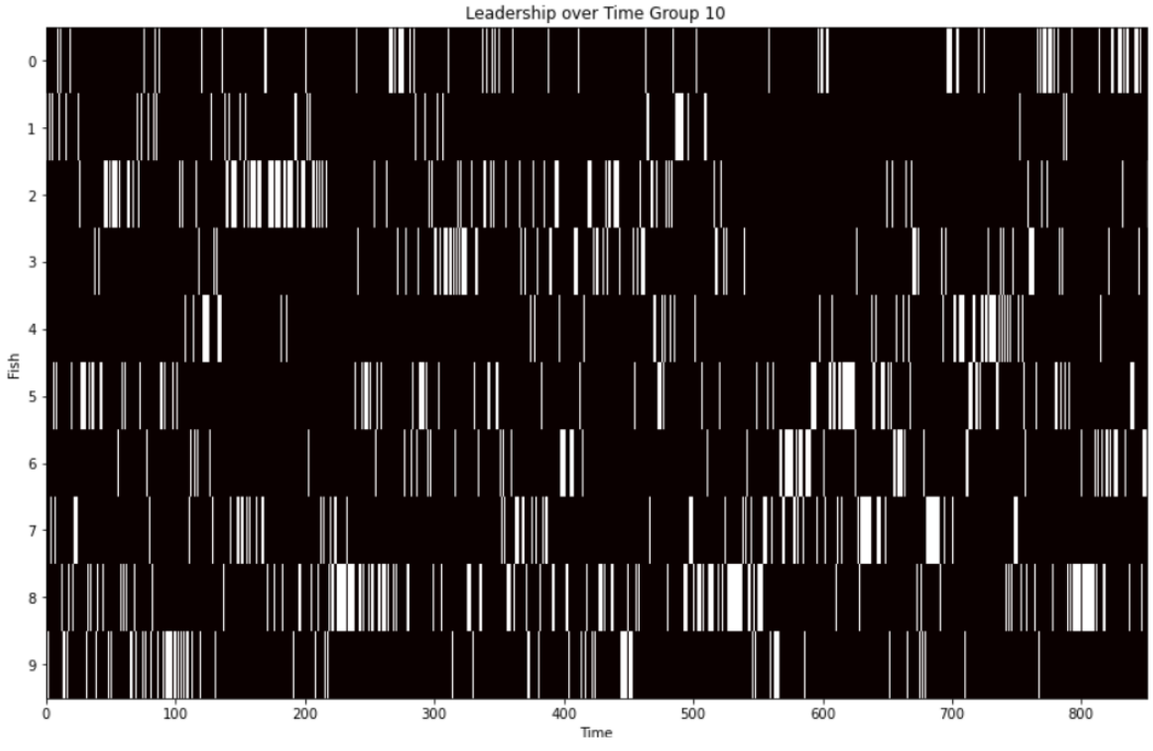


Figure 5.44: Evolution of the leader in group 10

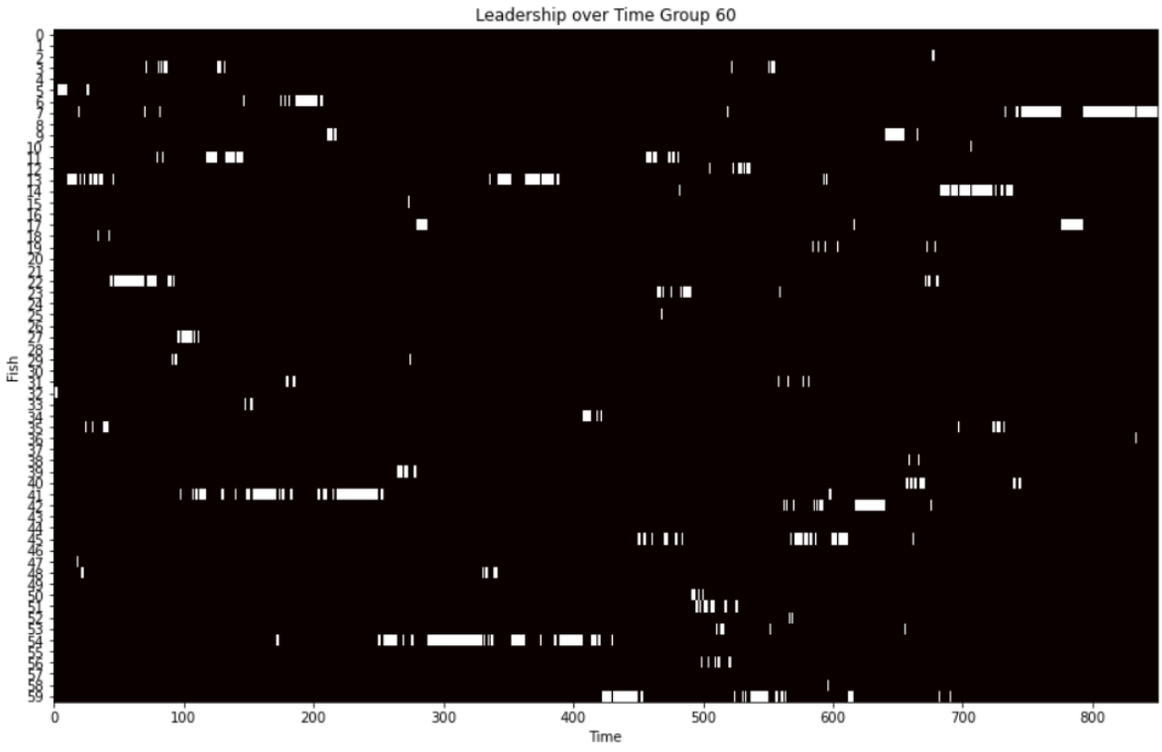


Figure 5.45: Evolution of the leader in group 60

The above show that the group of 60 fish, exhibit more stability in leadership than do groups 4 and 10, while 10 showing the most instability, which ties very well with the conclusions from figures 5.25,5.26,5.27.

The below table summarizes additional findings from the above plots:

Table 5.1: Summary of the Political Landscape of The Different Group Sizes

Group Size	**Number of Coups**	**Avg. Term (s)**	**Longest Term(s)**
4	327	0.082	1.156
10	545	0.049	0.375
60	241	0.11	1.313

For a more in depth analysis we take a look at the decay times of the autocorrelation functions for the in-degree, out-degree and betweenness centrality for each of the fish in each of the group sizes. To do so we obtain for each group size, the autocorrelation function for all of the mentioned time series, then we fit an exponential to them of shape $ae^{-b} + c$ and we record the decay time as the time step when the value of that function falls below 0.37 of it's maximum value. We show a few of these plots for reference, and then summarize the findings.

Here are examples of the decay time calculation:

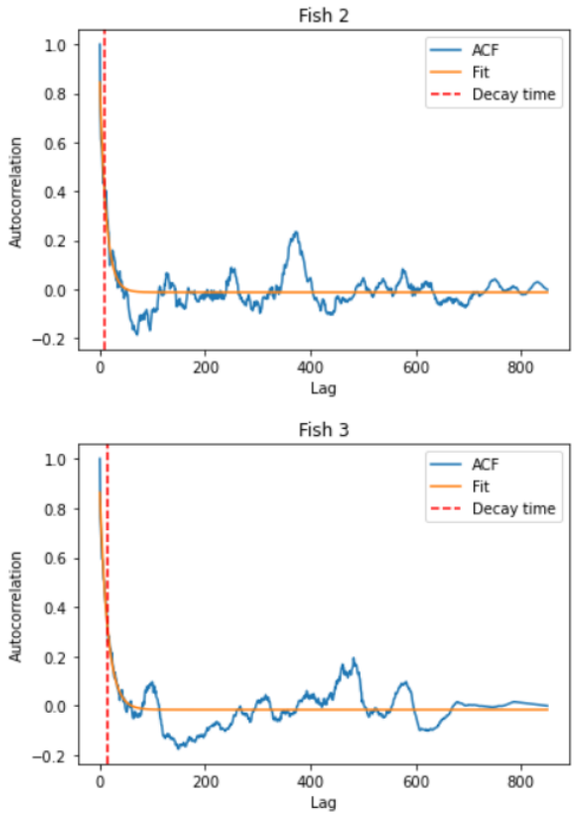


Figure 5.46: Example of Autocorrelation Functions and Decay Time Calculation by Exponential Curve Fitting

Below is a close up, to see the above fits on the first 50 frames, since the rest after that is considered as noise:

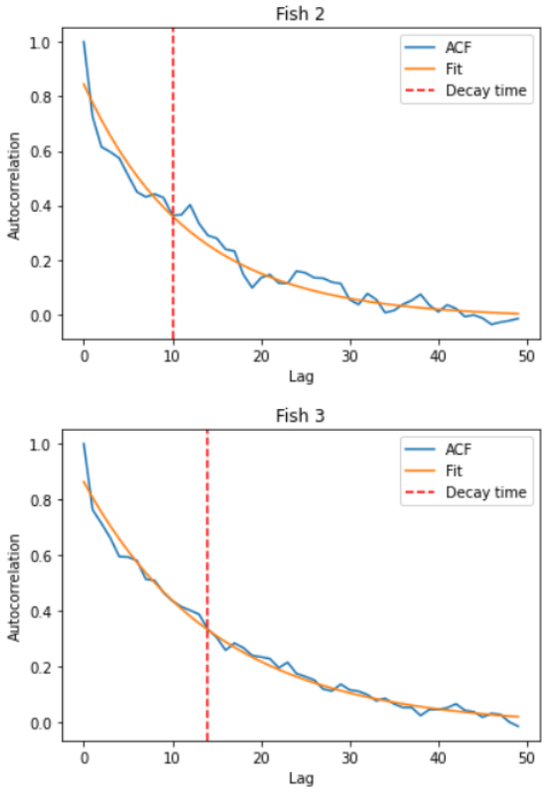


Figure 5.47: Close up Example of Autocorrelation Functions and Decay Time Calculation by Exponential Curve Fitting

We can now look at the decay time values for the different fish in each of the systems for the three types of time-series:

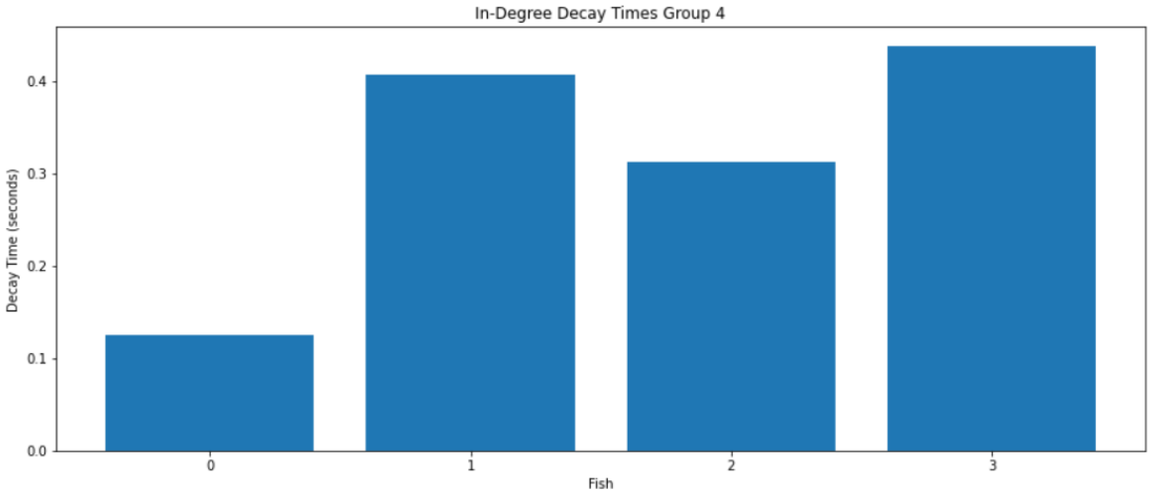


Figure 5.48: Values of the decay time for the In-Degree time series in the group of 4

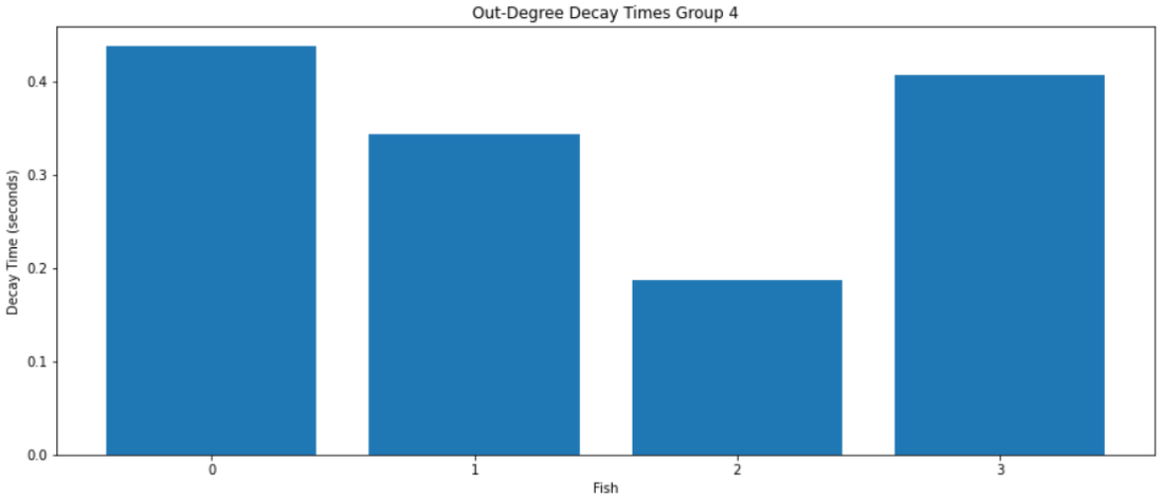


Figure 5.49: Values of the decay time for the Out-Degree time series in the group of 4

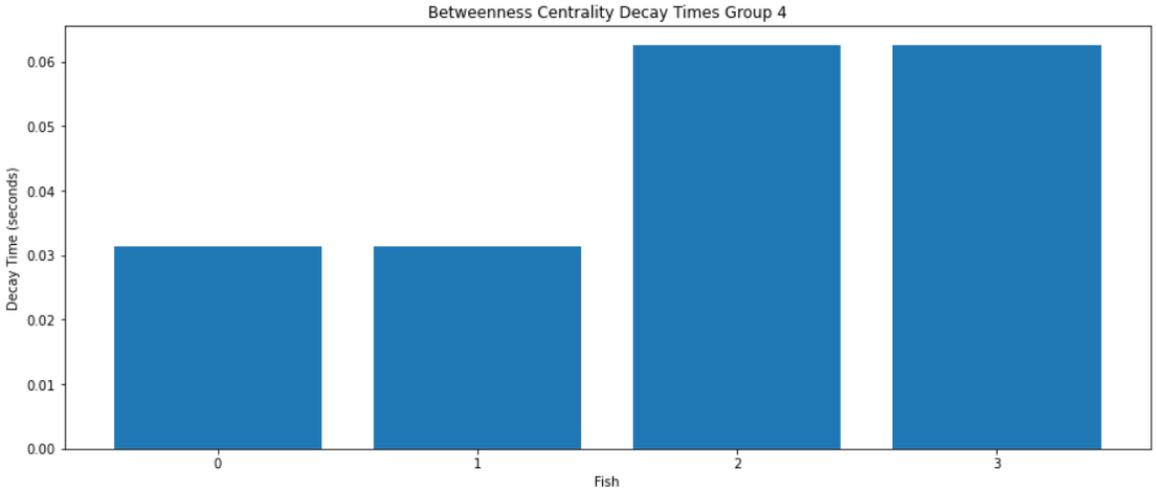


Figure 5.50: Values of the decay time for the Betweenness Centrality time series in the group of 4

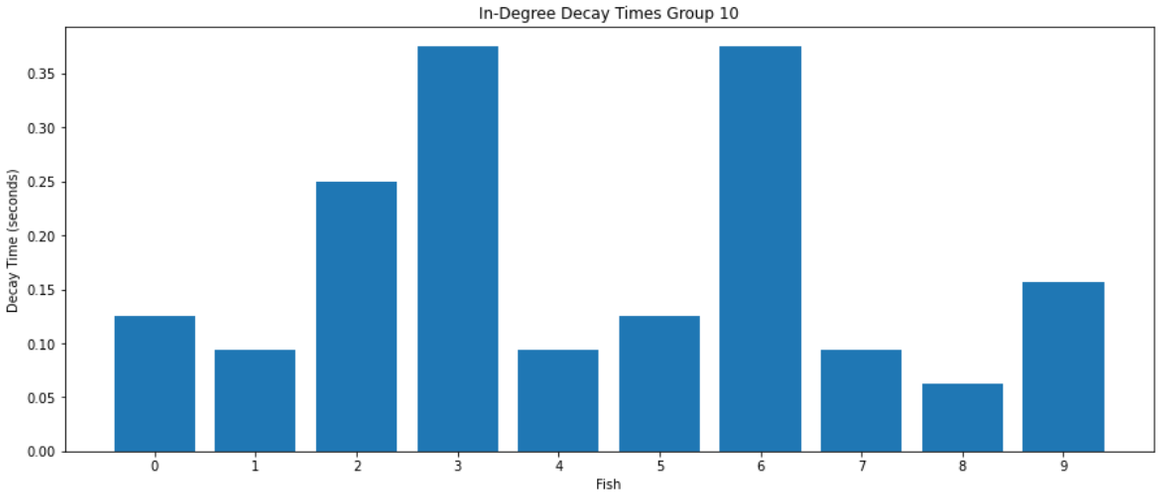


Figure 5.51: Values of the decay time for the In-Degree time series in the group of 10

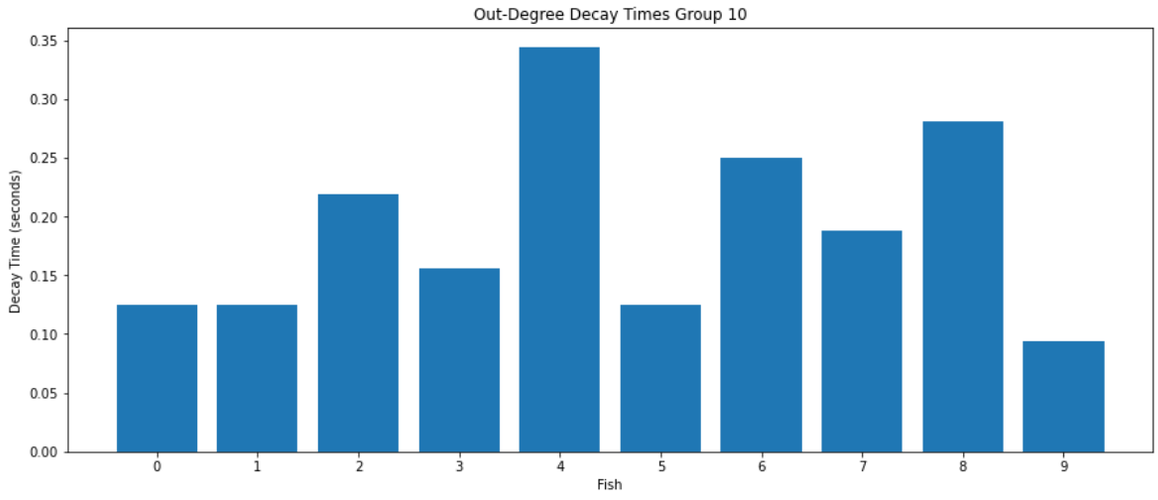


Figure 5.52: Values of the decay time for the Out-Degree time series in the group of 10

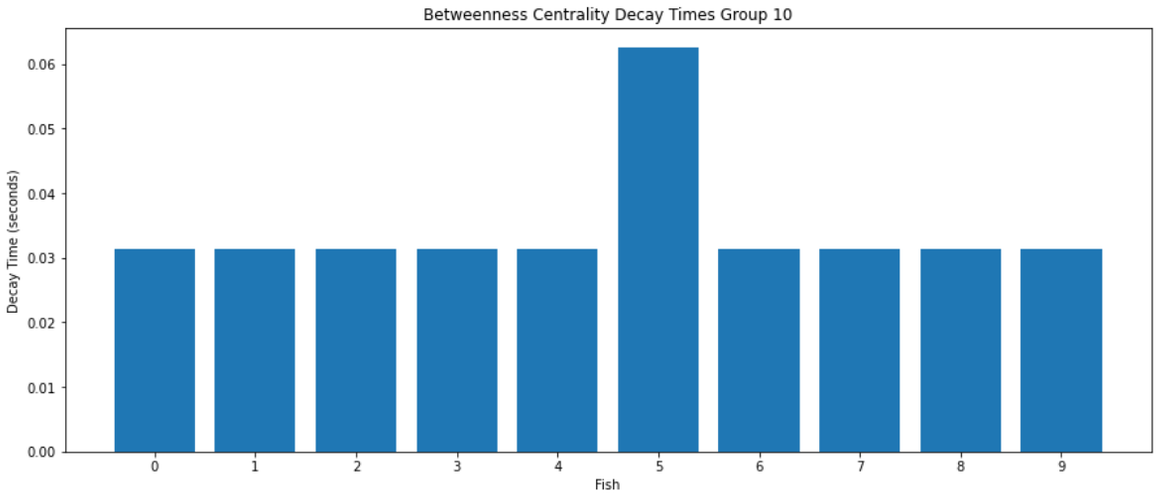


Figure 5.53: Values of the decay time for the Betweenness Centrality time series in the group of 10

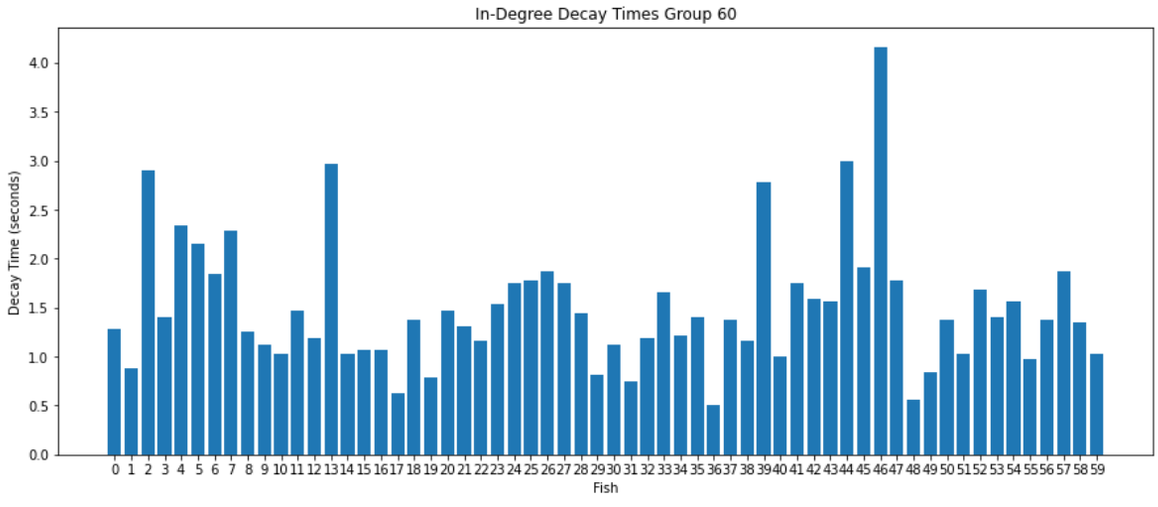


Figure 5.54: Values of the decay time for the In-Degree time series in the group of 60

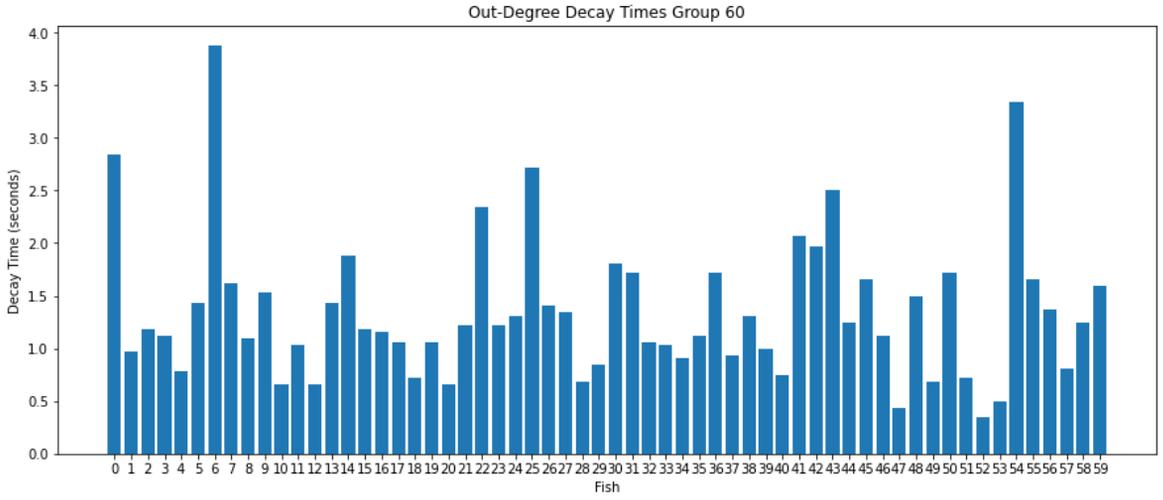


Figure 5.55: Values of the decay time for the Out-Degree time series in the group of 60

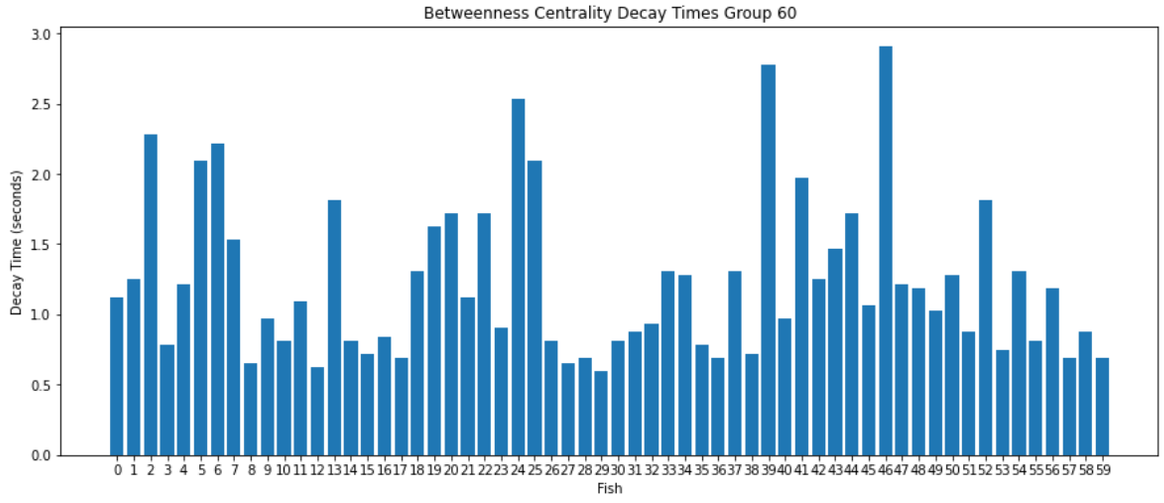


Figure 5.56: Values of the decay time for the Betweenness Centrality time series in the group of 60

To conclude this part of the analysis we take a look at the distributions of these decay values per group size:

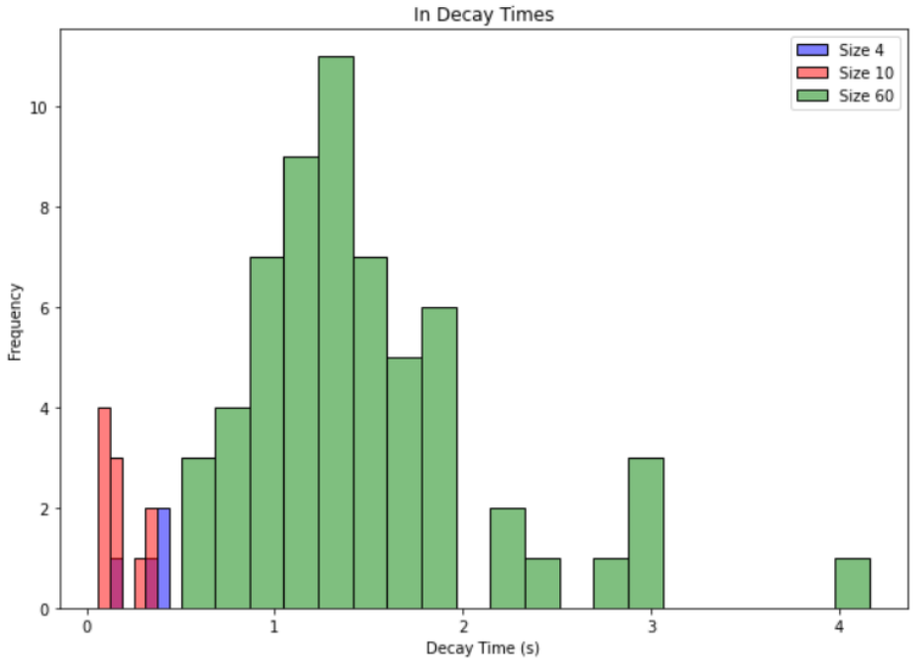


Figure 5.57: Distribution of the In-Degree decay times across the group sizes

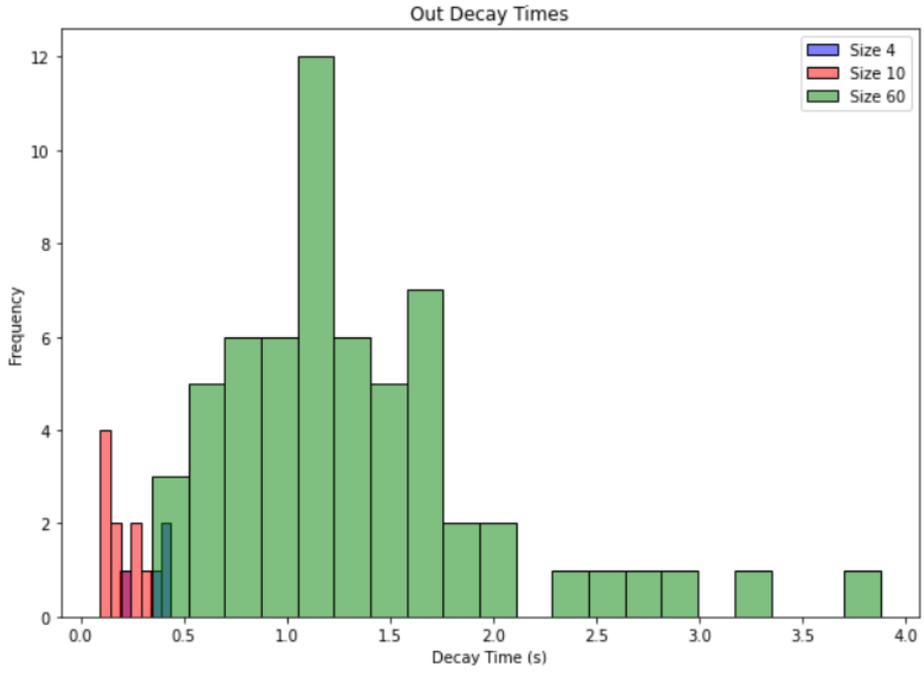


Figure 5.58: Distribution of the Out-Degree decay times across the group sizes

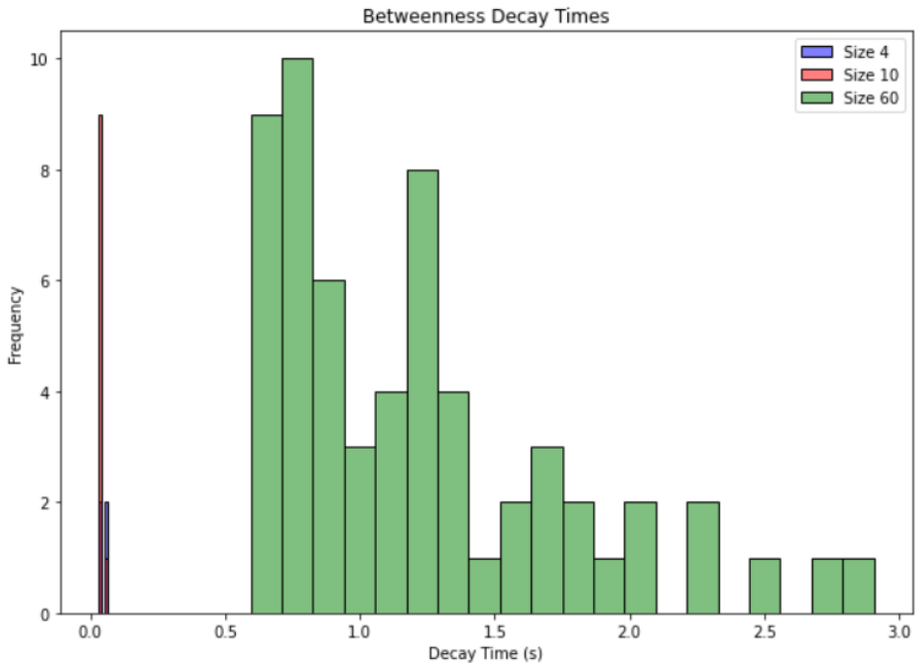


Figure 5.59: Distribution of the Betweenness decay times across the group sizes

One final part of the analysis on the autocorrelation functions of these metrics, consists of looking at the average autocorrelation function for in-degree, the out-degree, and the betweenness centrality time series. This is simply done by obtaining the autocorrelation function in each system for each individual fish, then taking the averages of these and computing the decay rate in the same way we've done at the beginning of this section. Below is an example of the fits, and three bar chart plots showing the values of the decay rates of the average autocorrelation functions for the three groups:

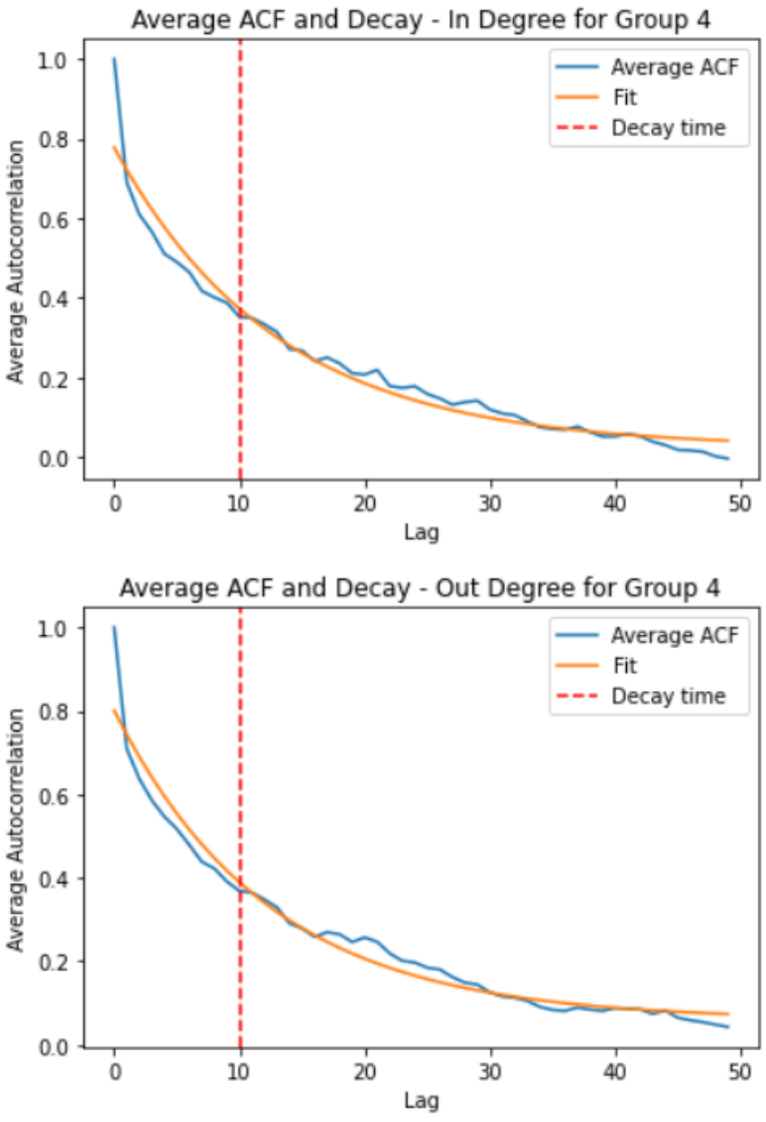


Figure 5.60: Example Showing the average autocorrelation functions and their corresponding fits and decay rates

And below are the values of the decay rates of the average autocorrelation function for each of these groups:

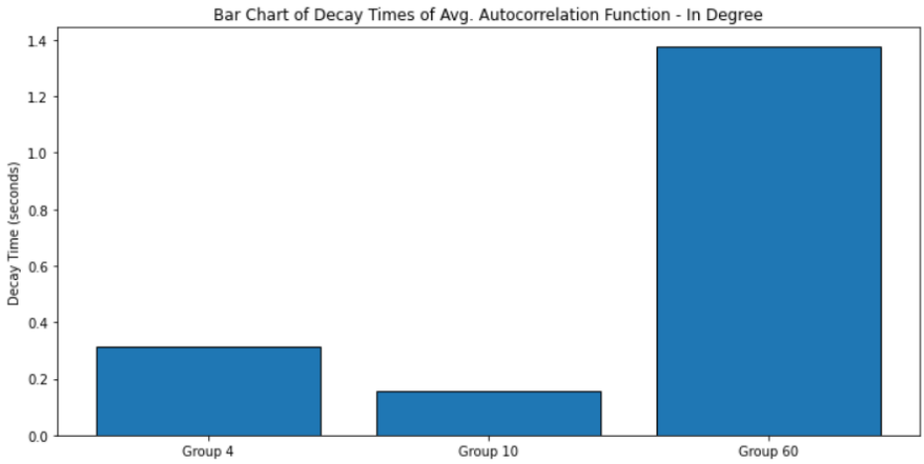


Figure 5.61: Decay times for the average autocorrelation function of the in-degree time series

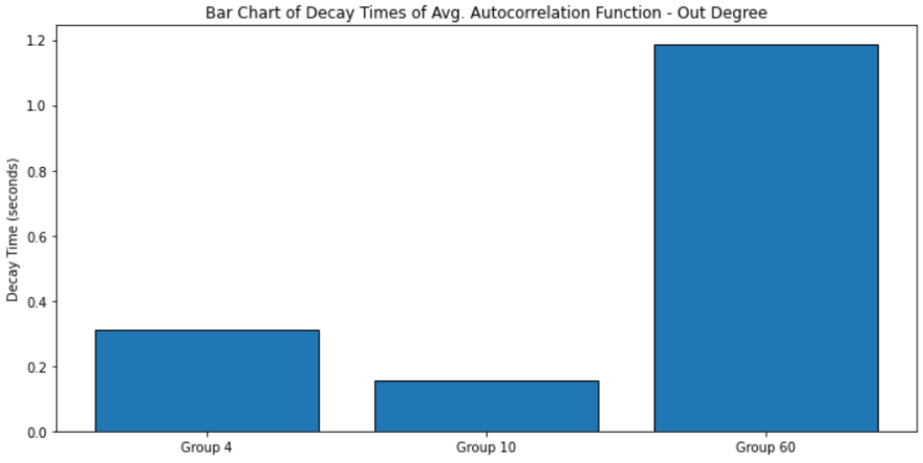


Figure 5.62: Decay times for the average autocorrelation function of the out-degree time series

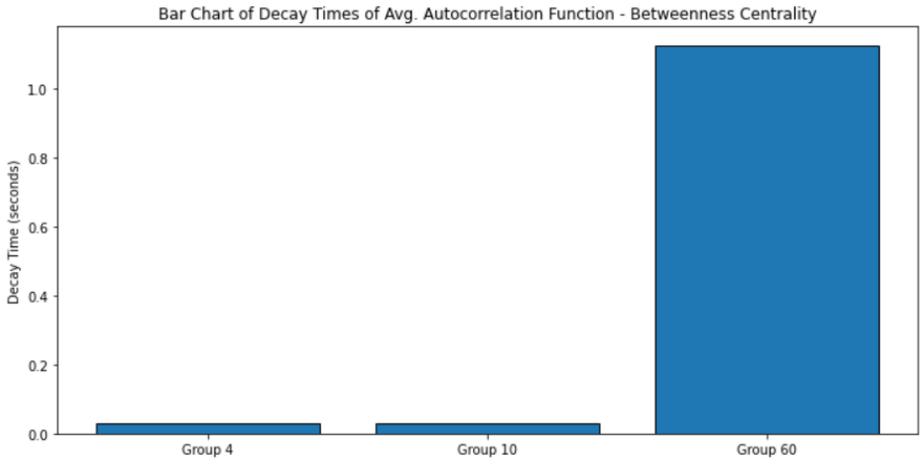


Figure 5.63: Decay times for the average autocorrelation function of the betweenness centrality time series

5.2.4 *Communication Efficiency and Information Flows*

In this final analysis on the causal networks, we take a look at three interesting measures: the average number of causal neighbors in each group, the average clustering coefficient which measures how causally tied of system is, and the global efficiency of the network, which measurse how easily information can be communicated in a network. The last two are graph theoretic tools which we use in our context to get insight as to why the larger groups seem to be more able to persist in rotational and/or orientational patterns as highlighted in the decay rates of the polarization and the rotation order paramaters in fig. 5.22.

Here are the results for the average number of causal neighbors in time:

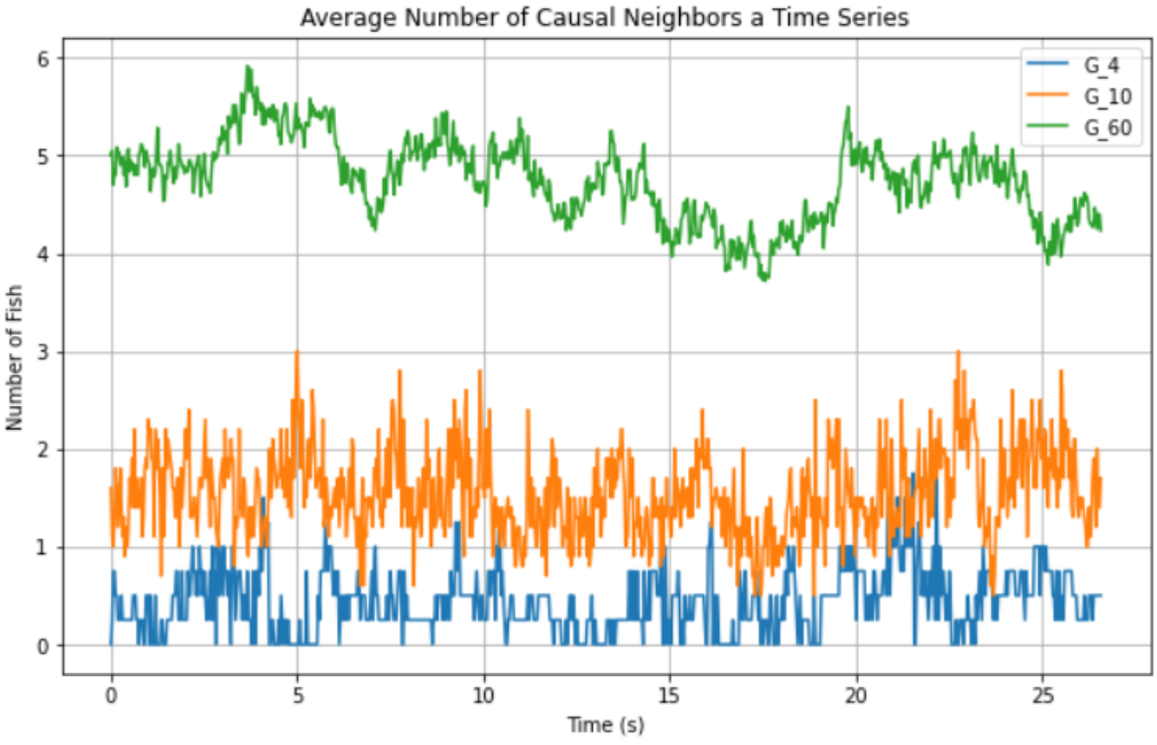


Figure 5.64: Average number of Causal Neighbors across the different group sizes

The average clustering coefficient and the global effiency of each system across in time:

Figure 5.65: Average Clustering Coefficient across the different group sizes

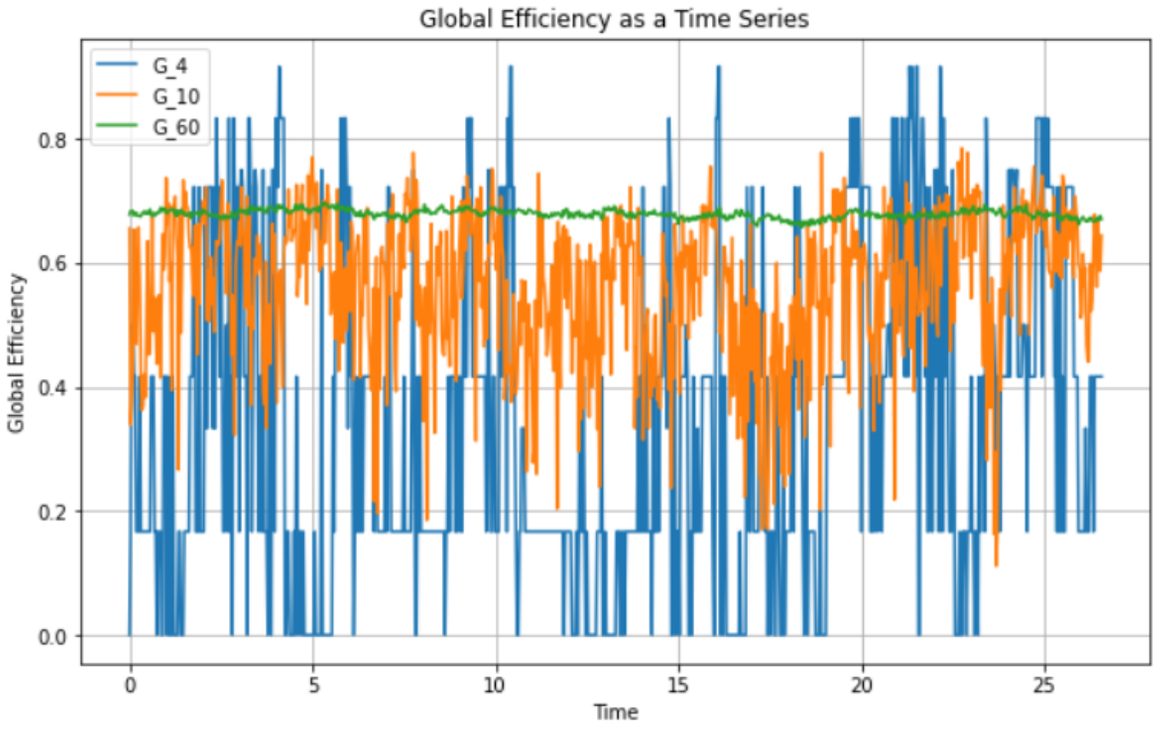


Figure 5.66: Average Global Efficiency across the different group sizes

Noticing the clear difference in the variability of the distribution in these different time series we show below the violin plot of the distribution for the Average Clustering Coefficient and of the Global Efficiency across the three groups:

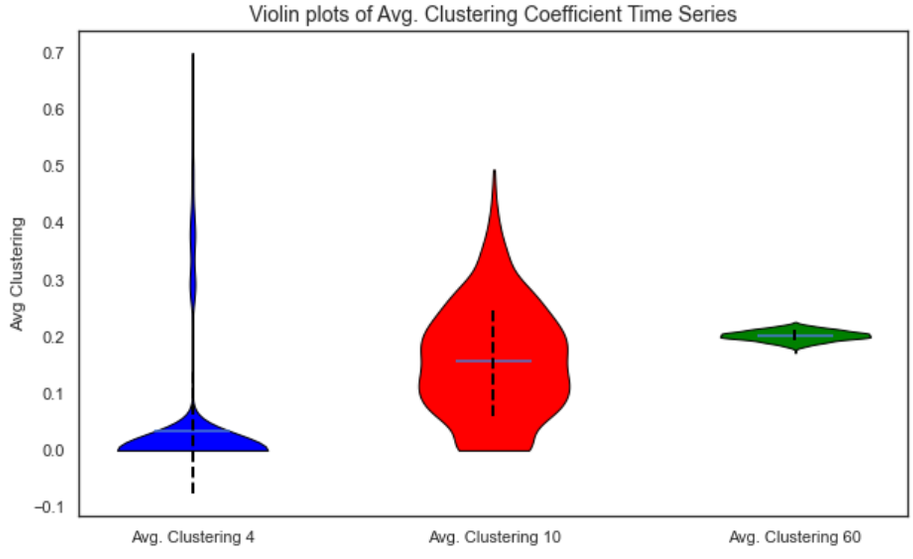


Figure 5.67: Violin plot of the Average Clustering Coefficient Across the Three Groups

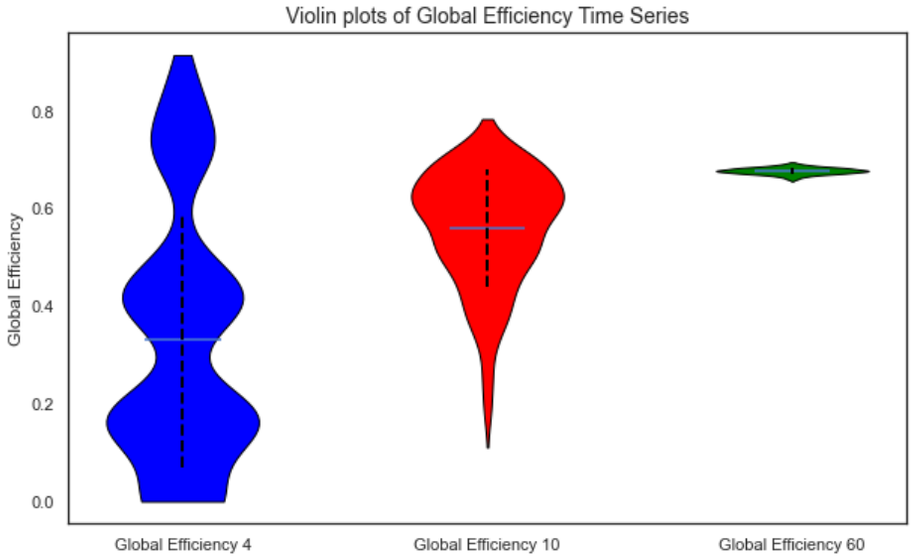


Figure 5.68: Violin plot of the Average Global Efficiency Across the Three Groups

Below is a summary of the findings:

Table 5.2: Average Number of Causal Neighbors (CN), Clustering Coeffient (CC) , And Global Efficiency (GE)

Group Size	**Avg. CN number**	**Avg. CC**	**Avg. GE**
4	0.44	0.033	0.331
10	1.55	0.158	0.561
60	4.75	0.203	0.680

Based on what we have learned so far, we think that the increased memory times for the time series of the order parameter, or just the increased level of coordination in the groups of zebrafish as the density goes up, has something to do with the causal topology of the shoal becoming less variable and more consistent; while the Global Efficiency of the group of 60 has an average that is slightly higher than that of the group 4 and 10, it's evident that the variance of it is much smaller than that of the other groups, indicating an interplay between the consistency of the causal structure and heightened levels of sustained coordination.

5.3 Reconstruction of Stochastic Differential Equations

5.3.1 *SDE Estimation for the Rotation Order Parameter*

We begin with the four fish system following the methodology described above, and ensuring that all diagnosis tests are done to understand when our estimations might fail.
Below are the part of the time series for the Rotation Order Parameter, the bin-averaged values for the drift and diffusion, the distribution of the time series and the auto correlation function.

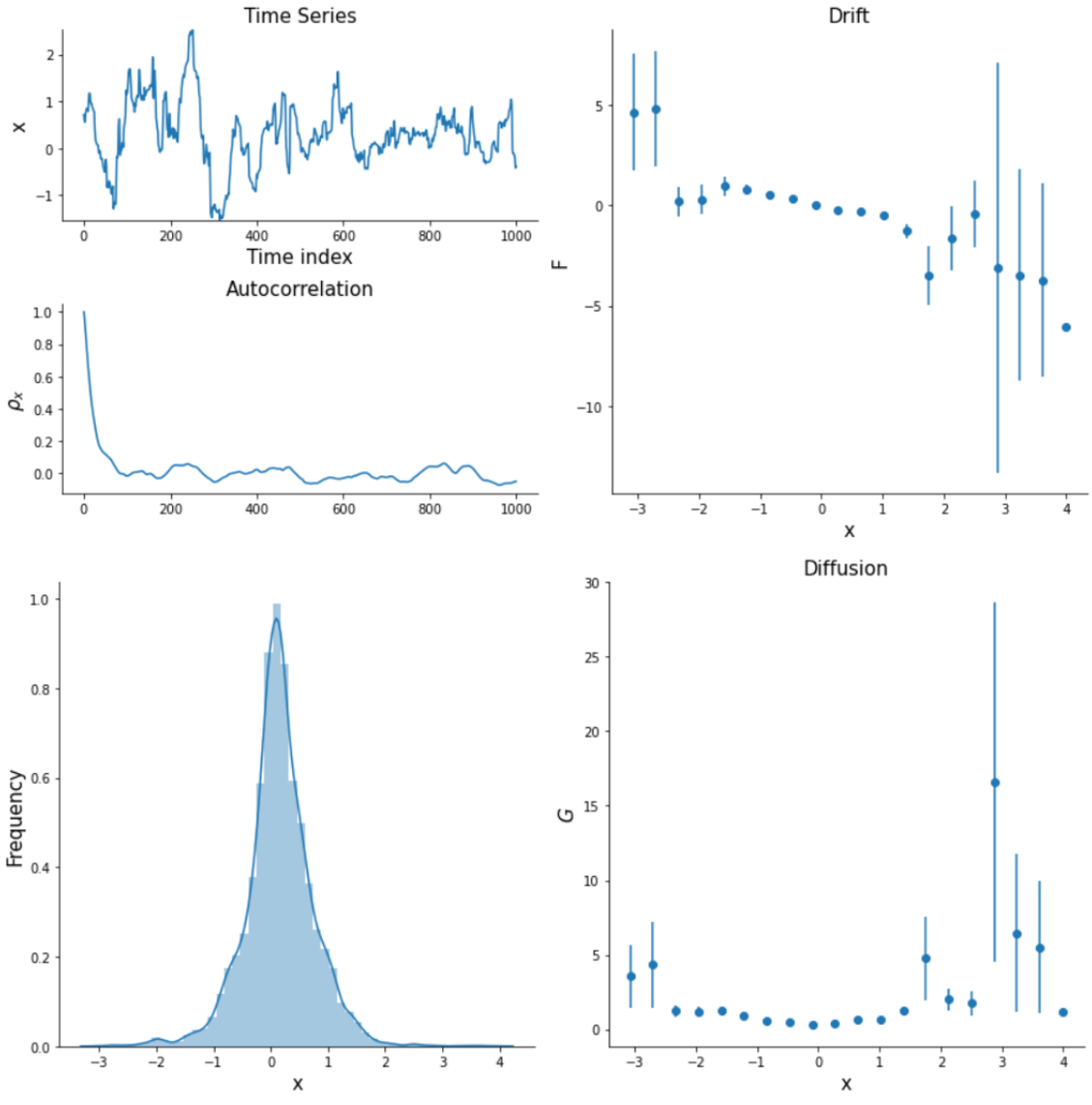


Figure 5.69: Time series of the Rotation OP, the scatter plot of the drift and diffusion, the distribution of the values of the time series and the autocorrelation function

Using Sparse Regression we fit for analytical expressions of the drift and the diffusion and obtain the following equations.
The drift:

$$F = (0.102 \pm 0.033) + (-0.692 \pm 0.061)x \quad BL^2/s^2 \tag{5.10}$$

And the diffusion:

$$G = (0.326 \pm 0.022) + (0.448 \pm 0.048)x^2 \quad BL^4/s^3 \tag{5.11}$$

Where x represents the Rotation OP. Below are the fits for the drift and the diffusion:

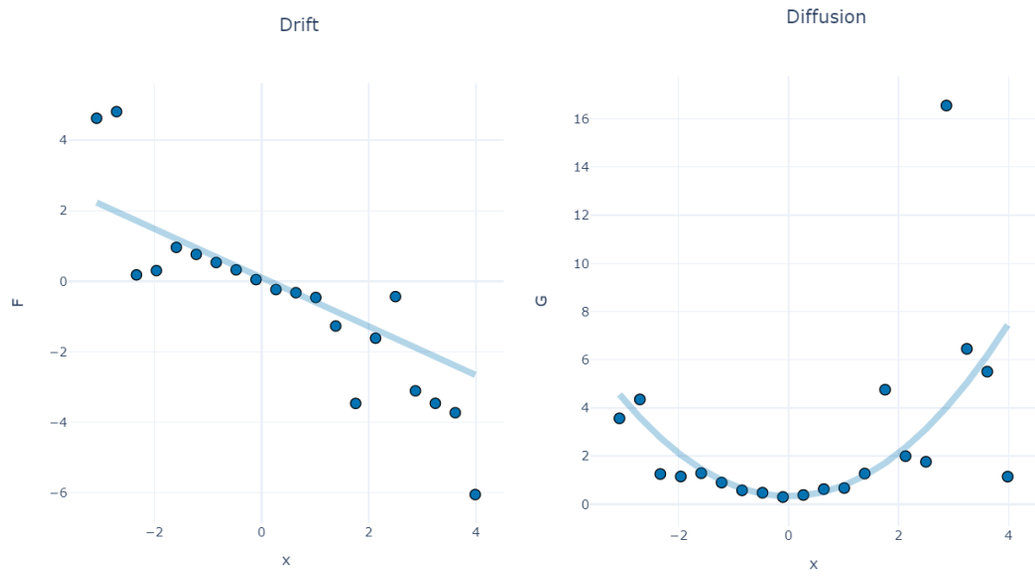


Figure 5.70: Drift and diffusion fits obtained by Sparse Regression

And in order to check the accuracy of the fit, we take a look at the distribution of the data compared with the distribution of 20 simulated time series from the stochastic equation that we recovered, and we see a very interesting match:

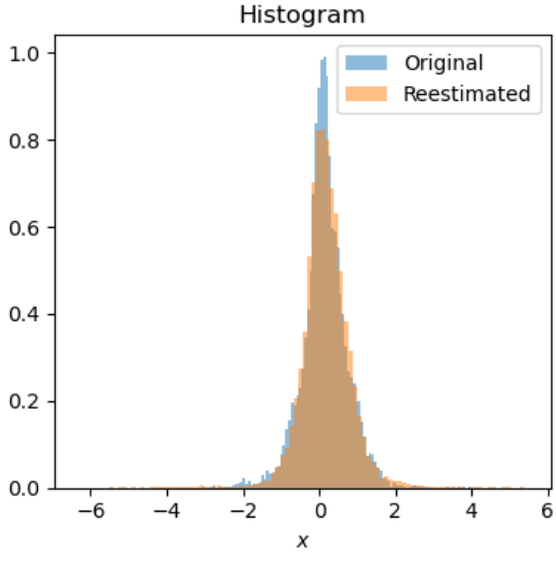


Figure 5.71: Comparison of the distributions of the original time series and the recovered time series from simulations of the recovered Stochastic Differential Equation

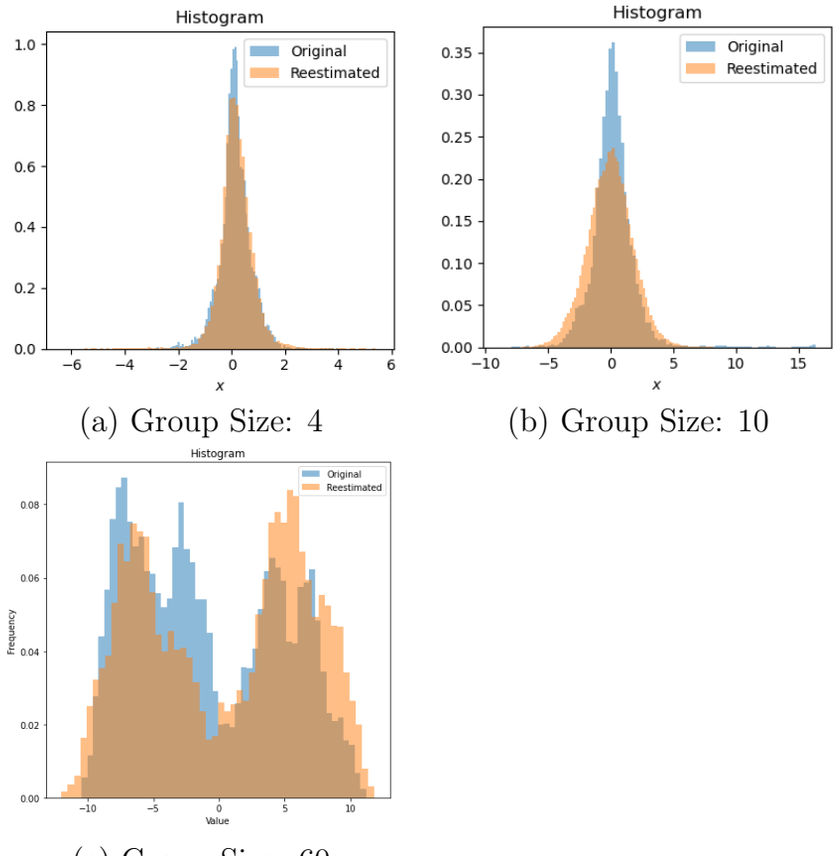
Below I will summarize the equations obtained from the different groups for the Rotation Order Parameter denoted as R, along with their histogram distributions. Note: When the errors-bars on the coefficients create overlapping intervals we combine the terms (by replacing them with their average):

Table 5.3: Summary of Derived Stochastic Differential Equations for the Rotation Order Parameter

Group Size	Equation
4	$\dot{R} = 0.1 - 0.7R + \sqrt{0.39(1+R^2)}\eta(t)$
10	$\dot{R} = -0.25R + \sqrt{1.6 + 0.05R^2}\eta(t)$
60	$\dot{R} = 1.5(sin(0.25R) + sin(0.125R) + \sqrt{1 + cos(0.125R)}\eta(t)$

Moreover, we analyzed the distributions of these groups' rotation order parameters. Histograms representing these distributions are shown in the following figure:

Figure 5.72: Distribution of Rotation Order Parameters of Original and Re-estimated time-series



(a) Group Size: 4

(b) Group Size: 10

(c) Group Size: 60

Analysis of Smaller Groups

Combining the obtained equations with the decay time obtained in section 5.1 we notice the following:

For smaller group sizes (4 and 10), the rotation order parameter oscillates around zero. This behavior is reflected in the SDEs, where the deterministic part contains

negative terms proportional to R (-0.7R for group size 4 and -0.25R for group size 10). The negative R term in the deterministic part of the SDEs acts as a damping force, driving the system back to zero whenever it deviates. This indicates a lack of long-term coordination among the fish in smaller groups, as the rotation order parameter does not maintain a constant value but instead tends to return to zero.

The noise term $\eta(t)$ in these equations is amplified by a term under the square root that increases with the square of R. This implies that the stochastic forces increase as the system deviates further from equilibrium (R=0). Thus, larger deviations from the mean value are associated with larger uncertainties. This contributes to a feedback mechanism that promotes the system's return to zero.

This behavior is confirmed by the calculated memory times for these groups, which are relatively short (1.4375s for group size 4 and 1.90625s for group size 10). This shows that the rotation order parameter quickly forgets its past values, consistent with the oscillatory behavior around zero.

Analysis of Larger Groups

In contrast, large groups exhibit distinct dynamics. The rotation order parameter in these groups tends to stabilize around certain values for extended periods, indicating a higher level of coordination among the fish. This is reflected in the significantly longer memory times of 25.75s, 34.1875s, and 38.875s for group sizes 60, 80, and 100, respectively.

The SDE for the 60 groups involves trigonometric functions, which suggest complex, oscillatory behavior. These equations imply a balance between forces that encourage the rotation order parameter to oscillate and forces that stabilize the parameter around certain values.

The noise term $\eta(t)$ in that SDE is modulated by a combination of constant and trigonometric terms under a square root. This suggests that the influence of stochastic forces on the system dynamics varies in a more complex manner compared to the smaller groups. This may be interpreted as a reflection of the increased complexity of interactions and coordination within larger groups of fish.

Our analysis reveals distinct differences in the dynamics of the rotation order parameter across different group sizes. Smaller groups exhibit a tendency for the rotation order parameter to fluctuate around zero, and larger deviations from zero lead to stronger stochastic forces that push the system back to zero. Larger groups, in contrast, show a stabilization of the rotation order parameter around certain values, indicative of sustained coordinated behavior.

5.3.2 *Coupled SDE Estimation for the Polarization Vector Order Parameters*

In the previous sections, we have investigated several aspects of the polarization phenomena. Now, we will focus our attention on the polarization vector, more specifically, the Polarization Vector Order parameter $M(M_x, M_y)$. The polarization vector, $M(t)$, represents the overall direction and degree of alignment of the velocities in our system. It is computed as follows: Let $r_i(t)$ represent the position of the i^{th} fish at time t. Then, the velocity, $v_i(t)$, of the i^{th} fish is computed as:

$$v_i(t) = r_i(t+1) - r_i(t) \tag{5.12}$$

for each fish i and time point t. After obtaining the velocities, we normalize each velocity vector to obtain a unit vector. The normalized velocity $\hat{v}_i(t)$ is given by:

$$\hat{v}_i(t) = \frac{v_i(t)}{\|v_i(t)\|} \tag{5.13}$$

where $\|\cdot\|$ denotes the Euclidean norm. Next, we sum all the normalized velocity vectors at each time point to get a single vector known as the "polarization vector" $M(t)$:

$$\mathbf{M(t)} = \frac{1}{N}\sum_{i}^{N} \hat{v}_i(t) \tag{5.14}$$

where the sum is over all fish i at a specific time point t. The group polarization vector, $\mathbf{M(t)} = (M_x, M_y)$, encapsulates the overall direction and degree of alignment of the velocities of the group at each time point. And the goal of this section is to obtain an equation of the form: Our goal in this subsection is to recover the Stochastic Differential Equations (SDEs) that govern the evolution of this order parameter. The equation we aim to derive has the following form:

$$\begin{pmatrix} dM_x(t) \\ dM_y(t) \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} dt + \begin{pmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{pmatrix} \begin{pmatrix} dw_1 \\ dw_2 \end{pmatrix}, \tag{5.15}$$

where the first term on the right-hand side represents the drift, and the second term corresponds to the diffusion (along with the off-diagonal cross-diffusion terms). And here are the results:
First for the group of four fish:

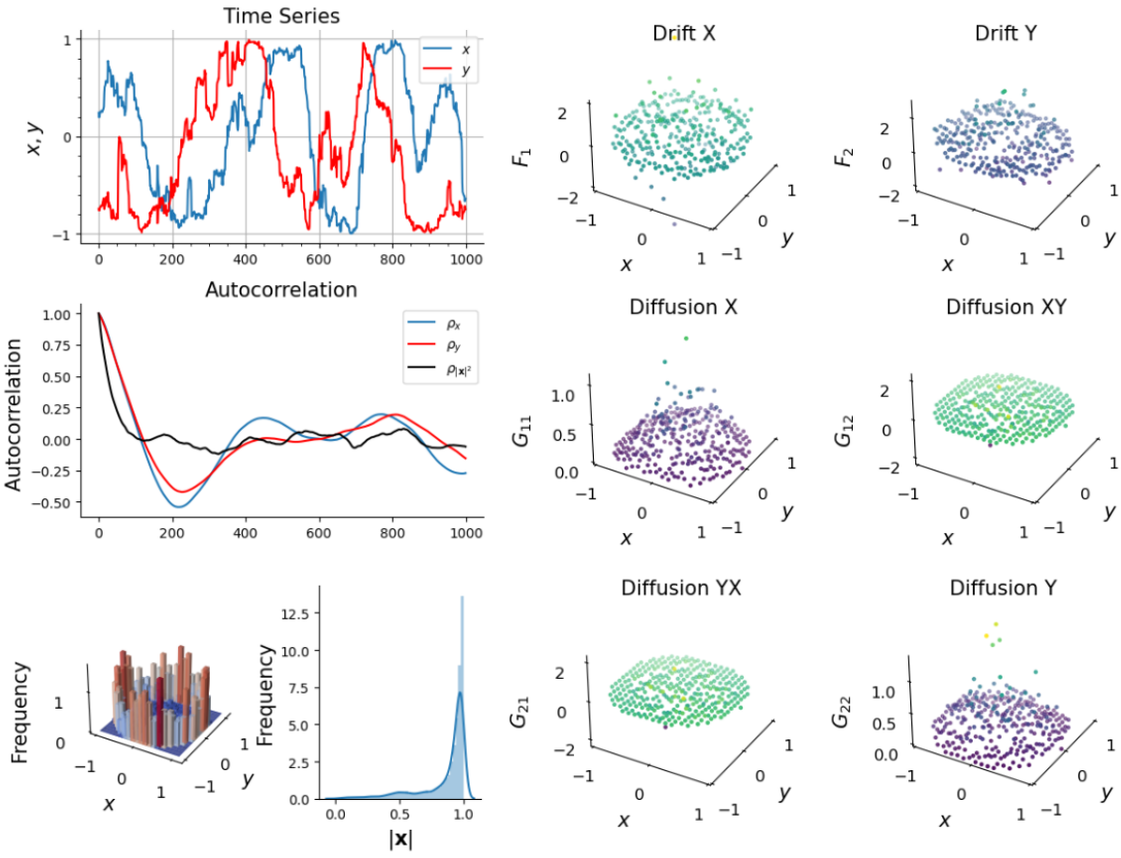


Figure 5.73: Time series of the group polarization vector components, the scatter plot of the drift and diffusion, the distribution of the values of the time series and the autocorrelation function

And here are the best fit surfaces for the drift and the diffusion, extending the work done in 1D in fig. 5.24 to two dimensions:

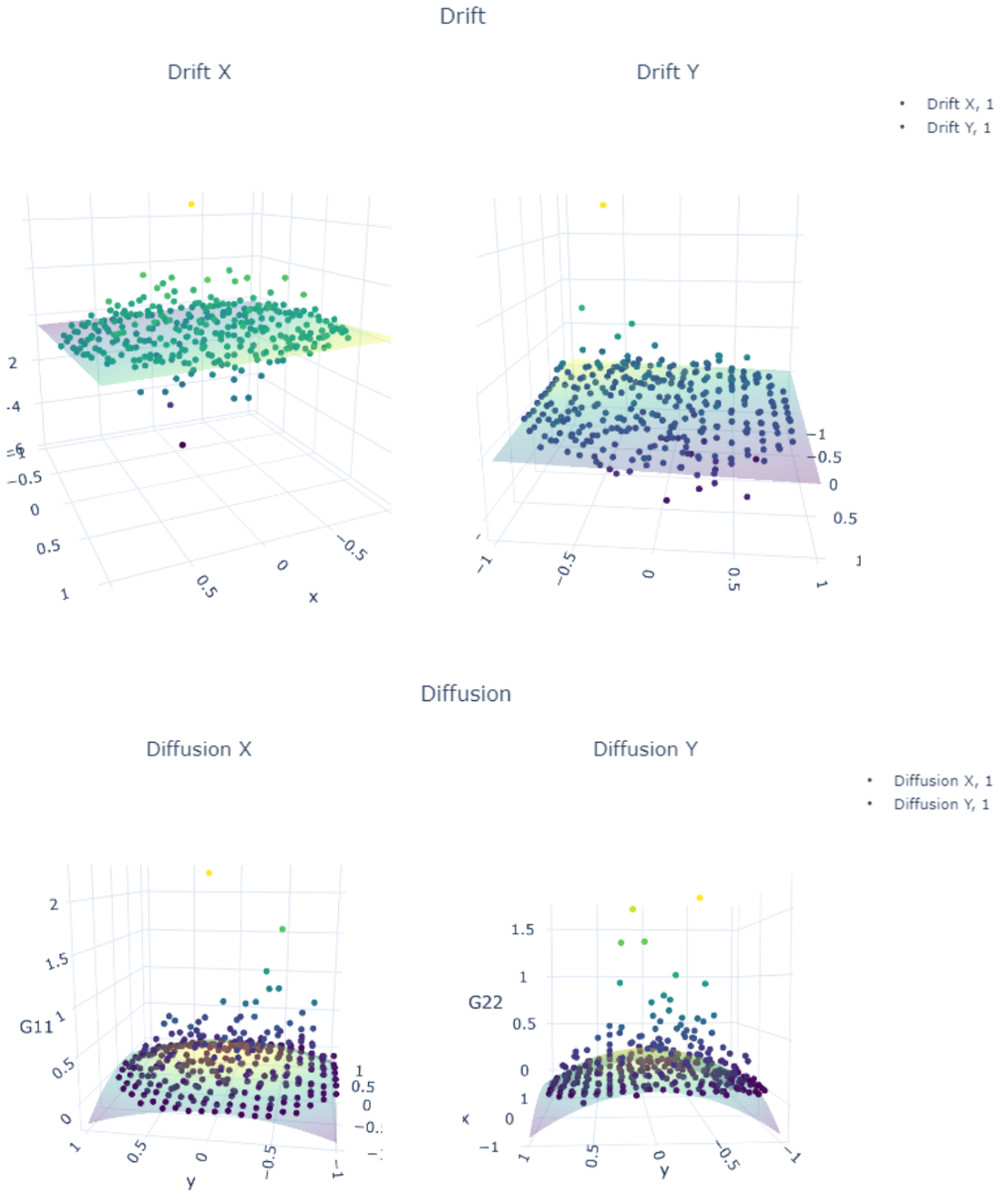


Figure 5.74: Drift and Diffusion fits obtained by Sparse Regression in the 2D case, neglecting the cross diffusion terms

In the above, the fits for the diffusion XY and diffusion YX (the cross diffusion terms) are neglected since the best fits for them are practically 0. The obtained equations for the drift and the diffusion are the following:
First for the drift components:

$$F1 = (-0.125 \pm 0.023)M_x + (0.256 \pm 0.024)M_y \tag{5.16}$$

$$F2 = (-0.247 \pm 0.024)M_x + (-0.146 \pm 0.025)M_y \tag{5.17}$$

And for the diffusion:

$$G11 = (0.302 \pm 0.011) + (-0.295 \pm 0.014){M_x}^2 + (-0.281 \pm 0.014){M_y}^2 \tag{5.18}$$
$$G22 = (0.324 \pm 0.012) + (-0.261 \pm 0.015){M_x}^2 + (-0.329 \pm 0.015){M_y}^2 \tag{5.19}$$
$$G12 = 0 \tag{5.20}$$
$$G21 = 0 \tag{5.21}$$

The stochastic equation for the group polarization vector is therefore:

$$\dot{\mathbf{m}} = \begin{pmatrix} -0.125 & 0.256 \\ -0.247 & -0.146 \end{pmatrix} \mathbf{m} + \sqrt{0.3\,(1 - |\mathbf{m}|^2)} \cdot \boldsymbol{\eta}(t) \tag{5.22}$$

Here are the results of the simulated distribution of this equation in comparison with the original distributions of $|\mathbf{m}|$ as well as the comparison of their respective auto-correlation functions. We also show the simulation of the m_x in comparison with the original. We look at m_x alone only to avoid crowdedness in the comparison plot.

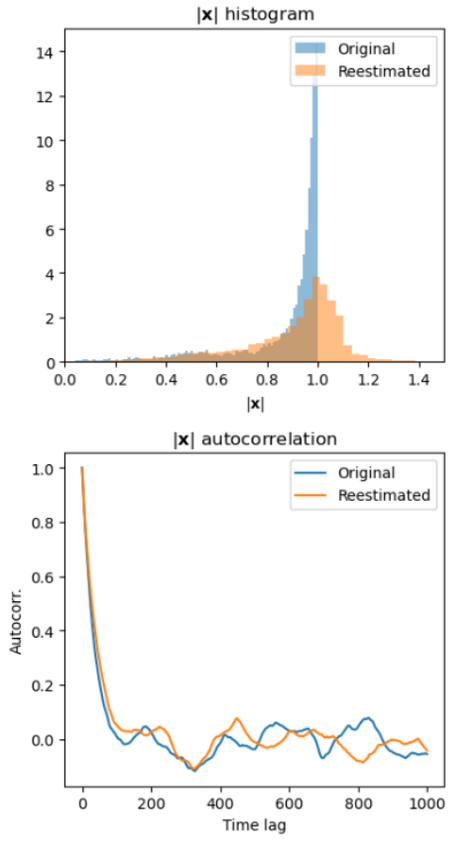


Figure 5.75: Comparison between the original and re-estimated distributions for $|\mathbf{m}|$ as well as the auto-correlation functions of each

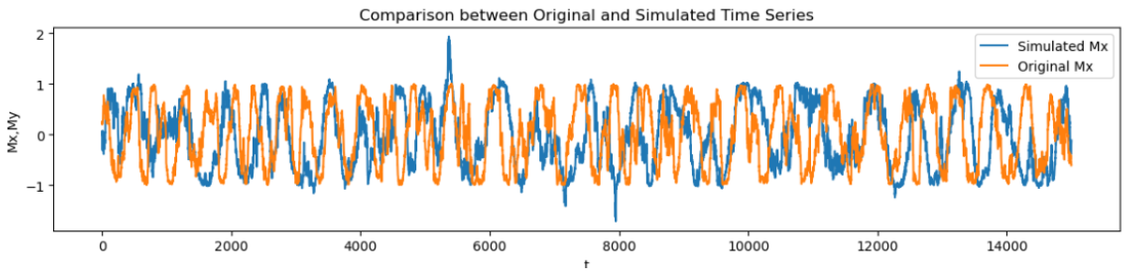


Figure 5.76: Simulated M_x time series versus the original M_x for the four fish group

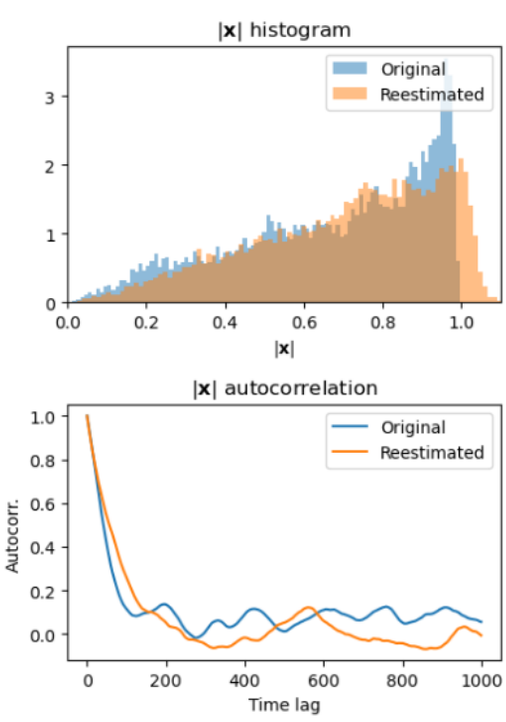
Below I will summarize the equations obtained from the different groups for the group Polarization Order Parameter denoted as $\mathbf{m}$, along with their histogram distributions, auto-correlation functions, and simulated vs actual time series. Note: When the errors-bars on the coefficients create overlapping intervals we combine the terms as we had done previously with the equations obtained for the rotation order parameter:

Table 5.4: Summary of Derived Stochastic Differential Equations Group Polarization Vector

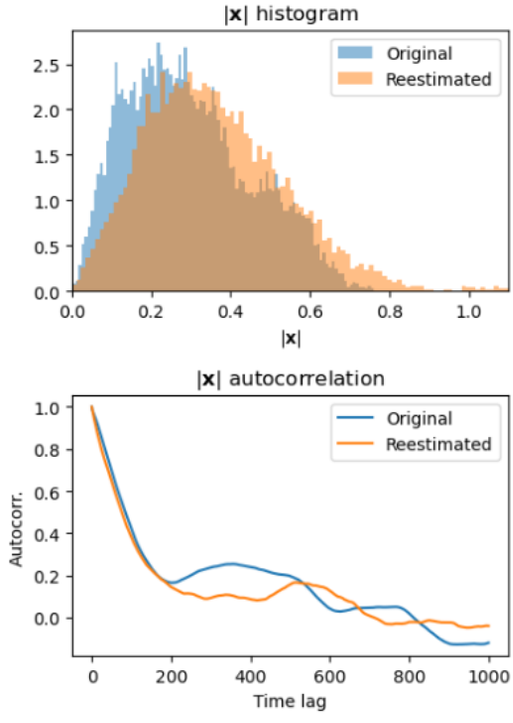
Group Size	Equation
4	$\dot{\mathbf{m}} = \begin{pmatrix} -0.125 & 0.256 \\ -0.247 & -0.146 \end{pmatrix} \mathbf{m} + \sqrt{0.3\,(1 - \lvert\mathbf{m}\rvert^2)} \cdot \boldsymbol{\eta}(t)$
10	$\dot{\mathbf{m}} = -0.1\mathbf{J}_2 \begin{pmatrix} m_x \\ m_x \cdot m_y \end{pmatrix} + \sqrt{0.1\,(1 - \lvert\mathbf{m}\rvert^2)} \cdot \boldsymbol{\eta}(t)$
60	$\dot{\mathbf{m}} = \begin{pmatrix} -0.84m_x^3 \\ -0.13m_x^2 m_y - 0.4m_x m_y^2 \end{pmatrix} + \sqrt{0.02} \cdot \boldsymbol{\eta}(t)$
80	$\dot{\mathbf{m}} = \begin{pmatrix} -0.3m_x^2 m_y - m_x m_y^2 - 0.26m_y^3 \\ -0.16m_y + 0.76m_x^2 m_y + 0.1m_y^3 \end{pmatrix} + \sqrt{0.009} \cdot \boldsymbol{\eta}(t)$
100	$\dot{\mathbf{m}} = \begin{pmatrix} -0.1m_x - 0.45m_x^2 m_y + 0.35m_x m_y^2 + 0.27m_y^3 \\ -0.1m_y - 0.43m_x m_y^2 + 0.1m_x^3 + 0.2m_y^3 \end{pmatrix} + \sqrt{0.006} \cdot \boldsymbol{\eta}(t)$

Moreover, we analyzed the distributions of these groups' polarization vector order parameters. Histograms and auto-correlation functions are shown in the following figures:

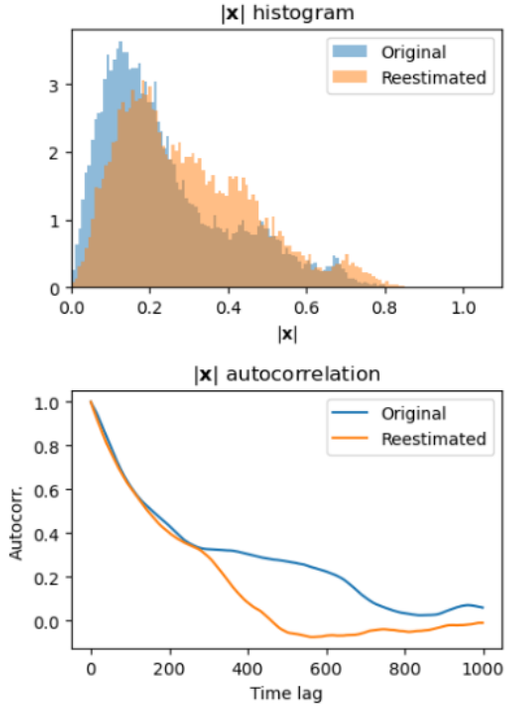
Figure 5.77: Distribution of Group Polarization Vector Order Parameter of Original and Re-estimated time-series



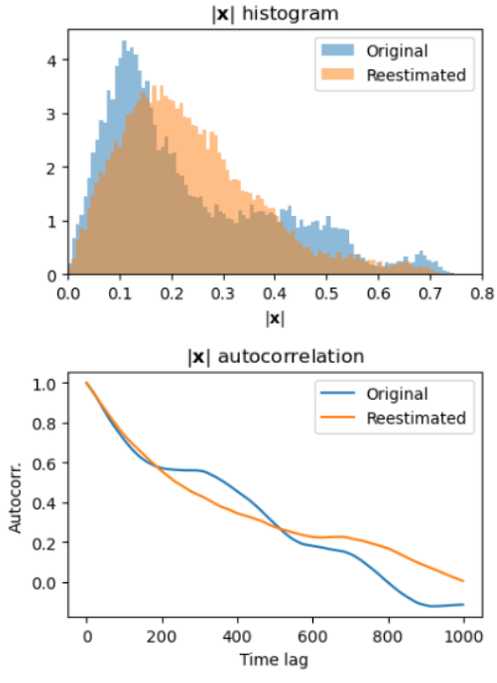
(a) Group Size: 10



(b) Group Size: 60



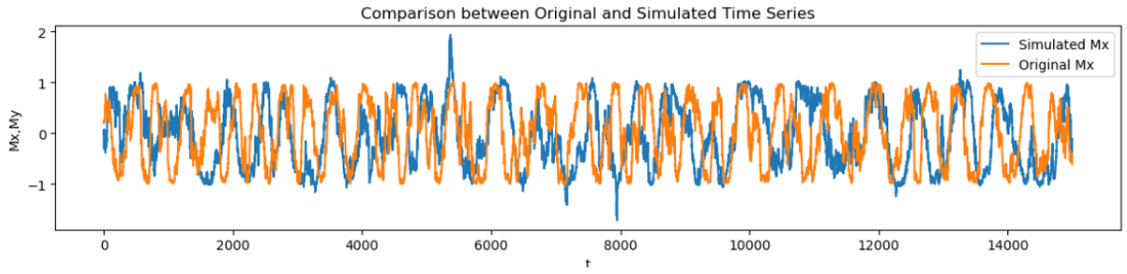
(c) Group Size: 80



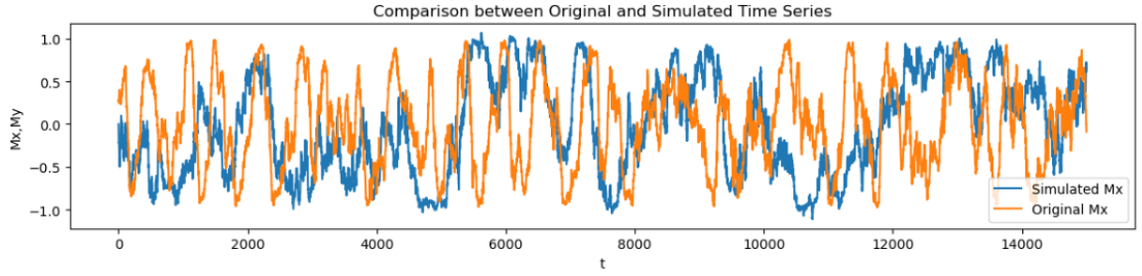
(d) Group Size: 100

And here are the simulated m_x from our equations compared with the original time series:

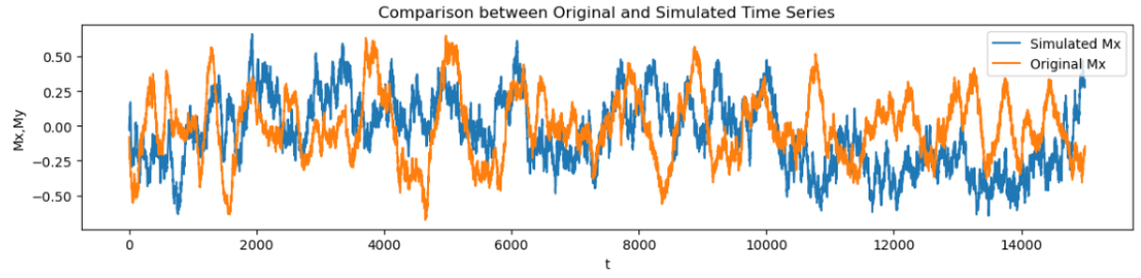
Figure 5.78: Simulated time series for m_x from our estimated SDEs compared with the original time series



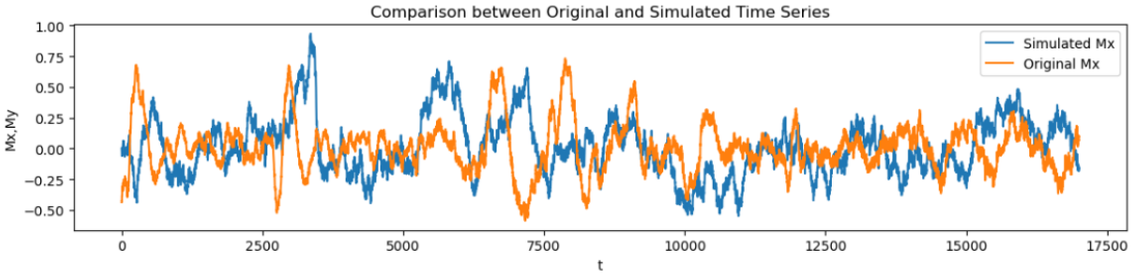
(a) Group Size: 4



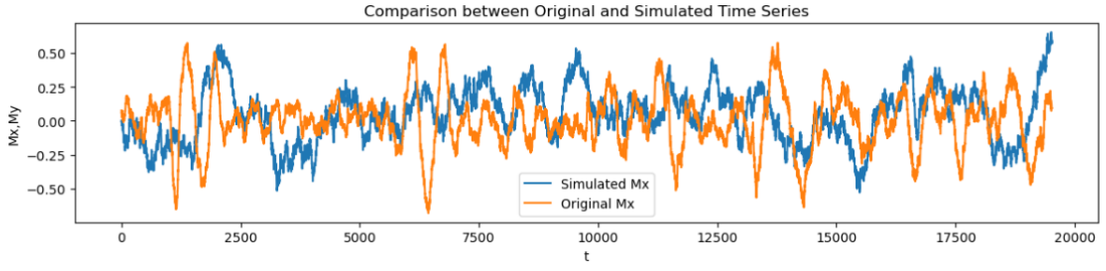
(b) Group Size: 10



(c) Group Size: 60



(d) Group Size: 80



(e) Group Size: 100

Following in the footsteps of the analysis of rotation order parameter, the analysis of the group behaviors can be separated into two main clusters: the smaller groups, with sizes of 4 and 10, and the larger groups, with sizes of 60, 80 and 100.

Analysis of Smaller Groups

For the deterministic parts: for the group size of 4, we have a linear deterministic system and for the group size of 10, it seems we are dealing with a nonlinear, yet quite simple system of equations where the dynamics of $\mathbf{m}$ are regulated by m_x and the interaction between m_x and m_y. The $\sqrt{0.3\,(1-|\mathbf{m}|^2)}$ and $\sqrt{0.1\,(1-|\mathbf{m}|^2)}$ terms in the equations represent the diffusion coefficients which are modulated by the magnitude of $\mathbf{m}$.

While the histograms for these groups show that the magnitude $|\mathbf{m}|$ is densely centered around its maximum value of 1, which indicates a strong directional preference within the group, the clear volatility marked by the increased randomness for larger values of $|\mathbf{m}|$ in the diffusion functions proves that the increased coordination in terms of polarization is spurious for these systems and is only a byproduct of their small number, since consistent polarized coordination is unsustainable. Furthermore, the short decay times of 1.5 and 2.5625 seconds for these small groups implies rapid memory loss, suggesting a high reactivity to changes in their environment, this ties well with the results for the rotation OP.

Analysis of Larger Groups

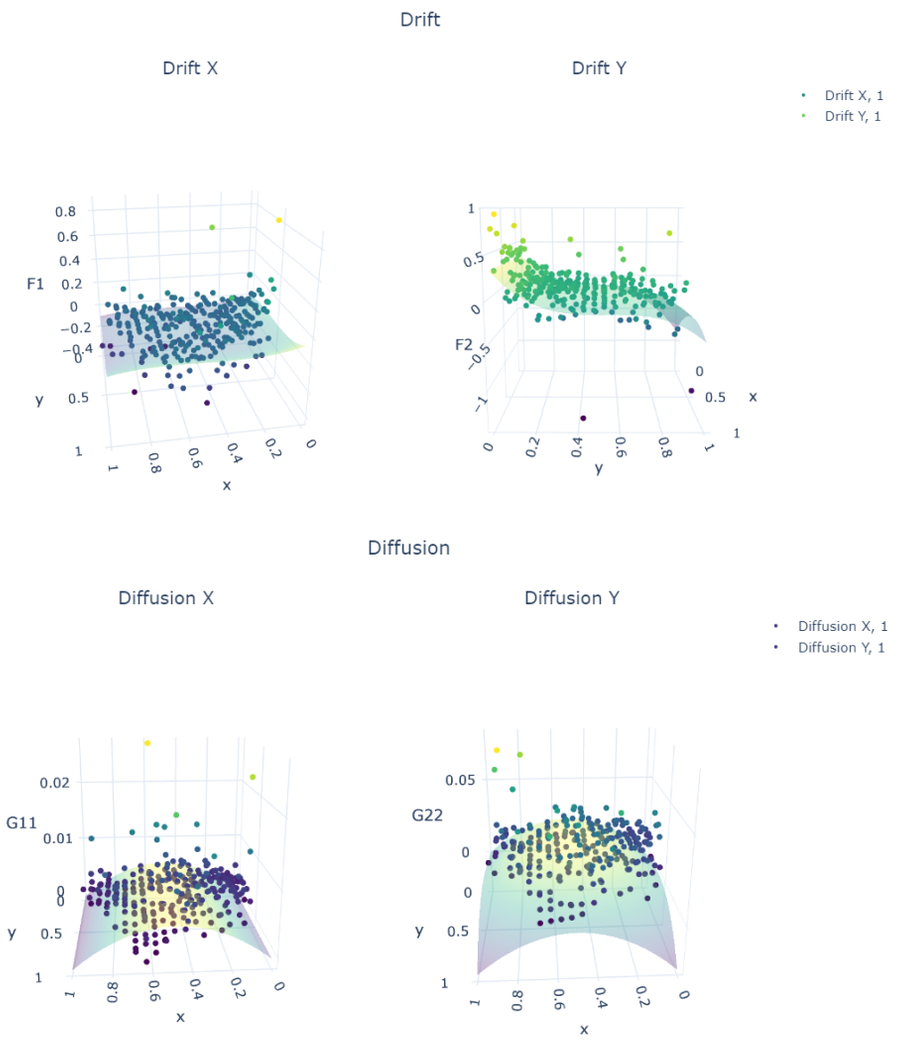
For the larger groups (60, 80, 100), the deterministic behavior is governed by a more complex nonlinear system, characterized by cubic and quadratic terms. As the group size increases, these nonlinearities become more prevalent, implying that the interactions within larger groups are more intricate.

The diffusion coefficients decrease for larger groups as shown by the $\sqrt{0.02}$, $\sqrt{0.009}$, and $\sqrt{0.006}$ terms in the respective equations. This suggests that as group size increases, the effect of stochasticity on the group's movement lessens, making the deterministic behavior more pronounced.

However, contrary to small groups, the $|\mathbf{m}|$ values in larger groups are concentrated around much smaller values, indicating a weaker average directional preference. This weaker alignment is consistent with the presence of the more complex interactions introduced by the higher order terms in the deterministic equations, and as we've already mentioned, in larger groups, the emergent rotational behavior can cause group members to have differing instantaneous directions even though they could be following the same overall circular path

5.3.3 *Coupled SDE of the Rotation and Polarization Order Parameters*

Finally, we look at a the rotational order parameter, and the magnitude of the polarization order parameter, aiming to look at the coupled stochastic equation that describes their evolution, but before doing that we pre-process the data, namely, we do a min-max re-scaling on the rotation order parameters, preserving the original shape of the data but now having the ranges of both parameters in [0,1]. Following the same methodology outlined above, we extract the 2D surfaces that best fit the drift and the diffusion:



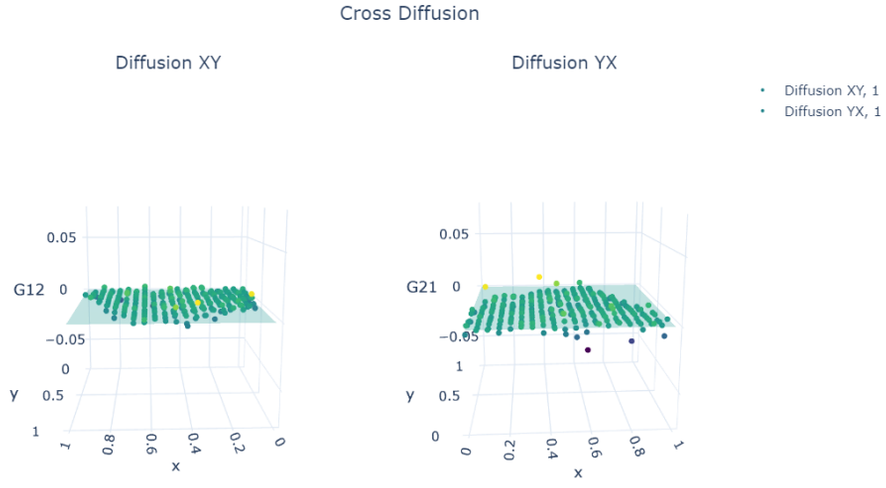


Figure 5.80: Drift and Diffusion and Cross Diffusion fits obtained by Sparse Regression in the 2D case, for the Coupled SDE of the rotation and the polarisation order parameters

And here our generated equations:

$$
\begin{aligned}
F_1 &= (-0.362 \pm 0.149)x + (-0.454 \pm 0.219)x^3 + (0.424 \pm 0.151)y \\
&\quad + (-1.024 \pm 0.343)y^2 + (0.748 \pm 0.271)y^3 \\
F_2 &= (0.293 \pm 0.055) + (-2.165 \pm 0.335)y + (-1.021 \pm 0.712)x^2 y \\
&\quad + (4.156 \pm 0.761)y^2 + (0.852 \pm 0.718)xy^2 \\
&\quad + (-3.118 \pm 0.601)y^3 \\
G_{11} &= (0.022 \pm 0.003)x + (-0.023 \pm 0.008)x^2 \\
G_{12} &= 0 \\
G_{21} &= 0 \\
G_{22} &= (0.095 \pm 0.015)x + (-0.094 \pm 0.034)x^2 \\
&\quad + (0.055 \pm 0.035)y^2 + (-0.080 \pm 0.028)y^3
\end{aligned}
$$

The full equation is thus:

$$
\begin{aligned}
\begin{pmatrix} \dot{R} \\ \dot{P} \end{pmatrix} &= \begin{pmatrix} -0.362R - 0.454R^3 + 0.424P - 1.024P^2 + 0.748P^3 \\ 0.293 - 2.165P - 1.021R^2P + 4.156P^2 + 0.852RP^2 - 3.118P^3 \end{pmatrix} \\
&\quad + \begin{pmatrix} 0.023R(1-R) & 0 \\ 0 & 0.095R(1-R) + 0.055P^2 - 0.080P^3 \end{pmatrix} \cdot \boldsymbol{\eta}(t) \qquad (5.23)
\end{aligned}
$$

Below are our checks for self-consistency, by looking at the re-estimated distribution of the magnitude of the vector $\rho = \sqrt{R^2 + P^2}$, its autocorrelation, and the time series of R and P compared with the original time series:

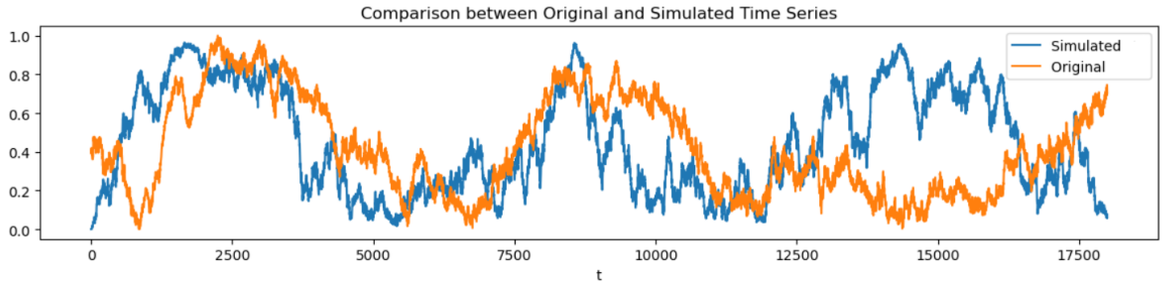


Figure 5.81: Comparison of the Original and Re-estimated Time Series for the Rotation OP

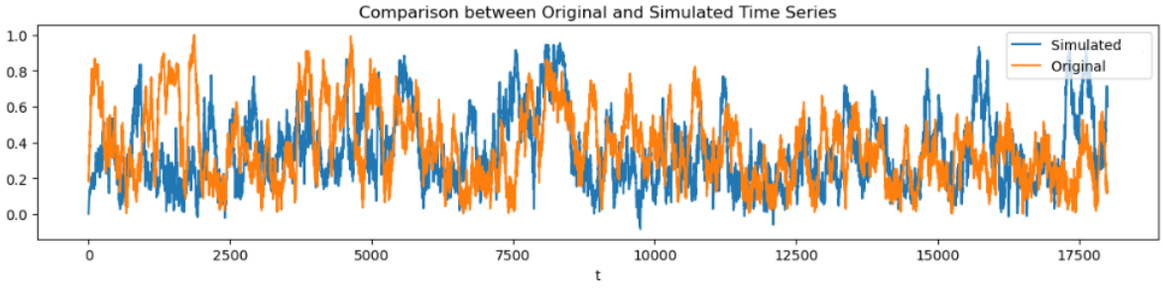


Figure 5.82: Comparison of the Original and Re-estimated Time Series for the Polarization OP

And the Distributions and Autocorrelation:

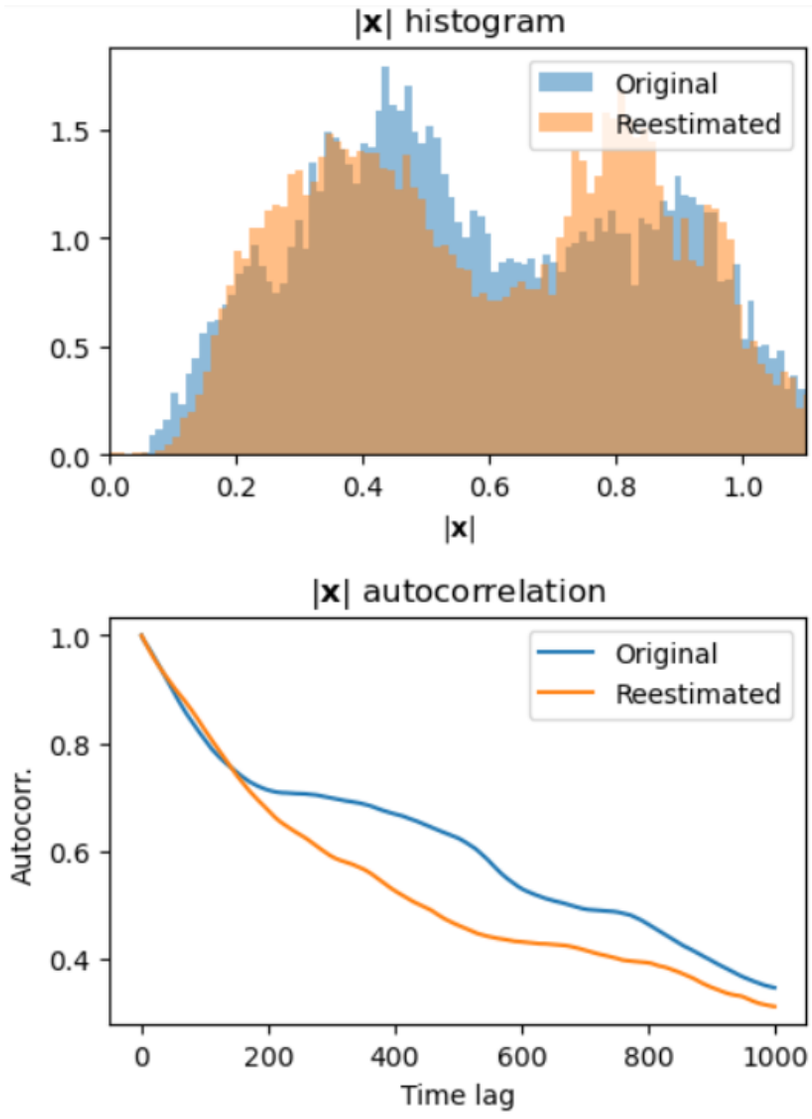


Figure 5.83: Comparison of the Original and Re-estimated Distribution and Autocorrelation of ρ

Chapter 6

Conclusion and Future Work

In conclusion, this study has demonstrated the potential of a multifaceted approach, combining physics, information theory, graph theory, and stochastic analysis to gain a deeper understanding of the collective behavior and dynamics of juvenile zebrafish shoals of varying group sizes.

The Optimal Causation Entropy principle (oCSE) proved to be an effective method to construct dynamic, causality-weighted networks, thereby offering a new lens through which we can analyze the interplay of coordination and density within these biological systems. With this new analysis tool, we've managed to explore the emergence of coordination and decipher the complex interactions within these causal networks across different group sizes, essentially reducing them to time-evolving pairwise interactions weighted by the interaction strength. We've been able to demonstrate how dense systems' increased memory times for maintaining their structure, particularly in their rotational motion, aligns with how larger group sizes have significantly less variability in the causal structure of the group. We've also empirically obtained the average number of interacting neighbors using real data, thus opening the door for applications to other species and in different environments. This approach permits an analysis of individual interactions, surpassing the confines of standard group analysis, which often focuses on mean-field behaviors, contrary to standard models of swarming and group behaviors our results hint at the existence of long-range interactions, thus showing that causal neighbors are not necessarily spatial neighbors. This is a direction that we wish to explore in the future, believing that understanding these long-range causal interactions is integral to shaping the behavior of groups, especially in water-borne systems where information could be transmitted through the medium and not only by sight.

In the last part of our study, we used the Kramers-Moyal equation and sparse regression techniques to get analytical expressions of the stochastic differential equations that describe how order parameters change over time. This permits the study of interacting systems by minimizing the number of assumptions required to model active matter, allowing for more accurate simulations if needed, and potentially providing a model-free approach to understanding active matter.

One immediate goal after that would be to build jump-diffusion stochastic equations [116] that describe the turn rates of the fish but which would require going in Kramers-Moyal expansion to much higher order terms, since the time series of the turn rates of our fish show clear jumps that could be explained by an additional jump-diffusion term to the stochastic equation (possibly modeled by a Poisson process). The reason for doing that would be to challenge available models in literature that assume that the functional form of the turn rates in the drift of the stochastic equation describing it depends on the local neighborhood of the fish [117], and potentially show that instead of summing over spatial neighbors, we could get better results by summing over causal neighbors.

Other potential areas of research are to expand the variables considered in our models beyond acceleration, to include other metrics such as position, velocity, or a combination thereof. This could potentially offer a more comprehensive understanding of the collective dynamics at play. Additionally, future research should also consider the variance in memory mechanisms and response times across different species. This would entail experimenting with different time delays in our calculations to discern whether there might be an optimal choice specific to each species.

Another promising avenue is the development of improved estimators of CSE or conditional mutual information. Given the rapidly evolving experimental conditions and data collection methods, it is expected that more data will become available in the future, likely leading to more accurate inferences. Moreover, developing estimators that minimize inference errors, even if the estimator itself might not be optimal in inferring individual CSE values, is a worthy pursuit. Alongside this, the development of an exact test for causality inference remains a potential issue. Despite these challenges, it is crucial to remember that the meaning and inference of causality still require a level of assumption and careful interpretation.

Appendix A

Kraskov-Strögbauer-Grassberger (KSG) estimator extension for mutual information

The KSG estimator can be effectively extended to estimate conditional mutual information. This extension has been recently proposed by Frenzel and Pompe [118] and independently by Vejmelka and Paluš [119]. Consider n independent samples $\{w_1, w_2, \ldots, w_n\}$ of the joint random variable $W = (X, Y, Z)$ where $w_i = (x_i, y_i, z_i)$. The estimate of $I(X;Y|Z)$ is given by:

$$I(X;Y|Z) = \psi(k) - \langle \psi(n_{xz}+1) + \ \psi(n_{yz}+1) - \psi(n_z+1) \rangle \tag{A.1}$$

In the equation, $\langle \cdot \rangle$ denotes the average over the samples and $\psi(t) = \frac{\Gamma'(t)}{\Gamma(t)}$ is the digamma function.
For a fixed value of k, let $\delta^{(i)}$ denotes the distance from $w_i = (x_i, y_i, z_i)$ to its k-th nearest neighbor, where distance is measured as the max-norm in the joint space,

$$||w_i - w_j||_{xyz} = \max \left\{ |x_i - x_j|_x, |y_i - y_j|_y, |z_i - z_j|_z \right\}, \tag{A.1}$$

and the norms used in the subspaces can be arbitrary but oftentimes the max norm is used (which is our choice for this paper) as well. From this we obtain

- $n_{xz}(i)$: number of points (x_j, z_j) ($j \neq i$) with
  $||(x_j, z_j) - (x_i, z_i)||_{xz} = \max \left\{ |x_j - x_i|_x, |z_j - z_i|_z \right\} < \delta^{(i)}$
- $n_{yz}(i)$: number of points (y_j, z_j) ($j \neq i$) with
  $||(y_j, z_j) - (y_i, z_i)||_{yz} = \max \left\{ |y_j - y_i|_y, |z_j - z_i|_z \right\} < \delta^{(i)}$
- $n_z(i)$: number of points z_j ($j \neq i$) with
  $\left| z_j - z_{i_z} \right| < \delta^{(i)}$

Appendix B

Kernel Density Estimation using Gaussian Kernel

Kernel Density Estimation (KDE) is a non-parametric method for estimating the probability density function of a given random variable. It is called 'non-parametric' because it does not assume any underlying distribution for the data. One of the popular implementations of KDE is available in the SciPy library, which uses the Gaussian kernel.

B.1 KDE in One Dimension

B.1.1 *Gaussian Kernel*

A kernel is a weighting function used in the estimation of the PDF. The Gaussian kernel is a popular choice due to its properties. The Gaussian kernel $K(x)$ for a one-dimensional input is defined as:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

where x is the distance between the data point and the location where we want to estimate the density.

B.1.2 *Kernel Density Estimation*

The kernel density estimate $f(x)$ at a point x for a given dataset $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is given by:

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where:

- n is the number of data points
- $K(x)$ is the kernel function

- h is the bandwidth which controls the smoothing parameter

h a critical factor in KDE that influences the bias-variance trade-off of the estimation. A small value of h may lead to a high-variance (over-fitted) estimate, while a large value may lead to a high-bias (under-fitted) estimate.

B.1.3 *Bandwidth Selection*

In Scipy python library, the bandwidth h is chosen automatically using Scott's Rule if not specified by the user:

$$h = n^{-\frac{1}{d+4}}$$

where n is the number of data points and d is the number of dimensions. This rule is a rule of thumb for bandwidth selection.

B.2 Gaussian Kernel Density Estimation in Two Dimensions

When we extend the KDE to two or more dimensions, the calculations become slightly more complex, but the principle remains the same. We'll specifically look at the two-dimensional case here.

B.2.1 *Two-Dimensional Gaussian Kernel*

The Gaussian kernel for a two-dimensional input vector $\mathbf{x} = [x_1, x_2]$ is defined as:

$$K(\mathbf{x}) = \frac{1}{2\pi} e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}}$$

where $\mathbf{x}^{\mathrm{T}}$ is the transpose of the vector $\mathbf{x}$, and $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ is the dot product of $\mathbf{x}$ and its transpose.

B.2.2 *Kernel Density Estimation in Two Dimensions*

The kernel density estimate $f(\mathbf{x})$ at a point $\mathbf{x}$ for a given dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is given by:

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|\mathbf{H}|^{1/2}} K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right)$$

where: n is the number of data points, K is the two-dimensional Gaussian kernel function, $\mathbf{H}$ is the bandwidth matrix, which replaces the scalar bandwidth h from the one-dimensional case. $\mathbf{H}$ controls the shape of the kernel, and $|\mathbf{H}|$ denotes the determinant of $\mathbf{H}$

B.2.3 *Bandwidth Matrix Selection*

In the Scipy python library, if the bandwidth matrix $\mathbf{H}$ is not specified by the user, it is chosen automatically using a generalization of Scott's Rule:

$$\mathbf{H} = n^{-1/(d+4)}\mathbf{\Sigma}$$

where n is the number of data points, d is the dimension of the data (in this case, $d = 2$, $\mathbf{\Sigma}$ is the covariance matrix of the data).

Bibliography

[1] T. Vicsek and A. Zafeiris, "Collective motion," *Physics reports*, vol. 517, no. 3-4, pp. 71–140, 2012.

[2] S. Ramaswamy, "Active matter," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, no. 5, p. 054 002, 2017.

[3] D. S. Calovi, U. Lopez, S. Ngo, C. Sire, H. Chaté, and G. Theraulaz, "Swarming, schooling, milling: Phase diagram of a data-driven fish school model," *New journal of Physics*, vol. 16, no. 1, p. 015 026, 2014.

[4] M. C. Marchetti, J.-F. Joanny, S. Ramaswamy, *et al.*, "Hydrodynamics of soft active matter," *Reviews of modern physics*, vol. 85, no. 3, p. 1143, 2013.

[5] W. M. Lord, J. Sun, N. T. Ouellette, and E. M. Bollt, "Inference of causal information flow in collective animal behavior," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 107–116, 2016.

[6] J. G. Puckett, R. Ni, and N. T. Ouellette, "Time-frequency analysis reveals pairwise interactions in insect swarms," *Physical review letters*, vol. 114, no. 25, p. 258 103, 2015.

[7] J. K. Parrish and L. Edelstein-Keshet, "Complexity, pattern, and evolutionary trade-offs in animal aggregation," *Science*, vol. 284, no. 5411, pp. 99–101, 1999.

[8] R. Ni, J. G. Puckett, E. R. Dufresne, and N. T. Ouellette, "Intrinsic fluctuations and driven response of insect swarms," *Physical review letters*, vol. 115, no. 11, p. 118 104, 2015.

[9] M. Tennenbaum, Z. Liu, D. Hu, and A. Fernandez-Nieves, "Mechanics of fire ant aggregations," *Nature materials*, vol. 15, no. 1, pp. 54–59, 2016.

[10] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet, "Inferring individual rules from collective behavior," *Proceedings of the National Academy of Sciences*, vol. 107, no. 28, pp. 12 576–12 580, 2010.

[11] J. E. Herbert-Read, A. Perna, R. P. Mann, T. M. Schaerf, D. J. Sumpter, and A. J. Ward, "Inferring the rules of interaction of shoaling fish," *Proceedings of the National Academy of Sciences*, vol. 108, no. 46, pp. 18 726–18 731, 2011.

[12] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel type of phase transition in a system of self-driven particles," *Physical review letters*, vol. 75, no. 6, p. 1226, 1995.

[13] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987, pp. 25–34.

[14] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks, "Collective memory and spatial sorting in animal groups," *Journal of theoretical biology*, vol. 218, no. 1, pp. 1–11, 2002.

[15] N. T. Ouellette, "Empirical questions for collective-behaviour modelling," *Pramana*, vol. 84, pp. 353–363, 2015.

[16] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[17] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[18] F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. Heras, and G. G. De Polavieja, "Idtracker. ai: Tracking all individuals in small or large collectives of unmarked animals," *Nature methods*, vol. 16, no. 2, pp. 179–182, 2019.

[19] L. E. Rocha, J. Ryckebusch, K. Schoors, and M. Smith, "The scaling of social interactions across animal species," *Scientific Reports*, vol. 11, no. 1, p. 12 584, 2021.

[20] L. F. Hughey, A. M. Hein, A. Strandburg-Peshkin, and F. H. Jensen, "Challenges and solutions for studying collective animal behaviour in the wild," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 373, no. 1746, p. 20 170 005, 2018.

[21] H. Ling, G. E. Mclvor, K. van der Vaart, R. T. Vaughan, A. Thornton, and N. T. Ouellette, "Local interactions and their group-level consequences in flocking jackdaws," *Proceedings of the Royal Society B*, vol. 286, no. 1906, p. 20 190 865, 2019.

[22] A. Nabeel, A. Karichannavar, S. Palathingal, J. Jhawar, V. Guttal, *et al.*, "Pydaddy: A python package for discovering stochastic dynamical equations from timeseries data," *arXiv preprint arXiv:2205.02645*, 2022.

[23] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[24] B. Bollobás, *Modern graph theory*. Springer Science & Business Media, 1998, vol. 184.

[25] U. Brandes, *Network analysis: methodological foundations*. Springer Science & Business Media, 2005, vol. 3418.

[26] P. J. Carrington, J. Scott, and S. Wasserman, *Models and methods in social network analysis*. Cambridge university press, 2005, vol. 28.

[27] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," 1994.

[28] P. R. Monge, N. S. Contractor, and P. S. Contractor, *Theories of communication networks*. Oxford University Press, USA, 2003.

[29] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *Annual review of information science and technology*, vol. 37, no. 1, pp. 179–255, 2003.

[30] B. Cronin and H. B. Atkins, *The web of knowledge: A festschrift in honor of Eugene Garfield*. Information Today, 2000.

[31] A.-L. Barabasi and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature reviews genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[32] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[33] M. Buchanan, "Small world: Uncovering nature's hidden networks," 2002.

[34] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, "Metric structure of random networks," *Nuclear Physics B*, vol. 653, no. 3, pp. 307–338, 2003.

[35] R. Pastor-Satorras and A. Vespignani, *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2004.

[36] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," *American Journal of sociology*, vol. 105, no. 2, pp. 493–527, 1999.

[37] S.-H. Yook, H. Jeong, and A.-L. Barabási, "Modeling the internet's large-scale topology," *Proceedings of the National Academy of Sciences*, vol. 99, no. 21, pp. 13 382–13 386, 2002.

[38] S. N. Dorogovtsev and J. F. Mendes, *Evolution of networks: From biological nets to the Internet and WWW*. Oxford university press, 2003.

[39] J. Park and M. E. Newman, "Origin of degree correlations in the internet and other networks," *Physical review e*, vol. 68, no. 2, p. 026 112, 2003.

[40] L. Euler, *Mechanica sive motus scienta analytice exposita... instar supplementi ad Commentar. Acad. scient. imper.* ex typographia Academiae scientarum, 1736, vol. 2.

[41] E. RENYI, "On random graph," *Publicationes Mathematicate*, vol. 6, pp. 290–297, 1959.

[42] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[43] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the national academy of sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.

[44] R. Guimera, S. Mossa, A. Turtschi, and L. N. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 22, pp. 7794–7799, 2005.

[45] G. Fagiolo, "Clustering in complex directed networks," *Physical Review E*, vol. 76, no. 2, p. 026 107, 2007.

[46] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Physical review letters*, vol. 87, no. 19, p. 198 701, 2001.

[47] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[48] V. Red, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM review*, vol. 53, no. 3, pp. 526–543, 2011.

[49] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge university press, 2008.

[50] G. Craciun and M. Feinberg, "Multiple equilibria in complex chemical reaction networks: Semiopen mass action systems," *SIAM Journal on Applied Mathematics*, vol. 70, no. 6, pp. 1859–1877, 2010.

[51] F. Dorfler and F. Bullo, "Synchronization and transient stability in power networks and nonuniform kuramoto oscillators," *SIAM Journal on Control and Optimization*, vol. 50, no. 3, pp. 1616–1642, 2012.

[52] M. Golubitsky, I. Stewart, and A. Török, "Patterns of synchrony in coupled cell networks with multiple arrows," *SIAM Journal on Applied Dynamical Systems*, vol. 4, no. 1, pp. 78–100, 2005.

[53] A. Pomerance, E. Ott, M. Girvan, and W. Losert, "The effect of network topology on the stability of discrete state models of genetic control," *Proceedings of the National Academy of Sciences*, vol. 106, no. 20, pp. 8209–8214, 2009.

[54] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing multiple nonlinear time series with extended granger causality," *Physics letters A*, vol. 324, no. 1, pp. 26–35, 2004.

[55] J. Sun and A. E. Motter, "Controllability transition and nonlocality in network control," *Physical review letters*, vol. 110, no. 20, p. 208 701, 2013.

[56] N. Chen, "On the approximability of influence in social networks," *SIAM Journal on Discrete Mathematics*, vol. 23, no. 3, pp. 1400–1415, 2009.

[57] J. Kleinberg, "The small-world phenomenon: An algorithmic perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 2000, pp. 163–170.

[58] T. Nishikawa and A. E. Motter, "Network synchronization landscape reveals compensatory structures, quantization, and the positive effect of negative interactions," *Proceedings of the National Academy of Sciences*, vol. 107, no. 23, pp. 10 342–10 347, 2010.

[59] J. Pearl *et al.*, "Models, reasoning and inference," *Cambridge, UK: CambridgeUniversityPress*, vol. 19, no. 2, 2000.

[60] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[61] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[62] T. P. Peixoto, "Network reconstruction and community detection from dynamics," *Physical review letters*, vol. 123, no. 12, p. 128 301, 2019.

[63] D. Knoke and S. Yang, *Social network analysis*. SAGE publications, 2019.

[64] A. A. Ganin, M. Kitsak, D. Marchese, J. M. Keisler, T. Seager, and I. Linkov, "Resilience and efficiency in transportation networks," *Science advances*, vol. 3, no. 12, e1701079, 2017.

[65] S. Athey, "Machine learning and causal inference for policy evaluation," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 5–6.

[66] J. F. Hair Jr and M. Sarstedt, "Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing," *Journal of Marketing Theory and Practice*, vol. 29, no. 1, pp. 65–77, 2021.

[67] J. Grimmer, "We are all social scientists now: How big data, machine learning, and causal inference work together," *PS: Political Science & Politics*, vol. 48, no. 1, pp. 80–83, 2015.

[68] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[69] B. Russell, "On the notion of cause," in *Proceedings of the Aristotelian society*, JSTOR, vol. 13, 1912, pp. 1–26.

[70] J. Pearl *et al.*, "Models, reasoning and inference," *Cambridge, UK: CambridgeUniversityPress*, vol. 19, no. 2, 2000.

[71] W. Norbert and E. Beckenbach, "The theory of prediction," *Modern Mathematics for Engineers; Beckenbach, EF, Ed.; McGraw-Hill: New York, NY, USA*, vol. 1, 1956.

[72] C. W. J. Granger, "Time series analysis, cointegration, and applications," *American Economic Review*, vol. 94, no. 3, pp. 421–425, 2004.

[73] J. Geweke, "Inference and causality in economic time series models," *Handbook of econometrics*, vol. 2, pp. 1101–1144, 1984.

[74] U. Triacca, "Is granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature?" *Theoretical and applied climatology*, vol. 81, pp. 133–135, 2005.

[75] D. Bell, J. Kay, and J. Malley, "A non-parametric approach to non-linear causality testing," *Economics Letters*, vol. 51, no. 1, pp. 7–18, 1996.

[76] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing multiple nonlinear time series with extended granger causality," *Physics letters A*, vol. 324, no. 1, pp. 26–35, 2004.

[77] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[78] M. Paluš, V. Komárek, Z. Hrnčiř, and K. Štěrbová, "Synchronization as adjustment of information rates: Detection from bivariate time series," *Physical Review E*, vol. 63, no. 4, p. 046 211, 2001.

[79] I. Mokhov and D. Smirnov, "El niño–southern oscillation drives north atlantic oscillation as revealed with nonlinear techniques from climatic indices," *Geophysical Research Letters*, vol. 33, no. 3, 2006.

[80] T. Katura, N. Tanaka, A. Obata, H. Sato, and A. Maki, "Quantitative evaluation of interrelations between spontaneous low-frequency oscillations in cerebral hemodynamics and systemic cardiovascular dynamics," *Neuroimage*, vol. 31, no. 4, pp. 1592–1600, 2006.

[81] M. Chávez, J. Martinerie, and M. Le Van Quyen, "Statistical assessment of nonlinear causality: Application to epileptic eeg signals," *Journal of neuroscience methods*, vol. 124, no. 2, pp. 113–128, 2003.

[82] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial granger causality—eliminating exogenous inputs and latent variables," *Journal of neuroscience methods*, vol. 172, no. 1, pp. 79–93, 2008.

[83] J. Sun and E. M. Bollt, "Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings," *Physica D: Nonlinear Phenomena*, vol. 267, pp. 49–57, 2014.

[84] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of computational neuroscience*, vol. 30, pp. 45–67, 2011.

[85] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj, "Geometric denoising of protein-protein interaction networks," *PLoS computational biology*, vol. 5, no. 8, e1000454, 2009.

[86] G. Stolovitzky, D. Monroe, and A. Califano, "Dialogue on reverse-engineering assessment and methods: The dream of high-throughput pathway inference," *Annals of the New York Academy of Sciences*, vol. 1115, no. 1, pp. 1–22, 2007.

[87] D. A. Smirnov, "Spurious causalities with transfer entropy," *Physical Review E*, vol. 87, no. 4, p. 042 917, 2013.

[88] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Physical review letters*, vol. 99, no. 20, p. 204 101, 2007.

[89] J. Runge, J. Heitzig, N. Marwan, and J. Kurths, "Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy," *Physical Review E*, vol. 86, no. 6, p. 061 121, 2012.

[90] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[91] T. Cover and J. Thomas, "Elements of information theory: Wiley online library," 1991.

[92] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066 138, 2004.

[93] M. Vejmelka and M. Paluš, "Inferring the directionality of coupling with conditional mutual information," *Physical Review E*, vol. 77, no. 2, p. 026 214, 2008.

[94] S. H. Strogatz, *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[95] A. J. McKane and T. J. Newman, "Stochastic models in population biology and their deterministic analogs," *Physical Review E*, vol. 70, no. 4, p. 041 902, 2004.

[96] H. Cheng, N. Yao, Z.-G. Huang, J. Park, Y. Do, and Y.-C. Lai, "Mesoscopic interactions and species coexistence in evolutionary game dynamics of cyclic competitions," *Scientific reports*, vol. 4, no. 1, pp. 1–7, 2014.

[97] C. A. Yates, R. Erban, C. Escudero, *et al.*, "Inherent noise can facilitate coherence in collective swarm motion," *Proceedings of the National Academy of Sciences*, vol. 106, no. 14, pp. 5464–5469, 2009.

[98] A. J. Black and A. J. McKane, "Stochastic formulation of ecological models and their applications," *Trends in ecology & evolution*, vol. 27, no. 6, pp. 337–345, 2012.

[99] S. Majumder, A. Das, A. Kushal, S. Sankaran, and V. Guttal, "Finite-size effects, demographic noise, and ecosystem dynamics," *The European Physical Journal Special Topics*, vol. 230, no. 16, pp. 3389–3401, 2021.

[100] T. Biancalani, L. Dyson, and A. J. McKane, "Noise-induced bistable states and their mean switching time in foraging colonies," *Physical review letters*, vol. 112, no. 3, p. 038 101, 2014.

[101] S. Leyk, A. E. Gaughan, S. B. Adamo, *et al.*, "The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use," *Earth System Science Data*, vol. 11, no. 3, pp. 1385–1409, 2019.

[102] R. Nathan, C. T. Monk, R. Arlinghaus, *et al.*, "Big-data approaches lead to an increased understanding of the ecology of animal movement," *Science*, vol. 375, no. 6582, eabg1780, 2022.

[103] J. Jhawar, R. G. Morris, U. Amith-Kumar, *et al.*, "Noise-induced schooling of fish," *Nature Physics*, vol. 16, no. 4, pp. 488–493, 2020.

[104] K. Tunstrøm, Y. Katz, C. C. Ioannou, C. Huepe, M. J. Lutz, and I. D. Couzin, "Collective states, multistability and transitional behavior in schooling fish," *PLoS computational biology*, vol. 9, no. 2, e1002915, 2013.

[105] N. C. Stenseth, W. Falck, O. N. Bjørnstad, and C. J. Krebs, "Population regulation in snowshoe hare and canadian lynx: Asymmetric food web configurations between hare and lynx," *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 5147–5152, 1997.

[106] R. E. Lenski, "Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations," *The ISME journal*, vol. 11, no. 10, pp. 2181–2194, 2017.

[107] J. Gradišek, S. Siegert, R. Friedrich, and I. Grabec, "Analysis of time series from stochastic processes," *Physical Review E*, vol. 62, no. 3, p. 3146, 2000.

[108] J. Gradišek, S. Siegert, R. Friedrich, and I. Grabec, "Analysis of time series from stochastic processes," *Physical Review E*, vol. 62, no. 3, p. 3146, 2000.

[109] R. Tabar, *Analysis and data-based reconstruction of complex nonlinear dynamical systems*. Springer, 2019, vol. 730.

[110] F. Dietrich, A. Makeev, G. Kevrekidis, *et al.*, "Learning effective stochastic differential equations from microscopic simulations: Linking stochastic numerics to deep learning," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 33, no. 2, p. 023 121, 2023.

[111] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations," *Science advances*, vol. 3, no. 4, e1602614, 2017.

[112] L. Boninsegna, F. Nüske, and C. Clementi, "Sparse learning of stochastic dynamical equations," *The Journal of chemical physics*, vol. 148, no. 24, p. 241 723, 2018.

[113] J. L. Callaham, J.-C. Loiseau, G. Rigas, and S. L. Brunton, "Nonlinear stochastic modelling with langevin regression," *Proceedings of the Royal Society A*, vol. 477, no. 2250, p. 20 210 092, 2021.

[114] Y. Huang, Y. Mabrouk, G. Gompper, and B. Sabass, "Sparse inference and active learning of stochastic differential equations from data," *Scientific Reports*, vol. 12, no. 1, p. 21 691, 2022.

[115] C. Gardiner, *Stochastic methods*. Springer Berlin, 2009, vol. 4.

[116] V. Mwaffo, R. P. Anderson, S. Butail, and M. Porfiri, "A jump persistent turning walker to model zebrafish locomotion," *Journal of The Royal Society Interface*, vol. 12, no. 102, p. 20 140 884, 2015.

[117] J. Gautrais, F. Ginelli, R. Fournier, *et al.*, "Deciphering interactions in moving animal groups," 2012.

[118] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Physical review letters*, vol. 99, no. 20, p. 204 101, 2007.

[119] M. Vejmelka and M. Paluš, "Inferring the directionality of coupling with conditional mutual information," *Physical Review E*, vol. 77, no. 2, p. 026 214, 2008.