Data-driven methods for social complex systems: social media, data assimilation and how they connect

Blas Kolic

Mansfield College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Abstract

The explosion of big, granular social data has enabled us to observe society from a microscopic perspective. With major events driving and reshaping our social systems, it is essential to exploit these data so that we can help drive these systems to a state of social well-being and sustainability. This thesis develops data-driven methods to study social systems from a bottom-up perspective. We divide it into two parts: one where granular, individual-level social data are available to analyze and one where they are not, so that we have to infer them.

In the first part, we analyze massive social media data containing Twitter discussions around Covid-19 and climate change. We study each of these discussions from distinct but complementary perspectives. For Covid-19, we quantify the public risk perception and emotion during the pandemic by exploiting the natural language used in the tweets. We find evidence of psychophysical numbing: Twitter users increasingly fixate on mortality, but in a decreasingly emotional and increasingly analytic tone. For climate change, we quantify the polarization dynamics based on the interaction structure between Twitter users. We find two stable, highly polarized groups: climate believers and climate skeptics, whose polarization drops significantly during the "FridaysForFuture" strikes of September 2019. In the second part, we develop methods for inferring individual-level data of complex system models when the data available are noisy, aggregated, and incomplete. Assuming our model is a dynamical system, we investigate under which conditions we can infer accurate initial conditions using incomplete data. The way the data are aggregated, their levels of noise, and the model's complexity highly influence the quality of the inference. We thus propose methods to estimate individual-level data on the fly as observations become available. We validate these methods for several chaotic systems and an agent-based model of social opinion dynamics. We hope this work helps bridge the gap in designing models that better predict possible societal pathways.

Acknowledgements

I want to express my gratitude and admiration for my supervisor Doyne Farmer. Doyne, you have been an infinite source of inspiration throughout my DPhil, not only in your academic brilliance and creativity but also in your life philosophy. Thanks for always being on my side, putting yourself in others' shoes, and always triggering such stimulating discussions and stories.

Many experienced minds have shaped how I think, consume, value, and discuss science. Thanks to Cameron Hepburn for co-supervising my DPhil and always giving me helpful and timely advice, to Renaud Lambiotte for always being so friendly and engaging with me in great discussions about networks and mathematics, and to Francois Lafond, whose opinion and experience I value a lot. Massive thanks to Renaud Lambiotte and to Sylvain Barde for examining my thesis thoroughly and providing me with very useful feedback and discussion. I am also grateful to the INET and Maths admin teams: Dorothy Nicholas, Sandhya Patel, Camilla Burmeister, Susan Mousley, and Elizabeth Rozeboom. You have almost literally saved my skin hundreds of times.

Special thanks to my co-authors Joel Dyer, Juan Sabuco, Doyne Farmer, Fabián Aguirre, Sergio Hernández, and Guillermo Garduño. Advancing science should be a collaborative process. Working with you has been my most wonderful academic experience.

I am grateful to the sources that have funded me to pursue this DPhil: the Mexican Conacyt-SENER scholarship and the Smith's award. Studying is a privilege I would not have been able to do without their financial support.

The DPhil is a long journey. I cannot imagine doing it if it were not for the amazing people in my life for the past four years. Today I celebrate with and for you:

Salud to Rita María del Río for inviting me to apply to Oxford, introducing me to Doyne, and being almost like a sister to me. Salud to the Tortuga team: Fabían Aguirre, Leonardo Castro, Rodrigo Leal, Santiago Martínez, and Carlota Segura. You have kept my heart warm and my head sharp all the time. Salud to my former flatmates: Rodrigo Leal, Garima Jaju, Lorena Valdivia, Luca Mungo, Giulia Bernardini, Diana Avadanii, and Andrius Vaicenavicius. You have made my experience in Oxford as fun as it can get, and I have grown with and because of you. Salud to my musician friends: Daniel Patiño, Andrius, Uwade Akhere, and Miguel Ángel Fernández. Reconnecting with music has been very important to me. I cannot even describe how much I enjoyed all the shows. Salud to all the people I have met in and through the

Oxford Mexican Society. You have created an oasis that tastes like Mexico in the middle of Great Britain.

Salud to the INET crew: Rita María del Río, Luca Mungo, Kieran Murray, Joris Bucker, Nils Rochowitz, José Morán, Joel Dyer, Andrea Bacillieri, Anton Pichler, Julian Winkler, Valentina Semenova, Karolina Bassa, Marco Pangallo, Torsten Heinrich, Will Scarrold, Penny Mealy, Rupert Way, Yangsiyu Lu, Sam Wiese, Maarten Scholl, Anna Berryman, Donovan Platt, Matt Ives, Lucas Kruitwagen, and more. I have learned so much from you. I am very grateful to have shared these four years of making science, having fun, and growing as students, academics, and friends.

There is more. Salud to the great friends I have not mentioned yet but that have had a profound impact on my life: Juan Rafael Álvarez, Sofía Lykou, Ismael Rodríguez, Giancarlo Antonucci, Giulia Anselmi, Bato Balvanera, John and Clem Pougué-Biyong, Valentina Rosas, Amy Ross, Tahnee Ooms, and many others. Wow, guys, you are incredible. Thanks for taking me and accepting me into your lives. Salud to the friends I knew from before and, for some reason or another, were in the UK during my DPhil: Elena and Claudio Pierard, Miguel Quetzeri, Elena Luciano, Alejandro Corinto, Iker de Icaza, and Daniela Sclavo. It is crazy how many new stories we have lived together that we would have never imagined a couple of years back. Salud to my scientific mentors which transmitted and endorsed me with the love for science, Maths, and Physics: Sergio de Regules and Luis Benet. Salud to my extended family: Alejandro Guerrero, Evelyn Gross, Shuki Zaragoza, Giancarlo Castello, and Nachito Loaiza.

I would like to thank all my family in Argentina, especially my aunts Sandra Sarancone and María Ramonda. They have been the warmest and most generous duo and the ones responsible for reuniting everyone together countless times.

To my mom, Claudia Mierez, and my dad, Daniel "Chicho" Kolic. Mamu, you are the backbone of who I am. Your infinite and selfless love, support, and wisdom throughout my life have made me grow into a person that makes me proud. I admire you. Weysoncho, you are the strongest wind in the sea and the brightest light on the horizon. I miss you, and I recognize myself in you.

Last and foremost, I want to thank my girlfriend, Karen Olivera. I could not have made a better decision than to share my life with you. You are the most beautiful, sensitive, and intuitive human I know, and for that, you are my favorite. Distance does not exist when you build strong bridges, and we have built the strongest one. I love you. I dedicate this thesis to you. *Y aquí estoy esperando el momento, al fin, de poderte ver.*

Contents

Chapter 1

Introduction

The increasing availability of granular individual-level data and the use of complex systems models for the social sciences has revolutionized our understanding of social phenomena [127, 114]. By harnessing the power of big data and computational techniques, we are now able to study social phenomena at unprecedented levels of detail and accuracy [12].

"Social science is the science of social interactions and their implications for society" [216]. Complex systems models are particularly well-suited for studying social phenomena, as they allow us to take into account the many different factors that can affect social interactions and outcomes. By understanding how the various elements of a social system interact with each other, we can develop more accurate models that can be used to make predictions about future behavior [134]. Interactions are a dynamic process between pairs of entities, so we can model them as a dynamical system taking place in a (possibly) evolving network [193].

Why, then, has the development of social science been primarily qualitative and descriptive if complex systems provide a quantifiable framework for studying it? Data. Or, to be more precise, the lack thereof. Only until recently, social data have been recorded, handled and analyzed in massive and granular quantities with the ongoing growth and widespread use of social media and the Internet [216]. While publicly-available granular social data are still limited to a few platforms, studying and validating social complex systems from a big data perspective has become an active and relevant area of research [32, 114].

In this thesis, we develop data-driven methods for understanding complex social systems. Our goal is to use data to develop more accurate models of how various elements of a social system interact with each other in the light of new information. We provide a theoretically sound mathematical framework motivated by theories from the social sciences, and our quantitative methods are mainly derived from network

science, natural language processing, dynamical systems, statistics, and data assimilation. We follow Lindgren's interdisciplinary approach to the relationship between data analysis and social theory, where he argues that social theory needs data analysis as much as data analysis needs social theory [147].

This thesis consists of two main parts, and we outline its structure in what follows.

Part I: Twitter as a social complex system. Twitter is an online platform that allows for real-time sharing of information and experiences. The tweets that users post can be viewed as a dynamic and complex network. In this context, developing data-driven methods for Twitter as a social complex system can help us understand how the public is reacting to major ongoing events such as Covid-19 and climate change, and what the potential implications of those reactions might be.

In Chapter 2, we make an overview of when and why researchers have followed a complex systems approach towards social media data. We discuss state-of-the-art methods and modeling techniques in the context of massive Twitter data. We then introduce the main concepts from the fields of network science and natural language processing, which conform the basis of the methods introduced in this part.

In Chapter 3, we investigate the risk perception and emotions of Twitter users during the first wave of the Covid-19 pandemic. To do so, we collect and analyze a massive dataset of Covid-19-related tweets from twelve Spanish- and English-speaking countries. We analyze the *content* of the tweets, which we quantify and map into multidimensional time-series using a word-based dictionary curated by psychologists [174]. We explore the evolution of the tweets' emotional framing using a word co-occurrence network analysis based on several linguistic categories. This study showcases how we can monitor and guide public communication using our methods. We construct such methods based on the theories of *risk perception* [93] and *psychophysics* [98].

Chapter 4 studies the interaction structure of the Twitter climate change conversation. We collect a massive dataset of climate-related tweets during 2019, when critical climate-related movements occurred. We analyze the temporal retweet network formed by Twitter users (nodes) and the interactions between them (edges). We quantify the system's emergent ideological structures - such as echo chambers - based solely on the retweeting patterns of Twitter users. We explore the dynamics of ideological polarization and analyze that Twitter-exogenous events, such as the "Fridays for Future" strikes [214], have in the polarization. Further, we uncover the ideological position of previously unobserved high-impact users based on the information channels of their audiences. We explain our results by drawing from the theories of homophily [158] and confirmation bias [227].

Part II: Connecting aggregated data with complex systems models. One of the great challenges of computational social science is understanding complex systems from incomplete data. This part focuses on developing data-assimilation methods to infer the unobserved variables of complex systems models when individual-level data are not available. In general, we assume we possess an accurate model of a complex system, but where the data is aggregated, noisy, and incomplete. These methods can provide insights into the underlying mechanisms of the systems studied.

In Chapter 5, we present an overview of the latent state inference methods, their extensive development in the atmospheric and oceanographic sciences, and how they are now being applied to social systems. Then, we formulate the latent state estimation problem mathematically and develop, from Bayesian statistics, a general form of most data-assimilation techniques.

In Chapter 6, we estimate the latent initial conditions of complex dynamical systems from incomplete information. We propose a methodology to initialize high-dimensional (possibly chaotic) systems based on noisy, aggregated time series. Our method is robust to how we aggregate the microstate as long as this aggregation does not destroy unrecoverable information[1] about the latent states. We analyze the amount of information needed to recover the ground-truth initial condition on several chaotic systems. We also analyze how the observational noise affects our ability to initialize a system [175].

In Chapter 7, we adapt the Ensemble Kalman filter (EnKF) [82], a method heavily used in the numerical weather prediction community, to estimate the latent states from noisy, aggregated data for models embedded in network topologies. To do so, we introduce a technique that incorporates our knowledge of the model's network topology and improves the quality of the EnKF estimations. Additionally, we incorporate constraints into the filtering equations when the state space is bounded [201], thus making the EnKF suitable for high-dimensional complex systems. Our method is efficient with respect to state-of-the-art latent state inference techniques and is easily extendable to other settings. We validate our method with synthetic data in chaotic systems and a social agent-based model of opinion dynamics [113, 230]

Finally, in Chapter 8 we conclude.

[1] By unrecoverable information, we mean that no matter how large and clean the data is, it is impossible to distinguish the ground-truth initial conditions from other states.

1.1 Statement of originality

A great portion of DPhil research was collaborative. In the following points, I specify which chapters involved collaborations and to what extent.

- **Chapter 3: Public risk perception and emotion during the Covid-19 pandemic**. This chapter was a joint collaboration with Joel Dyer in equal contribution. This work is publicly available at [78]. For most of this work, Joel and I kept constant online communication, and we both wrote the manuscript in equal proportions. Therefore, it is hard to pinpoint exactly who did what. In general terms, Joel proposed the word co-occurrence analysis, and compared the performance of the Weber-Fechner's law and the power law models. I proposed using the Weber-Fechner's law model in the first place and proposed to connect our results with the field of psychophysics. I suggested to take the word co-occurrence analysis to a more quantitative realm. We both cleaned and gathered the data, designed the linguistic scores, and wrote the manuscript in equal proportions.

- **Chapter 4: Quantifying the structure of the climate change conversation with unsupervised methods**: This chapter was a joint collaboration with Fabián Aguirre-López and, in minor proportion, with Sergio Hernández-Williams and Guillermo Garduño-Hernández. The latter two authors gathered and cleaned the data. This work is publicly available at [136]. The main bulk of this work was made by Fabián and me. Fabián and me conceptualized this work in equal proportions. Fabián came up with the derivation of the configuration model expected overlap distribution, and he suggested several ideas on how to quantify polarization and echo chambers. While we both helped write the manuscript, I wrote the majority of it. I came up with the idea of the *chamber*, and wrote all of the code. I also made the literature review and connected our quantitative results to the sociological theories of confirmation bias and homophily. I came up with the measure of polarization that we use on the paper.

- **Chapter 6: Estimating initial conditions from incomplete information**: The main bulk of this work is mine. Juan Sabuco and Doyne Farmer helped conceptualizing and supervising this project. They also helped me correct the first versions of the manuscript. The work is now published at [137].

- **Chapter 7: Latent state estimation of models with network topologies**: The main bulk of this work is mine. Doyne Farmer helped conceptualizing and supervising this project.

Enjoy!

Part I

Twitter as a social complex system

Chapter 2

Overview & preliminaries

2.1 Overview

Social media provides an essential platform for shaping and sharing opinions and consuming information in a decentralized way. Alongside social media platforms such as Facebook, Instagram and TikTok, Twitter has become one of the most popular social networking sites on the internet, with over 237 million daily active users [18, 17]. While some people use Twitter simply to stay in touch with friends and family, others use it as a platform to share news and information. In recent years, Twitter has become an important tool for studying social systems, from political elections [131], to the evolution of language online [206], to collective action and social movements [79].

Twitter users share posts that spread through their social network of followers and followees. Moreover, users interact by either reading, favoring, quoting, replying, or retweeting to each other's tweets. Thus, on top of the followers social networks, we can identify interaction networks characerized by the type and intensity of the interactions. These networks are coupled and co-evolving, which allows for analyzing real-time information sharing on a massive scale. This makes Twitter an ideal candidate for performing data-driven analyses of human behavior as a social complex system [15].

Many authors have analyzed Twitter a social-complex system to study human behavior for a wide range of topics and approaches. These approaches include analyzing the properties of social interaction Twitter networks [191, 23], linguistic-content dynamics [97], sentiment dynamics [35], the creation and evolution of topics in climate change conversations [68], or measuring echo chambers [43, 60] and polarization [200] in conversations such as climate change [58] or Covid-19 [126]. Other studies have

investigated how Twitter non-human accounts such as bots affect the users' behavior. For instance, Stella et al. [207] have investigated the effect that automated bots have in polarized social networks in the Catalunian (anti)independentist movements. Moreover, Ferrara et al. [89] have examined the effect that bots have in promoting misinformation and conspiracy theories in health emergencies. On a similar note, Beguerisse et al. [27] studied Twitter to promote health communication about diabetes by clustering temporal retweet networks, uncovering popular topics surrounding diabetes, and analyzing the users for different topical communities. These studies have found that Twitter is a valuable platform for studying collective behavior, and that users exhibit a variety of patterns of collective action.

However, new events can cause human behavior and opinions to change. For example, this is seen with Covid-19 and climate change. Covid-19 started as a sudden exogenous shock that has changed the way we live drastically. Therefore, it is important to track the public reaction and opinion around Covid-19. Authors have developed massive Covid-19 Twitter datasets [21], which have been explored to analyze the (mis)information spreading using epidemiological models [94] or uncover political narratives using dynamic topic modeling [197]. On the other hand, climate change is accelerating rapidly, and we are increasingly seeing its consequences. Yet, even when the scientific community has reached a consensus on the existence of anthropogenic climate change and its threats for the world [11], a huge portion of society has not [157]. Climate change conversations on Twitter have been studied for a variety of reasons, such as understanding information sharing behavior in the light of scientific evidence [223], quantifying the sentiment of different communities towards climate topics [63], or exploring the echo chambers and open forums of the conversation [232]. For these reasons, this part of the thesis aims to develop data-driven methods on massive Twitter data to advance our understanding of the polarization mechanisms in the climate change conversation and the perception of risk and emotion during Covid-19.

We organize the remainder of this part as follows. In Section 2.2, we review the main concepts used throughout this part. Section 2.2.1 introduces *networks* as mathematical objects and describes their primary uses. Section 2.2.2 introduces the field of *natural language processing*, its main computational techniques, and its application areas. In Chapter 3, we study the public risk perception and emotion of the Covid-19 Twitter conversation during the first wave of the pandemic. We analyze the linguistic content of tweets using natural language processing and semantic networks

to understand the dynamics of emotion and compare it with Covid-19 epidemiological data in several countries. In Chapter 4, we introduce unsupervised methods to study the polarization and echo chamber structures of the climate change conversation during 2019, when important climate-related movements flourished. We measure the ideological overlap between the leading users of the conversation and measure the polarization dynamics and the effect that exogenous events have on polarization.

2.2 Preliminaries

In the following sections, we introduce the main concepts from natural language processing and network theory, which will help us develop methods to analyze Twitter as a social complex system throughout this part of the thesis.

2.2.1 Networks

A *network* $\mathcal{G} = (V, E)$ is a tuple consisting of a set $V = (1, \ldots, N)$ of $N = |V|$ nodes (or agents) and a set of $E \subseteq V \times V$ of $M = |E|$ links (or interactions) that connect pairs of nodes. We say that a network is *complete* if $E = V \times V$, so that every node is adjacent to every other node. For every link $(i, j) \in E$, we assign a weight $w_{ij} > 0$ that accounts for the strength of the interaction. We say that a network is *undirected* if for every link $(i, j) \in E$, there exists $(j, i) \in E$ and $w_{ij} = w_{ji}$ and is *directed* otherwise. Moreover, we say that a network is *unweighted* if all the weights are one, i.e., if $w_{ij} = 1$ for all $(i, j) \in E$. Otherwise, the network is *weighted*. The *degree* of a node $i \in V$ is the combined weight of all the links connected to it. For directed networks, we distinguish between the *out-degree* and the *in-degree* as $k_i^{out} = \sum_j w_{ij}$ and $k_i^{in} = \sum_j w_{ji}$, respectively. For undirected networks, the out-degree and in-degree of node i coincide, so we simply call k_i the degree of i. If the network is unweighted, the degree counts the number of links that any node is attached to. We could interpret unweighted and/or undirected networks as special cases of general networks. However, a lot of early research on networks considered only these special cases providing many results that, although they can be generalized for any network sometimes, sometimes they cannot. Networks that are unweighted *and* undirected are known as *simple* networks.

We can also model *dynamic processes* on networks. One way to do this is by using walks and paths. A *walk* in a network $\mathcal{G}$ is a succession of adjacent nodes. A *path* is a walk where no node is repeated on the succession. An exception to this are *closed paths*, where the first node of a path is the same as the last one, and no

other node is repeated. For a complete network of nodes $V = (1, 2, 3)$ the successions $(2, 1, 2, 3, 1)$, $(1, 3, 2)$, and $(1, 3, 1)$ are examples of a walk, a path, and a closed path, respectively. The most basic dynamic process on a network is the *random walker* in which, starting from an initial node i, the walker visits its adjacent node j with probability w_{ij}/k_i^{out} and repeats the process but now starting from j. Paths and walks are also useful to determine important structural properties of networks. Examples range from determining if a network is *connected*, i.e., whenever there exists a path between any two nodes in the network, *ergodic*, i.e., whenever the distribution of random walkers becomes stationary for long walks, or *modular*, i.e., whenever walkers stay in local regions of the network for long periods of time.

Unlike ergodicty or connectedness, the notion of a modular network is not clearly defined. Roughly speaking, a network is *modular* - or is said to have a *community* structure - if we can *partition* its node set V into subsets that are densely connected, i.e., if the link weights within the subsets are significantly higher than between them. Identifying communities on networks is an active area of research, and its main algorithms range from considering the network structure explicitly [37], to analyzing the properties of dynamic processes (such as random walkers) [186], to generating random networks with stochastic block models [172].

Another active area of research is that of *random graph theory*, where a network is *sampled at random* from the ensemble of possible networks having a certain set of constraints. The simplest random network model is the *Erdos-Renyi model*, $\mathcal{G}_{ER}(N, p)$, where a network of N nodes is constructed by including each of the possible links with probability p, independently from every other link. the expected *mean degree* of an Erdos-Renyi realization is $\langle k \rangle = p(N - 1)$. Given an empirical unweighted network $\mathcal{G} = (V, E)$, its mean degree is $\langle k \rangle = |E|/|V|$, so realizations of the Erdos-Renyi model, $\mathcal{G}_{ER}(|V|, |E|/(|V|^2 - |V|)$, have the same expected mean degree than the empirical network. If the empirical network structure is similar to those sampled from the Erdos-Renyi model, one could conclude that the edges in the empirical network *were formed* independently of each other, with a similar probability for any pair of nodes in the network. However, most empirical networks are very different from a typical Erdos-Renyi realization, so better random network models have been developed in the literature.

Another simple yet useful random network model is the *configuration model*, $\mathcal{G}_c(N, \mathbf{k})$, which is the set of simple networks with N nodes and degree sequence $\boldsymbol{k} = (k_1, \ldots, k_N)$ [161]. Sampling a network from the configuration model consists

of first creating k_i *stubs*, or half-edges, for each node $i \in (1, \ldots, N)$, and then connecting each possible pair of stubs at random excluding stubs that go from a node to itself. A node i is connected to j with probability $k_i/(2|E|)$, where $2|E|$ is the total number of stubs in the network. The configuration model is often used as a *null model* to explain processes on empirical networks, because it considers the effect of heterogeneous degree sequences without any other constraints in the structure of the network.

For a better and deeper review on networks, we recommend Mark Newman's book [167] and Renaud Lambiotte's lecture notes for the Oxford's Networks course [139] and his book with Michael Schaub [140].

2.2.2 Natural language processing

Natural language processing (NLP) is the discipline of extracting information from the natural language, either written or spoken, using computational methods [183]. Its main goal is generating computer code that interprets the contents of a corpus sensibly. A *corpus* is a collection of documents, and a *document* is a set of structured text or audio signals of natural language.

Natural language is, roughly speaking, a way of communicating ideas by word associations functioning under specific sets of rules. A *word* is a minimal symbolic representation that, in isolation, carries objective or practical meaning. Some words can be joined together into *compound words*, which stem from those of the individual words. Sequences of words form *phrases* or *sentences* that expand the meaning of the words in isolation under the set of rules of the language. Sequences of phrases and sentences form documents. The meanings that emerge from word associations can be quite complex: they depend on the grammatical rules of the language, the cultural and historical context of the entities creating the word associations, and the context of words within their position and function in word sequences.

Researchers use NLP to extract meaning from word associations for various tasks using different approaches. Often, the approach depends on the task. For instance, a common task in NLP is *transcribing* spoken language from audio into text, where machine learning methods dominate [162]. Another common task is *assigning intention*, such as sentiment, to documents in a corpus or sentences in a document. A related but different task is *classifying* written text into linguistic categories, or topics, based on the word associations inside a corpus. The methods for approaching the last two tasks can be coarsely divided into knowledge-based and statistical methods [49].

Knowledge-based methods are supervised methods. They incorporate our understanding of words and n-grams[1] inside a vocabulary into an explicit set of rules that dictate how to classify or assign intention to those words. For instance, VADER (Valence Aware Dictionary for sEntiment Reasoning) is a ruled-based model for assigning sentiment focused on social media microblogs [122]. It rates the sentiment of a given document by comparing its word content to a human-curated list of lexical features sensitive to the polarity and intensity of the given words. These features are combined with five heuristic rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. Such rules are the use of exclamation mark(s), capitalization of words, degree modifiers such as adverbs, the use of the contrasting conjunction "but" as a shift of sentiment polarity, and negation flips (e.g., "not good"). Another example is the Linguistic Inquiry and Word Count (LIWC) algorithm [174]. LIWC is a text analysis program that reports the number of words in a document belonging to a set of predefined linguistically and psychologically meaningful categories and *classifies* the documents based on the proportion of words in each category.

Statistical methods can be supervised and unsupervised. They rely on the frequency distributions of words and word co-occurrences inside a document to assign importance and meaning to the words. They are often trained with machine learning models using large corpora. For instance, one successful approach is *embedding* the words in a corpus into a low-dimensional vector space using deep neural networks [239]. This approach is appealing because words close in the vector space[2] have similar semantic meanings. Words in this space can be added together, giving semantically meaningful results.

These approaches are relevant for analyzing social complex systems because they provide a way to quantify natural language. Thus, we can use the output of the NLP task and assign that output as the state of the entities forming the social system. For instance, [69] use a sentiment analysis algorithm designed for Twitter to assign a sentiment score between -1 and 1 to individual tweets. They construct an opinion dynamics model and validate it on the time series generated by those scores.

[1] An n-gram is a sequence of n words joined together as a single concept. For example, the sequence "climate change" is a 2-gram with a precise meaning that the words alone on their own do not capture.

[2] The low-dimensional vector space is Euclidean, so one can define a distance on the space and evaluate if two words are close or not

Chapter 3

Public risk perception and emotion during the Covid-19 pandemic

Disclaimer

This chapter is heavily motivated by the work I co-authored with my colleague Joel Dyer and is published in [78]. See Section 1.1 for details on the division of labor for this work.

3.1 Introduction

The Covid-19 pandemic has brought about widespread disruption to human life. In many countries, public gatherings have been broadly forbidden, mass restrictions on human movement have been introduced, and entire industries have been paralyzed in attempting to lower the peak stress on healthcare systems [109]. However, the degree to which these restrictions have been enforced by law has varied over time and by location, and their success in mitigating public health risks depends on the extent of cooperation on the part of the public.

A key determinant of the public's behavior and their cooperation with state-imposed social restrictions is the public's emotional response to, and their perception of the the risk presented by, the pandemic. However, the evolution of emotions and risk perception in response to disasters is not well-understood, and there is a need for more longitudinal data on such responses with which this understanding can be improved [48]. Our goal is thus to contribute to bettering this understanding, and we do so by exploring the empirical relationships present between the progression of the Covid-19 pandemic and the public's perception of the risk posed by the pandemic.

We explain the finding of this chapter in terms of the existing body of literature surrounding public perception of risk, disasters, and human suffering in cognitive psychology. In particular, we draw from psychophysics, the field that studies the relationship between stimulus and subjective sensation and perception [98]. The search for psychophysical "laws" of perception has existed since at least the mid-19th Century with the proposing of the Weber-Fechner law [88], which posits that the smallest perceptible change $\mathrm{d}s$ in a physical stimulus of magnitude s is proportional to s. Thus, the perceived magnitude p of such stimuli follows

$$\mathrm{d}p \propto \frac{\mathrm{d}s}{s}. \tag{3.1}$$

In the continuum limit, this implies that p grows logarithmically with the physical magnitude s of the stimulus. More recently, empirical studies by S. S. Stevens [209] supported, instead, a power law relationship between human perception of a stimulus and the physical magnitude of the stimulus:

$$p \propto s^{\beta}. \tag{3.2}$$

Summers *et al.* [210] extended this concept to human sensitivity to war death statistics and found that a power law with exponent $\beta = 0.32$ best fit the data. Note, however, that these psychophysics models are sensitive to how we measure the stimulus and its perceived signal in the first place. On the one hand, Different instruments may exhibit different impulse responses, making the exponent β, or even the functional form $p(s)$ different instrument dependent. Thus, the relevant signal, rather than the exponent per se, is the qualitative relationship between stimulus and perception, which in all the past experiments has been sub-linear, indicating a numbing effect. On the other hand, assessing what corresponds to an stimulus is not straightforwards when generalizing these models. Authors should be careful on asking what on the stiumulus is changing for it to generate a change in perception.

A number of further studies have corroborated the extension of these psychophysical laws describing the subjective perception of physical magnitudes to the subjective evaluations of human fatalities [203, 90, 92]. In all of these, perception is a concave function of the stimulus, meaning that the larger the stimulus magnitude, the more it has to change in absolute terms to be equally noticeable. Thus, perception is considered relative rather than absolute, implying that our judgments are comparative in nature. This observation has been shown to account for deviations from rationality in economic decision-making [229].

These proposed psychophysical laws of human perception present an opportunity for monitoring a population's response to a disaster scenario such as the Covid-19 pandemic. By evaluating the goodness of fit of these models to data on the perception of the progression of the pandemic, and determining the parameter values of such fits, we can describe the sensitivity of populations to the state of such crises, with important implications for risk communication and disaster management.

To this end, we make use of a massive Twitter dataset consisting of user-posted textual data to study the public's emotional and perceptual responses to the current public health crisis. Twitter provides convenient access to the conversation amongst members of the public across the globe on a plethora of topics, and many authors are studying several aspects of the public's response to the pandemic with it. Twitter is a particularly appropriate tool under conditions of physical distancing requirements and furlough schemes, where online communication has become more than ever a central feature of everyday life. Moreover, results from psycholinguistics and advances in natural language processing techniques enable the extraction of psychologically meaningful attributes and the reconstruction of cognitive structures (e.g. semantic networks) from textual data. With this dataset, our general approach is to offer a quantitative, spatiotemporal comparison between indicators of the state of the pandemic and the topics and psychologically meaningful linguistic features present in the discussion surrounding Covid-19 on social media on a country-by-country basis, for a selection of countries.

3.1.1 Related work

To our knowledge, this work is the first to use a large social media dataset spanning multiple countries to model the perceptual response of countries' citizens to the pandemic in the context of risk perception. To date, empirical validation of the aforementioned psychophysical laws has largely taken place in controlled laboratory settings, in which decisions, actions, and scenarios are artificial or hypothetical. Our work thus contributes to the body of literature surrounding risk perception by investigating these laws in a naturalistic setting.

However, there have been numerous authors using social media to analyze the public response to the Covid-19 pandemic. This includes work that has focused on the psychological burden of the social restrictions. For instance, Stella et al. [205] use the circumplex model of affect [179] and the NRC lexicon [160] to give a descriptive analysis of the public mood in Italy from a Twitter dataset collected during the week following the introduction of lockdown measures. In addition, Venigalla *et al* [204]

has developed a web portal for categorizing tweets by emotion in order to track mood in India on a daily basis.

Others have instead focused on negative emotions, as in the work of Schild et al. [194], where they study the rise of hate speech and sinophobia as a result of the outbreaks. More specifically on perception, Dryhust et al. [76] measured the perceived risk of the Covid-19 pandemic by conducting surveys at a global scale ($n \sim 6000$) and compared countries, finding that factors such as individualistic and pro-social values and trust in government and science were significant predictors of risk perception. de Bruin and Bennett [70] perform similar work in the United States. The closest work we have been able to find to our own are those of Barrios and Hochberg [25] and Aiello et al. [13], where both research pieces focus on the current pandemic using data from the United States. In the former, they combine internet search data with daily travel data to show that regions in the United States with a greater proportion of Trump voters exhibit behaviors that are consistent with a lower perceived risk during the Covid-19 pandemic. In the latter, they assess the epidemic psychology using Covid-19 Twitter data in the United States according to several linguistic features present in the tweets. They identify three psychological phases consistent with the refusal-suspended reality-acceptance stages of grief. Despite the above, we have been unable to find work that combines large-scale social media data with linguistic analysis to offer a spatiotemporal, quantitative analysis of emotion and risk perception during the Covid-19 pandemic across multiple countries.

Beyond the Covid-19 pandemic, our work is related to a small but growing body of literature on the use of data science in understanding human emotion and risk perception. In such work, natural language analysis has succeeded in supporting established linguistic theories such as the importance of the distribution of words in a vocabulary as a proxy for knowledge [110], and regarding the relation between the uncertainty of events and the emotional response to their outcome [87, 224]. For instance, using textual data from Twitter, Bhatia found that unexpected events elicit higher affective responses than those which are expected [35]. In another instance, the same author conducted experiments with 300 participants and predicted the perceived risk of several risk sources using a vector-space representation of natural language, concluding that the word distribution of language successfully captures human perception of risk [34]. Similar work has been conducted by Jaidka *et al.* [123] in the area of monitoring public well-being, in which they compare word-based and data-driven methods for predicting ground-truth survey results for subjective well-being of US citizens on a county-level basis using a 1.5 billion Tweet dataset constructed from 2009 to 2015.

The remainder of this chapter is laid out as follows. In Section 3.2, we present the data set used in the subsequent analysis. In Section 3.3, we provide further details on the approach followed to explore the relationships between indicators of the state of the pandemic and the public's perception of the pandemic, and discuss possible explanations for our observations by drawing on psychological literature. In Section 3.4, we summarize and offer concluding remarks, along with a discussion of the limitations of the current work and suggestions for avenues of future work.

3.2 Data

3.2.1 Twitter dataset

In the following analysis, we make use of the set of tweets gathered by J. Banda et. al [21], which are obtained and maintained using the Twitter free Stream API[1]. At the time of writing, this data set consists of $\sim$ 80 million *original* tweets spanning from March 11, 2020 to June 14, 2020. By original we mean that we do not consider retweets, which is standard for natural language processing [102, 21]. Data is collected according to the following query filters[2]: "COVID19", "Coronavirus-Pandemic", "COVID-19", "2019nCoV", "CoronaOutbreak", "coronavirus", "Wuhan-Virus", "covid19", "coronaviruspandemic", "covid-19", "2019ncov", "coronaoutbreak", "wuhanvirus".

For our analysis, we consider only the English and Spanish tweets with a non-empty self-reported location field. We process every self-reported location using OpenStreetMaps [169] and remove non-sensical locations (e.g. "Mars", "Everywhere", "Planet Earth"). This allows us to group the remaining tweets by country and proceed with our analysis on a country-by-country basis. To assure the statistical significance of our analysis, we keep the countries with the highest number of tweets for each language, resulting in a geolocated Twitter dataset of $\sim$ 20 million original tweets posted by $\sim$ 4 million users on 12 different countries, which we summarize in Table 3.1.

[1] The free Stream API randomly samples around 1% of the total tweets for the given queries

[2] A number of publicly available Twitter datasets have emerged in relation to the pandemic. We chose to work with this dataset since it used the most generic query terms among all the publicly available datasets we considered, and we wanted the least amount of bias possible for our analysis.

	Language	Number of tweets	Unique users
Argentina	Spanish	846,706	194,818
Australia	English	701,072	97,027
Canada	English	1,209,712	195,507
Chile	Spanish	342,013	60,235
Colombia	Spanish	466,477	103,845
India	English	1,806,685	344,894
Mexico	Spanish	1,133,350	187,064
Nigeria	English	754,152	133,797
South Africa	English	354,613	78,447
Spain	Spanish	1,697,049	274,010
United Kingdom	English	3,490,703	631,017
United States	English & Spanish	6,297,720	1,397,410
Total		19,072,850	3,699,071

Table 3.1: Per-country summary of the Twitter dataset constructed from the repository maintained by Banda et al. [21]. All tweets are original, i.e. all retweets are removed.

3.2.2 Epidemiological data

We measure the progression of the pandemic with the number of Covid-19 confirmed cases and deaths for all the countries in our analysis. The data was made publicly available by Our World in Data repository [154]. In particular, we take the daily Covid-19 cases and deaths, both in linear and logarithmic scale, since these are four epidemiological indicators that are most frequently used to summarize the state of the pandemic, and are therefore frequently encountered by the public.

We scrapped the epidemiological data in near real-time, meaning that the data was not subject to any post-processing - including extra data cleansing or data corrections. This is important because our object of study is the perception of the stimulus provided by the Covid-19 epidemiological data. Therefore, any signal provided by the data - accurate or not - will be part of the stimulus perceived by the per-country populations.

3.3 Analyzing the public's perception of the pandemic

In this section, we study the public's perception of the pandemic on a country-by-country basis, using the countries with the highest number of tweets in the observation

period (see Table 3.1). We do this on a country-by-country basis since the pandemic has often evoked nation-level responses, making nation-level analysis the most natural geographic scale. Our broad approach is to inspect and compare the linguistic features of the tweets released by users in the Twitter dataset described in Section 3.2.1 with the epidemiological data described in Section 3.2.2.

3.3.1 Defining perception from linguistic inquiry

Our goal is to explore the public's perception of the pandemic. To do this, we analyze the linguistic features present in the textual data generated by Twitter users, and map these features to psychologically meaningful categories that are indicative of the Twitter users' perception. Here, we are assuming that the words used by these Twitter users are indicative of their internal cognitive and emotional states [213], which is supported in [34] where they predict the perception of risk using text data. Thus, we quantify the linguistic content of each tweet using the Linguistic Inquiry and Word Count (LIWC) program [174]. LIWC has been widely adopted in several text data analyzes, and it has proven successful in applications ranging from measuring the perception of emotions [237] to predicting the German federal elections using Twitter [220]. Moreover, it has recently been used to successfully identify the early-epidemic psychological stages of grief in the current pandemic [13].

LIWC operates as text analysis program that reports the number of words in a document belonging to a set of predefined linguistically and psychologically meaningful categories[3] [213]. For our purposes, a document is a tweet d_i^t posted on date t and from a user based in country i. LIWC represents documents as an unordered set of words, and a LIWC category l is similarly a set of words associated with concept l. For a given document d_i^t, the *linguistic score* p^l for category l is the percentage of words in d_i^t that belong to l:

$$p^l(d_i^t) = \frac{|d_i^t \cap l|}{|d_i^t|} \cdot 100. \tag{3.3}$$

There are many such categories l, including Family, Work, and Motion. We capitalize such category titles, and use the titles to refer to either the set of words associated with that category or to refer to the category itself. Linguistic scores from Eq. (3.3) for individual tweets will be noisy, as they are short documents. Moreover, we are interested in the average response of the population of a country. For this

[3]For the English-language tweets, we make use of the 2015 English dictionary. For the Spanish-language tweets, the most recent dictionary is the 2007 edition, which has fewer categories than the 2015 English dictionary.

reason, we group the tweets by country i and by date t, and denote these sets of tweets as $D_i^t = \{\ d_{i'}^{t'} \mid i' = i,\ t' = t\ \}$. We then compute the *National Linguistic Score* (NLS) for category l as the average of the linguistic scores over documents in D_i^t relative to an empirically observed Twitter base rate p_B^l:

$$p_i^l(t) = \frac{100}{|D_i^t|} \sum_{d \in D_i^t} \frac{p^l(d) - p_B^l}{p_B^l}. \tag{3.4}$$

The base rates p_B^l for the use of words on Twitter associated with category l are given in [174]. Using Eq. (3.4) for all the selected linguistic categories, we construct multidimensional country-level time series that represent the evolution of the public perception of the pandemic, similar to the linguistic profiles introduced by Tumasjan *et al.* [220]. These perception dynamics are influenced by each user in our dataset, which may include bots and institutional or public relations accounts. We discuss the possible implications of this aspect of our data in Section 3.4.

In Figure 3.1, we show the collection of NLSs for a selection of relevant linguistic categories. We observe clear trends that, in most cases, are synchronized between countries and languages. In particular, most categories associated with emotion - notably Affect, Anger, Anxiety, Positive emotion, Negative emotion, and Swear words (swearing is associated with frustration and anger [125]) - have their highest scores in mid-to-late March, when the World Health Organisation (WHO) announced the pandemic status of Covid-19 and most Western countries introduced more stringent social restrictions [109]. These scores decay thereafter, indicating a relaxation of the emotional response in the conversation. This is consistent with results reported by Bhatia regarding the affective response to unexpected events [35] and with those of Aiello et al. [13] where the Death NLS of the United States rises from late March on. A qualitatively similar trend can be seen in the Social processes panel, the category involving "all non-first-person-singular personal pronouns as well as verbs that suggest human interaction (talking, sharing)" [174].

We also observe that health-related categories such as Death and Health show an overall rising trend, with Death rising most rapidly throughout March. These categories, with the exception of Positive Emotion and Health, peak again in the United States at the end of May, coinciding with the murder of George Floyd and the subsequent Black Lives Matter protests. Such universal trends are not apparent by visual inspection in the Money, Risk, and Sadness panels. An additional feature of these plots is the absolute scale of these values: in all cases, there is a significant percentage change from their baseline values, with large percentage increases observed

initially in the use of words associated with Anxiety and later with Death, and a moderate percentage increase in the use of words associated with Risk.

3.3.2 Comparing the public's perception with epidemiological data

In this section, we explore the relationship between the NLSs described in Section 3.3.1, which we use as a proxy for the public's perception, and the intensity of the pandemic, which we assume is the stimulus triggering this perception. Our measure of the intensity of the pandemic is the number of Covid-19 cases and deaths from the data described in Section 3.2.2.

A straightforward way of approaching this relationship is by computing the correlations between the NLSs and the epidemiological data in a per-country basis, and we show the average across countries of these per-country correlations in Figure 3.2. On the one hand, we observe significant negative correlations in emotionally charged categories (eg. Swear words, Anger, Anxiety, Affective processes), indicating a decay in emotion as the pandemic intensifies. Conversely, categories related with health and mortality (Death, Health) and analytical thinking (Analytic) show significant positive correlation[4].

3.3.2.1 Psychophysical numbing

We believe the trends we observe in Fig. 3.1 and the correlations we observe in Fig. 3.2 are consistent with the notion of **psychophysical numbing**. This term was introduced by Robert Jay Lifton [146], and developed by Paul Slovic [203, 90] in the context of human perception of genocides and their associated death tolls, to describe the paradoxical phenomenon in which people exhibit growing indifference towards human suffering as the number of humans suffering increases. By inspecting the correlations between the NLSs and the epidemiological indicators, we find that as the pandemic intensifies - in the sense of an increasing number of cases and deaths reported daily - our emotional response diminishes, as expected from a psychophysical numbing phenomenon.

Specifically, we observe negative correlations between almost all components of the NLSs associated with affect - Affective processes, Anger, Anxiety, Negative emotion,

[4]When analyzing these correlations, we found that, overall, the cumulative cases and deaths correlate better with most linguistic categories than the daily data. However, while this is sensible in the early stages of the pandemic, it is unlikely to remain the case over a long time horizon due to humans' finite memory. We therefore proceeded with our comparison using the daily epidemiological data alone for this reason.

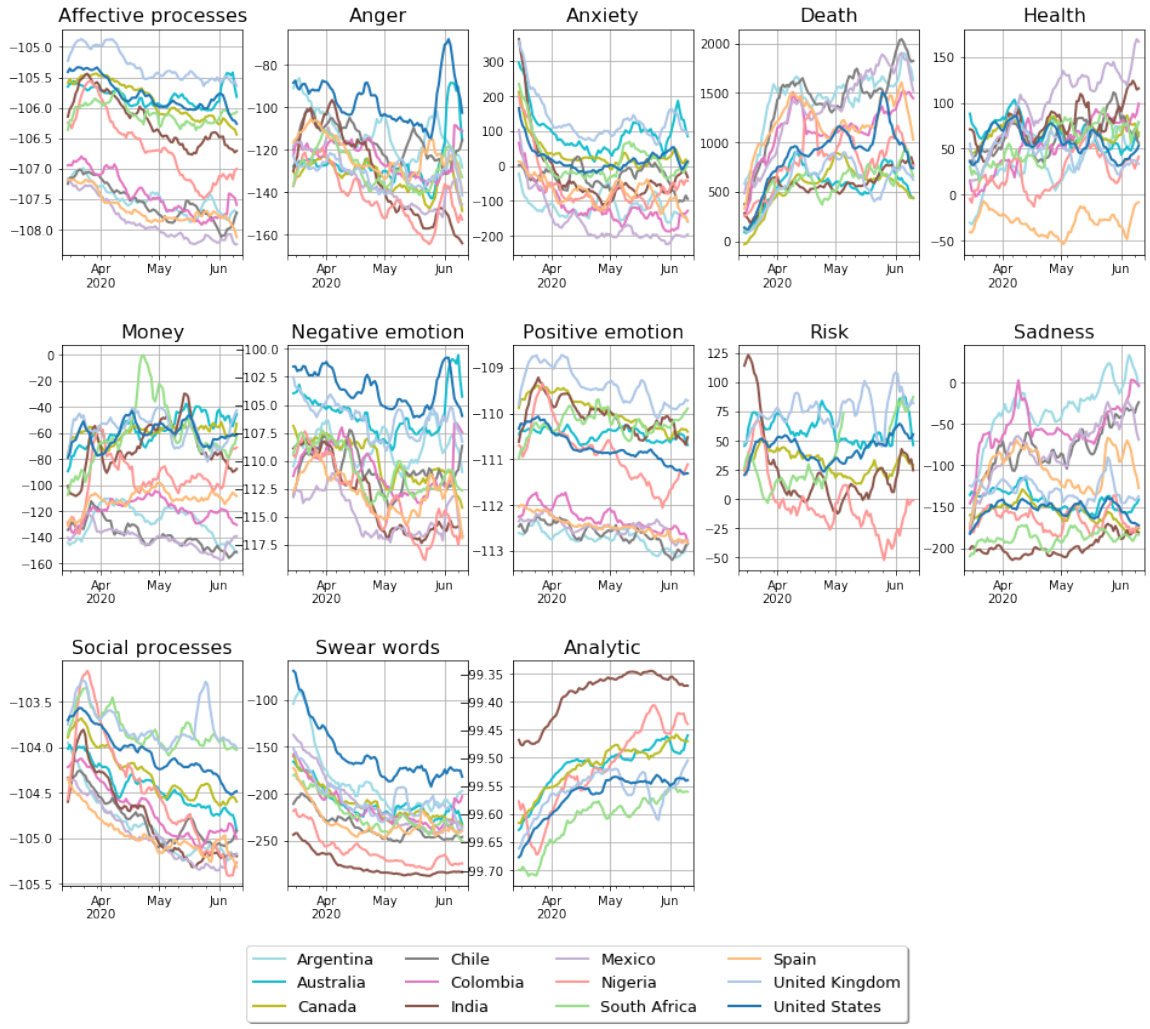


Figure 3.1: Time series for the NLSs for the countries as indicated by the legend. Each panel shows the individual linguistic categories. The units on the y-axis represent the percentage change of the National Linguistic Scores (NLS) on our data with respect to the LIWC baselines for Twitter (see Eq. (3.4)).

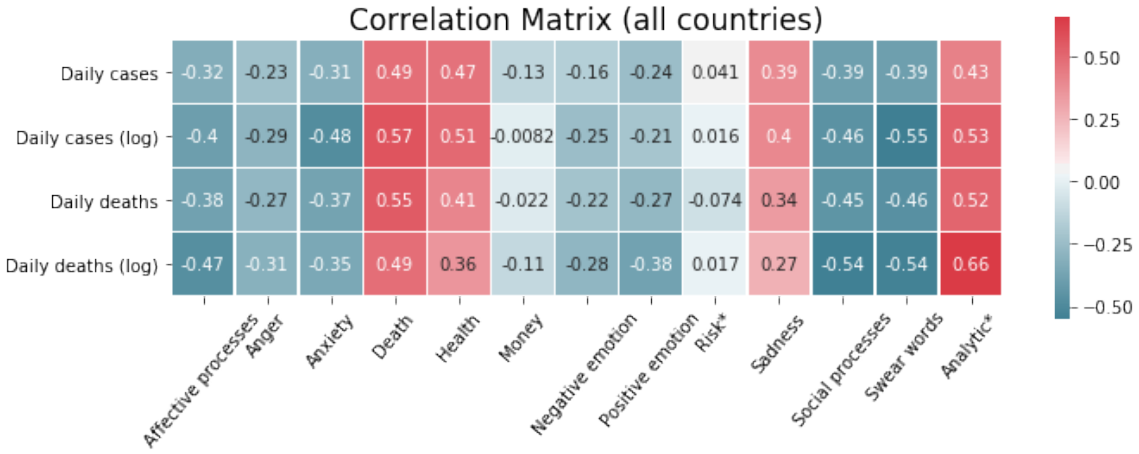


Figure 3.2: Correlation coefficients between epidemiological indicators and national linguistic scores (NLSs) averaged across all countries. *"Risk" and "Analytic" are only available for the English-language LIWC. These two categories are thus averages across English-language countries only.

Positive emotion, and Swear words - and the epidemiological data[5]. By inspecting Figure 3.1, we see that every country exhibits similar downward trends in these components and, with the exception of Anxiety, are all significantly lower than their baseline values throughout the observation period.

This unusually low and decreasing Affect word count is accompanied, conversely, with a growing awareness of the morbidity of the situation in that we observe significant positive correlations between the Death NLSs and the daily national cases and deaths, indicating that the decrease in affect occurs simultaneously with and *despite* an attentional shift towards Covid-19 related mortality. We also observe a simultaneous increase in the Analytic component of each English-language dataset[6] over this same period, indicating a movement towards more logical and analytical, rather than intuitive and emotional, thinking.

The potential implication of this is that the public is less perceptive of the risk that the pandemic poses to public health, since their emotional response is reduced and reducing [192]. For example, Van Bavel *et al.* [222] and Loewenstein *et al.* [148] describe that risk perception is driven more by association and affect-based processes than analytic and reason-based processes, with the affect-based processes

[5]The only exception is the cross-country average of the Sadness component of the NLSs, which is positively correlated with the epidemiological indicators and appears to be driven only from Argentina's, Chile's, and Colombia's increasing use of words related to Sadness. The remaining countries remain stationary at a lower-than-baseline value for this component.

[6]Unfortunately, the Spanish LIWC dictionary does not yet have an Analytic category.

typically prevailing when there is disagreement between the two modes of thinking. The negative correlations between the intensity of the pandemic and affective processes, together with its positive correlation with the prevalence of analytic processes, suggests that public risk communication could be adjusted to re-balance the degree of affective and analytic thinking amongst members of the public to achieve favorable risk avoidance behavior and, consequently, favorable public health outcomes.

3.3.3 Analyzing the emotional framing of Covid-19 casualties with semantic networks

To support our claim that these observations are attributable to psychophysical numbing, we construct word co-occurrence networks using tweets in our dataset. Word co-occurrence networks are a class of linguistic networks, in which nodes are words appearing in a body of text and an edge is placed between a pair of words with a weight given by some function of the number of co-occurrences of that pair in the text. Empirical word co-occurrence networks have been used in cognitive network science as approximate reconstructions of the author's latent cognitive structures, e.g. semantic or conceptual networks [199], with a given corpus deemed to be an empirical manifestation of such structures.

For example, Kenett *et al.* [133] reconstruct participants' internal semantic networks on the basis of their responses in a free word association task, reporting that participants that were found independently to have lower creativity scores also had less well-connected semantic networks - specifically, a higher modularity, average shortest path length, and diameter, and a lower small-world-ness [121] - than participants scoring more highly in creativity.

In the context of the Covid-19 pandemic, Stella *et al.* [205] use word and hashtag co-occurrence networks in conjunction with word-to-emotion mappings to uncover complex emotional profiles amongst Twitter users posting from Italy during the first week of lockdown. More generally, a plethora of models for inferring semantic relationships between words in natural language processing tasks are based on some notion of word co-occurrence [128]. The semantic proximity of a pair of words in such models has also been shown to possess predictive power regarding the subjective probability participants assign to hypothetical real-world events involving that pair of concepts [33]. For a more complete review of the use of linguistic networks in the study of human cognition, we refer the interested reader to [199].

Given the well-established utility of word co-occurrence analysis in providing a view of authors' internal cognitive structures, we employ such an approach on $\mathcal{W} =$

Death ∪ Affect - the set of words in either the Death or Affect categories - in an attempt to approximate the Twitter users' internal semantic relationships between these two concepts[7]. Specifically, we hypothesize that, if the psychophysical numbing effect is legitimate, the modular structure of these networks will separate Death-related and Affect-related words more decidedly at larger daily death counts than at lower death counts. This would indicate that conversation regarding Covid-19-related mortality evokes a weaker emotional response at higher daily death counts.

Given a set $\mathcal{T}$ of tweets, the word co-occurrence network $G(\mathcal{T})$ is represented by a weighted adjancency matrix $A(\mathcal{T})$ in which the nodes are words belonging to the Death and Affect LIWC dictionaries. Entry $A_{ij}(\mathcal{T})$ counts the number of co-occurrences between words i and j across all tweets in $\mathcal{T}$, and is computed as

$$A_{ij}(\mathcal{T}) = \left(B(\mathcal{T})^T B(\mathcal{T})\right)_{ij}, \tag{3.5}$$

where $B_{tk}(\mathcal{T})$ counts the number of instances of word k in tweet $t \in \mathcal{T}$. We ignore self-edges by imposing $A_{ii} = 0$, since it is the relationship between distinct words that is of interest. (See Appendix A.3.1 for further details on the construction of these networks.)

3.3.3.1 Qualitative overview of the Death-Affect partition

We identify three main periods for which we construct network snapshots of word co-occurrences (see Figures 3.3a to 3.3c). The first period spans **11th March to 9th April 2020**, in which the WHO declared Covid-19's pandemic status and governments generally imposed social restrictions. The second period spans **10th April to 23rd May**, during which most Covid-19 cases either underwent exponential growth or flattened out for some countries in Europe. The final period spans **24th May to 13th June**, during which most countries were at the peak daily rate of Covid-19 cases or where in a stage of decreasing number of daily cases. Moreover, the Black Lives Matter protests were triggered by the murder of George Floyd in the USA in this period. In constructing these networks, we weight each country equally by taking a random sample of approximately 300,000 tweets from each country.

In Figures 3.3a to 3.3c, we visualize these three snapshots for the English-language tweets. From these we observe that two clusters emerge in all cases: a left-hand cluster consisting mainly of Death-related words and a right-hand cluster consisting primarily

[7] The Affect category contains all the words related with affective processes. This includes the words in Anger, Anxiety, Positive and Negative emotion, and Swear words, which are all significantly correlated with Death and daily deaths.

Word	Figures	Interpretation
positive	3a to 3c	Used in reference to the number of people that have tested positive for Covid-19
isolat*	3a, 3b	Used in discussion surrounding symptomatic and at-risk individuals self-isolating
care	3b,3c	Used in relation to: the health care system; the death care industry; the admission of Covid-19 patients to intensive care units; and deaths occurring in care homes for the elderly
panic	3a	Panic-buying of household goods, e.g. toilet paper, hand-sanitiser
protests	3c	George Floyd's death and subsequent Black Lives Matter protests

Table 3.2: Words belonging to the Affect LIWC category that appear in the primarily Death-based clusters in the three snapshot word co-occurrence networks shown in Figures 3.3a to 3.3c. The middle and right columns indicate in which snapshots they are most prominent, and the likely explanation for their association with the concept of death during the pandemic.

of Affect-related words. We also observe that the relative sizes of these clusters vary over time: the Death-cluster grows in size as the pandemic progresses, and remains separated from the Affect-based cluster. This indicates that the evolving structure of these networks may be consistent with our hypothesis of psychophysical numbing: throughout, Covid-19 casualties appear not to evoke a strong emotional response.

However, we find that a number of the most highly connected nodes in these Death clusters are Affect-related words: in the first network, the Affect-related words "panic", "positive", and "isolat*" appear; in the second, the words "care", and "fail*" also appear; and in the third, "protests" appears[8]. While such words are normally associated with affective processes, we argue that some of these are more readily understood in terms of their association with Covid-19-specific topics that are less indicative of an affective experience in this context than they might be more generally. For example, "positive" is used very frequently in the context of the pandemic in relation to individuals "testing positive" for the virus. In Table 3.2, we address five of these words, providing what we believe are the most plausible explanations for their association with conversation surrounding mortality during the Covid-19 pandemic.

Altogether, this initial examination indicates that words associated with a subjective emotional/affective experience and words related to death may be well-separated in this Twitter data, which is consistent with the notion of psychophysical numbing

[8]Words ending with the wildcard "*" denote that any word starting with the suffix before the wildcard will be considered. For instance, the words "isolation" and "isolated" all count as valid instances of 'isolat*".

as an explanation for the trends and correlations observed in Figures 3.1 and 3.2. For completeness, we include the equivalent co-occurrence graphs for the Spanish-language tweets in Appendix A.3.2, about which similar statements can be made.

Our discussion has so far been qualitative given that the aforementioned network snapshots (i) vary considerably in size, (ii) represent the aggregate conversation of the tweets across countries in our dataset, and (iii) involve crude aggregation over large time periods. In the next section, we address these issues by investigating the change in a number of network measures over time and discuss the extent to which they support our hypothesis of psychophysical numbing.

3.3.3.2 Quantitative analysis of the Death-Affect partition

To further probe this hypothesis, we seek network measures that describe the strength of association between the concept of death and affective processes. Since the primary tenet of psychophysical numbing is that "the more who die, the less we care", our investigation is focused on the degree to which conversation around Covid-19 mortality evokes the use of affective language, which is our proxy for "degree of caring". In particular, we are interested in whether the emotional framing of such conversation changes as the daily death rates change in each country, where a less emotional conversation at higher daily death rates would support the hypothesis of psychic numbing.

For this purpose, we investigate the dynamics of the following network measures over a sequence of comparable snapshots for each country:

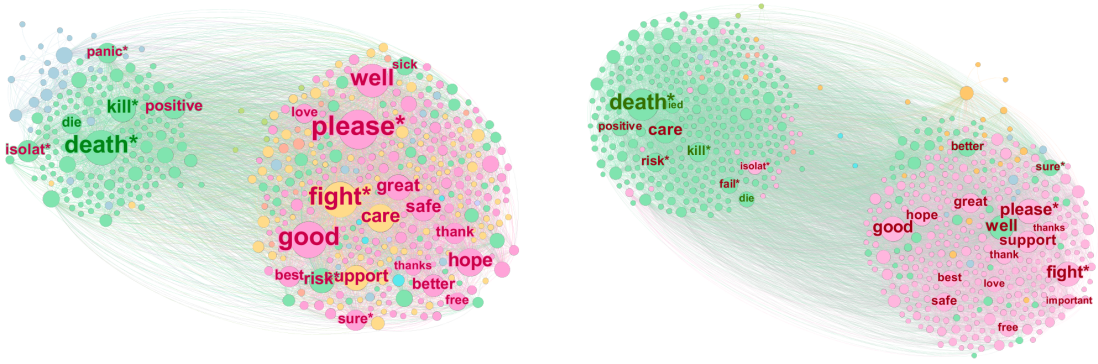
1. the **weighted modularity** for the partition $\mathcal{P}_{\text{LIWC}}$ induced by assigning nodes to their respective LIWC categories, i.e. Death or Affect. We define the weighted modularity following Newman [166] as

$$Q_{\text{LIWC}}(t) = \frac{1}{2m(t)} \sum_{ij} \left(A_{ij}(t) - \frac{k_i(t)k_j(t)}{2m(t)} \right) \delta(c_i, c_j), \tag{3.6}$$

   where $A_{ij}(t)$ is the weighted adjacency matrix of a network at snapshot t, $k_i(t) = \sum_j A_{ij}(t)$ is the strength of node/word i, $m(t) = \frac{1}{2}\sum_{ij} A_{ij}(t)$ is the total strength of the network, $c_i \in \{\text{Death, Affect}\}$ represents the community assigned of node i under partition $\mathcal{P}_{\text{LIWC}}$, and $\delta(\cdot,\cdot)$ is the Dirac delta function.

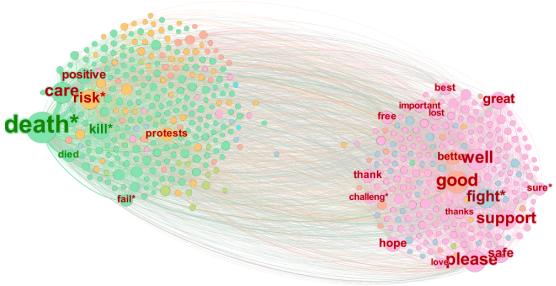
2. the **fraction of the total strength of node "death*" ("muert*")** that can be attributed to its connections with other nodes in the Death category:

$$f_{\text{Death}}(t) = \frac{1}{k_{\text{death}^*}(t)} \sum_j A_{\text{death}^*,j}(t)\, \delta(c_{\text{death}^*}, c_j). \tag{3.7}$$



(a) March 11th to April 9th, 2020.

(b) April 10th to May 23rd, 2020.



(c) May 24th to June 13th, 2020.

Figure 3.3: Snapshots of the word co-occurrences associated with Death (green labels) and Affect (red labels) for English-language tweets aggregated across all analyzed countries in three different time windows (see sub-captions). The nodes are colored according to their community label as obtained by maximising modularity with the Louvain algorithm [38]. We filtered edges with weight below 20 co-occurrences for visualisation purposes.

The range of f_{Death} is bounded between 0 and 1, being 0 when all of the neighbors of node "death*" ("muert*") are in the Affect category and 1 when all of its neighbors are in the Death category.

We henceforth omit the explicit time-dependence of f_{Death} and Q_{LIWC} to simplify notation. The first measure tracks the quality of separation of the Death and Affect categories in these semantic networks. A larger Q_{LIWC} indicates a better separation between Death and Affect in the empirical word co-occurrences. This is relevant to our investigation of psychophysical numbing in the following sense: if the numbing effect is genuine, we should expect that Q_{LIWC} is larger at larger values of the daily number of deaths and lower at lower values of the daily number of deaths. This would indicate that conversation around Covid-19-related deaths evokes affective responses less strongly for larger death rates. If a weakening association between the concept of death and affective processes is an accurate measure of growing apathy and indifference - of the "collapse of compassion" [50] - then observing a positive correlation between Q_{LIWC} and the daily national number of deaths would provide evidence supporting our hypothesis of psychophysical numbing.

The second is a local measure of the strength of association between the concept of Covid-19 deaths - represented with the word "death*" ("muert*") in a tweet - and the affective processes within those tweets. A high f_{Death} value suggests a weak evocation of affective responses during conversation around Covid-19-related deaths.

To perform this analysis, we compute a sequence $(G_t)_t$ of higher-frequency snapshots than those in Figures 3.3a to 3.3c, where $t = 1, \ldots, T$ labels each of the T snapshots for a given country. Each snapshot represents, on average, the tweets contained in 3 consecutive days and, for each country, each snapshot has roughly the same number of tweets (see Appendix A.3.1 for details on the construction of these networks). With this construction, each network contains approximately the same number of nodes, edges, and network total strength, enabling a fair comparison of the network measures, Eqs. (3.6) and (3.7), over time.

In Figure 3.4, we plot the z-scores of these network measures and of the log of the daily number of deaths $\log s(t)$ for each country, and report the Pearson correlation coefficients ρ_{Q} and ρ_f of Q_{LIWC} and f_{Death} with $\log s(t)$, respectively, in parenthesis above each plot. In general, we observe similar dynamics for both Q_{LIWC} and f_{Death}. This is sensible, since both are measures of the relative strength of association within the two communities induced by the Death and Affect word sets. Furthermore, we observe a number of instances - most notably, Canada, Colombia, Mexico, the United Kingdom, and the United States - in which there is a relatively strong correlation

Figure 3.4: Panel plot time series for the network measures Q_{LIWC} and f_{Death} (see Eqs. (3.6) and (3.7) respectively).

between these network measures and $\log s(t)$. These correlations are, however, weaker in other countries to varying degrees.

To verify that the observed ρ_{Q} and ρ_f can be attributed to the empirical word co-occurrences, we compute the same correlations for corresponding sequences of null network models. For our null model, we take the weighted version of the configuration model described in [41]. Here, a realisation $G^{\mathrm{null}}_{j,t}$ of the null model at random seed j involves assigning node i D_i stubs, where

$$D_i \sim p(d) \tag{3.8}$$

and $p(d)$ is the empirical degree distribution. The kth stub for node i is then assigned

Figure 3.5: Correlation coefficients between the daily national death counts and the network measures Q_{LIWC} and f_{Death} (see Eqs. (3.6) and (3.7) respectively).

a weight

$$W_{ik} \sim p(w|D_i = d), \tag{3.9}$$

where $p(w|D_i = d)$ is the empirical distribution of weights W for nodes with degree w. Stubs with the same weight are then joined with uniform probability.

As a baseline, we compute a sequence $\left(\left[G^{\text{null}}_{j,t}\right]_j\right)_t$ of null model ensembles for each country at each snapshot t, where $j = 1, \ldots, J$ labels each of the J realisations of the null model. Here, we take $J = 100$ realisations per snapshot. We then compute the average network measure over each ensemble for both Q_{LIWC} and f_{Death}, here denoted $Q^{\text{null}}_{\text{LIWC}}$ and $f^{\text{null}}_{\text{Death}}$ respectively. Similarly, we write the correlation coefficients as $\rho^{\text{null}}_{\text{Q}}$ and ρ^{null}_f. We report these coefficients, along with the correlation coefficients for the empirical networks, in Figure 3.5.

We find that, in most cases, the correlation coefficients are higher for the empirical word co-ocurrences than for the null model counterparts. In particular, each of Australia, Canada, Colombia, South Africa, the United Kingdom, and the United States have $\rho_{\text{Q}} \gg \rho^{\text{null}}_{\text{Q}}$. This difference is also present for Nigeria, although it is smaller in this instance. For the remaining countries, however, the differences are either negligible or in the opposite direction. Overall, nonetheless, we see that the average across countries of $\rho^{\text{null}}_{\text{Q}}$ is low, whereas the average across countries of ρ_{Q} is almost three times larger, and that $\rho_{\text{Q}} > \rho^{\text{null}}_{\text{Q}}$ for nine cases out of twelve.

A similar pattern is observed for ρ_f and ρ^{null}_f. In some instances - namely Australia, Colombia, South Africa, the United Kingdom, and the United States - we observe an increase > 0.2 in the correlation coefficients of the original sequence of snapshots

relative to the corresponding sequences of null snapshots. This indicates that these increases in ρ_f can be attributed to the empirical word co-occurrences. The difference is smaller but nonetheless in the correct direction for Chile, Mexico, and Nigeria. For the remaining countries, the difference is negligible.

3.3.3.3 Discussion and summary

Overall, this analysis provides evidence in favor of our hypothesis of psychophysical numbing, although this evidence is not definitive. We have seen that, for most countries, the separation between words associated with Death and Affect in our approximate semantic networks - as measured by Q_{LIWC} and f_{Death} - becomes more pronounced as the national daily deaths rise, and that this relationship is generally weaker in the null model realisations.

There are nonetheless some exceptions to this statement. In particular, we find for Chile and Mexico that the difference between ρ and ρ^{null} is marginal, but that both versions of the correlation coefficients are high. We also report low correlations between these network measures and the time series of daily deaths for Argentina, India, and Spain. For the case of Spain, however, there are two exogenous death-related events contributing to this anomalous behavior and low correlation values, see Appendix A.2 for details. For the case of India, there is evidence suggesting that Twitter users posting from India have a strong preference for using Hindi in the expression of negative sentiment and emotion, but English in the expression of positive emotion [190]. Our use of an English-language dictionary for evaluating the emotional content of such tweets may therefore bias our results, and a more thorough analysis including tweets and dictionaries in both Hindi and English (or in "Hinglish", the blending of the two [153]) should be performed in future. This is a specific case of a more general problem regarding the use of a single dictionary to analyze texts from different world regions, which typically differ in dialect.

For the remaining countries in our dataset, however, the empirical co-occurrences yield stronger correlations between the network measures and the national daily deaths than in the case of the baseline models, providing support for our psychophysical numbing hypothesis. Our observations thus indicate that psychophysical numbing may be a genuine effect for many Twitter users, but that other factors are possibly contributing to our results. Some of these factors are methodological issues with this work. First, we saw in Figures 3.3a to 3.3c that LIWC is unable to account for context, and that there are a number of words that are classically associated with affective processes that are more appropriately associated with concepts surrounding

mortality in the context of the pandemic. Second, in analyzig word co-occurrences, we only retain tweets that contain at least two distinct words in the set Death ∪ Affect by construction. We have evaluated separately the proportion of tweets in each snapshot that contribute to our word co-occurrence networks, and have seen that this usually corresponds to between 10-20% of tweets for each snapshot, with between 20-30% of tweets involving the use of only one word in Death ∪ Affect. As such, this potentially leads to a systematic overestimation of the relative strength of association between words in Death ∪ Affect. Finally, as with most studies of organic social media data, it is hard to control for exogenous factors that form part of the Covid-19 conversation (e.g. Black Lives Matter protests, death-related news). It is thus important to treat such evidence as complementary to classical laboratory-based, controlled psychological experiments.

3.3.4 Modeling attention to Covid-19 casualties

In the previous section, we demonstrated our finding that as the pandemic intensifies, the proportion of words that appear in the set of Tweets posted in each country that indicate emotion diminishes over time. This indicates that the actual emotional response to the pandemic diminishes as the intensity of the pandemic increases, implying a psychophysical numbing effect. We supported this explanation by showing that the word co-occurrence networks induced by our set of tweets host a community structure that separates words in the Death and Affect dictionaries, suggesting that people do not talk about Covid-19 deaths in a highly emotional tone. We built on this analysis by tracking a number of measures of this supposed separation in higher-frequency sequences of snapshots for each country, observing that these network measures behaved consistently with our hypothesis of psychophysical numbing for a number of countries.

The following sections model the relationship between the progression of the Covid-19 pandemic and the Twitter users' perception using grounded theories of psychophysical numbing. Until this point, we have used the emotional framing of the conversation around Covid-19 mortality as an indication of the degree of concern or indifference towards these casualties. However, one could argue that attention itself is equally indicative of the degree of concern experienced by individuals regarding such casualties. Indeed, both are recognized as key components to risk perception and the perception of threats [203]. For this reason, we investigate the relationship between the typical perceptual response of individuals to a stimulus, in this case the daily number of reported deaths nationally, and seek to describe this relationship using

established psychophysical laws, as in previous lab-based psychological experiments e.g. [208].

3.3.4.1 The Weber-Fechner law

Our analysis suggests that the public's perception of the progression of the pandemic is logarithmic or, at least, sublinear. From Figure 3.2, we observe that the correlation magnitudes between NLSs and epidemiological data are generally larger in absolute value whenever the latter are taken in logarithmic scale. To exemplify this observation, we show in Figure 3.6 the z-scores[9] of the Death NLSs and of the logarithm of the daily number of deaths and cases within each country.

The general correspondence between all three normalized features in each country is striking[10] We propose that this can be explained in terms of the **Weber-Fechner law** [88], which is a quantitative statement with its origins in psychology and psychophysics regarding humans' perceived magnitude p of a stimulus with physical magnitude s. It states that a human's perception of the magnitude of a stimulus varies as the logarithm of the physical magnitude s of the stimulus, meaning we are more sensitive to ratios when comparing different physical magnitudes than we are to absolute differences. In the continuum limit, Eq. (3.1) gives the following functional form for the Weber-Fechner law:

$$p(t) = k \log \frac{s(t)}{s_0} + R(t), \tag{3.10}$$

where k and s_0 are real-valued parameters and $R(t)$ the residual. Parameter k determines the sensitivity of perception to changes in the stimulus s, while s_0 determines the minimum threshold that the stimuli s must overcome in order to be perceived. The residual term $R(t)$ is a random variable representing noise not directly captured by the stimulus. For instance, exogenous events can trigger abrupt peaks in the Death score. This is the case, for example, with the murder of George Floyd in the United States, or the peak in Nigeria around April 17th 2020, triggered by a number of prominent African figures dying from Covid-19 around that day, including the Nigerian President's top aide (see Appendix A.2 for details in these peaks).

[9]Recall that the z-score of a sequence of observations $\mathbf{Y} = (y_1, \cdots, y_T)$ is given by $\mathbf{Z} = (\mathbf{Y} - \mu_Y)/\sigma_Y$, where μ_Y and σ_Y are the mean and standard deviation of $\mathbf{Y}$, respectively.

[10]We note that the correspondence is weaker for Australia, Nigeria, and South Africa due to the relatively low number of cases in these countries (see Fig. A.4 in the Appendix for reference). The correspondence is also weaker in Spain because it contains two exogenous peaks not related to psychophysical numbing. See Appendix A.2 for a discussion of these peaks for Spain and other countries, which we remove from the time series.

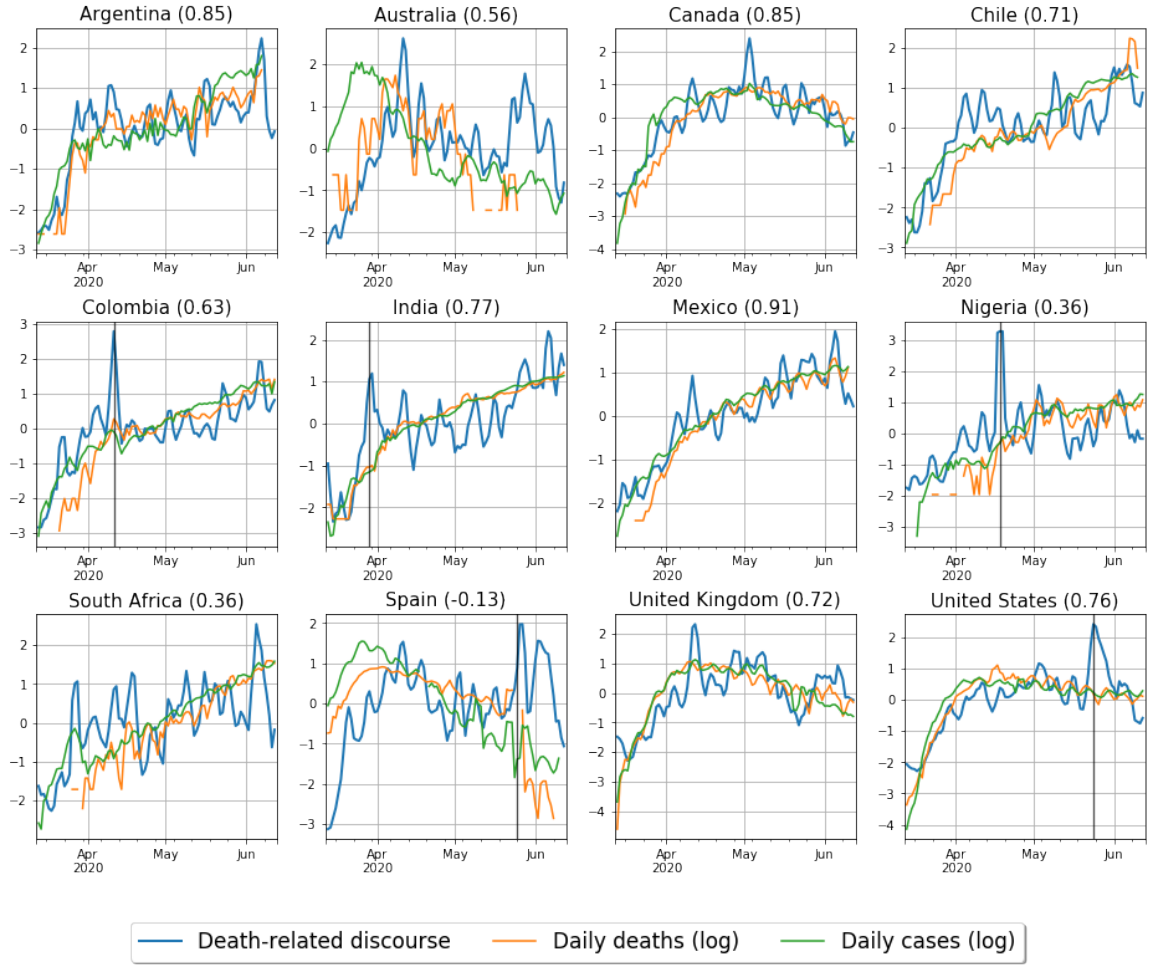


Figure 3.6: Panel time series for $p_i^{\text{Death}}(t)$ (blue), the logarithm of the daily deaths (orange), and the logarithm of the daily cases (green). Each panel presents a different country, with the country name provided in the subplot title. The correlation between $p_i^{\text{Death}}(t)$ and the national daily death rate is given in parentheses for each country. Data is smoothed with a 3-day moving average and standardized with their z-score to make them visually comparable. Vertical lines represent peaks in the death discourse caused by exogenous events not related to psychophysical numbing (see Appendix A.2 for details) which we remove from the time series.

Weber-Fechner law **Country**	k	s_0	95% CI (k)	t (k)	$P > \|t\|$ (k)	R^2	NRMSE	n
Argentina	1.044	0.0080	0.758 – 1.329	7.29	0.0	0.421	0.113	75
Australia	1.042	0.1047	0.508 – 1.576	3.94	0.0003	0.275	0.171	43
Canada	0.596	0.4477	0.508 – 0.683	13.53	0.0	0.683	0.105	87
Chile	0.575	0.0001	0.412 – 0.737	7.05	0.0	0.395	0.149	78
Colombia	0.604	0.0011	0.436 – 0.772	7.18	0.0	0.414	0.144	75
India	0.332	0.0061	0.264 – 0.4	9.69	0.0	0.543	0.128	81
Mexico	0.846	0.0300	0.742 – 0.95	16.2	0.0	0.775	0.101	78
Nigeria	0.457	0.0003	0.168 – 0.747	3.16	0.0025	0.147	0.222	60
South Africa	0.282	0.0001	0.127 – 0.436	3.64	0.0005	0.171	0.183	66
Spain	-0.016	inf	-0.198 – 0.165	-0.18	0.8593	0	0.198	82
United Kingdom	0.752	5.241	0.61 – 0.894	10.54	0.0	0.555	0.143	91
United States	0.788	4.2478	0.672 – 0.905	13.43	0.0	0.677	0.126	88

Table 3.3: Results from the fit of the Weber-Fechner law to the observed relationship between the Death NLS and the logarithm of the daily number of deaths in each country (see Figure 3.6). Overall, this model best describes the relationship between the daily number of deaths local to each country and the Death NLS.

In order to test the Weber-Fechner law, we fit a linear regression model to $p_i^{\text{Death}}(t)$, the Death NLS time series in country i, and $\log s_i(t)$, the daily number of deaths in the same country, and summarize the results of these fits in Table 3.3. We find that Eq. (3.10) accurately models the data, with significant coefficients (p-value < 0.01) for all countries except Spain. The sensitivity parameter k has the same order of magnitude for all significant countries. However, the country with the lowest k is ~ 3 times less sensitive than the highest, indicating that Twitter users in different countries may react differently to the evolution of the pandemic. The minimum stimuli threshold s_0, in the other hand, is always small: most countries, except for the United States and the United Kingdom, need only one Covid-19 death in a given day in order to be perceived. Conversely, the United States and United Kingdom need approximately 5 and 6 deaths to be perceived, which is small compared to the thousands of daily deaths registered in these countries during the observation period.

3.3.4.2 Power-law perception

An alternative functional form for the relationship between human perception p of a stimulus and the physical magnitude s of the stimulus is a power law relationship

$$p(t) = \nu \cdot s(t)^{\beta} + \tilde{R}(t), \tag{3.11}$$

where ν and β are parameters determining the perception from a stimulus of unit magnitude and the growth rate of the perception as a function of the stimulus magnitude,

Power law **Country**	β	ν	95% CI (β)	t (β)	$P > \|t\|$ (β)	R^{2*}	NRMSE	n
Argentina	0.164	2.21	0.121 – 0.208	7.59	0.0	0.411	0.114	75
Australia	0.363	0.99	0.181 – 0.546	4.02	0.0002	0.259	0.173	43
Canada	0.288	0.37	0.252 – 0.323	16.29	0.0	0.678	0.106	87
Chile	0.085	2.47	0.06 – 0.109	6.97	0.0	0.382	0.151	78
Colombia	0.112	1.81	0.083 – 0.142	7.57	0.0	0.425	0.143	75
India	0.126	0.77	0.101 – 0.15	10.33	0.0	0.558	0.126	81
Mexico	0.141	1.52	0.126 – 0.157	18.04	0.0	0.78	0.1	78
Nigeria	0.104	1.56	0.037 – 0.172	3.09	0.0031	0.143	0.223	60
South Africa	0.087	1.11	0.037 – 0.136	3.52	0.0008	0.16	0.184	66
Spain	0.014	2.14	-0.03 – 0.059	0.64	0.5241	-0.042	0.202	82
United Kingdom	0.356	0.16	0.302 – 0.409	13.21	0.0	0.514	0.149	91
United States	0.309	0.21	0.279 – 0.339	20.54	0.0	0.608	0.139	88

Table 3.4: The results from the fit of a power law to the relationship between the Death NLS and the national daily death count. This is the best model in some cases, though is outperformed by the Weber-Fechner law most times. *While we fit this model assuming a log-log relationship between p and s, we compute R^2 with linear p to make it comparable to the model implied by the Weber-Fechner law (see Eq. (A.2) in Appendix A.1 for details). This may cause negative values of R^2 as is the case for Spain.

and $\tilde{R}(t)$ is a residual term. This form has been shown to outperform the Weber-Fechner law in characterising human perception in a number of empirical studies [209]. We also therefore report the results of this model fit to the relationship between the Death NLS $p_i^{\text{Death}}(t)$ and national daily death counts $s_i(t)$ for each country i, reporting our results in Table 3.4.

In all cases, we observe sublinear exponents β for the perception of the daily deaths data, with significant exponents (p-value < 0.01) ranging between 0.085 and 0.36. These exponents are of the same order of magnitude as the β of 0.32 reported in [210], where in several laboratory experiments they measure psychophysical numbing in participants' perception of death statistics. As discussed previously, the data for Spain is unusual for a number of reasons, thus the model does not accurately describe the data in this instance. These results suggest that Twitter users in certain countries are more sensitive to change in the number of deaths than others.

3.3.4.3 Model comparison

Both the Weber-Fechner law and power-law relationships between the Death NLS and the daily number of reported deaths accurately model the data. Each captures the phenomenon in which "the first few fatalities in an ongoing event elicit more concern

NRMSE **Country**	Power law	Weber-Fechner law	Linear relationship
Argentina	0.114	**0.113**	0.116
Australia	0.173	**0.171**	0.175
Canada	0.106	**0.105**	0.117
Chile	0.151	**0.149**	0.17
Colombia	**0.143**	0.144	0.145
India	0.126	0.128	**0.125**
Mexico	**0.1**	0.101	0.133
Nigeria	0.223	0.222	**0.218**
South Africa	0.184	**0.183**	0.188
Spain	0.202	0.198	**0.193**
United Kingdom	0.149	**0.143**	0.166
United States	0.139	**0.126**	0.179
Mean	0.151	**0.149**	0.16
Proportion of best fits	16.7 %	**58.3 %**	25 %
Proportion of second-best fits	**66.7** %	33.3 %	0 %

Table 3.5: Comparison of the normalized root mean squared error (NRMSE) (see Eq. (3.12)) between the power law model of Eq. (3.11), the Weber-Fechner model of Eq. (3.10), and a linear relationship between variables, which we use as a benchmark model. Lower values indicate better-fitting models. Note that, overall, the Weber-Fechner law outperforms the other models. For further details, see Figs. A.1 and A.2 in Appendix A.1.

than those occurring later on" [168]. By way of comparison, we present in Table 3.5 the normalized root mean squared errors (NRMSE), defined as

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_t^n e(t)^2}}{p_{max} - p_{min}}, \tag{3.12}$$

for these models, in addition to a linear model between $p_i^{\text{Death}}(t)$ and $s_i(t)$ as a baseline "null" model. Here, $e(t) = p(t) - \hat{p}(t)$ is the model residual, and n is the sample size. The models are directly comparable in this sense, since each involves only two parameters. Bhatia [33] performed a similar model comparison to test psychophysical laws for subjective probability judgements of real-world events, in that case finding that the linear relationship was the best. In our case, however, a linear relationship between s and p is significantly worse than the present concave models of perception (see Appendix A.1 for the results of the linear model), reinforcing our hypothesis of psychophysical numbing.

While the Weber-Fechner law is better than the power law model overall, the difference in their goodness of fit - as measured by the NRMSE - is marginal. Both

are reasonable descriptions of the observed relationship, and similar conclusions can be drawn from both.

In particular, the parameters k and β from the Weber-Fechner law and power law, respectively, are analogous in their interpretation as the measure of the sensitivity of the nation's Twitter users to changes in the national Covid-19 daily death rate. To illustrate this, we rank the countries in our dataset in order of sensitivity to changes in the local death rate, as measured separately by these two parameters, and plot the correlation between the countries' ranks in Figure 3.7. Here, low rank indicates high sensitivity to changes in the number of daily deaths nationally. The correlation between the two methods of ranking - according to k, the Weber-Fechner law slope parameters, and according to β, the power law model exponents - is high, with correlation coefficient 0.77. This shows that the sensitivity of each country is relatively robust between models. By both measures, therefore, Twitter users tweeting in English and Spanish from Australia and Argentina, respectively, appear to be the most sensitive to changes in the national daily death rate, while Twitter users posting in English from South Africa, India, and Nigeria and in Spanish from Spain and Chile appear to be the least sensitive to these changes.

3.4 Discussion and conclusions

We explored the country-by-country relationship between the linguistic features present in a large set of tweets posted in relation to the Covid-19 pandemic, and the progression/intensity of the pandemic as measured by the daily number of cases and deaths in each country we consider. By considering the change, relative to a baseline, in the percentage of words present in each tweet that are associated with a number of psychologically meaningful categories - here called linguistic scores - we observed significant trends that we believe are indicative of a psychophysical numbing effect [203].

We found that the national linguistic scores (NLSs, see Eq. (3.4)) associated with emotion and affect decrease as the pandemic intensifies. This is in spite of a greater attentional focus on death and mortality and a simultaneous increase in use of words indicating analytic reasoning. We showed, by constructing word co-occurrence networks on different time periods of the pandemic, that words related to death co-occur more frequently with other words related to death than they do with words indicating affect and emotion. We constructed network measures of this separation between the concepts of death and emotion - namely the weighted modularity of the

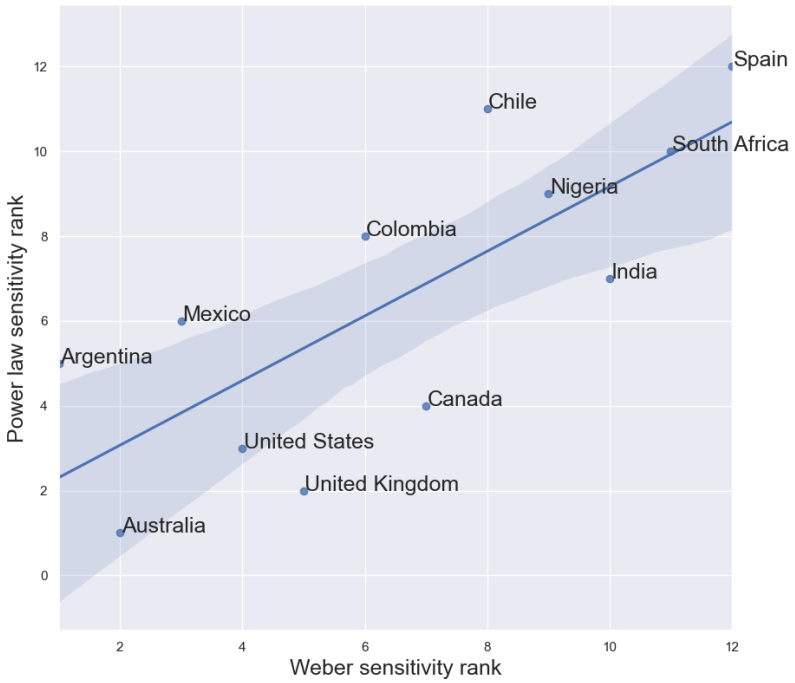


Figure 3.7: Comparison of the rank of each country as determined by their k and β parameters in the Weber-Fechner and power-law fits, respectively, which determine the sensitivity of Twitter users tweeting from each country to changes in the number of daily reported deaths. Low rank indicates high sensitivity relative to the remaining countries. The correlation between countries' ranks from both measures is high at 0.77.

partition induced by the Death and the Affect LIWC dictionaries, and the fraction of strength of the "death*" (muert*) node attributable to connections with other nodes in the Death category - and showed that this separation became more pronounced at larger daily death rates for a number of countries. This is consistent with the notion of psychophysical numbing, which we believe may explain these observations.

We also showed that the psychophysical laws of Weber-Fechner and of power law perception in humans accurately model the relationship between the frequency of words related to death and the actual daily number of Covid-19 deaths in each country. We estimated sub-linear exponents in the power law perception function that are of similar values to values previously estimated from psychological experiments [210]. These exponents, together with parameter k of the Weber-Fechner law (see Eq. (3.10)), tell us how sensitive the Twitter users in each country are to their national Covid-19 daily deaths, and were seen to vary by country, indicating inter-country differences in risk perception and sensitivity to death rates. Such sensitivities were consistent across models (see Fig. 3.7) suggesting that these measures of a nation's

Twitter users' sensitivities to changes in the national death rate are robust features of the data.

Overall, our results indicate that two key factors contributing to risk perception - attention and emotion [203] - may be evolving in line with that predicted by psychophysical numbing amongst members of the public. In general, both measures of the degree of concern towards Covid-19-related casualties expressed by the Twitter users in our dataset appear to decrease as the number of Covid-19-related casualties increases. This potentially reflects a collapse of compassion and a concavity in the value assigned to human lives as the number of potential casualties grows.

Our findings illustrate the signaling power of Twitter, and demonstrate its potential use as a tool for monitoring public perception of risk during large-scale crisis scenarios. With the modeling and visualization approaches we employ in this chapter, policy-makers and public officials could track in near-to-real-time the public's attitudes towards threats to public well-being and the prevalence of factors important to public perception of risk, including degree of outrage and relative attentional focus on the threat. Our findings also imply a functional form for agent perception of the system state in models of opinion dynamics. This will be instrumental for developing coupled opinion dynamics-epidemiological models, in which the bidirectional relationships between human perception, human behavior, and epidemic progression are modeled endogenously.

A natural extension to this work would involve nowcasting and/or forecasting of certain economic indicators. It has also been limited in that we assumed that only the national death rate is a significant predictor of perception. A more complete analysis should account for the effect of other countries' death statistics as a driver of local perception, or more broadly an advancement of a process-level explanation of the cross-cultural differences we observe in the sensitivity to death statistics. This analysis could also be enhanced by relating these measures of risk perception to behavioral data, which - since "people's behavior is mediated by their perceptions of risk" [229] - may be useful for understanding the role of emotions in driving behaviors that are conducive to public health during crises. Further, a deconstruction of the aggregate indicators we have developed to the state and regional level may be necessary to more accurately characterize the relationship between local crisis progression and human risk perception.

It is important to acknowledge that additional factors may be at play and contributing to our findings. In particular, we considered the relationship between perception and stimulus, where we took the stimulus to be the daily Covid-19 deaths. At

least in the short term, we believe this is a good approximation because we assume that people react to the change on daily Covid-19 deaths and not to the change of cumulative deaths. However, we this might not hold in the longer term because, for instance, the populations might grow increasing levels of fatigue, making the magnitude of perception a decreasing function of the number of waves that have passed.

On another note, our dataset is a large social media dataset in which non-human accounts - for instance, bots, institutional accounts, and companies' public relations accounts - coexist with human accounts. Such public relations and institutional accounts can be subject to editorial constraints on the kind of language used, and therefore may not reflect any true underlying subjective experience. The use of tweets from such non-human accounts may nonetheless be appropriate. Indeed, it is widely accepted that news media play a significant role in shaping public attention and opinion, e.g. via the Cultivation or Agenda-Setting theories of consumer-media relations [44, 156]. With almost half of all UK adults consuming news through social media in 2020 [8], for example, the inclusion of news and institutional accounts may act as a proxy for public attention and opinion at large.

With regard to bots: previous large-scale studies of Twitter data have demonstrated the influence bots can have on the exposure of human accounts to emotional content [207] and the extent to which they can distort the discussion on certain topics [30]. More recently and in the context of the current pandemic, bots have been shown to have a significant role in promoting political conspiracy theories [89]. By ignoring retweets and using unique original tweets only, we mitigate to some extent the potential effect of bots, which have previously been shown to engage in retweeting behavior significantly more frequently than they do the creation of original content [30]. It is nonetheless likely that, even if the hypothesised psychophysical numbing effect is genuine, our observations are partly attributable to the nature of content generated by these non-human accounts.

Furthermore, we stress that the results presented in this chapter may be indicative only of the responses of Twitter users posting from each of these countries in each of these languages, so extrapolating these results to the broader population will only be possible with a better understanding of the biases present in, and representativeness of, the dataset at hand. While the demography of Twitter users has been to some extent mapped for the United States (see e.g. [106]) and the United Kingdom (see e.g. [202]), it is difficult to find similar studies for the remaining countries in our dataset, and thus to interpret these country-level differences in terms of potentially

differing demographic representation on Twitter. We nonetheless advance this as a factor that possibly contributes to our results.

We also reiterate that our analysis has been crude in that we make use of a single dictionary for each language when extracting linguistic features from our data. This ignores important differences in dialect and language use between different nationalities and cultures, and can result in the systematic omission of certain linguistic features [190, 153] which may also contribute to the observed differences between countries. Further important differences between countries which may help to account for the observed results are differences in the importance of religion in each of the considered countries. The set of countries under consideration here span the full spectrum of importance assigned to religion [108], and attitudes towards death and the framing of mortality may vary accordingly by country. Despite these difficulties inherent to the empirical analysis of social media data, we nonetheless hope that our work inspires further investigations into the use of natural language processing and cognitive network science to investigate the prevalence of psychophysical numbing in naturalistic contexts

Chapter 4

Quantifying the structure of the climate change conversation with unsupervised methods

Disclaimer

This chapter is heavily motivated by the work I co-authored with Fabián Aguirre-López, Sergio Hernández-Williams, and Guillermo Garduño-Hernández. This work is now posted in the ArXiv [136] and currently under review. See Section 1.1 for details on the division of labor for this work.

4.1 Introduction

In the last chapter, we investigated the emotional framing of the Covid-19 conversation during the first wave of the pandemic in 12 different countries. We treated our Twitter dataset as a social complex system where its entities, the Twitter users, interacted with different levels of emotion that we quantified using natural language processing. In this chapter, we also take a complex systems approach of another massive Twitter dataset, but this time regarding the climate change conversation. Moreover, as we justify below, we focus on the interaction structure of the conversation using unsupervised methods.

Social media is crucial for information consumption and public opinion formation [60, 97, 95, 39]. It leverages communication channels between one-to-many and many-to-many in a decentralized way by enabling its users to choose whom to follow and interact with, thus democratizing information access and spreading [29, 57]. However, the decentralized nature of social media promotes that users consume in-

formation mostly aligned with their beliefs and interact with like-minded individuals and communication channels. Authors have studied these behaviors under the sociological frameworks of confirmation bias [227] and homophily [158]. Additionally, social media platforms are designed to maximize engagement time, with recommendation algorithms trained to show content similar to what users typically consume [80]. To a certain extent, the aforementioned mechanisms explain an empirical observation about controversial topics in social media: public discourse is polarized. Authors have studied polarization for several relevant topics, ranging from climate change [63, 232, 58], to politics [97, 39, 188], to Covid-19 [78, 126].

According to Caves' *Encyclopedia of the City* [56], social polarization is the segregation within a society that emerges from several socio-economic factors that differentiate social groups. In the context of public discourse in social networks, *echo chambers* reflect a parallel mechanism to social polarization. An echo chamber emerges, according to Bruns et al. [43], when a group of actors chooses to preferentially connect with each other, with the exclusion of outsiders. However, over which preferences would such actors choose to connect? Garimella et al. [95] state that opinions or beliefs stay inside communities created by like-minded people who reinforce and endorse each other's opinions. Therefore, agents prefer to connect to other actors with similar opinions and exclude those with contrasting ones, suggesting that *homophily* is a driving force in the creation of echo chambers [99, 158]. An actor inside an echo chamber will mostly receive information coherent with her beliefs, so she will reject any new information that does not align with those beliefs, i.e., actors in an echo chamber experience a *confirmation bias*. This bias creates a positive feedback mechanism that reinforces the echo chambers in the light of new information [200].

Echo chambers inhibit communication across groups with different ideologies. Thus, even when a scientific consensus about a topic is reached, it will hardly permeate all the echo chambers of a social group, as happens with climate change. According to the latest Intergovernmental Panel on Climate Change (IPCC) report [11], the cumulative scientific evidence is unequivocal: Climate change threatens human well-being and planetary health. Despite this, a significant fraction of the population in many countries express some form of doubt or skepticism about whether trends in climate change are human-made, if these trends will bring harmful consequences to society, or both [178, 138]. Such skepticism may be linked with factors such as the spread of misinformation online [219], nationalist and individualist identities [138], or, particularly in the United States, conservatism and conspiratorial beliefs [116]. The issue of climate change is not only polarizing but also a matter with deep social, economic, and

ecological impact. Shifts in public opinion might lead to significant behavioral and political changes that could face strong opposition from specific population sectors.

Several authors have studied the climate opinion landscape in social media, finding, in most cases, that a group of *climate believers*, i.e., actors that support the science about climate change, and a group of *climate skeptics* coexist and segregate into online *echo chambers* [232, 221, 124, 58]. Online social media has been a crucial driver of pro-climate movements, such as the "Fridays for Future" movement started by Greta Thunberg [214, 143]. Using Twitter data, Williams et al. [232] constructed several interaction networks from relevant climate-related hashtags and manually labeled the most active users as believers, skeptics, or neutral. They measured high levels of homophily in the followers and retweet networks, suggesting that the conversation is polarized. With a similar approach, Jang and Hart [124] found a polarized Twitter climate change conversation in the United States between Republicans and Democrats. They claim that "climate change" and "global warming" are meaningful query filters to study the climate conversation, with the latter being more common among climate skeptics. Chen et al. [58] studied Finland's Twitter climate change conversation by creating interaction networks from retweets. They found that the climate conversation was subject to partisan sorting and aligned with the universalist-communitarian dimension of European politics. Xia et al. [233] found that viral climate topics around the 2019 Nobel peace prize spread in different groups, enhancing in-group connections and repulsing out-group engagement, suggesting the presence of echo chambers.

In this chapter, we study the structure of the Twitter climate change conversation during 2019, when the "Fridays for Future" and "Extinction Rebellion" social movements flourished. We introduce unsupervised methods to 1) identify the leading users of the conversation throughout the year, 2) measure the ideological similarity between leading users by examining the many-to-many communication channels of the audience of a leading user, 3) determine the ideology of the leading users using their ideological similarities, and 4) present an operational definition of echo chamber based solely on the structure of the Twitter interaction networks for which we classify more than half of the total retweeting population. We construct these methods under the assumptions that retweeting is a good proxy for endorsement [97, 23], and that information flows in the opposite direction of retweets. Moreover, we acknowledge that Twitter conversations are highly heterogeneous in that a tiny number of users explain the majority of the retweets produced by the population in a given dataset [101]. These methods build upon other works that provide an operational definition

to detect echo chambers [60] and create a model to infer the ideological position of active users on Twitter [23].

We organize the remainder of this chapter as follows: In Section 4.2, we identify the *leading users* of the climate conversation, introduce the *chamber* associated with a leading user, and identify two polarized ideological groups based on the similarity of the chambers of the leading users using unsupervised methods. In Section 4.3, we introduce a definition of echo chamber and classify most of the users in the retweet network as either climate believers or climate skeptics. We inspect the properties of the echo chambers and discuss the communication within and across groups. Finally, in Section 4.4, we conclude and suggest future research directions. We code all the methods and analysis presented in this chapter using *Python* and is publicly available at `https://github.com/blas-ko/Twitter_chambers`.

4.2 The chambers of the Twitter conversation

Twitter users often consume information from various sources. They typically follow, audit, and interact with broadcasters that represent the one-to-many communication channels of the Twittersphere and with other lower-impact users that represent the many-to-many communication channels [145, 103]. In this chapter, we introduce quantitative methods that leverage the impact of the broadcasters, here called *leading users*, and the set of lower-impact users, here called *chambers*, of a given audience. Later, we introduce a quantitative definition of *echo chamber* based solely on the structure of the Twitter interaction networks. We focus in the climate change Twitter conversation during 2019, where, given the previous research [232, 221, 124, 58], we expect to find echo chambers of climate believers and climate skeptics.

Our aim is to construct a set of observables that characterizes the dynamics and structure of the conversation in an *unsupervised* way. Thus, we avoid missing certain features that could be ignored by our biases. Most importantly, the methods we introduce in this chapter are applicable to other datasets where we do not have clear expectation of what the structure is.

In this section, we define a *leading user* and its associated *chamber* and derive methods that identify polarized ideological groups in an unsupervised way. We validate *a posteriori* that such ideological groups correspond to those who believe in anthropogenic climate change and those who are skeptic about it.

4.2.1 Data

Alongside the company *Sinnia*[1], we create a dataset with 41.8 million climate-related tweets from 8.7 million users spanning 1^{st} March to 1^{st} December 2019, totalling 39 weeks. We collect *all* the tweets that satisfy any of the following queries: "*climate change*", "*global warming*", "*climatechange*" and "*globalwarming*". We choose these queries following the results from prior studies [223, 232, 124]. Of the total volume of tweets, 73% corresponds to *retweets*. We therefore focus on the retweets only, assuming that they provide significant information of what users are looking at or interested in. We decide to work with the retweets networks and not with the follower-followee network because, on the one hand, it is way harder to obtain through Twitter's API, and, on the other hand, the follower-followee network does not capture the direction of endorsement while the retweets network does, as we justify below.

4.2.2 High-impact and leading users

Discussions in Twitter are highly heterogeneous [191]. This means that a huge fraction of the information consumed and spread through Twitter interaction networks is generated by just a handful of users. However, these networks often exhibit community-like structures where the density of interactions is higher for users with similar features. Throughout this chapter, we assume that retweets are a good proxy for endorsement [23, 97, 58], so that we may study the retweet interaction network as a social network of endorsements. Thus, we study the dynamics of the retweet network, $\mathbf{W}^t$, which we construct following Beguerisse et al. [27] as

$$W_{ij}^t = \#\text{ of retweets of user } i \text{ originally posted by } j \text{ in a given week } t, \tag{4.1}$$

$$w_i^t = \sum_j W_{ji}^t = \#\text{ of retweets of tweets posted by user } i \text{ in week } t, \tag{4.2}$$

where i and j denote users from the complete set of retweeting users, $\mathcal{U}$, and t denotes time, where we choose the *week* as our temporal resolution.

In this context, we will refer to w_i^t as the *impact* of user i during week t. From a networks perspective, the impact is the same as the weighted in-degree of a given node. Using the impact vector, $\mathbf{w}^t = (w_i^t)$, we can quantify how unequal the discussion in Twitter is by treating w_i^t as a measure of the wealth of user i and computing $\mathbf{w}^t$'s Gini index. The Gini index would be 0 if every user received the same number of retweets and 1 if only one user is responsible for every retweet.

[1] http://www.sinnia.com/en/

The Gini index gives us a global picture of the impact heterogeneity, but we can also use $\mathbf{w}^t$ to identify important users in the conversation. We define the set of N *high-impact users* in week t as the N most retweeted users during that week, i.e.,

$$\mathcal{I}(t) = \{i \in \mathcal{U} \mid \text{rank}(w_i^t) \leq N\}. \tag{4.3}$$

Importantly, $\mathcal{I}(t)$ is a dynamic set, as $rank(w_i^t)$ typically changes weekly for all users. This contrasts with previous works, [101], where they fix the set of high-impact users a priori.

The set $\mathcal{I}(t)$ may vary significantly throughout the weeks, so, to characterize the most important actors in the conversation, we introduce the *persistence*, Δ_i, as the number of weeks where i is a high-impact user. The persistence lets us identify the high-impact users that are relevant in longer time windows. We consider users with high impact and high persistence as *leading users* because they function as one-to-many information channels that are relevant to focused audiences for extended periods [145]. We thus define the set of M *leading users* as

$$\mathcal{I}^{\Delta}(t) = \{i \in \mathcal{I}(t) \mid \text{rank}(\Delta_i) \leq M\}, \tag{4.4}$$

that have an associated vector of impact,

$$\mathbf{w}^t_{\mathcal{I}^{\Delta}} = (w_i^t)_{i \in \mathcal{I}^{\Delta}(t)}. \tag{4.5}$$

When we apply these definitions to our Twitter dataset on climate change, we observe that almost 50% of the weekly retweets are given to the top $N = 50$ users. This corresponds to a massive inequality, as the vast majority of interactions are produced by just a handful of tweets. We observe an average Gini index of 0.89 (± 0.02)[2], meaning that we should be able to understand a significant fraction of the Twitter conversation just by looking at its high-impact users.

We take the $M = 50$ leading users from the $N = 50$ high-impact weekly users. While the set of leading users, $\mathcal{I}^{\Delta}(t)$, is dynamic, in the analyses that follow we focus on the complete set of leading users, $\mathcal{I}^{\Delta} = \cup_t \mathcal{I}^{\Delta}(t)$ where $|\mathcal{I}^{\Delta}| = 50$. As a first approach to understanding the set of leading users, we *manually assign* each of them a category about climate change. We identify three main ideological currents: *climate believers* that support scientific consensus about anthropogenic climate change, *climate skeptics* that doubt that the trends in climate are anthropogenic or that such

[2]Unless otherwise stated, quantities with uncertainty bars represent the standard deviation around the mean across weeks.

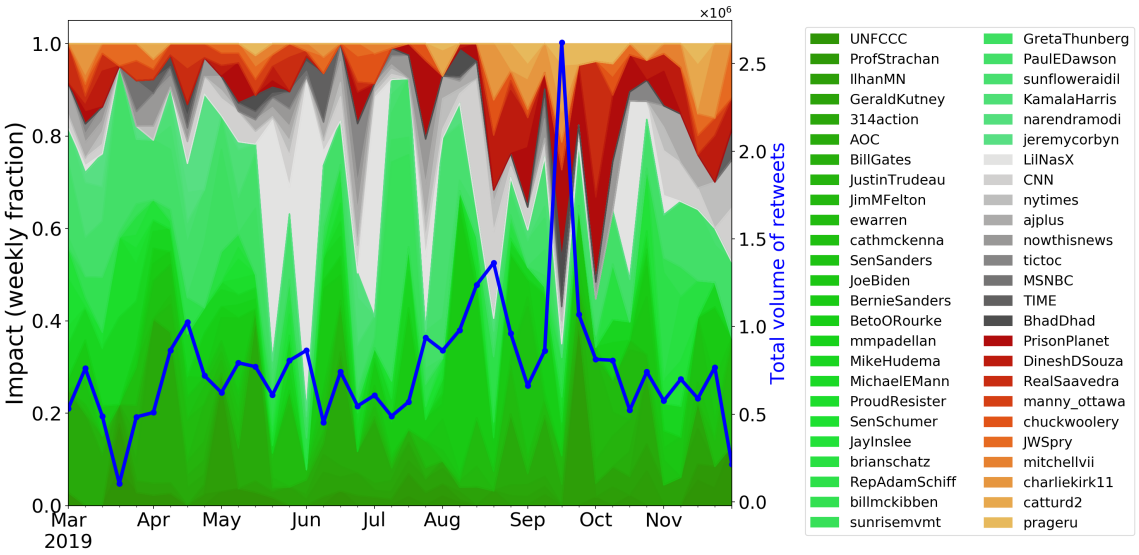


Figure 4.1: **Impact dynamics of the high-impact users** normalized to each week. We identify green users as *climate believers*, red users as *climate skeptics*, and gray users with other sources such as news media and artists. In blue, we show the total volume of retweets for the given week, where we identify two significant peaks: one on the 3^{rd} week of of August, and the other on the 3^{rd} week of the September, 2019.

trends will bring harmful consequences to society, and *news media channels and entertainers*. In Fig. 4.1, we show the impact dynamics of the leading users *per* week. Although we observe an overall dominance of the climate believers (green) throughout the year, we observe a clear irruption of the climate skeptics (red) in the second half of the year, with a specially pronounced peak in the third week of September, 2019. Additionally, in Appendix B.4, we present Table B.1 which enlists the $M = 50$ leading users and their main features.

4.2.3 The chamber: quantifying leading users' similarity

In the previous section, we proposed a way to identify the (small) set of *leading users* responsible for most of the Twitter discussion. We manually labelled them based on their ideological position about climate change by inspecting their Twitter bios, Wikipedia pages and related news stories. Manual labelling is often time consuming and requires external sources to retrieve information about the users. In this section, we introduce an unsupervised approach for identifying ideological (dis)similarities between leading users. This approach relies on analyzing the many-to-many communication channels of an audience, i.e., the chain of interactions stemming from members of the audience, with a behavior similar to viral spreading [145].

Each leading user has an associated *audience* who endorses her, where the content posted by the leading user flows in the direction of her audience, creating a one-to-many communication channel, similar to broadcasting [103]. Simultaneously, the audience consumes and endorses posts of other Twitter users, forming a many-to-many communication channel, which we call the *chamber* of such an audience. The audience and the chamber associated with a leading user are our main objects of study, and we introduce them precisely in what follows.

Definition 4.2.1 (Audience)**.** Given a user $i \in \mathcal{U}$, her *audience* at week t, $\mathcal{A}_i^t$, is the set of users that have retweeted i during t:

$$\mathcal{A}_i^t = \{j \in \mathcal{U} \mid W_{ji}^t > 0\}. \tag{4.6}$$

The set $\mathcal{A}_i^t$ indicates who endorsed i during week t. Thus, the ideologies of the members of $\mathcal{A}_i^t$ are coherent to that of i's. Moreover, the audience might interact with and endorse other lower-impact sources, creating an information flow from these sources to the audience. We call these collection of sources the *chamber* of the audience associated with the leading user i.

Definition 4.2.2 (Chamber)**.** Given an audience, $\mathcal{A}_i^t$, associated with a user $i \in \mathcal{U}$, the *chamber* of $\mathcal{A}_i^t$ during week t, $\mathcal{C}_i^t$, is the set of users retweeted by the audience of i excluding other high-impact users:

$$\mathcal{C}_i^t = \{j \in \mathcal{U} \mid W_{kj}^t > 0 \text{ for } k \in \mathcal{A}_i^t,\ k \notin \mathcal{I}(t)\}. \tag{4.7}$$

The audience is a collection of information consumers, while the chamber is a collection of sources. See Fig. 4.2, where we show a schematic representation of the audience and the chamber associated with a leading user. A leading user relates to its chamber through its audience, so we expect that the members of the chamber have a similar ideology to that of the leading user. We consider the (dynamic) audiences and chamber associated with the leading users, $\mathcal{I}^\Delta$, throughout the year.

The chambers transmit information from the sources to the audiences. Thus, we can estimate the information flowing between two audiences associated with the leading users i and j by comparing how similar their chambers are. Following from our assumption that retweets indicate endorsement, this similarity signals some notion of ideological distance between of i and j. Thus, we can compare the ideology two leading users by looking at their chamber overlaps, which we introduce in what follows.

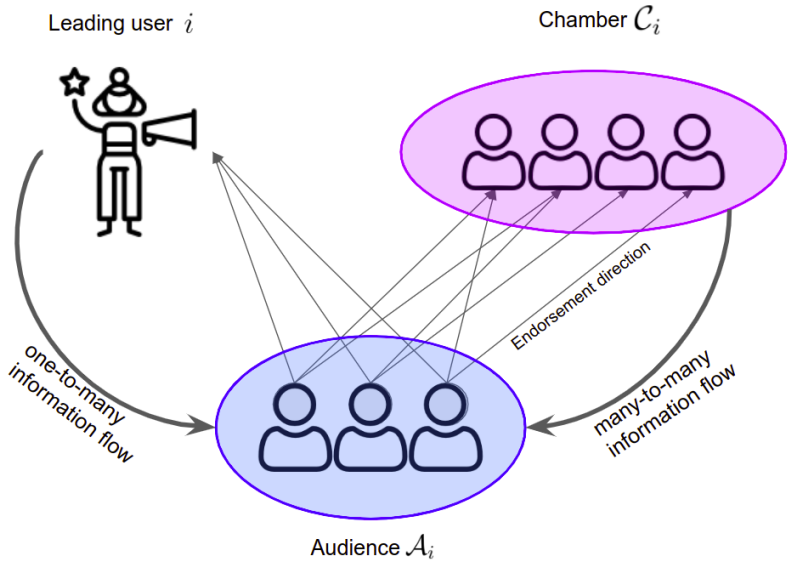


Figure 4.2: Schematic diagram of the *audience* and the *chamber* for some user $i \in \mathcal{I}^{\Delta}(t)$. The audience, $\mathcal{A}_i^t$, is the set of users that retweet i while the chamber, $\mathcal{C}_i^t$, is the set of users retweeted by every member of the audience excluding users from $\mathcal{I}(t)$, i.e. the *sources of information* of the audience other than the high-impact users.

Definition 4.2.3 (Chamber overlap)**.** Given two leading users i and j with respective chambers $\mathcal{C}_i^t$ and $\mathcal{C}_j^t$, the *chamber overlap*, q_{ij}^t, during week t is the Jaccard similarity between $\mathcal{C}_i^t$ and $\mathcal{C}_j^t$:

$$q_{ij}^t = \frac{\left|\mathcal{C}_i^t \cap \mathcal{C}_j^t\right|}{\left|\mathcal{C}_i^t \cup \mathcal{C}_j^t\right|} = \frac{\#\text{ of users in common of both chambers}}{\text{total }\#\text{ of users of both chambers}}. \tag{4.8}$$

The overlap between the chambers gives us a proxy of the ideological similarity between i and j *without* determining their ideological positions explicitly, meaning that q_{ij}^t is a relative measure *between* the two users. We could compare the *audience overlap* using the audiences $\mathcal{A}_i^t$ instead of the chambers in Eq. (4.8), but we prefer using the chambers instead for mathematical and conceptual considerations (See Appendix B.3 for details).

Our approach of taking overlaps is similar in spirit to the co-citation projections used in bibliometrics [167], where an edge between users i and j in the co-citation network indicates the number of common audience members between i and j. Beguerisse et al. [27] compute authority scores of important users in retweet networks based on the eigenspectra of their co-citation projections, which enables them to compute authority scores of the most important agents in social networks they analyze. Further, Becatti et al. [26] took the co-citation projection of the bipartite retweet network

between verified and non-verified Twitter users, and found communities of verified users based on their audience similarities. In this context, we could also construct bipartite networks between leading and non-leading users to determine communities of leading users based on the co-citation projection network. However, our approach involves taking *chamber overlaps*, which are second-order neighborhoods with more robust signals of ideological similarities - as we discuss in Appendix B.3.

Following from our analysis from Section 4.2.2, we expect for the climate change discussion to exhibit a clear separation between climate believers and climate skeptics [232, 63]. We quantify such a separation by looking at the aggregate chamber overlap distribution over the year - i.e., we consider $p(q)$ for $q \in (q_{ij}^t)_{t,i<j}$. If the separation exists, the chamber overlap between pairs of users from the same ideological group should be significantly larger than pairs from different groups.

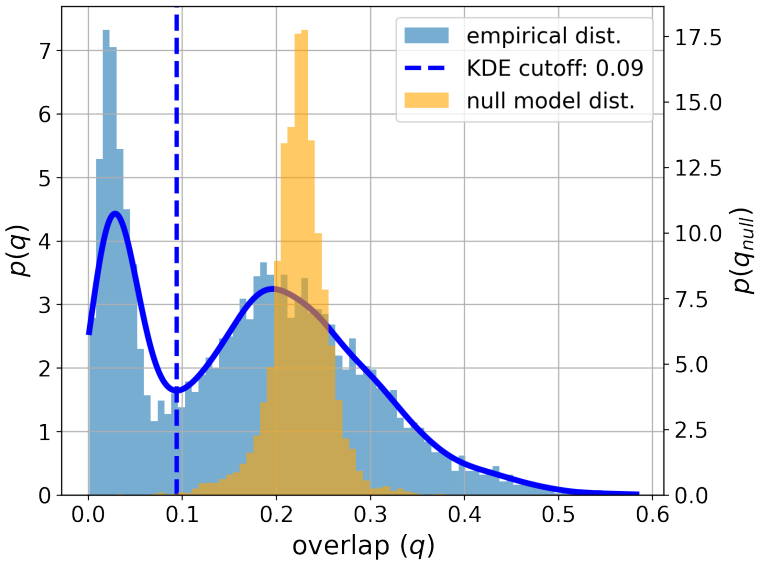


Figure 4.3: **Aggregate chamber overlap distributions for the empirical and null model networks.** We construct the aggregate empirical overlap distribution (blue, left y-axis) by concatenating the overlap pairs, q_{ij}^t, for every week t on the dataset. We observe a *bimodal* structure characterized by low-overlap peak, $q_{off} = 0.04 \pm 0.02$, and a high-overlap peak, $q_{in} = 0.23 \pm 0.08$ (see the main text for details on how we separate the peaks). We compute the null-model chamber overlap distribution (orange, right y-axis) using Eq. (B.8) over the empirical degree sequences for every week. Eq. (B.8) computes the expected overlap between pairs (i, j) of users of degree k_i and k_j, respectively, based on the configuration model (see Appendix B.1 for details). We observe that the configuration model predicts a *unimodal* overlap distribution characterized by $q_{null} = 0.22 \pm 0.03$ without any noticeable skew.

In Fig. 4.3, we show the overlap distribution, $p(q)$, aggregated over the whole

period, where we observe a clear bimodal structure with a sharp peak at $q_{off} = 0.04 \pm 0.02$ and a more spread-out peak at $q_{in} = 0.23 \pm 0.08$. We approximate the position and spread of the peaks using a Gaussian kernel density estimator with an optimal bandwidth parameter using Scott's rule [196] and separating the peaks according to the minimum between them. The observed bimodal structure is not trivial. To claim this, we compare the empirical structure with the expected chamber overlap distribution of the configuration null model. The configuration model consists of random graph ensembles where the edges are statistically independent, but the in-degree sequence is, on average, the same than that of the empirical network (see Appendix B.1 for details). For the null model, we find that the expected chamber overlap distribution is unimodal with a well-defined peak. We show the empirical and the expected null-model chamber overlap distributions in Fig. 4.3. We find that the peak of the null model coincides with the in-block empirical one but has a lower spread ($q_{null} = 0.22 \pm 0.03$). In Appendix B.1, we derive an explicit expression for the expected chamber overlap, $\langle q_{ij} \rangle$, between users i and j for a given degree sequence $\boldsymbol{k}$.

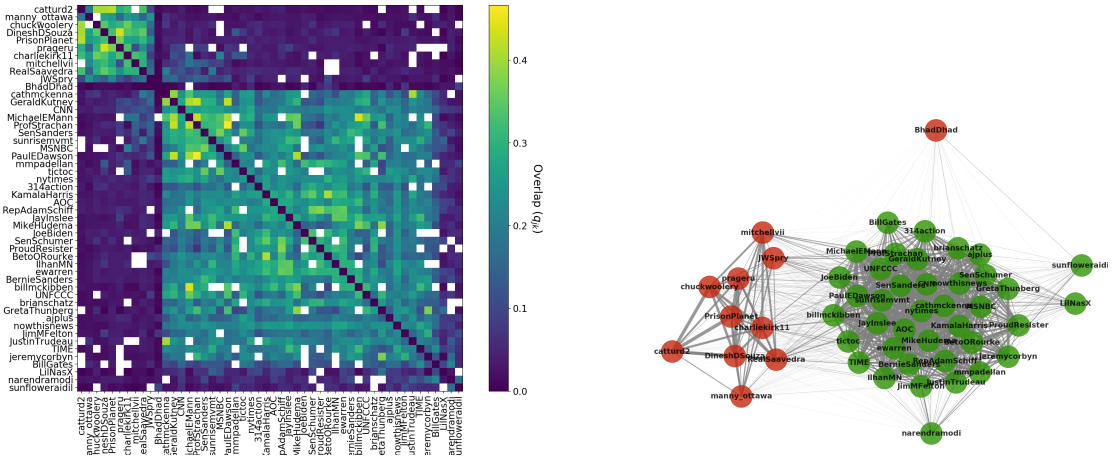
The overlap distribution indicates that the ideologies on the climate conversation are polarized. However, it does not identify which users are from what group nor how many groups are there. In what follows, we describe how to obtain the ideological position of the users based on their chamber overlaps, q_{ij}^t, and compare it with quantitatively with the manual labelling.

4.2.4 Classifying leading users by chamber overlap

In this section, we create a partition of the set of *all* leading users, $\mathcal{I}^{\Delta}$, based on the chamber overlap distribution, $p(q)$, which exhibits a bimodal structure (See Fig. 4.3). A partition corresponds to a collection of disjoint subsets of $\mathcal{I}^{\Delta}$, $\mathcal{P} = \{P_\alpha\}_{\alpha \in I}$ with $P_\alpha \subseteq \mathcal{I}^{\Delta}$ and $\cup_{\alpha \in I} \mathcal{P}_\alpha = \mathcal{I}^{\Delta}$. In general, we could create a partition with an arbitrary number of clusters, but, given that the Twitter climate discussion exhibits a natural ideological division into climate believers and climate skeptics, we expect to find a separation of $\mathcal{I}^{\Delta}$ into two clusters.

In order to create a partition of the leading users, we first consider the (temporal) chamber overlaps matrices between all the pairs of leading users, $\boldsymbol{Q}^t = (q_{ij}^t)_{i,j \in \mathcal{I}^{\Delta}(t)}$. By treating $\boldsymbol{Q}^t$ as weighted, undirected, adjacency matrices, we can dispose of a plethora of *community detection* algorithms to detect communities in networks. Thus, we classify the leading users *in an unsupervised way* by first considering the *aggregate* chamber overlap matrix , $\boldsymbol{Q} := \langle \boldsymbol{Q}^t \rangle_t$, where $\langle \cdot \rangle_t$ denotes average over t, and then partitioning $\boldsymbol{Q}$ into *two groups* using the spectral clustering algorithm described by

Mohar et al. [161] (see Appendix B.2 for details). We choose a spectral clustering algorithm because 1) it naturally partitions the network into two distinct groups, and 2) it ranks the nodes in the network (here, the leading users) according to how well-separated they are from the out-group. In general, we could consider other community detection algorithms [37, 187, 173] if we need to partition the network in more than two groups.



(a) Aggregate chamber overlap matrix (b) Aggregate similarity network

Figure 4.4: **Aggregate chamber overlaps between the leading users**. *a)* Aggregate chamber overlap matrix, $\boldsymbol{Q}$, of every leading user pair (see main text for details). White pixels represent leading users that were never present simultaneously during the same week. We order the users in $\boldsymbol{Q}$ according to the rank we obtain with the *unsupervised* spectral clustering algorithm (see Appendix B.2 for details). *b)* Weighted similarity network constructed from $\boldsymbol{Q}$. We color the nodes according to the partition obtained unsupervised spectral clustering. We identify a group of *climate believers* (green) and a group of *climate skeptics* (red) according to the users profiles within those groups.

In Fig. 4.4a, we present the aggregate chamber overlap matrix, $\boldsymbol{Q}$, where we sort the users according to the rank given by the spectral clustering algorithm (see Fig. B.1 in Appendix B.2). We observe a clear block structure that separates the climate skeptics (top left) from the climate believers (bottom right). Notice that the in-block overlaps are often two-digit percentages: a large fraction of the chamber is shared for users in the same community. Moreover, in Fig. 4.4b we show the resulting network constructed with $\boldsymbol{Q}$, where the node colors represent the partition found by the spectral clustering algorithm. We observe two well-separated groups that correspond to the climate believers and climate skeptics discussed before. While in our supervised

user classification we established a third group that included news channels and other media accounts, the leading users *BhadDhad* and *LilNasX* exhibit a significantly low overlap with every other leading user. However, all the news media channels that we classified as neutral - namely *CNN, nytimes, ajplus*, and *nowthisnews* - are statistically indistinguishable from any other climate believer, so the spectral clustering algorithm puts them into the climate believers group. Given this statistical similarity, other clustering algorithms would put news-media channels into the group of climate believers, while probably making isolated clusters for *BhadDhad* and *LilNasX*. We observe four satellite nodes: *BhadDhad*, *narendramodi*, *sunfloweraidi*, and *LilNasX* that have a significantly lower overlap than the rest of the in-group user pairs. In a few cases, there are cross-chamber overlaps that are higher than the average cross-chamber overlap, such as in the pair *cathmckenna*(believer)-*RealSaavedra*(skeptic). However, such cross-chamber overlaps are much smaller than the within-chamber overlaps.

4.2.4.1 Polarization dynamics of the leading users

The previous analysis suggests a big ideological polarization in the climate-related conversation on Twitter, which is a known fact for conversations about climate change [232, 124, 58]. However, our results are valuable in that 1) we detect these groups in an unsupervised way, and 2) we quantify the relative ideological similarity between leading users using the chamber overlaps, where we found significantly high (low) overlaps for pairs of users in the same (other) group. However, we have not yet quantified polarization, and the analysis so far has been static.

We define the *polarization* between two groups P_α and P_β as the difference between the probability that an edge exists within a group and the probability that an edge exist across groups. To do so, we consider the *adaptive E-I index*, following Chen et al. [58] and Bruns [43], as

$$\Phi(t|P_\alpha, P_\beta) = \frac{n^t_{\alpha\alpha} + n^t_{\beta\beta} - (n^t_{\alpha\beta} + n^t_{\beta\alpha})}{n^t_{\alpha\alpha} + n^t_{\beta\beta} + (n^t_{\alpha\beta} + n^t_{\beta\alpha})}, \tag{4.9}$$

where $n^t_{\alpha\beta} = \sum_{i\in P_\alpha, j\in P_\beta} q^t_{ij}$ is the total strength of the edges going from P_α to P_β.

Our measure of polarization is bounded such that $\Phi \in [-1, 1]$, where $\Phi = 1$ when all the connections happen within groups (total assortativity), $\Phi = -1$ when all connections happen between groups (total disassortativity), and $\Phi = 0$ when the connections between groups equal the connection within groups (no assortativity). We acknowledge that Eq. (4.9) is not the only way to quantify polarization. However,

most operational definitions of polarization involve comparing the out-group against the in-group interactions. See Table B.2 in Appendix B.4, where we present different notions of polarization and their mathematical quantification as described in the literature.

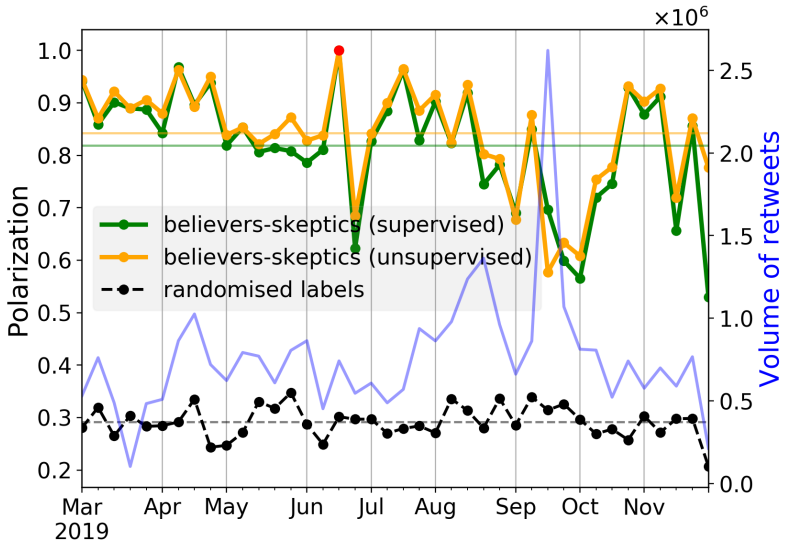


Figure 4.5: **Polarization dynamics** for the weekly chamber overlap matrices, $\boldsymbol{Q}^t$, of the leading users (see Eq. (4.8)) for the supervised partition of climate believers and skeptics (green, $\langle \Phi_{\text{sup}} \rangle = 0.83$), the unsupervised partition using spectral clustering (orange, $\langle \Phi_{\text{unsup}} \rangle = 0.84$) and the average over an ensemble of randomly reshuffled labels over the same networks (black, $\langle \Phi_{\text{null}} \rangle = 0.23$). The red marker indicates a week without leading climate skeptics.

In Fig. 4.5, we show the polarization dynamics of the leading users for both the supervised and unsupervised partitions of climate *believers* and *skeptics* (see Fig. 4.1). As a null model benchmark, we compute the polarization dynamics of the average over an ensemble of the same networks with randomly reshuffled labels.

We find a high polarization of $\Phi_{\text{sup}} = 0.83 \pm 0.11$ and $\Phi_{\text{unsup}} = 0.84 \pm 0.11$ for the supervised and unsupervised partitions, respectively, meaning that users with similar ideologies have similar (climate-related) chambers. Moreover, we show that the polarization dynamics between the supervised and unsupervised approaches are almost identical: even when we tried to separate the news media sources from the climate believers, the unsupervised clustering algorithm puts them in the same group. We find an average polarization of $\Phi_{\text{null}} = 0.23$ for 100 null model realizations of randomly reshuffling the labels. Such a low value indicates that neither the degree sequence nor the size of each ideological group can explain the high polarization of

the empirical network. We find no significant temporal trends in the polarization dynamics but observe that polarization decreases during the 3^{rd} and 4^{th} week of September, which coincides with the #FridaysForFuture biggest strikes organized by Greta Thunberg [143, 214]. We provide a deeper discussion about this and other signals that coincide with the strikes later in Section 4.3.1.

4.3 From chambers to echo chambers

Previously, we introduced the *chamber* as the many-to-many information sources of the *audience* of a *leading user*. In the case of the climate change retweet network, we found that the overlap distribution between all chamber pairs is bimodal (see Fig. 4.3), suggesting that such communication channels are divided, polarizing the climate change retweet network. Thus, we classified the chambers - following an unsupervised clustering approach - according to their overlap similarities and found two well-separated groups which we then identified as climate believers and climate skeptics.

While many authors have studied polarization and echo chambers in social networks, there is not yet a consensus of what an echo chamber is. However, some characteristics of echo chambers are transversal to all its definitions: homophilic interactions drive their formation [233, 99, 60], where actors in the system choose to preferentially connect with each other with the exclusion of outsiders [43], and attitudes and beliefs stay inside groups of like-minded people [95]. In particular for interaction networks in social media, we assume that information flows through the edges of the network and that bimodal structures indicate the presence of echo chambers [60].

Under this framework, the leading users associated with an ideological group (e.g. climate skeptics), together with their audiences and their chambers, *approximate* an *echo chamber*. Our rationale is that information flows through the opposite direction of retweets, as we schematize in Fig. 4.2, and such information flows mostly between high-overlapping chambers. The overlap between two chambers indicates the proportion of their common users. Therefore, a multimodal overlap distribution with high-overlap modes and a low-overlap mode (close to 0) suggests that users choose to preferentially connect with each other (high-overlap modes) *and* exclude outsiders (low-overlap mode). If a network exhibits such characteristics, then the set of leading users should be easily partitioned as described in the previous section, so we thus introduce a definition of an echo chamber.

Definition 4.3.1 (Echo chamber)**.** Given a partition of the leading users, $\mathcal{I}^{\Delta} = \cup_{x \in I} \mathcal{P}_x$, obtained by clustering the chamber overlap matrix, $\boldsymbol{Q} = (q_{ij})_{ij \in \mathcal{I}^{\Delta}}$, with the distribution $p(q)$ containing a well-defined peak near $q = 0$ and other well-defined peak(s) well-separated from $q = 0$, the *echo chamber* of the group $\mathcal{P}_x$ is the union of the leading users, chambers, and audiences associated with $\mathcal{P}_x$. Mathematically,

$$\mathcal{E}_{\mathcal{P}_x} = \bigcup_{i \in \mathcal{P}_x} (\{i\} \cup \mathcal{A}_i \cup \mathcal{C}_i) \tag{4.10}$$

where $\mathcal{A}_i$ is the audience of i and $\mathcal{C}_i$ her chamber.

An echo chamber is an ill-defined concept, so we remark that Def. 4.3.1 is an approximation that works well for retweet networks with multimodal overlap distributions, $p(q)$. In this chapter, we only consider a partition into two groups, but we can generalize this using other clustering algorithms.

Structurally, the climate change retweet network is well-separated, so we claim that we can observe ideological echo chambers. We construct two dynamic echo chambers - per week - based on the leading users, $\mathcal{I}^{\Delta}(t)$, and our partition of climate believers and skeptics. The believers echo chamber, $\mathcal{E}_B$, contains $14.8 \pm 7.8\%$ of the total users per week, while the skeptics echo chamber, $\mathcal{E}_S$, contains $2.6 \pm 2.1\%$. Together, they cover $17.4 \pm 9.9\%$ of the whole retweeting population per week. These echo chambers may have some overlapping users by construction, i.e., users that are classified both as believers and skeptics. However, we find that only $0.3 \pm 0.1\%$ users are in both echo chambers, indicating that the cross-communication between them is orders of magnitude lower[3].

4.3.1 Augmented echo chambers

In the last section, we obtained two echo chambers, $\mathcal{E}_B$ and $\mathcal{E}_S$, that, combined, cover a small minority ($17.4 \pm 9.9\%$) of the whole retweeting population because we only consider the leading users, $\mathcal{I}^{\Delta}(t)$, instead of the set of high-impact users, $\mathcal{I}(t)$. However, we can inspect the audiences of the remaining high-impact users and evaluate if they have a clear ideological position based on $\mathcal{E}_B$ and $\mathcal{E}_S$. To do so, we design an *ideology score*, similar to [73], that determines if the audience of a high-impact user is biased towards being believers, skeptics, or neither of the both. We compute the ideology score, s_i, for each high-impact user i as follows

$$s_i = \frac{n_B^i - n_S^i}{n_B^i + n_S^i}, \tag{4.11}$$

[3]It might be the case that if we inspected the reply network or the followers network instead of the retweet network, the intensity of cross-communication between echo chambers could be higher.

where $n_X^i = |\mathcal{A}_i \cap \mathcal{E}_X|$ is the number of users in the audience of user i that are also in the echo chamber $\mathcal{E}_X$. The score s_i is bounded between -1, when all the users in $\mathcal{A}_i$ are climate skeptics, and 1 when all of them are climate believers.

If the ideology score of a high-impact user is big in magnitude, we may associate her and her audience to an ideological group based on the sign of s_i. In particular, we recognize as climate believers to those high-impact users whose scores are close to 1 and as climate skeptics to those whose scores are close to -1. Thus, we can *augment* the echo chambers to consider the audiences of high-impact users with high-magnitude scores as we describe in the following definition. See Tables B.2 and B.3 in Appendix B.4 where we summarize different notions of polarization and echo chambers as proposed in the literature.

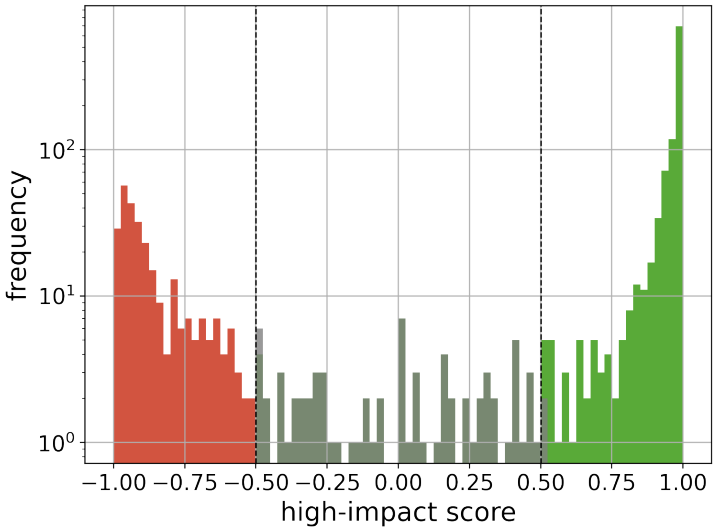


Figure 4.6: **Ideology scores distribution** for the high-impact users, $\mathcal{I}$, where we find a clear bimodal structure (see Eq. (4.11)). We choose a threshold $\eta = 0.5$ for which users with $s_i > \eta$ (green) are assigned as climate believers and those with $-s_i > \eta$ (red) are assigned as climate skeptics. Users with intermediate scores ($s_i \in [-\eta, \eta]$) are not included in the augmented echo chambers.

Definition 4.3.2 (Augmented echo chamber)**.** Given the echo chambers $\mathcal{E}_B$ and $\mathcal{E}_S$ and the collection of high-impact users for which $|s_i| \geq \eta$, with η a user-defined threshold, we define their corresponding *augmented echo chambers*, $\mathcal{E}_B^a$ and $\mathcal{E}_S^a$, as

follows

$$\mathcal{E}_B^a = \bigcup_{i:s_i \geq \eta} (\{i\} \cup \mathcal{A}_i) \cup \mathcal{E}_B, \tag{4.12}$$

$$\mathcal{E}_S^a = \bigcup_{i:-s_i \geq \eta} (\{i\} \cup \mathcal{A}_i) \cup \mathcal{E}_S, \tag{4.13}$$

where $\mathcal{A}_i$ is the audience of i.

In Fig. 4.6, we show the ideology score distribution for the remaining high-impact users, i.e., high-impact users who are not leading users. We find that the distribution is clearly bimodal, where the mode corresponding to climate believers (right) is significantly higher than the mode corresponding to climate skeptics (left). We choose $\eta = 0.5$ as our threshold to decide whether a high-impact user is included in the augmented echo chamber or not. Using this method, of the 1360 remaining high-impact users across all weeks, we classified 280 as climate skeptics, 1078 as climate believers, while 73 users remain unclassified. In addition to augmenting the number of users in the echo chambers, this method enables us to discover the ideological position of high-impact users in an unsupervised way. For instance, we identify users like *@RollingStone* ($s_i = 0.989$) or *@Greenpeace* ($s_i = 0.995$) as climate believers while users like *@DonaldJTrumpJr* ($s_i = -0.93$) or *@GovMikeHuckabee* ($s_i = -0.948$) as climate skeptics solely based on their scores. While we do not analyze users with scores below the threshold η thoroughly, we noticed that low-scoring users either stay neutral with respect to climate change, e.g., *@SkyNews* ($s_i = 0.17$), or are not a strong part of their content agenda, e.g., *@Imamofpeace* ($s_i = 0.03$).

The augmented echo chambers contain $53.1 \pm 9.8\%$ of the total retweeting population, with the error bars representing the standard deviation across weeks. This result is consistent with our initial observation that the top $N = 50$ high-impact users cover near the 50% of the total retweeting population. Moreover, this augmentation increases the sizes of the original echo chambers by a factor of 3.5. More specifically, the believers augmented echo chamber is bigger than its original counterpart by a factor of 3.7 ± 1.6, the skeptics by a factor of 2.6 ± 1.2, and the users in the intersection only by a factor of 1.4 ± 0.4.

In Fig. 4.7, we present the sizes of the augmented echo chambers, the number of users in their intersection and the total number of retweeting users each week. For most of the year, these sizes are stable, with the believers oscillating around 2.4×10^5 users and the users in the intersection oscillating around 1500 users. In contrast, the number of climate skeptics rises significantly from before the Fridays for Future

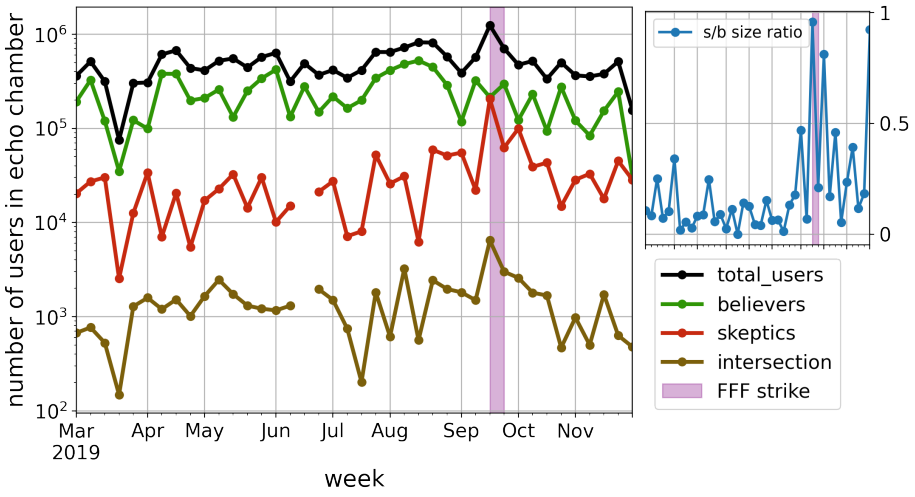


Figure 4.7: **Size of the augmented echo chambers** in logarithmic scale for every week in the dataset. We show, in green and red respectively, the number of believers and skeptics, while in brown we show the number of users in the intersection. We omit a data point in June, 2019 because the skeptics echo chamber, $\mathcal{E}_S$, was empty during that week. In black, we show the total number of users in the retweet network for that week. The inset plot shows the size ratio of skeptics against believers, where we observe a significant increase during late September, 2019, which we suspect was triggered by the Fridays for Future biggest strike [143, 214] (vertical pink interval).

September strikes [143, 214], oscillating from around 2.4×10^4 users, to oscillating around 5.6×10^4 users after the strikes. In the inset of the figure, we show the ratio between the number of skeptics against the number of believers, where we find a consistently higher ratio after the strikes, which we find counterintuitive in that the strikes were pro-climate events.

We believe that the correlation between the climate strikes and the increase in the skeptics' impact is not random. Besides such an increase, we already observed in Figure 4.5 a significant drop in polarization at the dates of the strikes that was not caused by changes in chamber sizes. There are several explanations for such a correlation with size and polarization, ranging from an endogenous social reaction from the climate skeptics to a coordinated response using bots [59], where the bots massively retweeted the skeptic *and* believers leading users during the dates of the strikes. While we cannot establish its causes with this analysis, we find it remarkable that we can observe such a signal from several angles purely from our unsupervised construction of echo chambers.

4.3.2 Renewal of users within ideological groups

Most of the statistical features in our analysis are stable over time, ranging from the bimodal structure of the overlap distribution to the number of users per week (with the exception of the peak in September that coincides with the Fridays for Future strikes), to the Gini index of the users' impact. Yet, Twitter is very dynamic, with users entering and leaving the conversation as different topics emerge and decay over time [65], as is the case of the climate change conversation [68].

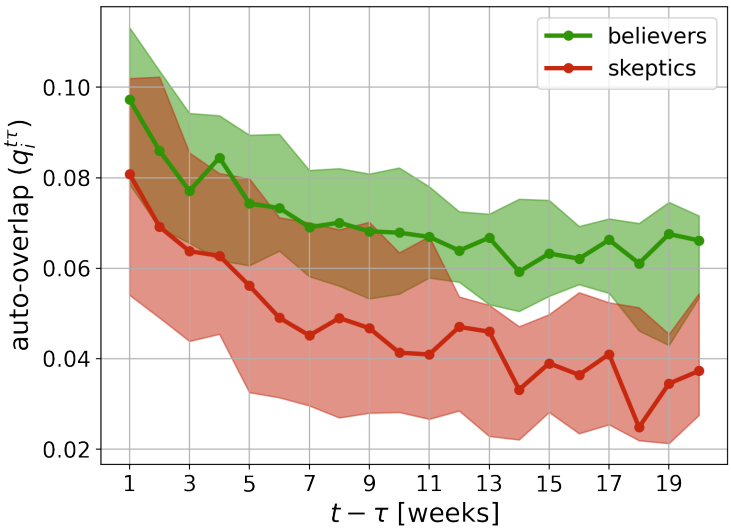


Figure 4.8: **Auto-overlap similarity decay for each augmented echo chamber** as a function of the time difference (in weeks) between them. The solid lines with markers indicate the median overlap similarity for each week while the spread correspond to the $25-75$ quantile spread.

Given these contrasting facts, we test the hypotheses of whether the overall conversation is sustained by the same people over the year or there is a big blackrenewal of users at the micro level that behave consistently at the macro level. To discriminate between these two explanations, we measure the flux of users within augmented echo chambers for different time windows. A small flux of users over time would favor the first hypothesis while a significantly high flux would favor the second one. We quantify such a flux - or renewal - by computing the *auto-overlap* of an augmented echo chamber $\mathcal{E}^a_X(t)$ of ideology group X between weeks t and τ as follows

$$q^{t\tau}_X = \frac{|\mathcal{E}^a_X(t) \cap \mathcal{E}^a_X(\tau)|}{|\mathcal{E}^a_X(t) \cup \mathcal{E}^a_X(\tau)|} \ . \tag{4.14}$$

This measure is similar to the chamber overlap of Eq. (4.8), but instead of comparing the chamber associated with two different leading users, we compare the augmented echo chambers of the same ideological group at different times. A *small* value of $q_X^{t\tau}$ would correspond to a *high flux* of users and viceversa. Note that this measure accounts for users that might have left the conversation from one week to the next, but that might have returned several weeks later. Thus, Eq. (4.14) can capture global cyclic patterns in the echo chambers, if any.

In Fig. 4.8, we show the auto-overlap of each augmented echo chamber as a function of the number of weeks between them. We present such auto-overlaps for the believers and the skeptics separately. In both cases, we find that the number of common users in the chambers decay monotonically with the number of weeks passed. We find that a higher proportion of believers stay in their echo chambers compared to the skeptics. Remarkably, we find that only $9.7 \pm 1.7\%$ of the believers and $8.1 \pm 2.4\%$ of the skeptics remain in the augmented echo chamber from one week to the next, i.e., at $t - \tau = 1$, indicating a massive flow of users in the Twitter conversation within both ideological groups. This result thus supports our second hypothesis: at the micro level, a lot of users enter and leave the conversation from week to week, but, at the macro level, the structural properties of the interaction networks are stable. This result suggests that users *select* which echo-chamber to join rather than *being put* into them, which can be a relevant feature when designing interventions to try to reduce polarization or the spread of misinformation in online social media.

4.4 Conclusion

In this chapter, we identified the echo chambers of the Twitter climate change conversation using *unsupervised* methods. Acknowledging that typically in Twitter, most retweets are given to just a handful of users [101], and assuming that retweets are a good proxy for endorsement [23, 97, 58], we analyzed the temporal structure of the climate-related retweet networks and measured the ideological similarities between *leading users*. We defined a leading user as that who is among the most retweeted users for several weeks throughout the conversation.

We introduced the notion of a leading user's *chamber*, which is the set of users retweeted by her audience. The chambers act as the many-to-many information sources associated with a leading user's audience [145], and we used them to measure the ideological similarities between every pair of leading users by computing their chamber overlap. We found that the chamber overlap distribution is clearly bimodal

with a low-overlap peak and a significantly higher-overlap peak. Thus, we classified the leading users into two groups using an unsupervised spectral clustering algorithm [161] and recognized one group as the *climate believers* and the other as the *climate skeptics*, validating previous findings [232, 221, 58].

Based on the bimodal structure of the chamber overlap distribution, we defined an *echo chamber* as the union of the same group's leading users with their associated audiences and chambers. Such a bimodality implies several features that characterize echo chambers: homophilic interactions drive their formation [233, 99, 60], actors inside them choose to preferentially connect with the exclusion of outsiders [43], and attitudes and beliefs stay inside groups of like-minded people [95]. We identified, on average per week, 15% of the total retweeting population as climate believers, 3% as climate skeptics, and only 0.3% of users classified as both believers and skeptics, suggesting that the cross-communication between echo chambers is negligible.

Furthermore, we designed an *ideology score* that uncovers the ideological position of any high-impact user, even when she has not been observed previously. This score depends on the proportion of her audience that belongs to either of the echo chambers, and we found that the ideology score distribution over all the high-impact users is bimodal, with modes at opposite extremes of the ideological spectrum. This finding reinforces our claim that the users inside the echo chambers are mostly like-minded. We validated the ideological positions of some of the high-impact users discovered with the ideology score. For instance, we uncovered the global campaigning network Green Peace (*@Greenpeace*) as a climate believer, the Republican governor Mike Huckabee (*@GovMikeHuckabee*) as a climate skeptic, and the British news channel Sky News (*@SkyNews*) as a neutral user.

Using the uncovered ideologies, we *augmented* our original echo chambers by combining them with the audiences of the high-impact users that share the same ideological position. Thus, the augmented echo chambers constitute more than half of the weekly retweeting population (46.2% climate believers, 6.5% climate skeptics, and only 0.3% of users classified in both ideological groups). We observed that, in most cases, the number of users in each echo chamber is roughly stable throughout the year. However, we found a strong positive correlation between the dates of the main #FridaysForFuture strikes and the skeptics' echo chamber sizes but not with the believers'. We find it remarkable that we could identify peak activity of certain parts of the population by using completely unsupervised methods. Moreover, we measured the flux of users within echo chambers as a function of time, where we found that most ($> 80\%$) users leave their echo chambers from one week to the next. This result

suggests that the stable properties of the echo chambers are an emergent feature of the system and not imposed by a fixed set of users.

We see several directions for future development of this method. First, we only assessed two ideological positions in our analysis, whereas Twitter discussions may have an arbitrary number of communities. Thus, we could consider more general community detection algorithms, ranging from inference methods able to fix the number of communities [173] to methods that maximize some notion of community and result in optimal partitions with an unsupervised number of clusters [37, 187]. With more than two ideologies, we should generalize our definition ideology score to a multi-class setting by, e.g., measuring the proportion of users in each echo chamber in a higher-dimensional simplex. Second, we do not distinguish between the nature of the Twitter users, so we did not analyze the different effects that validated accounts, non-human accounts, and bots may bring into the conversation. Previous large-scale studies of Twitter data have demonstrated the influence bots can have on the exposure of human accounts to emotional content [31], and how they can coordinate and distort the discussion in controversial conversations [59, 42]. Thus, we could incorporate bot-detection methods and create a taxonomy of users to analyze their effects.

We intentionally focused only on who and not on what the Twitter users consumed to perform our classification. Such approach is limited with respect of supervised approaches. Analyzing the tweets' content would give us more information about the ideological position of its users and help us reinforce the existence of echo chambers [60] and uncover relevant topics of discussions within and between ideological groups [39]. Nevertheless, we see our results as a proof of concept that the echo chamber structure can be observed clearly *even without* looking at the content of tweets. While the ideology of individual users posses a non-negligible degree of uncertainty, at the aggregate level, we clearly observed presence of distinct ideologies each with distinct structural features. In addition, we believe it is important to perform an initial analysis before biasing with external information. Such an initial unsupervised approach could also be useful in indicating starting points of subsequent more complicated analyses.

Overall, we believe our work highlights the importance and usefulness of exploiting the structural information present in social media networks, especially when looking at controversial conversations. Our methodology is computationally cheap, readily usable for other Twitter datasets, and does not suffer from the selection bias of supervised approaches. Furthermore, we showed that if the conversation is polarized

enough, we can identify echo chambers just by looking at the handful of users that produce the most important tweets, which is easier than deploying clustering algorithms on the whole network. From a social point of view, this condition shows that the climate-related Twitter conversation - and possibly most conversations in Twitter [101] - has very low complexity, meaning that we can identify large scale structures within the conversation just from the activity of the few leading users.

Part II

Connecting data with models of complex systems

Chapter 5

Overview & preliminaries

5.1 Introduction

In Part I, we took a *passive* approach for analyzing social complex systems. We aimed to identify nontrivial patterns and features of big, individual-level datasets using tools from complexity science rather than modeling the systems themselves. In this part, we assume that we have an accurate model of a complex system, but we possess data that are aggregated, noisy and incomplete. Our goal here is incorporating these data into the system and inferring what the state of the model should be, i.e., we want to *estimate* the *latent states* of the model that *best* describes the data.

The problem of specifying a model using incomplete information is not new. For instance, estimating the latent states in weather forecasting is an everyday problem. Ever since the sixties [132], researchers have developed tools to make accurate weather forecasts by merging their knowledge on weather physics (via the Navier-Stokes equations) with measurements obtained by satellites and other sensors. The process of merging measurements with a model of a system is known as *data assimilation*. See the reviews of Carrassi [52] and Houtekamer [117] for a historical overview of data assimilation methods. In a nutshell, data assimilation techniques employ several approximations of Bayes theorem to specify the model by combining the likelihood of generating the data given a model and our prior beliefs about the system.

However, estimating the latent state of complex systems has been rarely explored. Most complex systems models are agent-based, meaning that we typically define their basic entities, or *agents*, the states they can occupy, the actions they can perform, and the rules of how these agents interact. Data is rarely precise enough to describe agent-based models in detail, making *data assimilation* a natural candidate for merging our microscopical construction of complex systems with the data available. Previous researchers have already started to incorporate data assimilation techniques

into agent-based models, ranging from crowd and traffic dynamics [226, 61, 120] to epidemiology [62]. In every case, however, the states they estimated were the same ones they were measuring - except for the measurement noise -, so they were not latent. Other researchers have focused in estimating model parameters (rather than the states) using data assimilation, a process known as *calibration* in the agent-based modeling community (see [176] for a review of calibration methods). Most calibration methods make the simplifying assumption that model dynamics fluctuate around some steady state, so that the latent states do not matter [105], where normally they do [111]. Therefore, in this part we aim to expand the literature of latent state estimation for complex systems models.

The remaining of this part is divided as follows. First, we introduce the latent state estimation problem formally in Section 5.2. Then, in Chapter 6, we study the problem of finding the latent initial conditions of high-dimensional systems from sparse, aggregate time series data. We analyze how does the model complexity and the data quality affect our capacity to obtain accurate initial conditions and validate it on chaotic dynamical systems. Finally, in Chapter 7, we propose a data assimilation technique that estimates the latent state of complex systems. We validate it against a high-dimensional chaotic system and an agent-based model of nonlinear opinion dynamics.

5.2 Preliminaries

5.2.1 Latent state estimation problem

Throughout this part, we assume we can model the evolution of a complex system with a simulation model as follows

$$\boldsymbol{x}(t+\Delta t) = \boldsymbol{f}\left(\boldsymbol{x}(t); \boldsymbol{\theta}, \boldsymbol{\xi}(t)\right). \tag{5.1}$$

Here, $\boldsymbol{x}(t) \in \mathcal{X} \subset \mathbb{R}^{N_x}$ represents the (N_x-dimensional) state of the system at time $t \in \mathbb{R}^+$ living in some manifold $\mathcal{X}$, $\boldsymbol{\xi}(t) \in \mathbb{R}^{N_x}$ is a random variable from some distribution p_ξ that models the intrinsic stochasticity of the system, and $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of parameters. We will hereafter refer to the state $\boldsymbol{x}$ as either the *microstate* or the *latent state* to emphasize that $\boldsymbol{x}$ is not directly observable and is potentially high dimensional. In the context of agent-based models, $\boldsymbol{x}$ is a vector encodes the state of every agent. If $\Delta t > 0$ is finite, the model $\boldsymbol{f} : \mathcal{X} \times \mathbb{R}^d \to \mathcal{X}$ describes a discrete mapping that takes the microstate from time t to time $t + \Delta t$. Instead, if Δt

is infinitesimal, the system (5.1) describes a continuous-time dynamical system. We assume that that model is well calibrated, so that we know the parameters $\boldsymbol{\theta}$.

Although we know $\boldsymbol{f}$, we cannot observe sequences of latent states $\boldsymbol{x}_{T:0} = (\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \dots, \boldsymbol{x}_0)$. Instead, we are only able to observe a sequence of observations $\boldsymbol{y} = (y_T, y_{T-1}, \dots, y_1)$ measured at times $\{t_T, t_{T-1}, \dots, t_1\}$, where $y_k \in \mathbb{R}^{N_y}$ and $N_y < N_x$, i.e., we are interested in measurements that destroy information about the latent states by reducing its dimension from N_x to N_y. Moreover, these observations can be noisy. Thus, we can relate the observations to the model (5.1) as

$$y_k = \mathcal{H}\big(\boldsymbol{x}(t_k)\big) + \epsilon_k \ . \tag{5.2}$$

where $\mathcal{H} : \mathcal{X} \subset \mathbb{R}^{N_x} \to \mathbb{R}^{N_y}$ is a known (possibly) nonlinear *observation operator* and ϵ_k is a random variable from some distribution p_ϵ that accounts for *observational noise*.

Further, we assume that the observations are sampled at a uniform rate as such that

$$\Delta t_k = t_{k+1} - t_k = m\Delta t \ ,$$

where the positive integer m is the *sampling interval*[1]. We treat m as a parameter that controls how often we sample observations from the system. Thus, we can recast the model (5.1) as a discrete-time mapping

$$\boldsymbol{x}_{k+1} = \boldsymbol{\mathcal{M}}(\boldsymbol{x}_k) := \boldsymbol{f}^m(\boldsymbol{x}; \boldsymbol{\xi}_k) = \boldsymbol{f}^{mk}(\boldsymbol{x}_T; \boldsymbol{\xi}_{T:k}) \ , \tag{5.3}$$

where $\boldsymbol{f}^i(\cdot)$ is the composition of $\boldsymbol{f}$ with itself i-times, $\boldsymbol{x}_T = \boldsymbol{x}(t_T)$ the latent state at the start of the time assimilation window, and $\boldsymbol{\xi}_{T:k}$ the sequence of noise realizations from the distribution p_ξ. We introduce the notation $\boldsymbol{\mathcal{M}}(\cdot)$ to emphasize its difference from the model $\boldsymbol{f}$: the latter takes the state $\boldsymbol{x}(t_k)$ to $\boldsymbol{x}(t_k + \Delta t)$ while the former takes the state from the observation time stamp t_k to the next time stamp, t_{k+1}.

If the map (5.3) depends on the noise $\boldsymbol{\xi}$ we say the the dynamical system is *stochastic*, and we cannot longer determine their trajectories based on the single initial condition $\boldsymbol{x}_0$. Rather, each iteration of the mapping is determined by their *transition probabilities* $p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)$ which depends on the distribution p_ξ.

The *dynamical model* (5.1) and the *sequence of observations* $\boldsymbol{y}$ are two complementary but incomplete sources of information of the system. Data assimilation (DA) tackles the problem of estimating the latent state $\boldsymbol{x}$ based on $\boldsymbol{y}$.

[1]The procedures we introduce in the following chapters work with irregularly sampled observations as well.

5.2.2 Bayesian inference formulation

The random nature of both model stochasticity and observational noise motivates us the formulate the latent state estimation problem from a Bayesian perspective [52]. Bayes' rule states that

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}, \tag{5.4}$$

where $p(\boldsymbol{x}|\boldsymbol{y})$ is the probability density function (pdf) of the unknown underlying process $\boldsymbol{x}$ conditional on the collection of observations $\boldsymbol{y}$. The RHS of Eq. (5.4) contains the two main ingredients of data assimilation: 1) the *prior* $p(\boldsymbol{x})$, that incorporates whatever information we have about $\boldsymbol{x}$ before including the data $\boldsymbol{y}$, and 2) the *likelihood* $p(\boldsymbol{y}|\boldsymbol{x})$, that updates our knowledge of $\boldsymbol{x}$ given the observed data. The distribution $p(\boldsymbol{y}) = \int_{\mathcal{X}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ is independent of $\boldsymbol{x}$, so we treat it as a normalizing constant that assures that $p(\boldsymbol{x}|\boldsymbol{y})$ is an actual probability distribution.

5.2.2.1 General assumptions

Bayes rule is a general solution to the state estimation problem. However, it is extremely difficult to solve explicitly: we need to know the likelihood, prior and evidence distribution for the entirety of the sequences $\boldsymbol{x} = (\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_0)$ and $\boldsymbol{y} = (y_T, y_{T-1}, \ldots, y_1)$, which we rarely possess. To make the problem more manageable, the two following assumptions are often made on the data assimilation literature:

1. The random variables ξ_k and ϵ_k are independent and identically distributed (i.i.d.) in time. This is, given a sequence of latent states $\boldsymbol{x}_{T:0} := (\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_0)$ and observations $y_{T:1} := (y_T, y_{T-1}, \ldots, y_1)$ within the time interval $[t_0, t_K]$, we can rewrite the likelihood as

$$p(y_{T:1}|\boldsymbol{x}_{T:0}) = \prod_{k=1}^{T} p(y_k|\boldsymbol{x}_k) \ . \tag{5.5}$$

   Additionally, we assume that ξ_k and ϵ_k are mutually independent, so that $\mathbb{E}(\xi_k, \epsilon_k) = 0$ for all $k \in \{0, \ldots, T\}$.

2. The model (5.1) is a Markov process, meaning that the state $\boldsymbol{x}_k$ at time t_k conditioned in all its history $t < t_k$ depends only on the state at the most recent time t_{k-1}. Thus, we can factorize the prior of the sequence $\boldsymbol{x}_{T:0}$ as

$$p(\boldsymbol{x}_{T:0}) = p(\boldsymbol{x}_0) \prod_{k=1}^{T} p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \ . \tag{5.6}$$

Incorporating the assumptions of Eqs. (5.5) and (5.6) into Bayes' rule, we can rewrite the posterior distribution as

$$p(\boldsymbol{x}_{T:0}|y_{T:1}) = \frac{p(\boldsymbol{x}_0)\prod_{k=1}^{T} p(y_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})}{p(\boldsymbol{y}_{T:1})} \ . \tag{5.7}$$

The main advantage of this expression is that we can incorporate each observation in $\boldsymbol{y}$ independently to infer the latent states of the model. In the following chapters, we will start our analysis from Eq. (5.7) to construct suitable methods for *initializing* and *estimating* the latent states of complex systems from aggregate observations.

Chapter 6

Estimating initial conditions from incomplete information

6.1 Introduction

In the overview (see Chapter 5), we discussed the lack of methods to recover the unobservable, or *latent*, variables of complex systems models from incomplete data. We often model empirical processes as complex systems: they are composed of different entities that interact through simple rules but evolve in nontrivial ways. Applications of complex systems models range from opinion dynamics [181], to urban dynamics [226], to the human brain [46], to financial markets [86], to occupation and automation dynamics [72], to marine fisheries [100].

In this chapter, we study the interplay between the complexity of these kind of models and the information that the available data provides about them. More specifically, we focus on estimating the initial conditions that best describe the available data. Estimating and forecasting these systems accurately depend on 1) their inherent complexity, which depends on things like the systems' state space dimension, Lyapunov exponents, and attractors, 2) the sparsity and quality of the data, and 3) our ability to model them.

In low-dimensional systems with high-quality data, state-space attractor reconstruction techniques have succeeded as Packard *et al.* [170] and Takens [212] showed in their seminal papers. By reconstructing an attractor from data, we can make accurate predictions by choosing its closest points to the current state of the system and extrapolating them [84]. These techniques work well even without any modeling [236].

In high-dimensional systems, it is typically not possible to use time series models. It is nonetheless often possible to use a theoretical model, if one can only measure the

initial conditions that the model requires and compare them to the data. The task of estimating initial conditions that match observations is known as *initialization*. Similar to the state-space reconstruction techniques, we require to know the evolution function of the underlying dynamics of the system, or at least some approximation of it [85]. In particular, if the observations available are an aggregate of the dynamical system, then the process is known as *microstate initialization*, or latent state initialization. To do this we need to know how the microstate is aggregated, in addition to having a model of its dynamics.

In the fields of meteorology and numerical weather prediction (NWP), researchers have developed a framework for estimating the latent states of a system [175, 52], where a large number of observed states are available. This framework is known as data assimilation, and, although it is well justified from the Bayesian perspective, it incorporates several heuristics pertinent to the weather prediction field. For instance, modelers often incorporate Gaussian priors in the cost functionals involved with known statistics about the latent states [164]. Ideas from data assimilation have already permeated outside the NWP community, such as in urban dynamics [226]. Typically, data assimilation methods operate in a sequential matter: the microstate at time t gets nudged (or corrected) so it optimally approaches the empirical observation at t. Then, the modeler simulates the nudged microstate from time t to time $t + 1$, where the microstate is again nudged. Therefore, the resulting sequence of microstates is not a solution of the underlying dynamical system - the microstate was constantly modified throughout its trajectory. We, in contrast, seek to analyze real trajectories of the latent states, so this sequential approach does not suit our goal.

The initialization process is an optimization problem where we minimize a cost function that depends on some notion of distance between the observed and the model-generated data. Hence, in high-dimensional systems, it is essential to develop efficient algorithms to find initial conditions with high precision. Research has been done around the parameter estimation of stochastic dynamical systems in low-dimensional parameter spaces [105, 176]. Among other alternatives, gradient descent has proven to be superior in strongly nonlinear models [82, 130], but the main drawback is that it can get stuck in local minima. Other methods, like genetic algorithms, simulated annealing, and other meta-heuristics algorithms [235] usually find the global minimum in low dimensions, but they are likely to diverge in high-dimensional systems. We provide a thorough comparison of the state-of-the-art gradient descent methods [189] and discuss their performance in the context of microstate initialization.

We propose a gradient-based method for initializing chaotic systems where only aggregate observations of short observation windows are available. Using a combination of numerical simulations and analytical arguments, we study the conditions in which certain systems may be initialized with arbitrary precision. We explore the performance of several gradient descent algorithms and the effects of observational noise on the accuracy and convergence of the initialization process. Furthermore, we quantify the accuracy of our method numerically with out-of-sample forecasts. Under this framework, we offer a better understanding of what information the observations provide about a system's underlying dynamics, and, additionally, lay out the connections of our method to those in the data assimilation literature.

The remainder of this chapter is laid out as follows: In Section 6.2, we describe the initialization problem under the framework of dynamical systems and develop our initialization methodology accordingly. In Section 6.3, we test our method in two systems, namely the Lorenz and Mackey-Glass systems. Finally, in Section 6.4, we discuss our results and suggest further research directions.

6.2 Initialization procedure

Throughout this chapter, we assume we have a dynamical system, $\boldsymbol{\mathcal{M}}$, that models the time evolution of a state variable, $\boldsymbol{x} \in \mathbb{R}^{N_x}$, to an arbitrary time into the future. Additionally, we assume we possess a sequence of observations, $\boldsymbol{y} = (y_T, \ldots, y_0)$, where $y_i \in \mathbb{R}$ corresponds to the observation at time t_i. The state variable is mapped to the (lower-dimensional) space of the observations through an observation operator, $\mathcal{H}$. The *initialization problem* is that of obtaining the best representation of the present-time microstate $\boldsymbol{x}_0$ using the information of the past observations. We define what *best* means in terms of the Bayesian construction we did in Section 5.2.2.

We assume that the observations $\boldsymbol{y}$ are noisy, but that the underlying dynamics is deterministic, and our model perfectly specifies the system. In particular, we consider unimodal noise distributions, p_ϵ, with a well-defined variance, σ_y^2. Thus, we can approximate p_ϵ with a zero mean Gaussian distribution and variance σ_y^2. Under these assumptions, we can recast the posterior density (5.7) as

$$p(\boldsymbol{x}|y_{T:1}) = \frac{p(\boldsymbol{x})}{p(y_{T:1})} \frac{1}{\sqrt{2\pi}\sigma_y} \prod_{k=1}^{T} \exp\left(-\frac{(y_k - \hat{y}_k)^2}{2\pi\sigma_y^2} \right) , \tag{6.1}$$

where

$$\hat{y}_k(\boldsymbol{x}) = \mathcal{H}\Big(\boldsymbol{\mathcal{M}}^k(\boldsymbol{x})\Big) \tag{6.2}$$

is our estimate of observation y_k given $\boldsymbol{x}$. Now, the posterior only depends on the initial condition $\boldsymbol{x}$ rather than the whole sequence $\boldsymbol{x}_{T:0}$.

We thus consider that the best representation of $\boldsymbol{x}$ is the one that maximizes Eq. (6.1). Taking logarithms and rearranging terms, we can state the maximum posterior problem as

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \left[\sum_{k=1}^{T} (y_k - \hat{y}_k)^2 - \log p(\boldsymbol{x}) + K \right] , \tag{6.3}$$

where K is a constant. We assume that we have no available information about the latent states, so we set a uniform prior, so that $p(\boldsymbol{x})$ is also constant.

Following this analysis, we propose the following mean-squared *cost function* [1]:

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{T\sigma_y^2} \sum_{k=1}^{T} (y_k - \hat{y}_k)^2 , \tag{6.4}$$

where $(T\sigma_y)^{-1}$ is a normalizing constant that is arbitrary but will be of practical use later. We call the state $\hat{\boldsymbol{x}}_{-T}$ that minimizes $\mathcal{J}$ the *assimilated microstate* and the present-time state $\hat{\boldsymbol{x}}_0 = \boldsymbol{\mathcal{M}}^T(\hat{\boldsymbol{x}}_{-T})$ the *initialized microstate*. Our goal is to find an initialized microstate $\hat{\boldsymbol{x}}_0$ that is a good representation of the ground-truth microstate $\boldsymbol{x}_0$ [2].

The cost function $\mathcal{J}$ is known as a *filter* in the interpolation assimilation community, a *least-squares optimizer* in the optimization and machine learning communities, and *4D-Var* with infinite state uncertainty in the variational assimilation community. From a Bayesian perspective, the cost function $\mathcal{J}$ emerges naturally when we assume the prior $p(\boldsymbol{x})$ is uniform and observational noise is Gaussian: the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ is maximal when $\mathcal{J}$ is minimal [52].

If the observations in the time series $\boldsymbol{y}$ are noiseless, and given that we assume that the model $\boldsymbol{\mathcal{M}}$ perfectly describes the system, then $\mathcal{J}$ has a global minimum $\hat{\boldsymbol{x}}_{-T}$ for which $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) = 0$. However, even for the noiseless scenario, $\hat{\boldsymbol{x}}_{-T}$ is not necessarily unique. Take, for instance, the Lorenz system [150], which is symmetric around its z-axis, and take an observation operator of the form $\mathcal{H}(\boldsymbol{x}) = \boldsymbol{x}^\dagger\boldsymbol{x}$ with $\dagger$ denoting matrix transposition. Note that for any microstate $\boldsymbol{x} = (x, y, z)$ the microstate

[1] We may generalize this cost function for non-scalar observations as

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{T} \sum_k (\boldsymbol{y}_k - \hat{y}_k)^\dagger \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y}_k - \hat{y}_k) ,$$

where $\boldsymbol{\Sigma}_y$ is the covariance matrix of the observations.

[2] If we change our assumptions about the likelihood p_ϵ or the prior $p(\boldsymbol{x})$, then the cost function $\mathcal{J}$ would change accordingly. However, the methodology that we propose in what follows would not be significantly affected.

$(-x, -y, z)$ results in the same sequence observations, so (x, y, z) and $(-x, -y, z)$ are indistinguishable in terms of $\mathcal{J}$. We will refer to the set of indistinguishable microstates as the *feasible set* of solutions, and we will denote it as $\mathbf{\Omega}$.

An additional problem arises when the observations contain noise. In any finite time series, there might exist microstates with a lower value of the cost function than the ground-truth microstate - i.e., where $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) < \mathcal{J}(\boldsymbol{x}_{-T})$ for $\hat{\boldsymbol{x}}_{-T} \notin \mathbf{\Omega}$ - which we will call *dominating microstates*, following [129]. Judd *et al.* [129] suggest that using cost functions that minimize both the variance (as in Eq. (6.4)) and the kurtosis will do a better job of identifying the true microstate when the observational noise is Gaussian. While this is a good idea when there are many observations, we consider short time series in this work, which calls for other alternatives. Instead of modifying the cost function $\mathcal{J}$, we pre-process the data to reduce the probability of finding dominating trajectories outside of the feasible set $\mathbf{\Omega}$.

Following [175], we distinguish between the *error-free* and the *noise* contributions to the cost function $\mathcal{J}$. Isolating the contributions from the discrepancy between ground-truth and assimilated microstate and the observational noise will let us design a well-suited methodology to deal with noise and dominating trajectories in a separate manner. Recall that, given $\boldsymbol{x}_{-T}$, $\mathcal{H}(\boldsymbol{x}_k)$ is the error-free observation at time t_k, ϵ_k its associated noise and $\hat{y}_k$ our estimation of y_k. Thus, we can decompose Eq. (6.4) into three terms of the form

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{\sigma_y^2}\Big[\underbrace{\frac{1}{T}\sum_k (\mathcal{H}(\boldsymbol{x}_k) - \hat{y}_k)^2}_{\text{noise-free cost}} + \underbrace{\frac{1}{T}\sum_k \epsilon_k^2}_{\text{observational noise}} - \underbrace{\frac{2}{T}\sum_k (\mathcal{H}(\boldsymbol{x}_k) - \hat{y}_k)\epsilon_k}_{\text{error-noise covariates}}\Big]$$

The first term on the RHS is the noise-free cost, $\mathcal{J}_{\text{free}}(\boldsymbol{x})$, that we would obtain in the absence of observational error. The second term is the average contribution of the square of the noise, which converges to the noise variance, σ_n^2, for large T. Finally, the third term captures how the noise and the noise-free cost vary together, which goes to 0 for large T because we assume the observational noise and the system dynamics are uncorrelated. Considering a large number of observations, we can thus approximate

$\mathcal{J}$ as

$$\mathcal{J}(\boldsymbol{x}) \approx \mathcal{J}_{\text{free}}(\boldsymbol{x}) + \frac{\sigma_n^2}{\sigma_y^2} \,, \tag{6.5}$$

which shows the expected behavior of Eq. (6.4) in the presence of noise.

Now, note that if we take $\hat{y}_k = \mathbb{E}[\boldsymbol{y}]$ for all k [84], which is the best constant predictor for the observed time series, then $\mathcal{J}_{\text{free}} = 1$ and, therefore

$$\mathcal{J}(\boldsymbol{x} : \hat{y}_k = \mathbb{E}[\boldsymbol{y}] \ \forall k) = \mathcal{J}_{\text{const}} := 1 + \sigma_n^2/\sigma_y^2.$$

Naturally, we want to find an assimilated microstate $\hat{\boldsymbol{x}}_{-T}$ that performs better than a constant predictor, so that $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) \leq \mathcal{J}_{\text{const}}$. This motivates us to consider microstates $\boldsymbol{x}$ such as

$$\{\boldsymbol{x} : \mathcal{J}(\boldsymbol{x}) \leq \alpha + \frac{\sigma_n^2}{\sigma_y^2}\beta\} \,, \tag{6.6}$$

with α and β hyperparameters such that $\alpha \in (0, 1]$ accounts for the error-free tolerance about the global minimum while $\beta \in (0, 1]$ accounts for the noise tolerance. We explore these hyperparameters in the following sections.

The above discussion suggests that we should pay especial attention on handling dominating trajectories, which might be present by either the presence of observational noise or by microstates that live far away from the ground truth but that have low cost function values. Thus, we propose the following three stages to initialize the microstate from aggregate observations: 1) a *preprocessing* stage in which we reduce the noise of the observations, 2) a *bounding* stage in which we limit the region of the microstate space in which we search for an optimal solution, and 3) a *refinement* stage in which we minimize the cost function (6.4) in a small search space and estimate the optimal microstate given the observations.

6.2.1 Preprocess: noise reduction

First, we preprocess the observed time series to reduce the observational noise and thus lower the probability of obtaining dominating microstates. Casdagli *et al.* [54] showed that in the presence of observational noise, the distribution of local minima gets increasingly complex with increasing levels of noise, especially when the dynamics is chaotic. We handle time series with only a handful of data points (in the order of 100 data points or less), so reducing noise by orbit shadowing [85] or large window impulse response filters [107] are not suitable options. Instead, we find that the best

way to reduce the variance of the noise is using a low-pass moving average (LPMA) filter,

$$z_k := \begin{cases} \frac{1}{2}y_k + \frac{1}{2}y_{k+1} & k = -T \\ \frac{1}{2}y_k + \frac{1}{2}y_{k-1} & k = 0 \\ \frac{1}{2}y_k + \frac{1}{4}\left(y_{k-1} + y_{k+1}\right) & \text{otherwise} \end{cases}, \tag{6.7}$$

where z_k is the filtered data point at time t_k. We control the amount of noise reduction by repeatedly feeding the signal back into the LPMA filter of Eq. (6.7). Feeding the signal back into the filter q times, hereafter denoted as z_k^q, is equivalent to increasing the filtering window from three to $2q+1$ points, making the filtered signal smoother.

We expect for the resulting variance, $(\sigma_n^q)^2$, to be lower than the original noise variance σ_n^2. Thus, following [104] and assuming we can rewrite the filtered signal in terms of the microstates as $z_k^q = \mathcal{H}(\boldsymbol{x}_k) + \epsilon_k^q$, we can measure the performance of the LPMA filter with the increase of the signal-to-noise ratio

$$r_0 = \sqrt{\frac{\sum_k (\epsilon_k)^2}{\sum_k (\epsilon_k^q)^2}} \approx \frac{\sigma_n}{\sigma_n^q}, \tag{6.8}$$

where $r_0 > 1$ whenever $\sigma_n > \sigma_n^q$.

The resulting noise distribution of the filtered signal converges to a zero-mean Gaussian distribution if we have either many data samples or set q big enough. However, if q is too big, we may filter parts of the dynamics and mix them into noise, resulting in exotic noise distributions (see Fig. C.5 in Appendix C.3 for examples). What *big enough*, *many samples* and *too big* mean depend heavily on the dynamical system and the noise distribution, although we stress that the LPMA filter works optimally when the noise distribution has higher frequency spectrum on average than that of the dynamics of the system (see [163], Chapter 6.4.3.1.).

6.2.2 Bound: exploring the attractor

After the preprocessing stage, the next step is to bound the search space. Under no constraints in the cost function, the initialization procedure consists on searching through the whole microstate space $\mathcal{X}$ for a set of microstates that minimize Eq. (6.4), with no prior preference on where to start the search from. However, many real world systems are dissipative, meaning that their dynamics relax into an attractor, i.e. a subset manifold $\mathcal{S} \subset \mathcal{X}$ of the microstate space. Thus, it is safe to assume that the observations $\boldsymbol{y}$ derive from a sequence of microstates that live in or near $\mathcal{S}$, and, by the properties of dissipative systems, any microstate $\boldsymbol{x}$ in the basin of attraction will eventually visit every point in $\mathcal{S}$ [16]. This means that, if we wait for long enough,

then any point in the basin of attraction will get arbitrarily close to the ground truth microstate.

The bounding stage consists of exploiting the dissipative nature of real world systems and letting any arbitrary estimate of the microstate explore the basin of attraction until it *roughly* approaches the ground truth microstate. To be more precise, we say that the microstate $\boldsymbol{x}_{-T}^R \in \mathcal{X}$ *roughly approaches* $\boldsymbol{x}_{-T}$ if

$$\mathcal{J}(\boldsymbol{x}_{-T}^R) \leq \delta_R := \alpha_R + \frac{\sigma_n^2}{\sigma_y^2}\beta_R, \tag{6.9}$$

for some *rough threshold* $0 \ll \delta_R < 1$, where δ_R is determined by the hyperparameters α_R and β_R.. Thus, we let an arbitrary microstate evolve according to the model $\boldsymbol{\mathcal{M}}$ until Eq. (6.9) is satisfied.

At this stage, we want to obtain solutions with a cost value of the order of the unfiltered noise level σ_n^2/σ_y^2, so that $\boldsymbol{x}_{-T}^R$ is either near to the feasible set $\boldsymbol{\Omega}$ or to any of the dominating microstates driven by the noise. To achieve this, it suffices to set β_R close to unity and $0 \ll \alpha_R < \sigma_n^2/\sigma_y^2 \leq 1$. We find that our methodology is robust to the specific choice of these hyperparameters as long as they are of the order we propose (see Figure C.9 in Appendix C.3).

We note that whenever the time series $\boldsymbol{y}$ is noiseless, the bounding stage is only driven by α_R, the error-free tolerance of the points situated at the global minima of the cost function, which are exactly those in the feasible set $\boldsymbol{\Omega}$. Thus, our choice of α_R leverages how closely we approach to $\boldsymbol{\Omega}$. If we set α_R too close to 0, we would impose for $\hat{\boldsymbol{x}}_{-T}$ to lay near $\boldsymbol{\Omega}$; however, it would take too long simulation times to satisfy Eq. (6.9) for this approach to be practical. The idea, ultimately, is to set the lowest δ_R possible such that the time to satisfy Eq. (6.9) is *short*.

6.2.3 Refine: cost minimization

The final step of the initialization procedure is refining $\boldsymbol{x}_{-T}^R$. By this point, we expect that $\boldsymbol{x}_{-T}^R$, our estimate of $\boldsymbol{x}_{-T}$, has bypassed most of the high-valued local minima of the cost function landscape. Additionally, we preprocessed the observations $\boldsymbol{y}$ to reduce their observational noise, but we have not fully exploited such preprocessing yet. By reducing the variance of the observational noise, we lower the number of dominating trajectories of the cost function - i.e., trajectories for $\boldsymbol{x} \notin \boldsymbol{\Omega}$ such that $\mathcal{J}(\boldsymbol{x}) < \mathcal{J}(\boldsymbol{x}_{-T})$. Thus, starting from $\boldsymbol{x}_{-T}^R$, we can minimize $\mathcal{J}$ using any optimization scheme until

$$\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) \leq \delta_r := \alpha_r + \frac{\sigma_n^2}{\sigma_y^2}\beta_r \tag{6.10}$$

for some *refinement threshold* $0 < \delta_r \ll 1$ that is determined by the hyperparameters α_r and β_r, with $\hat{\boldsymbol{x}}_{-T}$ the assimilated microstate of the system. We then define $\hat{\boldsymbol{x}}_0 = \boldsymbol{\mathcal{M}}^T(\hat{\boldsymbol{x}}_{-T})$ to be the initialized microstate, hoping that $\hat{\boldsymbol{x}}_0$ is a good representation of $\boldsymbol{x}_0$.

We want for $\hat{\boldsymbol{x}}_{-T}$ to have *the lowest cost possible* at this stage. In terms of the error-free tolerance, we look for $\alpha_r \ll \alpha_R < \sigma_n^2/\sigma_y^2 \leq 1$ - i.e., as close to 0 as possible - but the actual magnitude of α_r is left for the modeler to choose. Regarding the contribution of the noise, recall from Eq. (6.8) that $r_0 \approx \sigma_n/\sigma_n^q > 1$, so the lowest *expected* cost is $\sigma_n^2/\sigma_y^2 r_0^{-2}$ (see Eq. (6.5)). Thus, we define the *refinement bound* of our initialization procedure as that of setting the hyperparameters $0 < \alpha_r \ll \alpha_R$ and β_r of the order of r_0^{-2}.

For our optimization scheme we explore a plethora of the most successful *gradient-based algorithms* in the literature [189]. These algorithms include Stochastic Gradient Descent [185], Momentum Descent [177], Nesterov [165], Adagrad [77], Adadelta [238], Rmsprop [217], Adam [135], and AdamX [182] and YamAdam [234].

In all cases we set the hyperparameters of the descent algorithms to those given in the literature. We then compute the gradient of $\mathcal{J}$ using centered finite differences of step size $\sqrt{\varepsilon_M} \approx 1.5 \times 10^{-8}$, where ε_M corresponded to double precision arithmetic. In the absence of observational noise, provided that the dynamics is not degenerate and that we have sufficient observations, we expect that the feasible set $\boldsymbol{\Omega}$ collapses to the ground-truth microstate only, so, at this stage, we expect to infer it with high precision from the data.

6.2.4 Validation

We validate the initialized microstate $\hat{\boldsymbol{x}}_0$ by comparing them with the present-time microstate $\boldsymbol{x}_0$ and making out-of-sample predictions of the observed time series. Recall that we take the convention in which the observations, $\boldsymbol{y} = (y_{-T}, \ldots, y_0)$, have non-positive time indexes, so we refer to the observation times t_k for $k < 0$ as *assimilative* while we refer to times for $k \geq 0$ as *predictive*. We measure the *discrepancy* between the real and the simulated observations using both the normalized squared error in the observation space and the normalized error in the model space, i.e.

$$\mathrm{NSE}_k^{obs} = \frac{(y_k - \hat{y}_k)^2}{\sigma_y^2}, \tag{6.11}$$

$$\mathrm{NSE}_k^{mod} = \frac{1}{N_x}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)^\dagger \Sigma_{\boldsymbol{x}}^{-1}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \tag{6.12}$$

where σ_y^2 is the variance of the data, $\Sigma_{\boldsymbol{x}}$ the covariance matrix of the microstates and $(\cdot)^\dagger$ denotes matrix transposition. In general, $\boldsymbol{x}_k$ and $\Sigma_{\boldsymbol{x}}$ are unknown to the modeller, but we use them to measure the performance of our initializations in the latent space of the microstates.

Note that if we let $T \to \infty$, then $\mathcal{J}(\boldsymbol{x})$ converges to $\mathbb{E}[NSE_0^{obs}]$. When k grows, the trajectories y_k and $\hat{y}_k$ diverge exponentially until they lose all memory about their initial conditions ($\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}_0}$ respectively). Given that such divergence is exponential, we therefore take the *median* of the NSE when comparing the performance over an ensemble of experiments as a better alternative to the mean.

On the same note, another way to validate the inferred microstate is by looking for how long our predictions accurately describe the system. Chaotic systems are by definition sensitive to initial conditions, meaning that microstates that are close to each other diverge exponentially over time. If these microstates live in a chaotic attractor, the distance between them is bounded by the size of the attractor [175]. Thus, we can assess the quality of our predictions by measuring how long the real and simulated observations retain memory about each other [84]. We refer to this limiting time as the *predictability horizon* of the system k_{max}, and we define it as the average number of steps before the separation between y_k and $\hat{y}_k$ is greater than the distance between two random points in the attractor of the system. Mathematically,

$$k_{max} := \mathbb{E}\Big[\arg\min_{k\geq 0}\big\{(y_k - \hat{y}_k)^2 \geq \mathcal{D}_{\mathcal{S}}\big\}\Big],$$

where $\mathcal{D}_{\mathcal{S}}$ is average squared distance between two random points in the attractor $\mathcal{S}$ normalized by the variance of the attractor. More specifically, if X, Y are two i.i.d. random variables such that $X \sim \mu(\mathcal{S})$ with $\mu(\cdot)$ denoting the natural measure, then

$$\mathcal{D}_{\mathcal{S}} := \frac{\mathbb{E}[\|X - Y\|^2]}{Var(X)} = 2.$$

Thus, given that $\mathbb{E}[\sigma_y] = Var(X)$, we can simplify k_{max} into an expression that only depends on Eq. (6.11) so that

$$k_{max} = \mathbb{E}\Big[\arg\min_{k\geq 0}\big\{\mathrm{NSE}_k \geq 2\big\}\Big]\ , \tag{6.13}$$

which is the formula we use to compute k_{max}.

Finally, we benchmark the inferred microstate by comparing k_{max} with a measure of the natural rate of divergence of the dynamics. As a measure of this, we use the *Lyapunov 10-fold time* t_λ, which indicates the average time for two neighboring

microstates to diverge from each other by one order of magnitude [19]. This is defined in terms of the inverse of the maximum Lyapunov exponent λ as

$$t_\lambda = \frac{\ln 10}{m\Delta t}\lambda^{-1}, \tag{6.14}$$

where the factor $\ln 10/(m\Delta t)$ lets us interpret t_λ in the units of the number of observations after which on average the dynamics causes the loss of an order of magnitude of precision. We obtain λ numerically using the two-particle method from Benettin *et al.* [28].

6.2.5 Initialization procedure summary

We summarize our *microstate initialization procedure* in the following steps:

1. **Preprocess:** Smooth the observed time series using the LMPA filter (see Eq. (6.7)) or any other suitable noise reduction technique. The smoothed signal will have fewer dominating trajectories [129] and a simpler distribution of local minima than the full noisy signal [54].

2. **Bound:** Make an arbitrary guess $\boldsymbol{x} \in \mathcal{X}$ of the microstate, and let it evolve under the model $\boldsymbol{\mathcal{M}}$ until the microstate *roughly approaches* the smoothed observations, i.e., until $\mathcal{J}(\boldsymbol{x}^R_{-T}) \leq \delta_R$ for $\boldsymbol{x}^R_{-T} = \boldsymbol{\mathcal{M}}^R(\boldsymbol{x})$ for some $R \geq 0$. (see Eq. (6.9)). If several attractors exist, make one arbitrary guess for each of the different basins of attraction in the system.

3. **Refine:** Minimize the cost function $\mathcal{J}$ starting from $\boldsymbol{x}^R_{-T}$ using Adam gradient descent or any other suitable optimization scheme until $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) \leq \delta_r$ (see Eq. (6.10)) and call $\hat{\boldsymbol{x}}_{-T}$ the *assimilated microstate* and $\hat{\boldsymbol{x}}_0 = \boldsymbol{\mathcal{M}}^T(\hat{\boldsymbol{x}}_{-T})$ the *initialized microstate*.

4. **Validate:** Compute the discrepancy between the real and simulated observations (see Eq. (6.11)) and the predictability horizon k_{max} (see Eq. (6.13)) on out-of-sample predictions of the system to evaluate the quality of the initialized microstate $\hat{\boldsymbol{x}}_0$. If possible, benchmark the predictability horizon with the Lyapunov 10-fold time (see Eq. (6.14)) of the system considered.

6.3 Results

In this section, we test the microstate initialization procedure on two paradigmatic chaotic systems: the well-known Lorenz system [150] and the high-dimensional Mackey-Glass system [152]. We approximate both systems using numerical integrators (described in each section that follows), and we take the approximated system as the real dynamical system. In all cases, we sample the ground-truth microstate $\boldsymbol{x}_{-T}$ from the attractor of the system considered.

We generate observations using the following nonlinear observation operator

$$\mathcal{H}(\boldsymbol{x}) = \sqrt[3]{\sum_{i=1}^{N_x} (\boldsymbol{x})_i^3}, \tag{6.15}$$

where $(\boldsymbol{x})_i$ is the i-th component of $\boldsymbol{x}$. Note that, while $\mathcal{H}$ is nonlinear, the mappings $x \to x^3$ and $x \to \sqrt[3]{x}$ are bijective, so there exists a diffeomorphism between $\mathcal{H}$ and any non-degenerate linear operator $\boldsymbol{H} : \mathcal{X} \subset \mathbb{R}^{N_x} \to \mathbb{R}$. In this sense, the operator $\mathcal{H}$ has a similar function to averaging the microstates, and, thus, enough measurements of $\mathcal{H}$ should provide sufficient statistics to obtain accurate estimations of the ground-truth microstate. Another common (nonlinear) observation the Euclidean distance of the microstate to the origin. However, the Euclidean distance destroys information because squaring the microstate components is not a bijective operation - if the state space includes negative numbers. Hence, some observation operators do not provide sufficient statistics, resulting in a bigger feasible set, $\boldsymbol{\Omega}$, that includes all the symmetries induced by $\mathcal{H}$.

Additionally to operator (6.15), in Appendix C.2 we explore different observation operators and the effect they have in the initialization procedure. Each of these operators have different levels of coupling between the microstate components and destroy different levels of information by aggregation. We find that the predictions we obtain in the observation space are almost identical regardless of what observation operator we use. However, given the symmetry about the z-axis of the Lorenz system, the operator with the maximum coupling recovers the right x component, but a reflection of the ground truth of the y and z components.

We test our initialization procedure on both noiseless time series and noisy time series. For the noisy series, we take a zero-mean Gaussian noise distribution $\epsilon_k \sim \mathcal{N}(0, \sigma_n^2)$ for all k. Here, σ_n represents the noise level of the observations, which we take to be 30% of the standard deviation of the observed data; i.e., $\sigma_n = 0.3\sigma_y$.

We always construct the noiseless and noisy time series from the same ground-truth microstate so that all our results are comparable.

Although our choice for the initial guess is arbitrary, we initialize our method with a microstate that matches the first observation of $\boldsymbol{y}$ exactly. This is, we take our initial guess at random from the set $\{\boldsymbol{x} \in \mathcal{X} : \mathcal{H}(\boldsymbol{x}) = y_{-T}\}$. This is straightforward to do for any homogeneous function, such as the one of Eq. (6.15).

Throughout the results, we use the parameters and system features described in Table 6.1 unless otherwise stated. However, we evaluate the performance of our method for various choices of the rough parameter δ_R (see Fig. C.9 in Appendix C.3) and the optimizers described in Section 6.2.3 (See Figs. C.3 and C.4 in Appendix C.3). Varying δ_R determines the effect of bounding the search space into finer-grained regions of the attractor, and we find that when observations are noiseless, the lower δ_R the better the initialization but the longer it takes to meet condition (6.9). For noisy observations, we find that if we set δ_R very small, the refinement stage yields no improvement over the initialized microstates obtained. In terms of the optimization schemes of the refinement stage, we find that Adadelta [238] and the various flavors of Adam [135, 182, 234] outperform all other alternatives, with significantly better results than vanilla Gradient Descent.

	$\boldsymbol{\alpha_R}$	$\boldsymbol{\alpha_r}$	$\boldsymbol{\beta_R}$	$\boldsymbol{\beta_r}$	$\boldsymbol{T}$	$\boldsymbol{N_x}$	$\boldsymbol{m}$	$\boldsymbol{\sigma_n/\sigma_y}$	$\boldsymbol{q}$	$\boldsymbol{r_0}$	$\boldsymbol{t_\lambda}$
Lorenz	0.05	10^{-4}	0.5	$0.8r_0^{-2}$	50	3	2	0.3	4	2.02	127
Mackey-Glass	0.05	10^{-5}	0.5	$0.2r_0^{-2}$	25	50	2	0.3	5	2.41	230

Table 6.1: **Parameters and other quantities**. α_R, α_r, β_R, and β_r are the parameters of our procedure. T is the number of data points in the time series $\boldsymbol{y}$, N_x is the dimension of the microstate space, m is the sampling interval between observations, q is the number of times we feed the signal back into the LPMA filter, r_0 is the increase in the signal-to-noise ratio, σ_n/σ_y is the noise level, and t_λ is the Lyapunov 10-fold time of the system,.

6.3.1 Low-dimensional Example: Lorenz System

We first test our microstate initialization procedure on the Lorenz system [150], described by the following differential equations

$$\begin{aligned} \dot{x} &= \sigma(y - x) \ , \\ \dot{y} &= x(\rho - z) - y \ , \\ \dot{z} &= xy - \beta z \ , \end{aligned} \tag{6.16}$$

where $\boldsymbol{x}_k = \big(x(t_k), y(t_k), z(t_k)\big)$ is the microstate of the system at time t_k. The dynamics exhibit chaotic behavior for the parameters $\sigma = 28$, $\rho = 10$, and $\beta = 8/3$. We solve the system using a 4-th order Runge-Kutta integrator with a fixed step size of 0.01 units. Thus, the model, $\mathcal{M}$, is the Runge-Kutta solution to the system (6.16). Under these settings, the system's Lyapunov 10-fold time is $t_\lambda = 127$ samples while its attractor has a Kaplan-Yorke dimension of $N_S = 2.06$.

We perform 1000 independent experiments. For each experiment we sample ground-truth microstates at random from the attractor of the system and generate time series of $T = 50$ observations with a sampling interval of $m = 2$ time steps per observation. We initialize the microstate for each time series following the steps presented in Section 6.2.5 using the parameters summarized in Table 6.1. We present our results in Fig. 6.1, where we show the median assimilation ($k < 0$) and prediction errors ($k \geq 0$) of the noiseless and noisy time series for both the model and observation spaces (see Eqs. (6.11)-(6.12)).

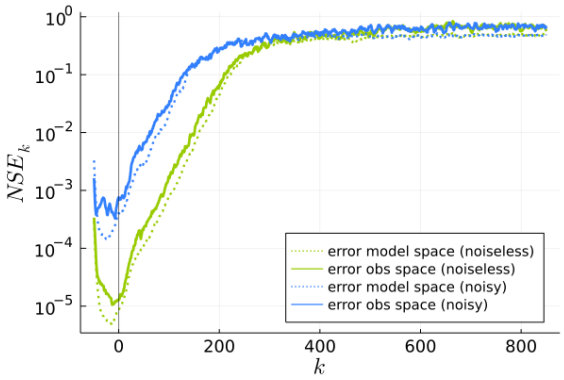


Figure 6.1: **Prediction error for the Lorenz system,** showing the median normalized squared error over 1000 experiments for the observation space (solid lines) and the model space (dotted lines) for the case of noiseless (green) and noisy (blue) observations. The solid vertical line separates the assimilative regime ($k < 0$) from the predictive regime ($k \geq 0$).

From the assimilation side, our estimations get progressively better the closer they are to the present time at $k = 0$. This means that the longer the time series - and the more information we have from the *past* -, the better the quality of the initialized microstate. Note that in the noisy case, the assimilative error plateaus near 10^{-3} in the observation space, marking the noise level of the observations. In contrast, the estimations keep getting better in the model space, indicating that even in the presence of noise, having more observations mitigates the probability of having dominating trajectories.

From the prediction side, we observe that the error diverges at essentially the same rate in both the noiseless and noisy cases. The main difference is that the error intercept at $k = 0$ is higher in the noisy case, thus making the predictions saturate earlier than when the observations are noiseless. Specifically, we find prediction horizons of $k_{max} = 171$ and $k_{max}^{noisy} = 113$ steps, which correspond to $1.35t_\lambda$ vs. $0.89t_\lambda$ for the noiseless and noisy time series (see Fig C.6 in Appendix C.3 for an alternative approach on the prediction horizons). Moreover, we find that the prediction errors on the model and observation spaces are almost identical throughout the whole prediction window. Thus, measuring how the errors diverge in the observation space gives us a good proxy of the out-of-sample behavior of the latent microstates of the system.

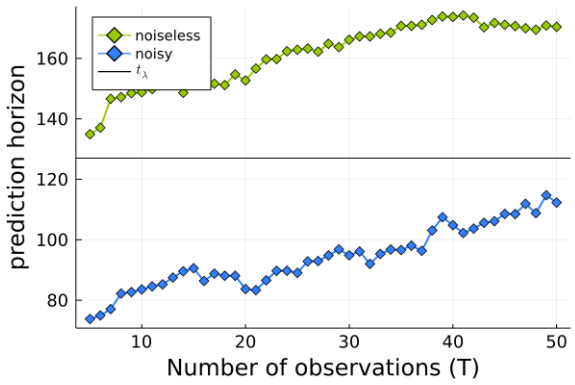


Figure 6.2: **Predictability vs. number of observations.** We show how the predictability horizon k_{max} for the Lorenz system changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series. The horizontal black solid line indicates the Lyapunov 10-fold time t_λ.

In Figs. 6.2 and 6.3 we analyze how the performance of our method depends on the length of the observation window, showing how the prediction horizons and the discrepancies between $\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}}_0$ change with the number of observations.

In Fig. 6.2 we find that the prediction horizon increases linearly with the number of observations available, with a similar slope for both the noiseless and noisy cases. Not surprisingly, the noise affects the time horizon over which one can make an effective prediction.

Additionally, we observe in Fig. 6.3 that the discrepancy decreases monotonically in both the noiseless and noisy cases. For the noiseless case, we observe a higher than exponential decrease in the discrepancy that ranges from $\text{NSE}_0^{model} \sim 10^{-3}$ for $T = 5$ to $\text{NSE}_0^{model} \sim 10^{-5}$ for $T = 50$ observations. While the change in discrepancy is less pronounced for the noisy time series, it decreases 2 orders of magnitude with $\text{NSE}_0^{model} \sim 10^{-1.5}$ for $T = 5$ to $\text{NSE}_0^{model} \sim 10^{-3.5}$ for $T = 50$ observations.

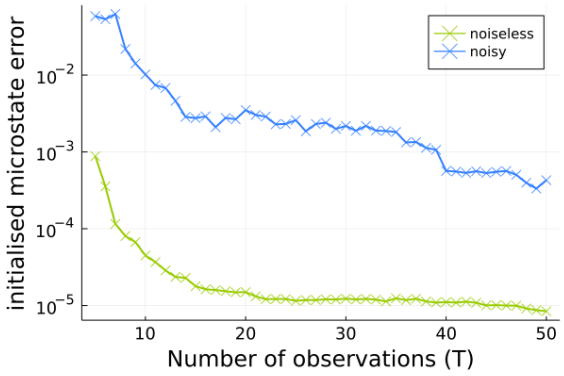


Figure 6.3: **Initialized microstate error for the Lorenz system.** We show how the average discrepancy NSE_0^{model} between the true present-time microstate $\boldsymbol{x}_0$ and the initialized microstate $\hat{\boldsymbol{x}}_0$ changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series.

The Lorenz equations are a low-dimensional system ($N_x = 3$) with a low dimensional attractor of dimension $N_{\mathcal{S}} = 2.06$. The number of observations in our experiments can be much larger than the dimension of the system. When combined with the fact that this system does not have any severe degeneracies, (see Appendix C.1), we recover the ground-truth microstate precisely with only a handful of noiseless observations. Every additional observation is, in theory, redundant for finding $\boldsymbol{x}_{-T}$ but, in practice, measurements can have several sources of error such as observational noise or the finite precision of numerical integration methods. Each additional observation thus further averages out these errors, which probably why k_{max} gets better proportionally to T.

6.3.2 High-dimensional System: Mackey-Glass

The Mackey-Glass system [152] describes the dynamics of the density of cells in the blood with the following delayed differential equation

$$\dot{x} = \mathscr{F}\left(x, x_{t_d}\right) = \frac{a x_{t_d}}{1 + x_{t_d}^c} - bx. \tag{6.17}$$

The state $x_{t_d} = x(t - t_d)$ is the density of cells delayed by t_d time units and a, b, and c are parameters. It exhibits chaotic dynamics for $t_d > 16.8$ with $a = 0.2$, $b = 0.1$ and $c = 10$ [83]. In terms of blood cell density, the chaotic regime represents a pathological behavior.

The evolution of Eq. (6.17) relies on knowing the state of x in the continuous interval $[t - t_d, t]$, making its state space infinite-dimensional. However, we can approximate such state by taking N_x samples at intervals of length $\Delta t = t_d/N_x$ and

constructing the N_x-dimensional microstate vector

$$\begin{aligned}\boldsymbol{x}_k &= ((\boldsymbol{x}_k)_1, \cdots, (\boldsymbol{x}_k)_{N_x}) \\ &= \big(x(t_k - \underbrace{N_x \Delta t}_{t_d}), \ldots, x(t_k - \Delta t), x(t_k)\big),\end{aligned} \tag{6.18}$$

where $(\boldsymbol{x}_i)_k = x(t - (N_x - i)\Delta t)$.

Using this vector, we can obtain trajectories of the Mackey-Glass system with any numerical integrator, for which we use the Euler method with a fixed step size of Δt for simplicity. We can thus recast this approximate system with the following N_x-dimensional deterministic mapping

$$\boldsymbol{x}_{k+1} = \boldsymbol{\mathcal{M}}(\boldsymbol{x}_k) = \begin{cases} (\boldsymbol{x}_t)_{N_x} + \Delta t \mathscr{F}\big((\boldsymbol{x}_t)_{N_x}, (\boldsymbol{x}_t)_1\big) \\ (\boldsymbol{x}_{t+1})_1 + \Delta t \mathscr{F}\big((\boldsymbol{x}_{t+1})_1, (\boldsymbol{x}_t)_2\big) \\ \vdots \\ (\boldsymbol{x}_{t+1})_{N_x-1} + \Delta t \mathscr{F}\big((\boldsymbol{x}_{t+1})_{N_x-1}, (\boldsymbol{x}_t)_{N_x}\big), \end{cases} \tag{6.19}$$

using $\mathscr{F}$ as defined in Eq. (6.17). The microstate-space dimension N_x of this mapping is determined by $t_d/\Delta t$, for which we take $N_x = 50$ and $t_d = 25$ so that the system exhibits chaotic dynamics. Under these settings, the system's Lyapunovs 10-fold time is $t_\lambda = 230$ samples while its attractor has a Kaplan-Yorke dimension of $N_{\mathcal{S}} = 2.34$. In this sense, the Mackey-Glass system is not really 50-dimensional in that its microstates live in a much lower-dimensional manifold. Therefore, although we expect to need at least 50 observations to disentangle the individual microstate components from the observation operator, the actual dynamics of the microstates is relatively simple - see, for instance, Fig. C.7 in the Appendix, where we show how the Mackey-Glass system has a rapidly decaying power spectra, meaning that it is very unlikely that the microstates exhibit high-frequency oscillations and thus easing the inital condition inference.

As before, we perform 1000 independent experiments in which, for each experiment, we sample ground-truth microstates at random from the attractor of the system and generate time series of $T = 25$ observations with a sampling interval of $m = 2$ time steps per observation (see Table 6.1 for details). In contrast to the Lorenz system, we consider time series containing fewer data points than the dimension of the microstate space (in this case $T = 25$ and $N_x = 50$, respectively), making the problem under-determined. We present our results in Fig. 6.4, where we show the median assimilation ($k < 0$) and prediction errors ($k \geq 0$) for the noiseless and noisy time series for both the model and observation spaces.

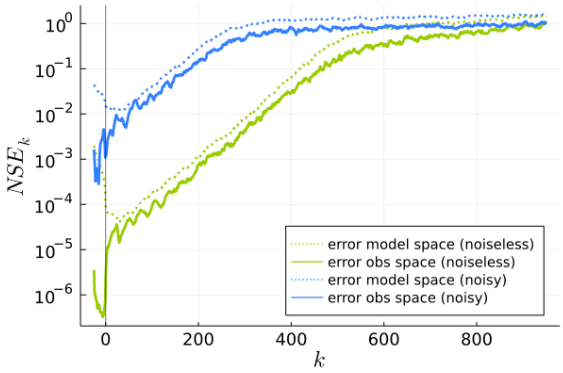


Figure 6.4: **Prediction error for the Mackey-Glass system.** We show the median normalized squared error over 1000 experiments for the observation space (solid lines) and the model space (dotted lines) for the case of noiseless (green) and noisy (blue) observations. The solid vertical line separates the assimilative regime ($k < 0$) from the predictive regime ($k \geq 0$).

Unexpectedly, our method yields more accurate initializations for the Mackey-Glass system than the Lorenz system, even though both the attractor and the microstate space of the former have a higher dimension than the latter. However, if we compare the power spectra of the two systems (see Fig. C.7 in Appendix C.3), we find that the Mackey-Glass system has a faster frequency decay and more frequency peaks than the Lorenz system, suggesting that the former is easier to initialize than the latter. For instance, the power amplitude for frequency 1/6 is more than 1000 higher in the Lorenz system than in the Mackey-Glass system. This further suggests that looking at the power spectra of the system is a better indicator of the initializability of a system than the dimension of its chaotic attractor.

From the prediction side, our results are qualitatively similar to what we saw for the Lorenz system. The error diverges at roughly the same rate in both the noiseless and noisy time series, with an error intercept at $k = 0$ that is higher for the noisy experiments. Additionally, we find a very close correspondence between the errors in the model and observation spaces, which supports our claim that measuring the error in the observation space gives us a good proxy of the out-of-sample behavior of the latent microstate dynamics.

In terms of their predictability horizons, we find that $k_{max} = 556$ and $k_{max}^{noisy} = 285$, corresponding to $2.4t_\lambda$ and $1.2t_\lambda$ for the noiseless and noisy ensembles respectively, (see Fig C.6 in Appendix C.3 for an alternative approach on the prediction horizons). We find it remarkable that with only 25 observations of the system, we obtain predictions that stay accurate for significantly longer than the Lyapunov 10-fold time of

the system.

From the assimilation side, the microstate estimations get progressively better the closer they are to the present time, similar to what we observed in the Lorenz system. However, the assimilation error is significantly lower in the observation space than in the model space, suggesting, misleadingly, that the initialized microstate is much more accurate than the error we observe in the model space. Nonetheless, the errors in model and observation spaces converge to each other as soon as the prediction window starts, meaning that the error in the observation space is still an accurate proxy of the out-of-sample behavior in the model space.

Interestingly, the first few out-of-sample predictions in the model space have a lower discrepancy than in the observation space in both the noiseless and noisy cases. This happens, we believe, because the time span of the observations $\boldsymbol{y}$ is not long enough for the initialized microstate to converge onto the attractor. With only $T = 25$ data points of a 50-dimensional chaotic system, we do not possess enough information to recover the present-time microstate precisely.

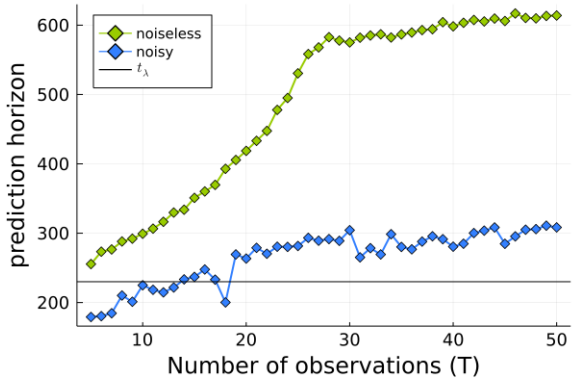


Figure 6.5: **Predictability horizon of the Mackey-Glass system.** We show how the predictability horizon k_{max} changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series. The horizontal black solid line indicates the Lyapunov 10-fold time t_λ. For the noiseless case, we observe a critical transition on the behavior of k_{max} for $T_c = 25$.

The previous discussion suggests that we need more observations to better initialize the system. To investigate this we perform a series of experiments in which we vary the number of data points of the observed time series and assess the quality of the predictions. Similar to what we did for the Lorenz system, we focus on the prediction horizon (see Fig 6.5) and the discrepancy between the present-time and the initialized microstate (see Fig 6.6).

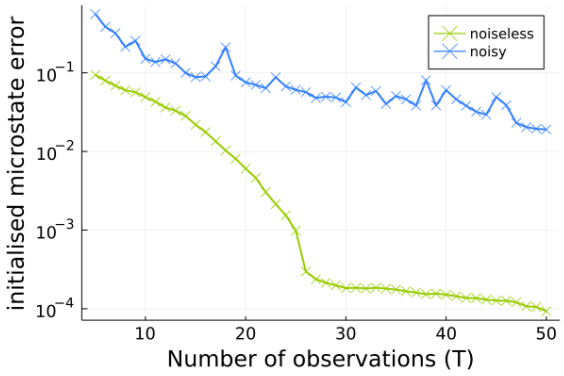


Figure 6.6: **Initialized model error for the Mackey-Glass system.** We show how NSE_0^{model}, the average discrepancy between the true present-time microstate $\boldsymbol{x}_0$ and the initialized microstate $\hat{\boldsymbol{x}}_0$, changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series. For the noiseless case, we observe a critical transition in the behavior of NSE_0^{model} for $T_c = 25$.

We find a contrasting behavior regarding the experiments between noisy and noiseless observations. When the observations are noisy, the prediction horizon increases linearly with the number of observations (see Fig. 6.5 blue), ranging from $k_{max} = 0.78t_\lambda$ when $T = 5$ to $k_{max} = 1.34t_\lambda$ when $T = 50$. We also find that, in general, the discrepancy between the initialized and ground-truth present microstate decreases monotonically with the number of observations (see Fig. 6.6 bottom). These results are qualitatively similar to what we found for the Lorenz system: the more observations the better.

When the observations are noiseless, we find a *critical change of behavior* at roughly $T = 25$ observations. In Fig. 6.5 (green), we find that the prediction horizon rises superlinearly for $T < 25$, with $k_{max} = 1.11t_\lambda$ for $T = 5$ to $k_{max} = 2.31t_\lambda$ for $T = 25$. Afterwards, the prediction horizon grows linearly and with a marginal increase, getting to $k_{max} = 2.67t_\lambda$ for $T = 50$. We note, however, that the increase in this linear regime is almost double of what we find in the noisy counterpart, with a slope of $\Delta k_{max} = 15.2$ steps per observation against $\Delta k_{max}^{noisy} = 8.3$, respectively. In parallel, we find that the discrepancy of the initialized microstate decreases abruptly for the time series of $T \geq 25$ observations, as we show in Fig. 6.6 (green). This sharp transition reflects that time series with more than 25 observations have enough information to pin down the present-time latent microstate precisely, meaning that we possess enough data points to uniquely separate the mixing of the microstate generated by the observation operator $\mathcal{H}$ into its individual components.

In short, having 25 (or more) noiseless measurements of the system gives us enough

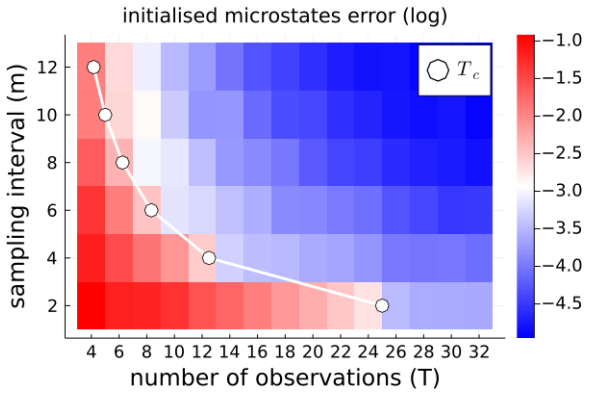


Figure 6.7: **Critical transition heatmap of the Mackey-Glass system:** In the z-axis, we show the (base-10 logarithm of the) initialized microstate discrepancy, NSE_0^{model}, as a function of the number of observations T and the sampling interval m for ensembles of noiseless time series. In white, we plot the $m = N_x/T$ curve for fixed $N_x = 50$. We find that the microstate discrepancies decrease abruptly before and after this curve.

information to precisely recover the present-time microstate, which is 50-dimensional. Recall that we are considering the discrete map (6.19) as the real system, so the only information we lose comes from either measuring the system with $\mathcal{H}$ or taking time series with a coarse-grained sampling frequency. Thus, for a fixed $\mathcal{H}$ and noiseless observations, recovering the initial microstates precisely should depend solely on how well the samples describe the latent trajectory of the system. Inspired by the Nyquist-Shannon sampling theorem [198], if we can establish a clear cutoff frequency on the power spectrum of the system, we could argue that if we observe the system with twice the frequency as the systems's cutoff frequency, then the observed signal would not lose any information with respect to the signal sampled for every update of map (6.19).

We thus claim that, if we can establish a clear cutoff frequency f_c and the dynamical system does not suffer from severe degeneracies, we can precisely obtain the initial conditions of an N_x-dimensional (possibly) nonlinear system observed every m updates with a scalar (possibly) nonlinear observation operator $\mathcal{H}$ if 1) $m^{-1} \geq 2f_c$ and 2) $T_c \geq N_x/m$. Having these conditions satisfied is equivalent to having an invertible observation-matrix $\boldsymbol{M}$ that determines the solution of the system $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}_0$ exactly (see Appendix C.1 for a deeper development of this discussion).

We make further experiments to check the validity of our claim, where we measure the initialized microstate discrepancy when varying both the number of observations T and the sampling interval m while leaving the dimension of the system fixed to

$N_x = 50$. If our claim about the Nyquist-Shannon theorem is a good approximation, we expect to find a $T_c \sim 1/m$ relation where, for $T \geq T_c$, the error in the initialized microstate becomes significantly lower than for $T < T_c$. In Fig 6.7, we show the results of these experiments, in which we observe such a change of regimes before and after the $T = N_x/m$ line, thus supporting the Nyquist-Shannon hypothesis.

6.4 Conclusions

Many natural and social processes can be accurately described by how their microstates interact and evolve into rich nontrivial dynamics with emerging properties. We often only possess aggregate noisy measurements of such processes, so it is of great interest to develop methods that let us extract information about the latent microstate dynamics from a given dataset.

In this chapter, we tackled the problem of initializing the latent microstate of a known (possibly) nonlinear dynamical system from aggregate (possibly) noisy and nonlinear observations of the system. We propose a three-step method to obtain such latent microstate that consists of 1) reducing the observational noise to mitigate possible dominating trajectories, 2) letting the system explore its attractor(s) and thus limiting the region in which we search for an optimal solution, and 3) minimizing the discrepancy between the simulated and real observations to obtain a refined estimation of the ground-truth microstate. We quantified the discrepancy between observations and simulations using a least-squares cost function in the observation space, similar to [175, 52]. We minimized the cost function using a plethora of gradient-based algorithms, for which we find that Adadelta [238] and Adam-oriented schemes [135, 182, 234] perform the best.

We tested our method on two chaotic paradigmatic examples: the Lorenz and a high-dimensional approximation of the Mackey-Glass systems. We obtained initialized microstates that accurate fit the data, with out-of-sample predictions that outperformed the systems' Lyapunov 10-fold times, even when the observed time series were very short. We found that good predictions in the observations space always implied good predictions in the space of the microstates. We considered nonlinear observation operators that aggregate all the microstate component into a real number in all cases, with robust result with all the operators considered. Surprisingly, we obtained better results for the Mackey-Glass system, which has a higher-dimensional model space and higher-dimensional attractor but faster-decaying frequency spectrum than the Lorenz system, suggesting that the frequency spectrum gives us a

better proxy of the initiability of the system than the observations to dimension of the model ratio.

In most experiments, the quality of the initialized microstate was proportional to the number of data points of the observed time series. However, when the dimension of the system was higher than the number of observations and these observations were noiseless, the quality of the initialized microstate grew superlinearly. This superlinear regime transitions into the more common linear regime in a nontrivial manner, and we explored the conditions for such a transition. We claim that as long as we can establish a clear cutoff frequency of the observed data and this data meets the Nyquist-Shannon sampling theorem conditions with respect to the cutoff frequency, we can recover the ground-truth microstate precisely with fewer observations than the dimension of the system, thus marking the transition between regimes. This implies that if we possess a dataset where observations are sampled at an optimal rate such that we lose the least possible information of the underlying system, we can obtain high-quality initializations with just a handful of samples.

How well we can initialize a system depends on the amount of information the observed data contains and on the intrinsic features of the system. On the one hand, the amount of observational noise, the mixing that results from aggregating the system, the number of observations, and the data sampling rate contribute significantly in estimating microstates that may fit the data well but extrapolate poorly into the future. On the other hand, the dimension of the dynamical system, the frequency spectrum of its attractor(s), and how chaotic the system is, determines the window in which the predictions stay accurate.

In this chapter, we provided a conceptual framework to understand the interplay between data quality, data quantity, and the complexity of high-dimensional models. In all cases, we estimated initial conditions that converged to the ground-truth state when the data was noiseless. However, this framework only works if we know 1) the model that perfectly specifies the system, 2) the observation operator, and 3) we have a good estimate of the observational noise level. It is also limited in that it tries to recover a single initial condition to describe the whole data, so this method suffers for very long time series or highly stochastic systems. In the next chapter, we present a data assimilation method that relaxes some of these limitations by sequentially estimating the latent state of the system for every observation separately instead of pinning down a single initial condition for the whole sequence. It makes a trade-off between the accuracy of what we presented in this chapter with a more flexible, computationally-cheaper alternative.

Chapter 7

Latent state estimation of models with network topologies

7.1 Introduction

In the last chapter (see Chapter 6), we analyzed when and to what extent we may recover the initial conditions of chaotic dynamical systems from incomplete data. We proposed an initialization method that, in summary, filters the noise out of the observations, explores the feasible space of solutions to obtain a rough estimate of the initial condition, and refines this estimate using descent-based algorithms. However, exploring the feasible space and using descent-based algorithms is often impracticable in complex systems models, particularly in agent-based models. Typically, these models are non-differentiable, high-dimensional, stochastic, and open to external perturbations. In this chapter, we propose a more flexible, computationally efficient method to infer latent states from incomplete data. Our method leverages accuracy for flexibility, and it is readily applicable to some agent-based models.

Hassan et al. [111, 112] have discussed the importance of injecting realistic initial conditions into agent-based simulations and the effect they have when compared to random initializations. They suggest modelers to include the distribution of surveys data and other samples of representative populations into the model's initial conditions. However, the authors do not provide a method to assimilate the data regarding real-world macro observations. Furthermore, Barde [24] showed that the information in the initial condition is conserved as certain systems - he shows this for the Schelling segregation agent-based model - evolve, so, in principle, one could directly reconstruct the final state distribution from the initial condition and vice versa using maximum entropy arguments. However, the class of models in which his approach works is yet too narrow. Finally, Ward et al. [226] and the subsequent paper by the same

group [61] take from the data assimilation literature and use ensemble Kalman filters (EnKF) and Unscented Kalman filters (UKF) to assimilate real-world data into an agent-based model that describes the footfall dynamics in cities. The papers in the group of Ward shed important contributions to state estimation in agent-based models: they work, they link the agent-based modeling formalism with the data assimilation formalism, and they propose a practical way to assimilate agent-based models using EnKFs and UKFs with a small simulation burden. Cocucci et al. [62] take a similar approach with an epidemiological agent-based model.

However, these papers have some fundamental limitations which we aim to improve in this chapter. First, the state variables in their models are all macroscopic, so there is not really an interplay between micro- and macro-dynamics. Second, the observations of their systems are the actual state variables - in the notation that follows, the observation operator $\mathcal{H}$ is the identity matrix -, so there is no information loss in observing the system. Finally, their model is a SIR-like model simulated with the Gillespie algorithm [66], so it is agent-based only statistically[1].

In this chapter, we address these limitations by developing a data assimilation technique that works well when the data is aggregate, noisy and incomplete. Essentially, we employ the ensemble Kalman filter described in [226, 61, 62] and expand it. In most agent-based models, agents interact through some sort of topology, so we incorporate such a topology in the data assimilation cycle to obtain more accurate estimations.

The remainder of this chapter is laid out as follows. In Section 7.2, we describe the state estimation problem as a general sequential filter, following the Bayesian formulation we describe in Section 5.2.2. In Section 7.2.2, we review and derive ensemble Kalman filter. In Section 7.2.2.4, we discuss the main drawbacks of the ensemble Kalman filter and propose methods to adapt it for complex systems with network topologies. We test our adapted ensemble Kalman filter in Section 7.3, where we validate the localized EnKF in two systems, namely, the chaotic Mackey-Glass system and the Hegselmann-Krause agent-based model of opinion dynamics. Finally, we discuss our results and suggest further research directions in Section 7.4.

[1]The Gillespie algorithm generates a statistically correct trajectory (possible solution) of a stochastic equation system for which the reaction rates are known.

7.2 Latent state estimation procedure

Throughout this chapter, we assume we have a dynamical system, $\boldsymbol{\mathcal{M}}$, that models the time evolution of a state variable, $\boldsymbol{x} \in \mathbb{R}^{N_x}$, to an arbitrary time into the future. Additionally, we assume we possess a sequence of observations, $\boldsymbol{y} = (y_T, \ldots, y_0)$, where $y_i \in \mathbb{R}$ corresponds to the observation at time t_i. The state variable is mapped to the (lower-dimensional) space of the observations through an observation operator, $\mathcal{H}$. The *latent state estimation problem* is that of inferring the sequence of latent states $\boldsymbol{x}_{T:0} = (\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_0)$ that best represents the data $\boldsymbol{y}$ given the model $\boldsymbol{\mathcal{M}}$.

From Section 5.2.2, we know that we can write the state estimation problem probabilistically using Bayes theorem as

$$p(\boldsymbol{x}_{T:0}|y_{T:1}) \propto p(\boldsymbol{x}_0) \prod_{k=1}^{T} p(y_k|\boldsymbol{x}_k) p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}), \tag{7.1}$$

where $p(\boldsymbol{x}_0)$ is the latent states prior, $p(y_k|\boldsymbol{x}_k)$ the likelihood of observing y_k given $\boldsymbol{x}_k$, and $p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}$ the transition probability from state $\boldsymbol{x}(t_{k-1})$ to state $\boldsymbol{x}(t_k)$. Expression (7.1) describes the *posterior probability* of obtaining the sequence $\boldsymbol{x}_{T:0}$ given the data $\boldsymbol{y}T : 1$, and it is of central importance for data assimilation because it can be solved recursively.

There are three flavors of assimilating data depending on *when* we want to estimate the state of the system:

1. *Smoothing*: estimate $p(\boldsymbol{x}_n|y_{k:1})$ for $t_0 \leq t_n < t_k$.
2. *Filtering*: estimate $p(\boldsymbol{x}_k|y_{k:1})$.
3. *Forecasting*: estimate $p(\boldsymbol{x}_n|y_{k:1})$ for $t_n > t_k$.

Smoothing is useful for estimating a state based on future observations, *filtering* estimates the present state using the observation history up to the present time, and *forecasting* estimates a state in the future using the observation history of the past.

7.2.1 General sequential filter

We tackle the latent state estimate problem using *sequential filtering*. This is a bit like nudging: we let the dynamical process (5.1) evolve and, whenever there is a new observation at, say, time t_k, we update the state at t_k with the information provided by the new observation.

Thus, sequential filtering consists of the two following steps:

1. **forecast**: We propagate the pdf associated with state $\boldsymbol{x}_k$ at time t_k to time t_{k+1}. Formally, we do this by solving the Chapman-Kolmogorov marginalization:

$$p(\boldsymbol{x}_{k+1}|y_{k:1}) = \int_X p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)p(\boldsymbol{x}_k|y_{k:1})d\boldsymbol{x}_k \ . \tag{7.2}$$

2. **analysis**: We update the state estimate $\boldsymbol{x}_{k+1}$ using the new observation y_{k+1}. Formally, we do this by applying the recursive version of the posterior of Eq. (5.7)

$$p(\boldsymbol{x}_{k+1}|y_{k+1:1}) \propto p(y_{k+1}|\boldsymbol{x}_{k+1})p(\boldsymbol{x}_{k+1}|y_{k:1}) \ . \tag{7.3}$$

The *forecast-analysis* combo forms one *data assimilation cycle*. We can then propagate the analysis pdf (7.3) using the forecast (7.2) to estimate the state at time t_{k+2}, and so on.

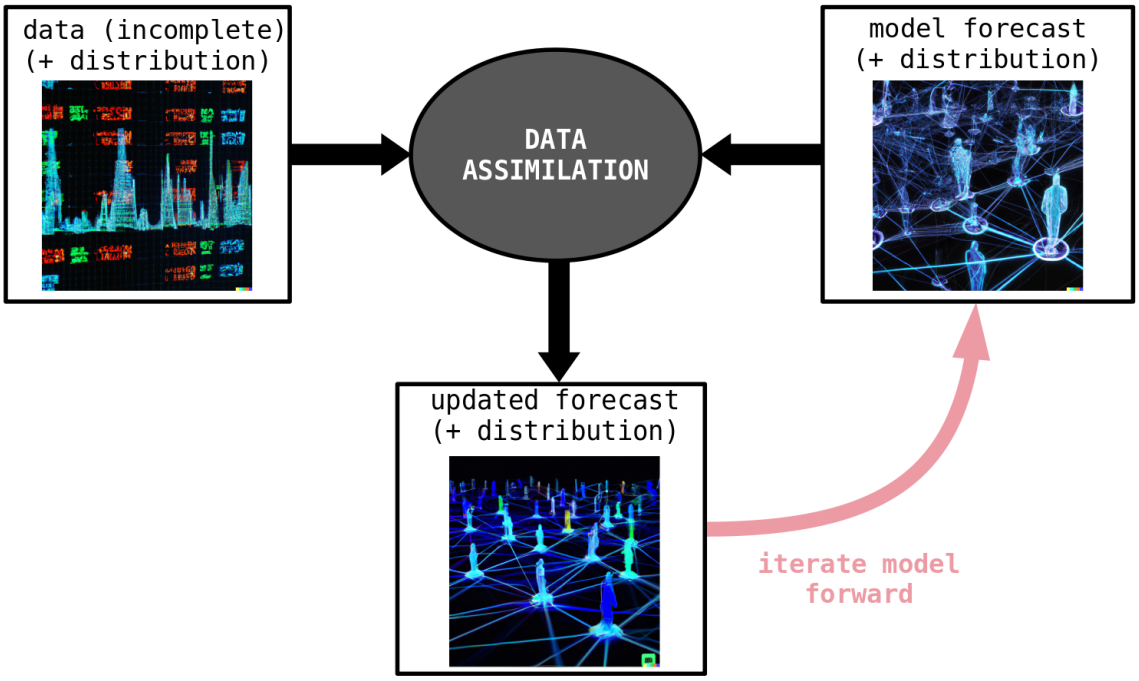


Figure 7.1: **Schematic diagram of a data assimilation (DA) cycle**. A DA cycle works by incorporating observations one at a time - or, in some cases, batches of observations [67] - into a state forecast - obtained with the model $\mathcal{M}$ and a state prior $p(\boldsymbol{x}_0)$ - to produce an updated forecast, or *analysis*, of the state variables. The analysis is then iterated forward in time so that it reaches the next timestamp of the data, and the cycle repeats.

Given the system formed by Eqs. (5.1) and (5.2), the only input of the general sequential filter is the prior information of the state, $p(\boldsymbol{x}_0)$, which we then update recursively with the available observations. In Figure 7.1, we present a schematic

diagram of the data assimilation cycle [2]: we consider the current model forecast plus the available data and update the posterior probability to obtain an updated forecast. Then, we iterate the model forward to match the time of the next observation, and repeat.

In practice, computing Eqs. (7.2) and (7.3) explicitly is intractable for most real-world applications. Thus, various types of approximations and simplifications are made to the *forecast* and *analysis* expressions.

7.2.2 Kalman filter preliminaries

The Ensemble Kalman Filter (EnKF) [81] is one of the most commonly used filters in numerical weather prediction. As we will make clear later, the EnKF has the desired advantages of 1) treating the dynamical process (5.1) as a black-box model, and 2) requiring a very small number of model simulations compared with the (often) high dimensionality of the state space. In its non-ensemble format, it has a closed-form optimal solution of the forecast and analysis update equations (see Eqs. (7.2) and (7.3)) when the system is linear and the errors are Gaussian.

7.2.2.1 The original Kalman filter

The Kalman filter [132] is the optimal sequential filter when the dynamical process and observations are linear and all random variables are Gaussian and additive.

In this scenario, we can rewrite Eqs. (5.1) and (5.2) as

$$\boldsymbol{x}_k = \boldsymbol{M}_k \boldsymbol{x}_{k-1} + \xi_k \ , \tag{7.4}$$

$$y_k = \boldsymbol{H}_k \boldsymbol{x}_k + \epsilon_k \ , \tag{7.5}$$

where $\boldsymbol{M}_k$ and $\boldsymbol{H}_k$ are $N_x \times N_x$ and $N_y \times N_x$ matrices, respectively. The random variables ξ_k and ϵ_k are iid, zero-mean, Gaussian distributions with known covariances $\boldsymbol{Q}_k$ and $\boldsymbol{R_k}$, respectively. Thus, the pdfs associated with the forecast and analysis update equations are also Gaussian, where

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \sim \mathcal{N}(\boldsymbol{M}_k \boldsymbol{x}_{k-1}, \boldsymbol{Q}_k) \ ,$$
$$p(\boldsymbol{x}_k|y_k) \sim \mathcal{N}(\boldsymbol{H}_k x_k, \boldsymbol{R}_k) \ .$$

[2] We generated the images in this diagram using *DALL.E2*, an artificial intelligence software from *openai* that generates images from text prompts. The exact prompts we used were, for the top left and top right images respectively: "*plot of stock market time series price data; digital cyberpunk*" and "*complex agent-based model embedded in a social network, digital futurist art*". The bottom image is a variation of the top-right one also generated by *DALL.E2*.

Taking an initial estimate $\boldsymbol{x}_0^a$ with covariance $\boldsymbol{P}_0^a$, and considering that the product of Gaussian distributions is again Gaussian, we can recast Eqs. (7.2) and (7.3) as

$$p(\boldsymbol{x}_k|y_{k-1:1}) \sim \mathcal{N}(\boldsymbol{x}_k^f, \boldsymbol{P}_k^f) \ ,$$
$$p(\boldsymbol{x}_k|y_{k:1}) \sim \mathcal{N}(\boldsymbol{x}_k^a, \boldsymbol{P}_k^a) \ ,$$

where

forecast:

$$\boldsymbol{x}_k^f = \boldsymbol{M}_k \boldsymbol{x}_{k-1}^a \ , \tag{7.6}$$
$$\boldsymbol{P}_k^f = \boldsymbol{M}_k \boldsymbol{P}_{k-1}^a \boldsymbol{M}_k^T + \boldsymbol{Q}_k \ . \tag{7.7}$$

analysis:

$$\boldsymbol{x}_k^a = \boldsymbol{x}_k^f + \boldsymbol{K}_k(y_k - \boldsymbol{H}_k \boldsymbol{x}_k^f) \ , \tag{7.8}$$
$$\boldsymbol{P}_k^a = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{H}_k)\boldsymbol{P}_k^f \ , \tag{7.9}$$
$$\boldsymbol{K}_k = \boldsymbol{P}_k^f \boldsymbol{H}_k^T \left(\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^T + \boldsymbol{R}_k\right)^{-1} \ , \tag{7.10}$$

where $(\cdot)^T$ denotes matrix transposition. The explicit derivation of these equations (see, eg. [231]) involves some matrix algebra and using the explicit Gaussian density functions. The quantity $\boldsymbol{K}_k$ - known as the *Kalman gain* - is an $N_x \times N_y$ matrix that minimizes the analysis covariance.

Note that the *forecast* step depends only on the linear dynamical process (7.4) while the *analysis* step integrates the observation y_k to the state estimation. The analysis state update (see Eq. (7.8)) is a linear combination between the forecast state $\boldsymbol{x}_k^f$ and the *innovation* $y_k - \boldsymbol{H}_k \boldsymbol{x}_k^f$ weighted by the Kalman gain $\boldsymbol{K}_k$. If the quality of the observations is good, then the observation error covariance $\boldsymbol{R}_k$ tends to zero and $\boldsymbol{K}_k \to \boldsymbol{H}_k^{-1}$. Thus, the analysis state will be determined from the observation y_k only. If, on the other hand, the observations are very noisy, then the denominator in $\boldsymbol{K}_k$ will be huge and $\boldsymbol{K}_k \to 0$. Thus, the analysis state $\boldsymbol{x}_k^a$ will be almost identical to the forecast $\boldsymbol{x}_k^f$.

While these update formulas are straightforward to implement and easily interpretable, the Kalman filter requires intensive computational resources to update the covariance $\boldsymbol{P}_k$, because its size is $N_x \times N_x$. Furthermore, this filter is rigid in that it assumes linear dynamics and Gaussian random variables. However, it turns out that that Kalman filtering can also handle nonlinearities both in model and observation

spaces with a little tweaking of its update equations. Additionally, we can overcome the computational burden of the original Kalman filter by approximating the covariance and noise matrices with (a few) Monte Carlo *ensemble* simulations. These solutions provide Kalman filtering-like techniques with such flexibility that they are still one of the main tools for nonlinear, high-dimensional data assimilation. The next section addresses these issues by introducing the *Ensemble Kalman filter* adapted to nonlinear dynamics.

7.2.2.2 Ensemble formulation

The idea of the Ensemble Kalman Filter (EnKF) (first developed by [81] and extended by [47]) is to approximate the state error covariance $\boldsymbol{P}_k$ of the original Kalman filter with an ensemble of states that are propagated through the dynamical process. Using such an ensemble loosens the rigidity imposed by the original Kalman filter of dealing with linear processes with Gaussian random variables. Nevertheless, its forecast and analysis equations are based directly on the original Kalman filter, as we describe in what follows.

Let $\boldsymbol{X}_k^f = \left[\boldsymbol{x}_k^{f,(1)}, \ldots, \boldsymbol{x}_k^{f,(N_e)}\right]$ be an ensemble of $N_e \ll N_x$ states, each sampled from $\mathcal{X}$. We can thus approximate the state covariance as

$$\boldsymbol{P}_k^f \approx \boldsymbol{P}_k^{f,e} = \frac{1}{N_e - 1}(\boldsymbol{X}_k^f - \overline{\boldsymbol{x}_k^f})(\boldsymbol{X}_k^f - \overline{\boldsymbol{x}_k^f})^T \ , \tag{7.11}$$

where

$$\overline{\boldsymbol{x}_k^f} = \frac{1}{N_e}\sum_{i=1}^{N_e} \boldsymbol{x}_k^{f,(i)} \tag{7.12}$$

is the best state estimate given by the average of the forecast. If the system is linear and Gaussian, the EnKF coincides with the original KF when $N_e \to \infty$, resulting in an optimal filter.

Under the ensemble formulation, however, we do not have the hard constraint of having linear operators or Gaussian random variables. Actually, we can compute the *forecast* update using the nonlinear model (5.1) for each ensemble member as follows

forecast:

$$\boldsymbol{x}_k^{f,(i)} = \boldsymbol{\mathcal{M}}(\boldsymbol{x}_{k-1}^{a,(i)}, \xi_{k-1}^{(i)}) \ , \tag{7.13}$$

$$\overline{\boldsymbol{x}_k^f} = \frac{1}{N_e}\sum_{i=1}^{N_e} \boldsymbol{x}_k^{f,(i)} \ , \tag{7.14}$$

$$\boldsymbol{P}_k^{f,e} = \tilde{\boldsymbol{X}}_k^f \left(\tilde{\boldsymbol{X}}_k^f\right)^T \ , \tag{7.15}$$

where we introduce the *ensemble anomaly matrix*

$$\tilde{\boldsymbol{X}}_k^f := \frac{1}{\sqrt{N_e - 1}}(\boldsymbol{X}_k^f - \overline{\boldsymbol{x}_k^f}) \ . \tag{7.16}$$

If the random variables in the model (5.1) are additive with zero mean and covariance $\boldsymbol{Q}_k$, then $\overline{\boldsymbol{x}_{k+1}^f} = \overline{\boldsymbol{\mathcal{M}}(\boldsymbol{x}_k^f)} = \boldsymbol{\mathcal{M}}(\overline{\boldsymbol{x}_k^f}) + \text{n.l.}$, where *n.l.* stands for the nonlinear contribution of $\mathcal{M}$ over the ensemble mean[3]. Moreover, if $\boldsymbol{M}_k$ is the linear approximation of $\boldsymbol{\mathcal{M}}$ at time t_k, then $\boldsymbol{P}_k^{k,e} = \boldsymbol{M}_k \boldsymbol{P}_{k-1}^{a,e} \boldsymbol{M}_k^T + \boldsymbol{Q}_k + \text{n.l.}$, which extends the covariance update of Eq. (7.7) with the nonlinearities of the model without computing $\boldsymbol{M}_k \boldsymbol{P}_{k-1}^{a,e} \boldsymbol{M}_k^T$ explicitly. The EnKF enables the estimation of nonlinear systems as long as the nonlinearities can be well-approximated by a power series near the current estimate at time t_k. This is, as long as the signal of the *n.l.* part does not overcomes the contribution of the linear part (and the Gaussianity assumption of priors and likelihoods are approximately accurate), the estimates of the EnKF will converge to the true mean and covariance with enough particles. Note that this is a milder condition than, e.g., assuming global Lipschitz continuity for the system's dynamics. However, if the system contains discontinuities or sharp nonlinearities, the EnKF will likely result in very biased estimates of the latent states [119]. In general, the stronger the nonlinearities, the shorter the time intervals we can take between observation to obtain accurate estimates.

Regarding the *analysis* update, we first need to adapt the Kalman gain matrix $\boldsymbol{K}_k$ to the ensemble formulation. We will present two adaptations of $\boldsymbol{K}_k^e$ for which we will exploit some properties in the next section. The first adaptation consists of linearizing the observation operator $\mathcal{H}$ about the current estimate and using the ensemble state covariance as follows

$$\boldsymbol{K}_k \approx \boldsymbol{K}_k^e = \boldsymbol{P}_k^{f,e} \boldsymbol{H}_k^T \left(\boldsymbol{H}_k \boldsymbol{P}_k^{f,e} \boldsymbol{H}_k^T + \boldsymbol{R}_k \right)^{-1}, \tag{7.17}$$

where $\boldsymbol{H}_k = \frac{\partial \mathcal{H}}{\partial \boldsymbol{x}}\big|_{\overline{\boldsymbol{x}}_k^f}$ is the linearized observation matrix and $\boldsymbol{R}_k$ is the observation error covariance. Our second adaptation is applying the full nonlinear operator $\mathcal{H}$ to each ensemble member (as done in [36]). Recalling that $\boldsymbol{P}_k^{f,e} = \tilde{\boldsymbol{X}}_k^f \left(\tilde{\boldsymbol{X}}_k^f \right)^T$, we recast the Kalman gain as follows

$$\boldsymbol{K}_k^e = \tilde{\boldsymbol{X}}_k^f \left(\mathcal{H}(\tilde{\boldsymbol{X}}_k^f) \right)^T \left(\mathcal{H}(\tilde{\boldsymbol{X}}_k^f) \left(\mathcal{H}(\tilde{\boldsymbol{X}}_k^f) \right)^T + \boldsymbol{R}_k \right)^{-1}, \tag{7.18}$$

where $\mathcal{H}(\tilde{\boldsymbol{X}}_k^f)$ is an $N_e \times N_y$ matrix. Note that if the observation operator is linear, then Eqs. (7.17) and (7.18) are equivalent.

[3]The same applies if the random variables are multiplicative with unit mean.

Besides adapting the Kalman gain, Burgers et al. [47] showed that observations should also be treated as random variables to obtain better analysis estimates. Thus, instead of updating the state analysis with the same observation y_k for every ensemble member, we should create an ensemble of perturbed observations $\boldsymbol{Y}_k = \left[y_k^{(1)}, \ldots, y_k^{(N_e)}\right] = \left[y_k + \epsilon_k^{(1)}, \ldots, y_k + \epsilon_k^{(N_e)}\right]$, where $\epsilon_k^{(i)}$ is zero-mean Gaussian noise with covariance $\boldsymbol{R}_k$.

Thus, the full analysis update of the EnKF is computed as follows

analysis:

$$\boldsymbol{x}_k^{a,(i)} = \boldsymbol{x}_k^{f,(i)} + \boldsymbol{K}_k^e \left(y_k^{(i)} - \mathcal{H}\big(\boldsymbol{x}_k^{f,(i)}\big)\right) \ , \tag{7.19}$$

$$\overline{\boldsymbol{x}_k^a} = \frac{1}{N_e}\sum_{i=1}^{N_e} \boldsymbol{x}_k^{a,(i)} = \overline{\boldsymbol{x}_k^f} + \boldsymbol{K}_k^e \left(y_k - \overline{\mathcal{H}\big(\boldsymbol{X}_k^f\big)}\right) \ , \tag{7.20}$$

$$\tilde{\boldsymbol{X}}_k^a = \left[\boldsymbol{x}_k^{a,(1)} - \overline{\boldsymbol{x}_k^a}, \ldots, \boldsymbol{x}_k^{a,(N_e)} - \overline{\boldsymbol{x}_k^a}\right] \ , \tag{7.21}$$

$$\boldsymbol{P}_k^a = \tilde{\boldsymbol{X}}_k^a (\tilde{\boldsymbol{X}}_k^a)^T \ . \tag{7.22}$$

In appendix D, we show that the analysis covariance of Eq. (7.22) is equivalent to the covariance of the original Kalman filter (see Eq. (7.9)) when one uses the linearized Kalman gain of Eq. (7.17). Moreover, we show that the analysis update $\overline{\boldsymbol{x}_a}$ lives in the space spanned by the forecast ensemble $\tilde{\boldsymbol{X}}_k^f$. Thus, if we believe that the true latent states live in a lower-dimensional manifold of the state-space $\mathcal{X}$, we can sample ensemble members from such a manifold partially mitigating the *curse of dimensionality* problem of high-dimensional models. However, the *forecast* and *analysis* equations of the EnKF are motivated by the Bayesian estimation of linear models with Gaussian random variables and, therefore, the resulting analyzes will tend to be Gaussian as well.

The advantages of using one adaptation of the Kalman gain over the other will be better discussed in the following section. The main problem arises from the rank deficiency of the approximate state error covariance $\boldsymbol{P}_k^{f,e}$. Note that the ensemble anomaly matrix is formed by N_e rows - each row corresponding to each ensemble member -, so then the rank of $\tilde{\boldsymbol{X}}_k^f (\tilde{\boldsymbol{X}}_k^f)^T$ is, at most, $N_e - 1$. Another problem arises when the state space $\mathcal{X}$ is bounded - e.g., agent states in opinion dynamics models are often described by opinions ranging between -1 (total disagreement) and 1 (total agreement) [55]. The EnKF and other filtering methods are *unconstrained*, meaning that that, in theory, the state estimates may result in any value in $\mathbb{R}^{N_x}$ and not necessarily inside the state space $\mathcal{X}$.

In the next two sections, we discuss plausible solutions to each of these two problems: imposing state constraints into the filtering equations and increasing the rank of the state covariance matrix and, therefore, of the ensemble Kalman gain. We propose different approaches to adapt the EnKF equations based on these solutions.

7.2.2.3 Constrained state space filtering

As we mentioned in the last section, most filtering methods are unbounded. This may result in state estimates that are not in the state space $\mathcal{X}$, so these estimates are wrong by default. Fortunately, we may rewrite the Kalman filtering equations taking into account the bounds imposed by $\mathcal{X}$.

Assume for simplicity that $\mathcal{X}$ is a hyper-rectangle in $\mathbb{R}^{N_x}$, i.e.,

$$\mathcal{X} = \left\{ \boldsymbol{x} \in \mathbb{R}^{N_x} : \boldsymbol{l} \leq \boldsymbol{x} \leq \boldsymbol{u} \right\},$$

where $\boldsymbol{l}$ and $\boldsymbol{u}$ are the lower and upper bounds of the hyper-rectangle, respectively. Notice that this imposes two constraints into the state variables: $\boldsymbol{l} \leq \boldsymbol{x}$ and $\boldsymbol{x} \leq \boldsymbol{u}$.

Thus, according to Simon [201], we may write the constrained estimate of the Kalman filter as a projection of the unconstrained estimate $\boldsymbol{x}^a$ into the constraint surface $\mathcal{X}$. This is equivalent to solving

$$\begin{aligned} \tilde{\boldsymbol{x}}^a \arg\min_{\boldsymbol{x} \in \mathbb{R}^{N_x}} \quad & (\boldsymbol{x} - \boldsymbol{x}^a)^T \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{x}^a)^T \\ \text{s.t.} \quad & \boldsymbol{l} \leq \boldsymbol{x} \\ & \boldsymbol{x} \leq \boldsymbol{u}, \end{aligned} \tag{7.23}$$

for $\boldsymbol{W}$ a positive-definite weighting matrix, which we take to be the identity matrix.

The optimization problem (7.23) can be decomposed into two sub-problems: one for each constraint. Without loss of generality, we focus on solving the sub-problem with inequality constraint $\boldsymbol{x} \leq \boldsymbol{u}$. Note that the components of $\boldsymbol{x}^a$ that are smaller than the corresponding components of $\boldsymbol{u}$ already solve the sub-problem. Thus, we only care about the *active* components of $\boldsymbol{x}^a$, i.e., those components greater than $\boldsymbol{u}$'s. Call $\hat{\boldsymbol{x}}^a$ and $\hat{\boldsymbol{u}}$ to the projections of $\boldsymbol{x}^a$ and $\boldsymbol{u}$ into the active components. The active components will only meet the optimality condition whenever $\hat{\boldsymbol{x}} = \hat{\boldsymbol{u}}$. Therefore, we can recast the sub-problem with inequality constraint $\boldsymbol{x} \leq \boldsymbol{u}$ into a problem with equality constraint $\hat{\boldsymbol{x}} = \hat{\boldsymbol{u}}$ while leaving the inactive components untouched. This argument is the same if we apply it for the $\boldsymbol{l} \leq \boldsymbol{x}$ constraint.

In other words, the formal solution to (7.23) is *clipping* the state estimate $\boldsymbol{x}^a$ inside $\mathcal{X}$ when the bounds of $\mathcal{X}$ are a hyper-rectangle. Doing so results in an unbiased

estimate that is closer to the true state than the unconstrained estimate at each time step [201]. This has the very practical advantage of improving the accuracy of the filtering equations at no cost. In general, even when $\mathcal{X}$ is not a hyper-rectangle, we may bound $\mathcal{X}$ with one and improve the overall state estimates.

7.2.2.4 Covariance localization: handling rank-deficiency

In the last section, we followed Simon [201] to show formally that whenever the state space $\mathcal{X}$ is bounded, clipping the state estimates within $\mathcal{X}$ improves them. However, clipping does not solve neither the rank-deficiency or the spurious correlations problems we mentioned at the end of Section 7.2.2.2. In this section, we overview *covariance localization*, a technique that deals with these latter two problems.

The EnKF approximates the *true* state covariance, $\boldsymbol{P}_k^f$, with an *approximate* state covariance, $\boldsymbol{P}_k^e$, that we compute using an ensemble of N_e states. While $\boldsymbol{P}_k^f$ is a full-rank $N_x \times N_x$ matrix, $\boldsymbol{P}_k^e$ is of rank $N_e - 1 \ll N_x$ at best. Therefore, the EnKF suffers from significant undersampling that leads to big spurious correlations, underestimation of the error covariance, and filter divergence [117]. However, most models in numerical weather prediction (NWP) lie in a geographical space, where the interaction between states is typically short-ranged. This means that the evolution of two states very far apart will be negligibly correlated. Thus, we may use information about the geography to increase the state covariance rank by filtering out the spurious correlations present in far-apart states. The problem of having rank-deficient covariance matrices is also common in econometrics [51], where often there exist spatial correlations of instrumental variables and, with too few samples, the instruments' covariance matrix is ill conditioned. To alleviate this problem, authors have applied regularization techniques [215] to increase the rank, or studied the conditions of the agents' interaction networks to identify such instruments [40].

If, on the one hand, sites i and j are apart by a great distance, d_{ij}, then the covariance between i and j should be close to 0, i.e., $(\boldsymbol{P}_k)_{ij} \to 0$ when $d_{ij} \gg 1$. On the other hand, if these sites are close to each other, then the low-rank approximation of the covariance should be a good one, i.e., $(\boldsymbol{P}_k^e)_{ij} \approx (\boldsymbol{P}_k)_{ij}$. One solution that authors in NWP have suggested [36, 118] is *tapering* the the approximate covariance $\boldsymbol{P}_k^e$ with a distance-dependent *correlation function* $\rho : \mathbb{R}^+ \to [0, 1]$, where $\rho(0) = 1$ and $\rho(\infty) = 0$, decaying monotonically. A common choice for ρ is the Gaspari-Cohn local-support approximation of a Gaussian decay [96] that depends on the distance d

as

$$\rho(d) = \begin{cases} 1 - \frac{5}{3}d^2 + \frac{5}{8}d^3 + \frac{1}{2}d^4 - \frac{1}{4}d^5 & \text{if } 0 \leq d < L \ , \\ 4 - 5d + \frac{5}{3}d^2 - \frac{1}{2}d^4 + \frac{1}{12}d^5 - \frac{2}{3d} & \text{if } L \leq d < 2L \ , \\ 0 & \text{if } d \geq 2L \ , \end{cases} \tag{7.24}$$

where L is a *localization radius* parameters that dictates how fast ρ decays with distance.

If we consider all the possible pairs of sites (i, j), we can construct a correlation matrix $\rho_{ij} = \rho(d_{ij})$ and define the *localized covariance* as follows

$$\boldsymbol{P}_k^{e,loc} := \rho \circ \boldsymbol{P}_k^e \ , \tag{7.25}$$

where $\circ$ denotes elementwise multiplication[4].

We substitute the localized covariance $\boldsymbol{P}_k^{e,loc}$ directly into the linear Kalman gain (see Eq. (7.17)) because it uses $\boldsymbol{P}_k^e$ explicitly. However, the nonlinear Kalman gain (see Eq. (7.18)) depends on $\tilde{\boldsymbol{X}}_k^f$, so introducing the correlation function ρ into the filtering equations is a harder challenge. Nevertheless, Lei et *al.* [144] did so by decomposing ρ, $\boldsymbol{P}_k^{e,loc}$, and $\boldsymbol{P}_k^{e,loc}$ into their square-root form as

$$\boldsymbol{P}_k^{e,loc} = \rho \circ \boldsymbol{P}_k^e =: (\boldsymbol{W}\boldsymbol{W}^T) \circ (\tilde{\boldsymbol{X}}_k^f (\tilde{\boldsymbol{X}}_k^f)^T) = \tilde{\boldsymbol{Z}}_k \tilde{\boldsymbol{Z}}_k^T , \tag{7.26}$$

and solving for $\tilde{\boldsymbol{Z}}_k$. Note that ρ may have several square roots, so $\boldsymbol{W}$ is an $N_x \times J$ matrix where each its columns correspond to the J leading eigenvectors of ρ, and J is a parameter tuned by the modeler. The solution for $\tilde{\boldsymbol{Z}}_k$ is given by

$$\begin{aligned} \tilde{\boldsymbol{Z}}_k = \boldsymbol{W} \triangle \tilde{\boldsymbol{X}}_k^f := \Big[& (\tilde{\boldsymbol{X}}_k^f)_1 \circ (\boldsymbol{W})_1 \ , \ldots , (\tilde{\boldsymbol{X}}_k^f)_1 \circ (\boldsymbol{W})_J \ , \\ & (\tilde{\boldsymbol{X}}_k^f)_2 \circ (\boldsymbol{W})_1 \ , \ldots , (\tilde{\boldsymbol{X}}_k^f)_2 \circ (\boldsymbol{W})_J \ , \\ & \vdots \\ & (\tilde{\boldsymbol{X}}_k^f)_{N_e} \circ (\boldsymbol{W})_1 \ , \ldots , (\tilde{\boldsymbol{X}}_k^f)_{N_e} \circ (\boldsymbol{W})_J \Big] \ , \end{aligned} \tag{7.27}$$

where the operator $\triangle$ denotes the *modulation product* and $(\boldsymbol{B})_i$ represents the i-th column of some matrix $\boldsymbol{B}$ (see proof in Appendix D.2). Note that matrix $\tilde{\boldsymbol{Z}}_k$ is of dimension $N_x \times JN_e$ and, therefore, $\boldsymbol{P}_k^{e,loc}$ is of rank $\min(N_x, JN_e - 1)$. The matrix $\tilde{\boldsymbol{Z}}_k$ is equivalent to the ensemble anomaly matrix, $\tilde{\boldsymbol{X}}_k$, but with modulated perturbations given by $\boldsymbol{W}$ such that $\boldsymbol{P}_k^{e,loc} = \tilde{\boldsymbol{Z}}_k \tilde{\boldsymbol{Z}}_k^T$. Therefore, we substitute $\tilde{\boldsymbol{X}}_k^f$ for $\tilde{\boldsymbol{Z}}_k$ in the nonlinear Kalman gain.

[4]By the Schur product theorem [[195], Theorem VII], if $\boldsymbol{P}$ is a covariance matrix and ρ is positive semi-definite, then $\rho \circ \boldsymbol{P}$ is also a covariance matrix.

The explicit localization for the linear and nonlinear Kalman gains are the following

localized linear Kalman gain:

$$\boldsymbol{K}_k^{e,loc} = (\rho \circ \boldsymbol{P}_k^{f,e})\boldsymbol{H}_k^T \left(\boldsymbol{H}_k(\rho \circ \boldsymbol{P}_k^{f,e})\boldsymbol{H}_k^T + \boldsymbol{R}_k\right)^{-1} , \tag{7.28}$$

localized nonlinear Kalman gain:

$$\boldsymbol{K}_k^{e,loc} = \tilde{\boldsymbol{Z}}_k \left(\mathcal{H}(\tilde{\boldsymbol{Z}}_k)\right)^T \left(\mathcal{H}(\tilde{\boldsymbol{Z}}_k) \left(\mathcal{H}(\tilde{\boldsymbol{Z}}_k)\right)^T + \boldsymbol{R}_k\right)^{-1} . \tag{7.29}$$

To perform covariance localization, we substitute equations (7.28) and (7.29) into the EnKF filtering equations. We will refer to the linear (nonlinear) Kalman gain by adding suffix *-lin* (*-nl*) to the EnKF notation. Note that the main difference between the EnKF-lin and the EnKF-nl stems from the former taking the Jacobian of the observation operator, $\boldsymbol{H}$, while the latter uses the full nonlinear observation operator $\mathcal{H}$. Therefore, as long as $\boldsymbol{H}$ is a good-enough approximation of $\mathcal{H}$ around the ensemble average, then both filters should perform similarly well, independent of the specific model $\boldsymbol{\mathcal{M}}$.

In the next section, we consider the case when our model's state variables interact in a network. Thus, we introduce how to perform covariance localization in *networks*. We will refer as *network ensemble Kalman filter* (NEnKF) to the EnKF adapted with network covariance localization. The NEnKF constitutes our main contribution to the latent state estimation literature for complex systems models.

7.2.3 Localization on networks

As we discussed in the previous section, using covariance localization should significantly improve the performance of the ensemble Kalman filter (EnKF) in high-dimensional systems. Localization operates by incorporating the information about the underlying geography and mitigates the spurious correlations created by the low dimension of the ensemble, which, in turn, increases the rank of the approximate covariance matrix.

However, we are mostly interested in complex systems models with states that interact through a network, and, to the best of our knowledge, no one has applied the covariance localization other than to geographical topologies. The closest we have found to network localization is the network regularization techniques found in econometrics [215, 40], where authors have identified the peer effects of instrumental variables stemming from the network formed by agent interactions. In this section,

we introduce the *network covariance localization*, which builds upon the ideas of geographical localization when the topology is a network and not geographical space. Note that everything we described in Section 7.2.2.4 is based on the assumption that interactions between states are typically short-ranged. So, if we define a notion of distance in networks, and if the approximation of short-ranged interactions still holds, then we should be able to apply the localization techniques in such a network.

Mathematically, a *network* $\mathcal{G} = (V, E)$ is a tuple consisting of a set V of $N = |V|$ nodes (or agents) and a set of $E \subseteq V \times V$ edges (or interactions) that connect pairs of nodes. We only consider networks that are finite, undirected, and connected (see Section 2.2.1 of Part I for an overview on networks).

We define the *distance* between nodes with a function $d : V \times V \to \mathbb{R}^+ \cup \{0\}$ that defines a metric and captures a notion of how far two nodes i and j are. In this chapter, we consider the *shortest-path distance*, where $d(i, j)$ is the length of the shortest path between nodes i and j and is indeed a metric [45].

The notion of distance in networks has several differences to the Euclidean or geographic distances, altering how we perform covariance localization in networks. First, network distances are discrete, while geographic distances are continuous. Thus, the way that correlation decays with distance is no longer smooth in networks, making the Gaspari-Cohn an unnatural choice for a correlation function. One could argue that for very big networks - e.g., online social networks -, one could approximate the network distance as a continuous quantity. However, complex (social) networks often exhibit a *small-world* property [228], where most or all node pairs in the network are separated by just a handful of neighbors. Put it another way, the average shortest path distance, $\langle d \rangle$, of small-world networks is *very* small. In the small-world network model of Watts and Strogatz [228], $\langle d \rangle \sim \log N$, with N the network size. Small-world networks have a very small *diameter* (or maximum distance), so the discrete nature of distance in this kind of networks is crucial for network localization.

Second, from a regularization perspective [215], if we take a localization radius (see Eq. (7.24)) of the order of the network's diameter, then the effect of regularization would be marginal. This happens because this would result in dense correlation matrices that would not help increase the rank of the original state covariance matrix. On the other extreme, if we take the correlation matrix to be the identity matix (i.e., if we take $L = 0$), then we obtain a full-rank localized covariance matrix, but with no information about the system's topology. This results in uninformative estimates of the latent states because the localized covariance would only contain the variance of the latent states in the diagonal, but not their covariances. Therefore, to perform

network localization in small-world complex networks, we need to find a sweet spot between taking localization radius of the order of the network's diameter, and taking no localization radius at all.

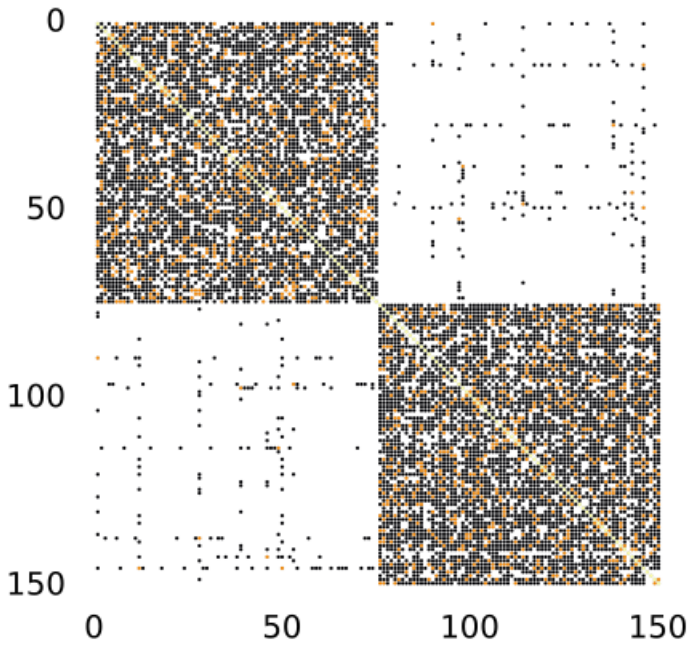


Figure 7.2: **Correlation matrix ρ of a clustered network**. We generate this network using a stochastic block model with two blocks of 75 nodes, an inner-block average degree of 10 links per node, and an inter-block average degree of 0.1 link per node. For the correlation matrix, we consider the first and second neighbors only ($L = 1$). White pixels indicate no correlation, yellow pixels indicate first neighbors, and black pixels indicate second neighbors.

For complex social networks with small-world characteristics such as the ones we have focused on in this thesis, we find it optimal to take correlation matrices that consider either the first or the first and second neighbors of any given node. More specifically, we consider correlation matrices of the form

$$\rho = \boldsymbol{I} + \lambda_1 \boldsymbol{A} + \lambda_2 \boldsymbol{A}_2, \tag{7.30}$$

with $0 \leq \lambda_1 \leq 1$ and $\lambda_2 \leq \lambda_1$ correlation parameters, $\boldsymbol{A}$ the adjacency matrix and $\boldsymbol{A}_2 = \boldsymbol{A}^2 - \boldsymbol{D}$ the second-neighbors matrix, where $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{ii}$ the degree of node i. Taking $\lambda_2 = 0$ (and $\lambda_1 \neq 0$) has the effect of only taking the first-order neighbors to tapper the covariance matrix, while taking $\lambda_2 \neq 0$ considers

the first- and second-order neighbors instead. In the results that follow, we consider $\lambda_1 = \lambda_2 = 1$, which is equivalent to taking a localization radius $L = 1$ with no correlation decay.

In Figure 7.2, we visualize the correlation matrix ρ for a clustered network that we generate using stochastic-block model [172]. Most of the entries of ρ are exactly zero (white pixels), and the only entries where ρ is 1 are for the diagonal, the first, and the second neighbors.

7.3 Results

In the previous sections, we derived the ensemble Kalman filter (EnKF) and adapted its equations to deal with the cases when the state variables are bounded (see Section 6.2.2) and when the model states interact in a network, so that we could exploit such a network topology to improve the quality of the estimations (see Section 7.2.3). We sketched out two ways of updating the analysis of the ensemble estimate: the first involves using the observation operator Jacobian to construct a linear Kalman gain, while the second uses the full observation operator to construct a nonlinear Kalman gain. For both cases, we introduced a network localization technique that increases the rank of the error state covariance matrices and overcomes the problems imposed by taking small ensembles. To the best of our knowledge, covariance localization had only been done in geographical topologies - which has been used successfully in the numerical weather prediction (NWP) literature [144] -, so we are the first to try the EnKF adapted with network localization.

In what follows, we test four variants of the EnKF to estimating latent states from aggregate observations: 1) the vanilla EnKF with a linear gain matrix, hereafter referred to as *EnKF-lin* (see Eq. (7.17)), 2) the vanilla EnKF with nonlinear gain matrix, referred to as *EnKF-nl* (see Eq. (7.18)), 3) the EnKF adapted with bound constraints, network covariance localization and a linear gain matrix, referred to as *NEnKF-lin* (see Eq. (7.28)), and 4) the analogous to the NEnKF-lin but with nonlinear gain matrix, referred to as *NEnKF-nl* (see Eq. (7.29)). Recall that the main difference between the linear and nonlinear filters is that the former approximates the observation operator $\mathcal{H}$ with its Jacobian while the latter one does not. Thus, if the Jacobian is a good approximation of $\mathcal{H}$ around the ensemble average, then both filters should perform similarly good, independent of the nonlinearity of the model $\boldsymbol{\mathcal{M}}$.

We benchmark our methods with an *importance sampling bootstrap particle filter* (PF) [74], a popular data-assimilation method that approximates the posterior with a

weighted average of particles sampled from our prior on the states variables. At each assimilation cycle, each particle is weighted by the likelihood of observing the data given forecast model estimate. In theory, the PF can approximate any distribution with enough particles, but the number of particle may get prohibitively large in high-dimensional systems [53]. We describe the PF algorithm in Appendix D.3.

We quantify the assimilation accuracy of our estimations by computing the normalized squared error (NSE) between the estimated states and the true states for both the observation and the model spaces. Mathematically, this is

$$\text{NSE}_k^{obs}(y) = \frac{(y_k - y)^2}{\sigma_y^2}, \tag{7.31}$$

$$\text{NSE}_k^{mod}(\boldsymbol{x}) = \frac{1}{N_x}(\boldsymbol{x}_k - \boldsymbol{x})^T \Sigma_{\boldsymbol{x}}^{-1}(\boldsymbol{x}_k - \boldsymbol{x}), \tag{7.32}$$

where σ_y^2 is the variance of the data, $\Sigma_{\boldsymbol{x}}$ the state covariance matrix and $(\cdot)^T$ denotes matrix transposition. In general, $\boldsymbol{x}_k$ and $\Sigma_{\boldsymbol{x}}$ are unknown to the modeler, but we use them to measure the performance of our methods in the space of the latent states.

We validate our methods with two informative models: 1) a high-dimensional approximation of the Mackey-Glass system [83], and 2) an opinion dynamics agent-based model embedded in a social network [113, 151]. In all cases, we add a small amount of stochasticity to the state variables to account for potential model misspecification. The former model serves as a paradigmatic example of assimilating high-dimensional chaotic systems, for which, in Chapter 6, we showed we could initialize with arbitrary precision if we possess enough noiseless observations and well-behaved observation operators. Notice, however, that the Mackey-Glass system is not embedded in a network, nor it is an agent-based model. We include this model to compare and validate the results we obtained on Chapter 6 with respect to the ensemble Kalman filter, given that the latter provides a more practical alternative for modelers. The latter model gives us an ideal test-bed for the NEnKF because its state space $\mathcal{X}$ is bounded, its agents interact through a social network, and it exhibits a plethora of stylized facts such as social polarization, consensus and fragmentation.

7.3.1 Mackey-Glass chaotic dynamical system

We first test the EnKF in the chaotic Mackey-Glass system, which we previously described in Section 6.3.2. In this section, we do not use the network localization techniques that we introduced in Section 7.2.3 because the states of the Mackey-Glass system are not embedded in a network topology. Recall that we test the EnKF

on this system because we want to compare the key findings from Chapter 6 with those that we get using the EnKF. From last chapter, we have that the approximated Mackey-Glass model (6.19) follows

$$\boldsymbol{\mathcal{M}}(\boldsymbol{x}_k) = \begin{cases} (\boldsymbol{x}_k)_{N_x} + \Delta t \mathscr{F}\big((\boldsymbol{x}_k)_{N_x}, (\boldsymbol{x}_k)_1\big) \\ (\boldsymbol{x}_{k+1})_1 + \Delta t \mathscr{F}\big((\boldsymbol{x}_{k+1})_1, (\boldsymbol{x}_t)_2)\big) \\ \vdots \\ (\boldsymbol{x}_{k+1})_{N_x-1} + \Delta t \mathscr{F}\big((\boldsymbol{x}_{k+1})_{N_x-1}, (\boldsymbol{x}_k)_{N_x}\big), \end{cases}$$

where $\mathscr{F}(x, y) = ay/(1 + y^c) - bx$ as in Eq. (6.17). The state-space dimension, N_x, is determined by $t_d/\Delta t$, for which we take $N_x = 100$ and $t_d = 25$ so that the system exhibits chaotic dynamics [83]. Moreover, we add noise to the state variables so that the system is stochastic. Thus, the update equations become

$$\boldsymbol{x}_{k+1} = \boldsymbol{\mathcal{M}}(\boldsymbol{x}_k) + \eta \boldsymbol{\xi}_k, \tag{7.33}$$

where $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and we choose $\eta = 0.2$.

We consider noisy observations of the form $y_k = \mathcal{H}(\boldsymbol{x}_k) + \nu \boldsymbol{\epsilon}_k$, where $\boldsymbol{\epsilon}_k \sim \mathcal{N}\left(0, \sqrt{|\Sigma_{\boldsymbol{x}}|}\right)$, with $\Sigma_{\boldsymbol{x}}$ the covariance of the state variables, and we consider $\nu = 0.2$. Similar to Chapter 6, we take the following nonlinear observation operator

$$\mathcal{H}(\boldsymbol{x}) = \sqrt[3]{\sum_{i=1}^{N_x} (\boldsymbol{x})_i^3}\ .$$

In our experiments, we construct time series of $T = 300$ observations, and we sample ground-truth states at random from the system's attractor. We initialize the EnKF and the PF with an ensemble of $N_e = 20$ particles that we sample at random from a multivariate Gaussian distribution with the mean and covariance of the attractor. We assume we know the noise level of the observations, so we set the observation error covariance as $\boldsymbol{R}_k = \eta^2 |\Sigma_{\boldsymbol{x}}|$ for all k.

In Fig. 7.3, we show typical realizations of the different filtering algorithms, where we plot the trajectories of the ground-truth latent states (black) against the estimated latent trajectories using the EnKF-lin (green), the EnKF-nl (red) and the PF (purple). We observe that both EnKFs are able to estimate the ground-truth states with remarkable accuracy for most of the observation window. However, the estimations around the first handful of observations (until $t_k \sim 30$) do not yet converge to the ground truth states. This is a consequence of sequential filtering methods in general: the state estimates for the first couple of observations are often inaccurate because

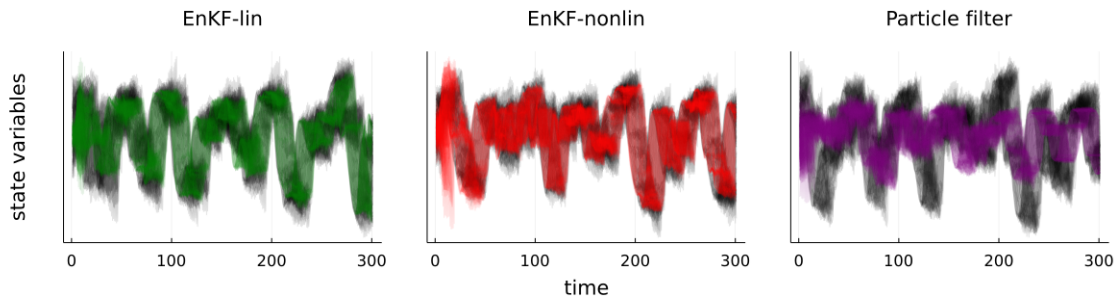


Figure 7.3: **Mackey-Glass ground-truth and estimated latent trajectories** using the EnKF with linear Kalman gain (left, green), the EnKF with nonlinear Kalman gain (center, red), and the importance sampling particle filter (right, purple). In all cases, we use $N_e = 20$ particles to estimate the latent trajectories.

these observations do not provide enough information. We also expected this adjustment phase based on the results that we obtained in Chapter 6: high-dimensional systems need several observations to have enough information to disentangle the state variables from the aggregate measurements, which we observe on the Mackey-Glass system both in Chapter 6 and this chapter. Both EnKFs exhibit a similar behavior and have a similar accuracy. On the other hand, we the PF performs significantly worse than both EnKFs. With only 20 particles, the PF does not have enough particles to approximate the posterior distribution accurately.

To be more robust, we perform 100 independent experiments with each of the methods, where in each experiment we vary the ground-truth initial conditions and the random seeds. As an additional benchmark, we run the PF with $N_e = 1000$ particles to assess how much it improves with respect to the PF with $N_e = 20$ particles. We summarize these results in Fig. 7.4, where we show the median error over all the experiments for time windows of $T = 300$ observations. The results are unequivocal: the EnKFs is significantly better than the PF for the chaotic Mackey-Glass system, even when the latter uses 1000 particles compared with the 20 particles of the EnKFs. Moreover, the EnKF-lin and the EnKF-nl show an almost identical performance, suggesting that the EnKF is a robust tool, at least with this model. Interestingly, the EnKFs converge to a stable level of error at around the 100th observation, which, considering that $N_x = 100$, is a result that agrees with our findings of Chapter 6. We observe that the EnKF observation space error is below the noise level, suggesting that the estimations cannot get any better for the given observational noise. The PF performs badly in this system. However, we observe a significant improvement when increasing the PF from 20 to 1000 particles. We believe that the PF struggles with chaotic systems because its importance sampling step does not nudge the particles at

each iteration but, instead, re-weights them.

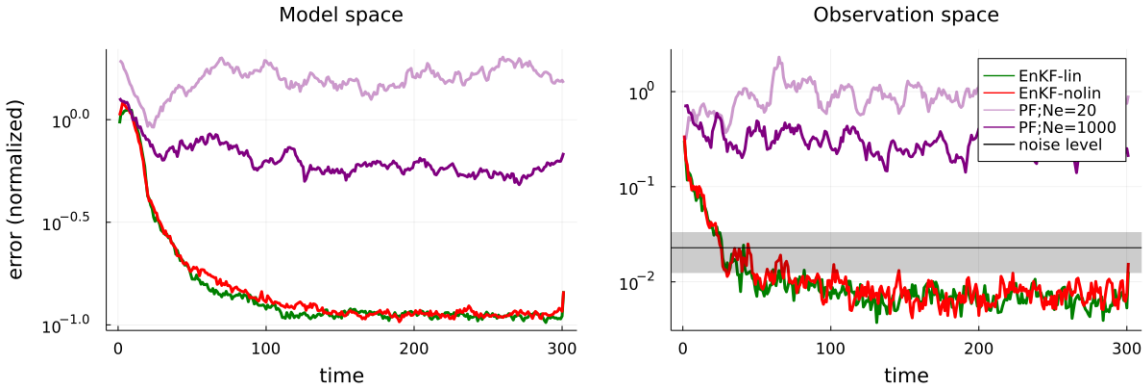


Figure 7.4: **Mackey-Glass median estimation errors** on the model (left) and observations (right) space for the EnKF-lin (green), the EnKF-nonlin (red), the PF with 20 particles (light purple), and the PF with 1000 particles (dark purple). The gray horizontal stripe in the right plot represents the range of observational noise.

These analyses show that the EnKF is a promising tool to estimate the latent states of high-dimensional systems from noisy, low-dimensional observations. While we cannot fine-tune the accuracy of the estimations with the EnKF as accurately as the method we proposed in the last chapter (see Chapter 6), the EnKF needs much fewer model simulations to operate, as it does not involve any optimization process that requires gradients. Moreover, we find that the EnKF performs significantly better than the PF in this system. We believe that this is the case because, for every observation, the EnKF nudges the particles to bring them closer to the observation. In contrast, the PF re-weights the particles to favor the ones closer to the observation without changing their state. We find that the EnKF-lin and the EnKF-nl perform almost identically well, which suggests that either form of the EnKF could be used in chaotic models as long as $\mathcal{H}$ is not heavily nonlinear. In the following section, we test of our network localization technique in an opinion dynamics agent-based model where agents interact through a social network. Further, we compare the performance of the EnKF with and without network localization.

7.3.2 Nonlinear social agent-based model

This section considers the Hegselmann-Krause agent-based model of opinion dynamics [113], where each agent is a node embedded in a social network [151]. We can describe the network $\mathcal{G} = (V, E)$ using its *adjacency matrix*, $\boldsymbol{A}$, where $A_{ij} = 1$ if agents j are i are connected, i.e., if $(i, j) \in E$, and is 0 otherwise. The state of the system is a vector

comprising the opinion of all agents. Opinions are bounded between 0 (complete disagreement) and 1 (complete agreement), so the state space is $\mathcal{X} = [0,1]^{N_x}$, where N_x is the number of agents in the network. This model is nonlinear in that the agents follow a *bounded confidence* rationale: an agent i only considers to interact - and to update her opinion - with another neighboring agent j if j's opinion is sufficiently close to the one of i. In this model, an agent i updates her opinion by considering the average opinion of all other agents she interacts with. Mathematically, the update equation for agent i is as follows

$$x_i(t_{k+1}) = \boldsymbol{\mathcal{M}}_i(\boldsymbol{x}(t_k)) = \frac{1}{|I_i^\alpha(t_k)||\boldsymbol{A}_i|}\left(x_i(t_k) + \sum_{j \in I_i^\alpha(t_k)} \boldsymbol{A}_{ij} x_j(t_k)\right), \tag{7.34}$$

where $x_i(t)$ denotes the opinion of agent i at time t, $|\cdot|$ denotes the number of nonzero elements, and

$$I_i^\alpha(t) = \Big\{ j \neq i : |x_i(t) - x_j(t)| \leq \alpha \Big\} \tag{7.35}$$

is the collection of agents in the system whose opinions differ from i's not more than a certain *confidence level* $\alpha \in [0,1]$.

The only parameter of the model, α, dictates the width of the window of interaction between agents. For any $\alpha \geq 0.4$, the system always reaches a consensus for ergodic social networks. In contrast, for $\alpha < 0.4$, the system may reach either consensus, polarization (two stable opinions), or multistability depending on the initial conditions [14]. In what follows, we set $\alpha = 0.3$.

As in the last section, we add some noise to the agents' opinions at each step to make the system stochastic. Thus, the update equations become

$$\boldsymbol{x}_{k+1} = \boldsymbol{\mathcal{M}}(\boldsymbol{x}_k) + \eta \boldsymbol{\xi}_k, \tag{7.36}$$

where $\boldsymbol{\xi}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and we choose $\eta = 0.1$. This noise can be interpreted as a case where agents are not being perfectly rational, so that their opinions change a little bit just by random chance.

For this system, we consider an observation operator that, instead of aggregating all the opinion states into a single number like in the last section, it samples the opinion states of a small fraction of the agent population. Mathematically, given $\mathcal{S} \subset \{1, \ldots, N_x\}$ a set of $|\mathcal{S}| < N_x$ indexes sampled at random without replacement, the observation operator is

$$\mathcal{H}(\boldsymbol{x}) = \{x_i \in \boldsymbol{x} : i \in \mathcal{S}\}, \tag{7.37}$$

such that we *observe* the state x_i of agent i if $i \in \mathcal{S}$. This is a case of observing incomplete information of the system. On top of this, observations are noisy, so that $y_k = \mathcal{H}(\boldsymbol{x}_k) + \nu\boldsymbol{\epsilon}_k$, where $\boldsymbol{\epsilon}_k \sim \mathcal{N}\left(0, \sqrt{|\Sigma_{\boldsymbol{x}}|}\right)$, with $\Sigma_{\boldsymbol{x}}$ the covariance of the opinion states, and we consider $\nu = 0.2$. Moreover, we assume we only observe 10% of the opinion states, so that $|\mathcal{S}| = \lceil 0.1N_x \rceil$, with $\lceil\cdot\rceil$ the *ceil* operation.

This model - and a broad class of agent-based models - can operate under a *synchronous* or an *asynchronous* updating scheme. Synchronous updating means that the states of the agents are updated all at once, i.e., every agent considers the same snapshot of the system regardless of which agent acts first. On the other hand, asynchronous updating means that the system is updated every time an agent acts so that agents acting later consider the states of those who acted before them. For the asynchronous case, agents are either activated at random or following their activation clock [141], bringing a natural layer of stochasticity to agent-based models. In the following experiments, we synchronously update the agents' opinions. We note, however, that the same final opinions are reached on average via the asynchronous updating scheme [171], so the results in what follows should work for both schemes.

Additionally, agents experience exogenous events that may alter their opinions [69]. For instance, an exogenous event can be a political party that organizes a campaign to sway the opinion of the agents towards the desired outcome [22] or by setting an agenda on mass media [155]. We model such events with random shocks that alter the system's state in the following way. At a given time $t = t_{shock}$, we shift the individual opinion of every agent by a certain amount described by a Gaussian random variable with mean μ_{shock} and variance σ^2_{shock}. The mean μ_{shock} describes the average opinion shift of each agent while the variance describes how widespread the shift will be. Mathematically, we can include the shock on top of the model as $\boldsymbol{x}(t_{k+1}) = \boldsymbol{\mathcal{M}}(\boldsymbol{x}(t_k)) + \boldsymbol{\psi}(t_k)\delta_{t_k,t_{shock}}$, where $\boldsymbol{\mathcal{M}}$ is the model and $\boldsymbol{\psi}(t) \sim \mathcal{N}(\mu_{shock}\mathbb{1}, \sigma^2_{shock}\boldsymbol{I})$ is the shock.

The agents in our model interact in a social network with two well-defined clusters: the edge density in each cluster is significantly higher than the edge density between clusters. We generate this social network using a stochastic-block model [172] with two blocks of 75 agents each, an inner-block average degree of 10 edges per agent, and an inter-block average degree of 0.1 edges per agent. We present the correlation matrix ρ of this network in Fig. 7.2, which clearly shows its cluster structure. In all cases, we simulate an exogenous shock at time $t_{shock} = 50$, where $\mu_{shock} = 0.2$ and $\sigma_{shock} = 0.2$; i.e., the exogenous shock drives the opinion state 20% towards *agreement* on average.

In the following experiments, we consider the linear and the nonlinear EnKF with and without network localization, namely *EnKF-lin*, *EnKF-nl*, *NEnKF-lin*, and *NEnKF-nl*. In all cases, we constraint the filtering equations to $[0, 1]^{N_x}$ using the clipping technique we described in Section 7.2.2.3. In the case of the NEnKF, we compute the network correlation matrices using the *shortest-path distance* (see Section 7.2.3) and the Gaspari-Cohn correlation function with a localization radius $L = 2$ (see Section 7.2.2.4). We consider time series of $T = 100$ observations constructed by following Section 7.3 and sampling ground-truth initial opinions uniformly at random for values between 0 and 1. We initialize each filtering method with an ensemble of $N_e = 20$ particles that we sample from a uniform distribution between 0 and 1. Finally, we set the observation error covariance as $\boldsymbol{R}_k = \eta^2 |\Sigma_{\boldsymbol{x}}|/2$ for all k.

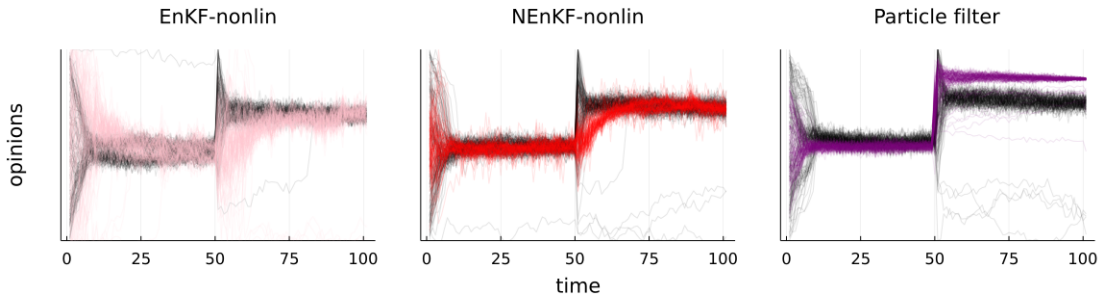


Figure 7.5: **Hegselmann-Krause ground-truth and estimated latent trajectories** for nonlinear EnKF (left, pink), the nonlinear NEnKF - which uses network localization - (center, red), and the importance sampling particle filter (right, purple). In all cases, we use $N_e = 20$ particles to estimate the latent trajectories. The sudden jump at time $k = 50$ corresponds to the exogenous shock to the opinion states. Notice how for EnKF and NEnKF, the filters take some time to adjust after the shock. As expected, the PF reacts immediately to the shock because we hard-coded the shock into the filtering equations - the PF only reweights the particles, it does not resamples them, so it is impossible for the PF to react to exogenous shocks.

In Fig. 7.5, we show typical realizations of the different filtering algorithms, where we plot the trajectories of the ground-truth latent states (black) against the estimated latent trajectories using the EnKF-nl (pink), the NEnKF-nl (red) and the PF (purple). The NEnKF is the most accurate of all filters. It quickly converges to the ground truth opinions and quickly reacts to the shock by driving the opinions into the new stable ground-truth state. The EnKF also accurately converges to the ground-truth opinions before and after the shock. However, it does so significantly slower than the NEnKF and with more noisy estimations. Finally, the PF converges to the correct average opinion before the shock but not after it[5]- notice that the

[5]In general, the PF does not behave well with exogenous shocks because it does not alter the

PF reacts immediately to the shock because we hard coded the shock into the PF equations. The PF estimates converge to opinion states with a very small spread (see how the purple trajectories are significantly thinner than the rest), which makes us believe that 1) the PF might have *degenerate* weights in this example, and 2) opinion states with small spreads might overshoot the opinion states after the shock. These results already suggest that using network localization helps the EnKF to find more accurate estimates of the state variables: incorporating the network topology into the filtering equations gives them additional information about the latent states.

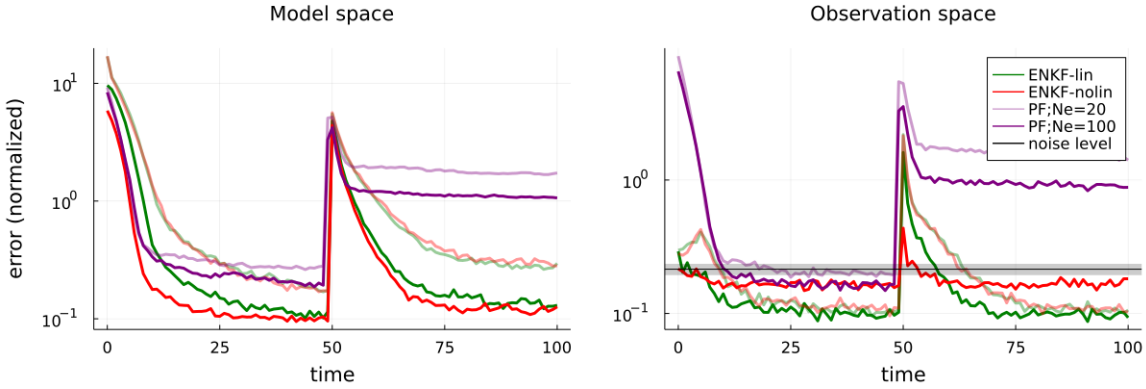


Figure 7.6: **Hegselmann-Krause median estimation errors** on the model (left) and observations (right) spaces. In each plot, we show the normalized squared error for the EnKF-lin (light green), the NEnKF-lin (dark green), the EnKF-nl (light red), the NEnKF-nl (dark red), the PF with 20 particles (light purple), and the PF with 100 particles (dark purple). The gray horizontal stripe in the right plot represents the range of observational noise. Recall that the PF reacts immediately to the exogenous shock - at $t_{shock} = 50$ - because we hard-coded the shock into the filtering equations.

In order to draw stronger conclusions on the performance of the NEnKF, we repeat 100 independent experiments for each of the methods, including the linear variants of the EnKF and the NEnKF which we did not include in Fig. 7.5. Moreover, we run the PF with $N_e = 20$ and $N_e = 100$ particles to assess the improvement of the PF with increasing number of particles. In each experiment we vary the initial opinions, the set $\mathcal{S}$ of observed agents and the random seed. We keep the social network fixed for all the experiments. We summarize these results in Fig. 7.6, where we show the median error over all the experiments for time windows of $T = 100$ observations. On the one hand, in the observations space all methods converge to below noise level before the shock, meaning that they easily find the stable opinion state. After the shock, however, the PF consistently fails to converge to the right

particle states but reweights them, so anything that the model does not strictly generate will be hard to assimilate. For this reason, we hard-coded the shock into the PF so that it reacts to it.

opinion state (possibly by over- or under-shooting it). Both versions of the EnKF and the NEnKF converge below noise level after the shock, but the NEnKFs converges faster than the EnKFs. On the other hand, in the model space we find that the NEnKFs are consistently better than the ENKFs and the PF: they estimate the latent opinions more accurately and faster than the other filtering methods. In both the observation and model spaces, the EnKF-lin and EnKF-nl have an almost identical behavior. This is something we expect because the observation operator (7.37) is a linear projection of the opinions to a lower-dimensional space, making the EnKF-lin and EnKF-nl equivalent in this case. In contrast, the NEnKF-lin is not equivalent to the NEnKF-nl because the latter uses the network correlation matrix ρ directly while the NEnKF-nl has to perform the square-root approximation we described in Section 7.2.2.4. Nevertheless, the performances of the NEnKF-lin and the NEnKF-nl are very similar, which suggests that we can potentially use either version depending on our computational constraints. All in all, the NEnKFs outperformed all the other filtering methods, which is promising. We hope these experiments pave the way for future research where we explore how using the network topology in estimation algorithms might improve performance with real-world data.

7.4 Discussion & Conclusions

We tackled the problem of estimating the latent states of complex system models with network topologies from aggregate observations using data assimilation techniques. The purpose of this chapter is twofold. On the one hand, we bridged the gap between the complex systems modeling and the data assimilation literature. We formulated the state estimation problem from a Bayesian framework and derived the ensemble Kalman filter (EnKF) equations under the assumptions of linear operators and Gaussian random variables. We discussed the advantages and the limits of the EnKF and discussed how to overcome them using constrained filtering [201] and model-space localization techniques [117], where we introduced localization for network topologies. On the other hand, we showed that the network localized EnKF (or NEnKF) is a promising tool for sequentially estimating the latent states of complex system models because 1) it works in our two test cases, 2) it requires a small amount of model simulations - compared with the dimension of the system -, 3) it uses the model as a black box, and 4) we modified the Kalman filtering equations to adapt to the needs of network-based models and justified these modifications formally.

We validated the EnKF and the NEnKF against a high-dimensional approximation of the Mackey-Glass chaotic system [152, 83] and the Hegselmann-Krause opinion dynamics agent-based model where agents interact through a social network [113, 151]. We considered observation operators that either aggregated the latent states or gave us incomplete information about them. Moreover, we benchmarked our methods with an importance sampling particle filter (PF) [74]. We found accurate and robust results in both systems that outperformed the PF in most cases, where each system provided us with valuable insights that deserve special attention.

The Mackey-Glass system is a chaotic dissipative system, so its dynamics relax into a chaotic attractor. Thus, we initialized ensembles of states directly from the attractor statistics, reducing the state space to a lower-dimensional space. Additionally, having chaotic dynamics let us make out-of-sample predictions of the system and compare them with its predictability horizon. We obtained results consistent with the previous chapter (see Chapter 6). As the Mackey-Glass model is not embedded in any topology, we did not use the localization techniques for the EnKF. We found that the linear and the nonlinear EnKF (EnKF-lin and EnKF-nl, respectively) perform similarly in this case. Both EnKFs significantly outperformed the PF even when the latter operated with significantly more particles, suggesting that the EnKF is an efficient alternative to estimate high-dimensional chaotic systems.

The Hegselmann-Krause model is a nonlinear opinion dynamics agent-based model where agents interact through a social network, so this model served as a testbed for other social complex systems models. In this case, we assimilated transient opinions of the agents as they reached equilibrium. Moreover, we tested our filtering algorithms in the presence of exogenous shocks and showed that the Kalman-based filtering equations significantly outperformed the PF, even when the latter used a larger number of particles. More importantly, the NeNKf outperformed the EnKF, suggesting that incorporating the network topology into the filtering equations can be beneficial for obtaining more accurate estimation of the latent states.

This work provides a practical framework for connecting simulation models with real-world aggregate data. The EnKF has the potential of becoming a standard method, but it comes with the following series of limitations that we should consider for future work.

First, the EnKF tends to Gaussianize the estimations by design while the distribution of latent states for many realistic systems is non-Gaussian. However, authors have extended the EnKF with a mixture of Gaussians [75] and, by using an expectation-maximization algorithm [184], they extract the means and variances of the mixture

modes. While this approach would have a higher computational complexity than the vanilla EnKF, it is much more efficient than particle filters.

Second, we considered state spaces that were subsets of $\mathbb{R}^{N_x}$, limiting the class of simulation models with binary or categorical latent states. Authors have already considered extending the Kalman filter with a binary filter to estimate state variables from mixed and binary observations [180].

Third, microscopic socio-economic models are often built on untested theories rather than physical laws, so they are often imperfect. The EnKF has successfully estimated the states of imperfect models in the numerical weather prediction communities [159, 117], so we should test the EnKF for such models.

Fourth, sequential filtering methods return estimations on the fly. Thus, it is likely that these estimations are inaccurate for the first handful of observations because there is not enough information to determine latent states that are close to the feasible set of solutions. This behavior happens whenever there is a big exogenous shock in the system - which we showed in the Hegselmann-Krause model - because the shock destroys the information from the past, so the filters have to adjust to the new system state. To speed up the adjustment time for exogenous shocks, we could introduce an adaptive Kalman gain which increases the weight on the observations near the shock, and decreases such a weight once the filter readjusts. Moreover, we could employ Kalman smoothers [53] to refine the estimated trajectories of the latent states at the end of the simulations. Smoothers have the desired advantage of using information of both the past and the future to improve the latent state estimations at the same computational cost of filters.

Finally, we believe that using surrogate models to estimate the latent states of a more complex model should be further explored. In the calibration literature, authors have proposed using machine-learning surrogates to estimate the parameters of complex agent-based models [142]. Based on our observations in the Hegselmann-Krause model (see Section 7.3.2), we propose to use models with synchronous updating schemes as a surrogate for the same model with asynchronous updating. Amelkin et al. [14] showed that models using asynchronous schemes have the same average behavior as their synchronous counterparts for certain opinion dynamics models. We should explore this interplay in a wider class of agent-based models The dynamics of ergodic models always converge to a measure-preserving attractor and thus are suitable for state estimation [105].

This chapter builds upon the methodology we proposed in Chapter 6, where we analyzed the relation between the number of observations in a time series with the

dimension of the latent states' state space by incorporating and extending the EnKF methodology proposed by Ward et al. [226]. Hopefully, we have shed some light for future research and discussions that bring us closer to connecting simulation models with real-world data.

Chapter 8

Conclusions

The interplay between data and social systems is a complex and dynamic one. Data mining and data assimilation techniques can be used to better understand how social systems work and how they are changing. The ability to access granular data at a large scale has allowed researchers to test theories and make inferences about social systems that were not possible before. This is particularly relevant now[1] that many global events shaping and destabilizing society are taking place.

This thesis studied social systems by focusing on the individuals: how they act, interact with each other, and connect to conform to a social system as a cohesive unit. Our approach was connecting social information at individual level - either as granular data or a complex systems model - with the emerging global properties of the system they conform. To do this, we split the thesis into two parts: one where we analyzed social systems when individual-level data were available, and one where we inferred the individual-level data when they were not.

In Part I, we quantified system-level structures, such as polarization or echo chambers, and studied the collective emotion and risk perception stemming from user-level data on massive social media platforms, such as Twitter. We analyzed data of users tweeting about topics such as Covid-19 and climate change. In Part II, we built upon the data assimilation literature to explore the connection between the individual-level states of a complex systems model and the data that describes them. We developed techniques that use incomplete data to help specify complex systems models and validated these methods in chaotic systems and a social agent-based model of opinion dynamics.

More in detail, Chapter 3 studied the emotional reaction from the Twitter public during the Covid-19 pandemic. We analyzed how the language employed in Covid-

[1] We are writing this thesis in 2022.

19-related tweets evolved from March 2020, which was the onset on the pandemic in most Western countries, to June 2020, when many countries were either at the peak or at the decay of the first wave. For this analysis, we created country-specific linguistic profiles from the tweets based on a word-to-linguistic-category dictionary curated by psychologists. We thus tracked the dynamics of these profiles and performed time-series regressions and word-co-occurrence network analyses. We found that Twitter users had a very strong emotional reaction to Covid-19's first wave that wore off over time. Simultaneously, users increasingly fixated on mortality, but in a decreasingly emotional and increasingly analytic tone. Our findings suggested that Twitter users experienced *psychophysical numbing* during the first wave of the pandemic. In simple terms, that means that people exhibited growing indifference towards human suffering as the number of humans suffering increased [203, 90].

This chapter has potential implications for how to deal with societies when they experience events that cause widespread suffering. However, the study was limited by its bias towards the specific Twitter users that posted about Covid-19, which might not be representative of the whole population. Moreover, we did not discriminate between human and non-human users - such as bots or institutional accounts - which might have influenced our results. For future investigations, we will analyze if the psychophysical numbing hypothesis holds for longer time periods and to what extent. While we found an intense emotional response and attention to Covid-19 during the first wave, such a response may not hold for subsequent waves. Could we react less intensely because the pandemic has stopped being novel? Do we develop an emotional resilience to threats we have experienced previously? Answering these questions will not only help us be better prepared for future pandemics, but it will also help guiding us in other - potentially bigger - threats such as climate change.

In Chapter 4, we quantified the interaction structure dynamics of Twitter users tweeting about climate change during 2019, when important climate movements flourished. We identified clear ideological poles based on the retweeting patterns between users. More specifically, we studied the information sources - here called *chambers* - retweeted by the audience of the conversation's leading users. Considering the chamber overlap between two users as a proxy of their ideological similarity, we identified *echo chambers* of climate believers and climate skeptics. Users with similar (contrasting) ideological positions showed significantly high (low)-overlapping chambers, resulting in a bimodal overlap distribution. In short, we found that the climate change Twitter conversation is highly polarized, and that such polarization was roughly constant throughout most of 2019. Moreover, we found an exception of this almost

constant behavior when the biggest *"FridaysForFuture"* occurred (20-27 September, 2019) [214], where the polarization decreased with respect to the rest of the year.

This chapter highlighted the importance and usefulness of exploiting the structural information present in social media networks, especially when looking at controversial conversations. Our methodology is computationally cheap, readily usable for other Twitter datasets, and does not suffer from the selection bias of supervised approaches. Furthermore, we showed that if the conversation is polarized enough, we could identify echo chambers just by looking at the handful of users that produce the most popular tweets, which is easier than deploying clustering algorithms on the whole interaction network. From a social point of view, this condition shows that the climate-related Twitter conversation –and possibly many other Twitter conversations [101]– has a low complexity, meaning that we can identify large scale structures within the conversation just from the activity of the few leading users. For future investigations, we will focus of the tweets' content - and not only on the interaction structures - in each echo chamber. This will help us, for instance, to find the causes behind the polarization decrease during the "FridaysForFuture" strikes. The methods we introduced in Chapter 4 helped us create a taxonomy of the users discussing climate change in Twitter, but they did not inform us about the discourse dynamics within and between climate believers and skeptics. Overall, we wished to understand what mechanisms drive the changes in polarization and use them to foster a better climate-change communication.

We believe that Part I's methods could be improved by discussing their design and applicability with sociologists, psychologists and other social scientists. Creating an interdisciplinary channel of communication would highly benefit the empirical study of social systems from the bottom-up in this era of increasing volumes of granular data [147]. However, we do not always possess granular data, and even when we do, they are often not representative of the entirety of the system. According to DeAngelis and Díaz [71], bottom-up models, such as agent-based ones, can simulate realistic individuals living in and reacting to an environment based on their internal states. Yet, if data at the individual-level are not available, then deciding what the internal, latent states of these individuals should be becomes a big challenge. This was the approach that we followed in Part II.

In Chapter 6, we studied the problem of inferring the latent initial state of a dynamical system under incomplete information, i.e., we assumed we observed aggregate statistics of the system rather than its state variables directly. Studying several model systems, we inferred the latent initial condition that best reproduce an

observed time series when the observations are sparse, noisy, and aggregated under a (possibly) nonlinear observation operator. We did this by first filtering the noise out of the observations, then exploring the feasible space of solutions, and finally estimating the initial condition using gradient-based algorithms. We validated our method on two chaotic dynamical systems, where we obtained accurate out-of-sample predictions. Further, we analyzed the predicting power of our method as a function of the number of observations available, the amount of information destroyed by the observation operator, and the properties of the dynamical system. We found that, the more high-frequency oscillations the dynamical system produces, the harder it is to initialize. Finally, we showed that the Mackey-Glass system [152] undergoes a critical transition for which, after a certain number of observations, we were able to initialize it with arbitrary precision.

This chapter provided a conceptual framework to understand the interface between aggregate data and microscopic interactions. However, it was limited to the case in which we knew the model that perfectly specifies the system, as well as the observation operator and the characteristics of the noise in the observations. It was also limited in that it required to compute the model's gradient to find the initial latent state, whereas many bottom-up models from the social sciences may have non-differentiable components. Finally, this method focused on finding the initial condition that describe the whole trajectory of the system. In more practical scenarios, it might be desirable to infer the latent variables as observations become available instead of trying to find a single, fits-all initial condition.

Finally, in Chapter 7, we introduced the network ensemble Kalman filter (NEnKF), a computationally-efficient method to estimate the latent states of complex systems models as observations become available. The NEnKF extends the ensemble Kalman filter [81], which is heavily used in the numerical weather prediction community [117], by incorporating the network topology of the underlying model into the filtering equations. In addition, we showed that when the model's state variables are bounded (e.g., an agent's opinion state could be between 0 and 1), clipping the inferred states inside the bounds is justified by imposing inequality constraints into the filtering equations [201]. By doing so, the NEnKF accounted for the model's network structure and the state-space boundaries, which led to improved estimates of the latent states. We validated the NEnKF in the chaotic Mackey-Glass system [83] and the Hegselmann-Krausse agent-based model of bounded-confidence opinion dynamics [113].

One of the key findings of the NEnKF was that providing more structure to the estimation algorithms leads to better results. By incorporating the network topol-

ogy and state bounds into the ensemble Kalman filter equations, we were able to significantly improve the quality of assimilated latent states. The advantage of using Kalman-based assimilation techniques is that they treat the model as a black box, require relatively few model simulations, work for nonlinear systems, and are simple to implement. Going forward, we plan to test the NEnKF on real social trace data. For instance, we could validate the effectiveness of the NEnKF by aggregating granular Twitter data using a nonlinear observation operator and inferring the granular data back using the NEnKF.

All in all, this thesis described social systems based on what information we have about them. We developed data-driven methods to better understand social complex system for two particular scenarios: when we possess high-quality, granular level data and when we do not. In the latter case, we additionally assumed we possessed an accurate model of the system of interest and devised methods to infer the granular data from incomplete information. As such, these scenarios describe two extremes of the same problem: we never have perfect data nor a perfect model. This fact leads us back to the original question we proposed in the title of this thesis: "*how do data assimilation and social media - and social complex systems - connect?*". More granular data allows us to develop a greater understanding of the mechanisms driving the evolution of social systems, which in turn lets us construct better bottom-up models. Data assimilation then allows us to use these models to draw inferences about individuals, even when data are not granular. These two forces - constructing better models and inferring individual-level information - could create a positive feedback loop for analyzing social systems and help find effective tipping points to guide policy and action that drives society toward becoming more equal, sustainable, and resilient.

Appendix A

Public risk perception and emotion during the Covid-19 pandemic

A.1 Further model comparison

In this section, we present further results of our models to give a more complete overview of their quality. Besides the Weber-Fechner law and power law models (see Eqs. (3.10) and (3.11)), we use the following linear relationship between s and p as our benchmark model

$$p(t) = a \cdot s(t) + b, \tag{A.1}$$

where a and b are parameters. We summarize our results for the linear model in Table A.1.

For all models, we compute the R^2 values

$$R^2 = 1 - \frac{\sum_{t=1}^{n} e(t)^2}{(n-1)\sigma_p^2}, \tag{A.2}$$

where $e(t) = p(t) - \hat{p}(t)$ is the model residual, $\sigma_p^2 = \sum_{t=1}^{n}(p(t) - \mu_p)^2/(n-1)$ is the variance of $p(t)$, and n is the sample size. The R^2 values for all models are summarized in Table A.2. (Note that as the power law model implies a log-normal residual, the R^2 values can be negative.) From this table we see that, once again, the Weber-Fechner law is generally a better fit to the data across all countries, but that the power law and Weber-Fechner models are often comparable and significantly better than the linear model.

We also show in Figures A.1 and A.2 scatterplots of the Death NLSs against the logarithm of the daily number of deaths in each country, with the y-axis in linear- and log-scales, respectively. Red lines indicate the line of best fit, with the slope equal to k and β in Eqs. (3.10) and (3.11), respectively.

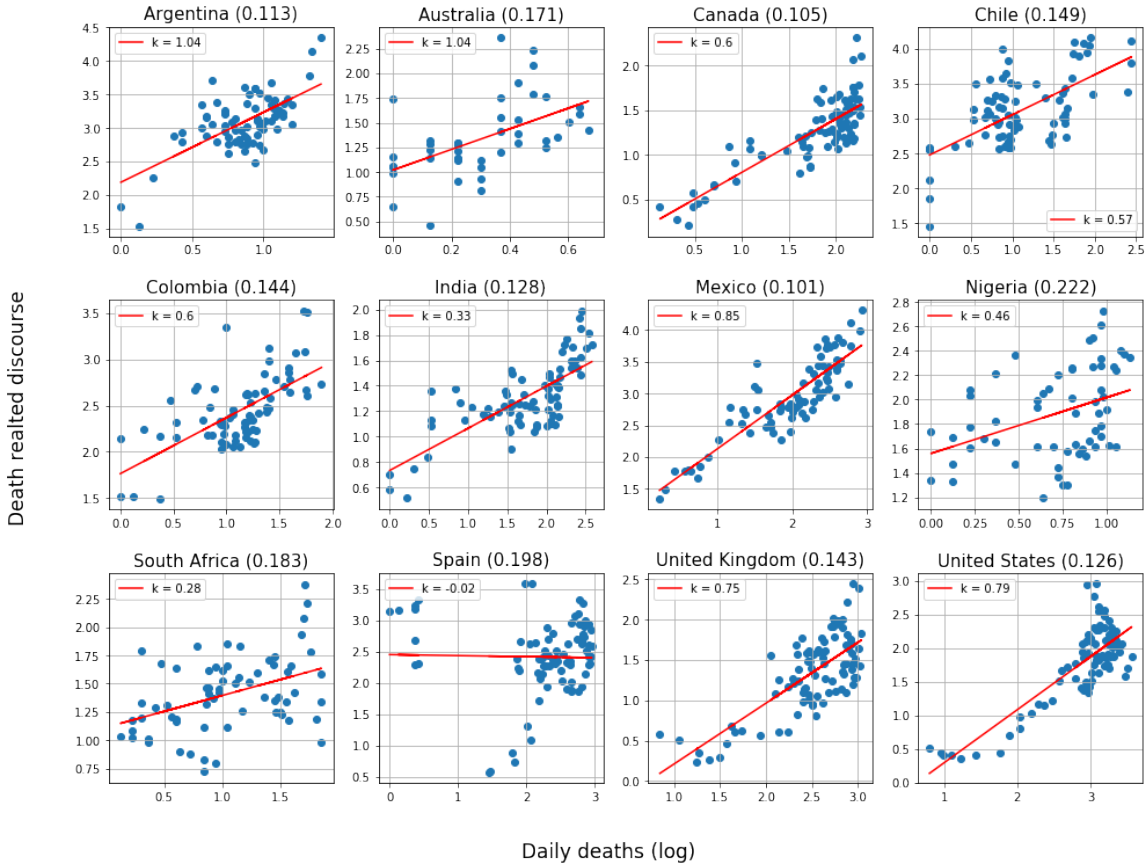


Figure A.1: Resulting scatter plot for the **Weber-Fechner law** model fit, where each panel shows a different country with their corresponding NRMSE in parenthesis (the lower the better).

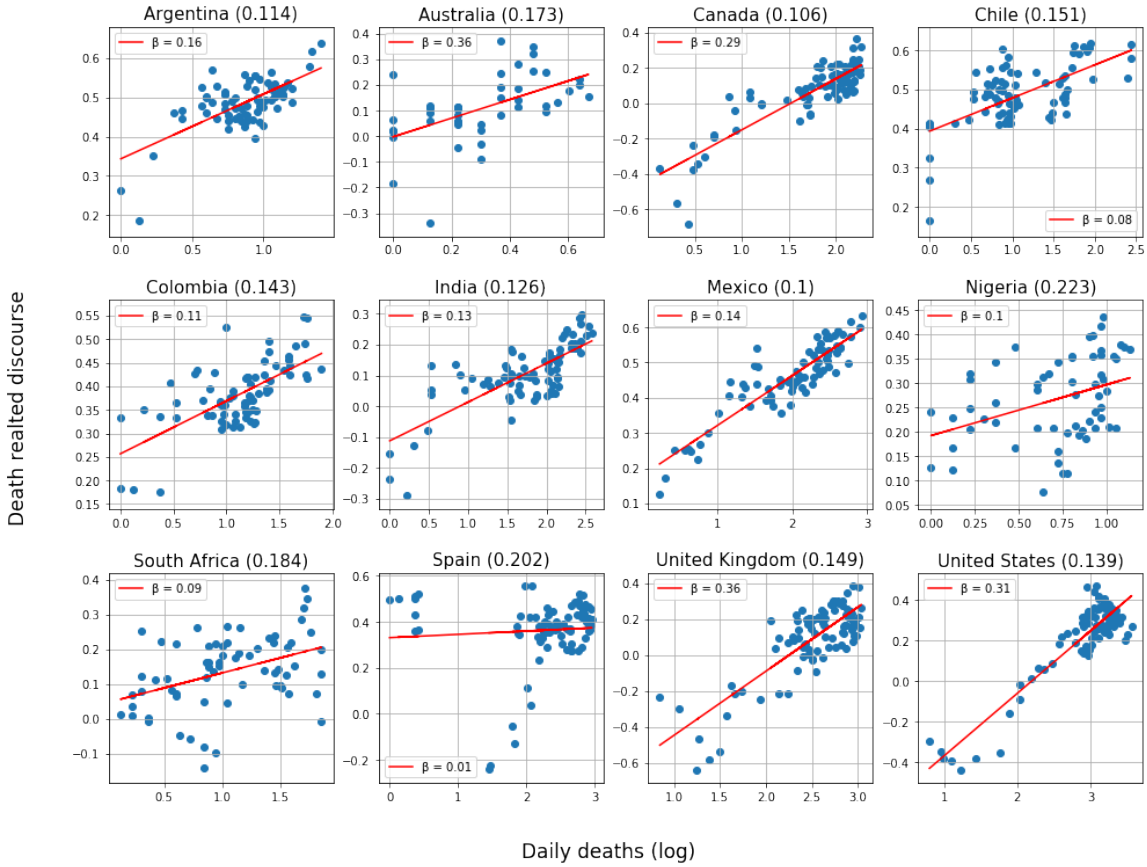


Figure A.2: Resulting scatter plot for the **power law** model fit, where each panel shows a different country with their corresponding NRMSE in parenthesis (the lower the better).

Country	a	b	95% CI (a)	t (a)	$P > \|t\|$ (a)	R^2	NRMSE	n
Argentina	0.057	2.603	0.04 – 0.074	6.79	0.0	0.387	0.116	75
Australia	0.192	0.912	0.087 – 0.298	3.67	0.0007	0.247	0.175	43
Canada	0.005	0.759	0.005 – 0.006	11.46	0.0	0.607	0.117	87
Chile	0.005	2.969	0.003 – 0.007	4.58	0.0	0.216	0.17	78
Colombia	0.016	2.147	0.011 – 0.02	7.11	0.0	0.409	0.145	75
India	0.002	1.074	0.002 – 0.003	10.14	0.0	0.566	0.125	81
Mexico	0.003	2.428	0.002 – 0.003	10.97	0.0	0.613	0.133	78
Nigeria	0.047	1.589	0.021 – 0.073	3.61	0.0006	0.184	0.218	60
South Africa	0.006	1.295	0.002 – 0.01	2.95	0.0045	0.119	0.188	66
Spain	0	2.248	-0.0 – 0.001	1.98	0.0506	0.047	0.193	82
United Kingdom	0.001	0.877	0.001 – 0.001	7.69	0.0	0.399	0.166	91
United States	0	1.235	0.0 – 0.001	6.84	0.0	0.352	0.179	88

Table A.1: Results for the linear model defined in Eq. (A.1).

A.2 National Linguistic Scores

A.2.1 Exogenous peaks in the National Linguistic Scores

In this section, we address significant deviations in the National Linguistic Scores from our proposal of psychophysical numbing as an explanation for their trends over the observation period, and suggest possible explanations for their occurrence, see Table A.3. We stress that the following Table might be prone to error although we double checked every peak.

A.3 Word co-occurrence analysis

A.3.1 Further technical details on co-occurrence network construction

In constructing the word co-occurrence networks presented in Section 3.3.2.1, we preform basic text preprocessing, including taking the lower-case form of all letters, removing URLs, removing punctuation, and removing the following small set of stop-words from the vocabulary:

`to, today, too, has, have, like.`

We retain hashtags, since LIWC also recognises hashtags and because hashtags are an essential aspect to communications on Twitter. It is also necessary to account for the fact that a number of "words" appearing in the LIWC dictionary are in fact regular

R^2 **Country**	Power law	Weber-Fechner law	Linear relationship
Argentina	0.411	**0.421**	0.387
Australia	0.259	**0.275**	0.247
Canada	0.678	**0.683**	0.607
Chile	0.382	**0.395**	0.216
Colombia	**0.425**	0.319	0.28
India	**0.558**	0.397	0.477
Mexico	**0.78**	0.775	0.613
Nigeria	**0.143**	0.004	0.007
South Africa	0.16	**0.171**	0.119
Spain	-0.042	0	**0.047**
United Kingdom	0.514	**0.555**	0.399
United States	**0.608**	0.556	0.24
Mean	0.36	**0.379**	0.303
Proportion of best fits	41.7 %	**50** %	8.33 %
Proportion of second-best fits	**66.7** %	33.3 %	0 %

Table A.2: Comparison of R^2 between the power law model of Eq. (3.11), the Weber-Fechner model of Eq. (3.10) and a linear relationship between variables, which we use as a benchmark model. Higher values indicate better models.

expressions to which many complete words in the Twitter dataset map. For example, the "word" "isolat*" appears in the English LIWC dictionary, to which each of the following words would map: "isolate", "isolated", "isolating". Thus, construction of the word co-occurrence networks G_i' involves a two-step procedure: first, constructing the raw word co-occurrence networks G_i, in which the nodes are words exactly as they appear in the Twitter dataset; and then reducing this to a quotient graph G_i' by contracting nodes in G_i that are matched by the same regular expression in the LIWC dictionary. More formally: the LIWC dictionary implies an equivalence relation $\sim$ on the vocabulary $\mathcal{V}$ implied by the Twitter dataset, such that $v \sim u$ for words $v, u \in \mathcal{V}$ if both v and u are matched by the same regular expression in the LIWC dictionary. The weights of edges between nodes $v' \subset \mathcal{V}$ and $u' \subset \mathcal{V}$ in G_i' are then taken to be

$$w_{G_i'}(u', v') = \sum_{u \in u', v \in v'} w_{G_i}(u, v), \tag{A.3}$$

where $w_G(x, y)$ is the weight of edge (x, y) in G. Note that $w_G(x, y) = w_G(y, x)$ and $w_G(x, y) = 0$ if (x, y) is not an edge in G.

To construct the higher-frequency sequences of snapshots, we impose a minimum document frequency of 5×10^{-3} (2.5×10^{-3} for Spanish tweets) for each term in the

Country	Date	Category	Description
United States	June 2, 2020	Anger	Donald Trump responding near the White House to the protests against the murder of George Floyd [64]
Australia, Nigeria, United Kingdom	June 6-7, 2020	Anger	Worldwide protests against the murder of George Floyd in the United States [9]
Chile	May 13-15, 2020	Anger	Chilean Health Ministry announces total lockdown [7].
Argentina	March 20, 2020	Death	President Alberto Fernandez announces total national lockdown [2].
Australia	May 20-26, 2020	Death	New Covid-19 deaths after several days without casualties. Moreover, the murder of George Floyd in the United States took place on May 25 [5].
Canada	May 2, 2020	Death	*Unknown*
Colombia	March 22, 2020	Death	Shootout in a prison triggered by prisoners demanding better hygiene conditions for Covid-19 results in 23 deaths. [6]
India	March 29, 2020	Death	*Unknown*
Nigeria	April 18, 2020	Death	Many important African political figures die from Covid-19 this day [1].
Spain	May 25, 2020	Death	Correction in Covid-19 data repositories show negative daily deaths [10].
Spain	June 1, 2020	Death	First day in Spain without Covid-19 deaths [4].
United Kingdom	April 10, 2020	Death	The United Kingdom surpasses 10,000 Covid-19 deaths [3].
United States	May 25, 2020	Death	Murder of George Floyd [5]

Table A.3: A list of plausible explanations for anomalous peaks observed in Figure 3.1. Most of these peaks arise from the murder of George Floyd and the consequent protests. We focused mainly on the Death and emotionally-charged categories as these are the ones that are most related to the psychophysical numbing effect we describe in the main text.

Country	**Argentina**	**Australia**	**Canada**	**Chile**
N_{tweets}	3.5×10^4	2×10^4	5×10^4	1×10^4
Country	**Colombia**	**India**	**Mexico**	**Nigeria**
N_{tweets}	1.5×10^4	5×10^4	4×10^4	2×10^4
Country	**South Africa**	**Spain**	**United Kingdom**	**United States**
N_{tweets}	1×10^4	5×10^4	1×10^5	1×10^5

Table A.4: The number of tweets taken per snapshot for each country.

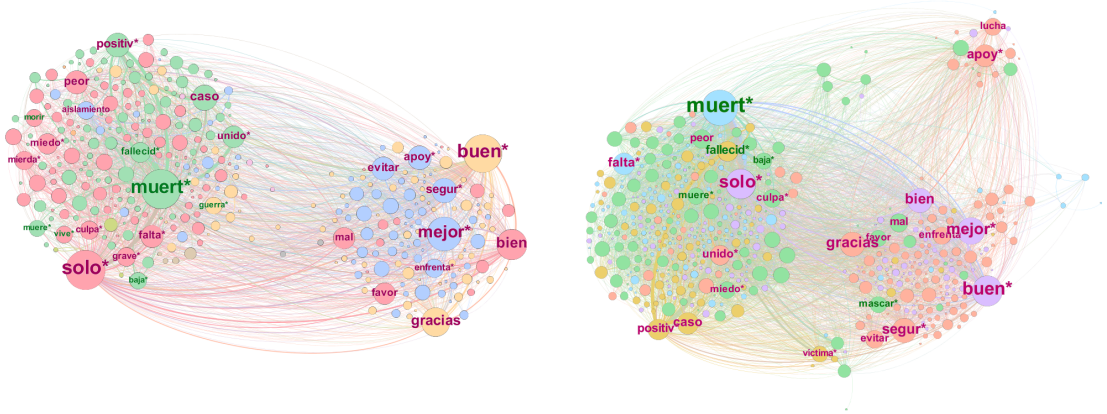
vocabulary in order to reduce the effect of noise. In Table A.4, we summarise the approximate number N_{tweets} of tweets per snapshot for each country. The number of tweets per snapshot for each country was chosen in order that each country had approximately the same number of data points separated by approximately 3 days, and such that edge effects did not yield a final snapshot with a disproportionately low number of tweets. While this ultimately led to some snapshots representing aggregation over longer periods than others, this yielded sequences of networks that are comparable in terms of their total strength and order, enabling reasonably fair comparison of the modularities of the partition induced by the Death and Affect LIWC categories.

A.3.2 Word co-occurrence networks for Spanish-language tweets

For completeness, we provide in Figure A.3 the word co-occurrence graphs for the Spanish language tweets. We omit a discussion of the results, since similar conclusions can be drawn from these as in the English counterparts.

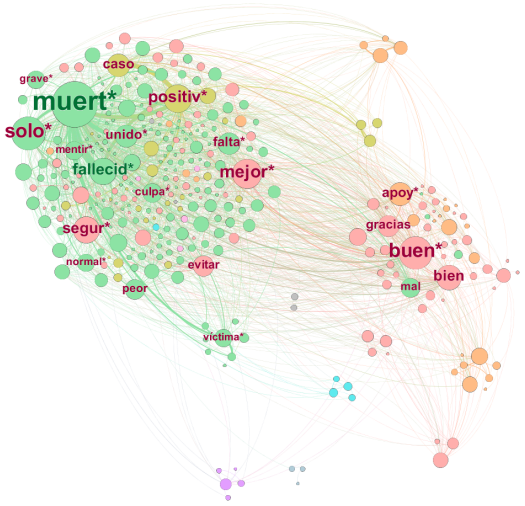
A.4 Covid-19 epidemiological data

We include this section as a reference for the actual number of deaths in each country for the period we analysed throughout the paper, which we present in Fig. A.4.



(a) March 11th to April 9th, 2020.

(b) April 10th to May 23rd, 2020.



(c) May 24th to June 14th, 2020.

Figure A.3: Snapshots of the word co-occurrences associated with death ("muerte", green labels) and affect ("afecto", red labels) for Spanish-language tweets aggregated across all analyzed countries in three different time windows (see sub-captions). The nodes are coloured based on the community labels obtained by maximising modularity using the Louvain algorithm [38]. We filtered edges with weight below 20 co-occurrences for visualisation purposes.

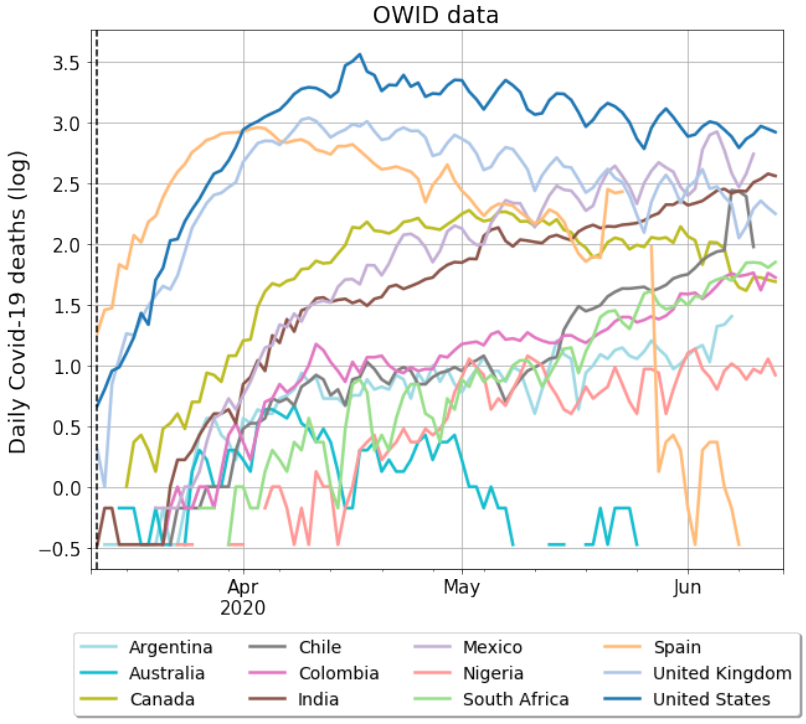


Figure A.4: Daily deaths related to Covid-19 for each of the countries in our analysis (see legend) from March 11 to June 14, 2020.

Appendix B

Quantifying the structure of the climate change conversation with unsupervised methods

B.1 Soft configuration model

In order to estimate the overlap distribution in a random network we analyze the degree sequence $\boldsymbol{k}$ of the empirical retweet network (with adjacency matrix $\boldsymbol{A}$) using the configuration model [[115], Chapter 7] with a soft constraint on the in-degree

$$p(\mathbf{A}|\mathbf{k}) = \prod_{i=1}^{N} \prod_{j \neq i} \left[\frac{k_i}{N} \delta_{A_{ji},1} + \left(1 - \frac{k_i}{N}\right) \delta_{A_{ji},0} \right]. \tag{B.1}$$

This choice of randomness results in networks with a locally tree-like structure, more realistic choices can be proposed but they are in general harder to control when trying to match with real networks, [149, 91]. The relevant observables are also harder to compute analytically, we therefore choose to work with the locally tree like structure of the configuration model.

We can then approximate the expected overlap of the configuration model as follows

$$\langle q_{ij}^{\mathcal{C}} \rangle = \left\langle \frac{\sum_\ell \mathbb{I}\left[\ell \in \mathcal{C}_i \vee \ell \in \mathcal{C}_j\right]}{\sum_\ell \mathbb{I}\left[\ell \in \mathcal{C}_i \wedge \ell \in \mathcal{C}_j\right]} \right\rangle \approx \frac{\sum_\ell \text{Prob}\left[\ell \in \mathcal{C}_i\right] \text{Prob}\left[\ell \in \mathcal{C}_j\right]}{\sum_\ell \text{Prob}\left[\ell \in \mathcal{C}_i\right] + \text{Prob}\left[\ell \in \mathcal{C}_j\right]}, \tag{B.2}$$

which is the sum of joint probabilities of a user ℓ being in both $\mathcal{C}_j$) and $\mathcal{C}_i$, assuming independence. We need to calculate this probability, which for the configuration model it is simply

$$\text{Prob}\left[\ell \in \mathcal{C}_i\right] = 1 - \text{Prob}\left[\sum_{j=1}^{N} A_{ji} A_{j\ell} = 0\right] = 1 - \left(1 - \frac{k_i k_\ell}{N^2}\right)^N, \tag{B.3}$$

where $\text{Prob}\left[A_{ji}A_{j\ell}=0\right]$ is the probability of user j retweeting i and not ℓ or viceversa.

We can do a similar analysis for the overlap similarity between audiences:

$$\left\langle q_{ij}^{\mathcal{A}}\right\rangle = \left\langle \frac{\sum_\ell \mathbb{I}\left[\ell\in\mathcal{A}_i \vee \ell\in\mathcal{A}_j\right]}{\sum_\ell \mathbb{I}\left[\ell\in\mathcal{A}_i \wedge \ell\in\mathcal{A}_j\right]}\right\rangle \approx \frac{\sum_\ell \text{Prob}\left[\ell\in\mathcal{A}_i\right]\text{Prob}\left[\ell\in\mathcal{A}_j\right]}{\sum_\ell \text{Prob}\left[\ell\in\mathcal{A}_i\right]+\text{Prob}\left[\ell\in\mathcal{A}_j\right]}, \tag{B.4}$$

where

$$\text{Prob}\left[\ell\in\mathcal{A}_i\right] = \frac{k_i}{N} \tag{B.5}$$

Substituting Eq. (B.3) into $q_{ij}^{\mathcal{C}}$ and Eq. (B.5) into $q_{ij}^{\mathcal{A}}$ we obtain

$$\left\langle q_{ij}^{\mathcal{C}}\right\rangle = \frac{\left\langle \left[1-\left(1-\frac{k_ik}{N^2}\right)^N\right]\left[1-\left(1-\frac{k_jk}{N^2}\right)^N\right]\right\rangle_{p(k)}}{\left\langle 2-\left(1-\frac{k_ik}{N^2}\right)^N-\left(1-\frac{k_jk}{N^2}\right)^N\right\rangle_{p(k)}} \tag{B.6}$$

$$\left\langle q_{ij}^{\mathcal{A}}\right\rangle = \frac{k_ik_j}{N(k_i+k_j)}, \tag{B.7}$$

so if $k_{max}\ll N$ and $\langle k^2\rangle > \langle k\rangle$, then it follows that

$$\left\langle q_{ij}^{\mathcal{C}}\right\rangle \approx \frac{k_ik_j}{N(k_i+k_j)}\frac{\langle k^2\rangle}{\langle k\rangle} > \frac{k_ik_j}{N(k_i+k_j)} = \left\langle q_{ij}^{\mathcal{A}}\right\rangle. \tag{B.8}$$

In most social networks we have that the degree variance, $\langle k^2\rangle$, is much larger than the average degree, $\langle k\rangle$. Therefore, we expect the chamber overlap to be significantly larger than the audience overlap.

B.2 Unsupervised clustering

We take an spectral clustering approach to obtain the communities of the leading users. We consider the similarity Laplacian matrix, $\mathbf{L}=\mathbf{D}-\mathbf{Q}$, where $\mathbf{Q}$ is the overlap similarity matrix aggregated over all the weeks in the dataset, and $\mathbf{D}$ is a diagonal matrix with the degree sequence of $\boldsymbol{Q}$ in the diagonal, i.e., we have that $\mathbf{D}_{ii}=\sum_j q_{ij}$ and $D_{ij}=0$ for $i\neq j$.

By the spectral properties of $\boldsymbol{L}$, we know that connected networks have a smallest eigenvalue of 0 followed by positive eigenvalues. The second smallest eigenvalue $\lambda_2>0$, also called the *algebraic connectivity*, reflects how well-connected the network is. Its associated eigenvector, $\boldsymbol{u}_2$, also called the *Fiedler vector*, is typically used to partition networks into two non-trivial communities that minimize that cut size, i.e., it minimizes the sum of weights between the groups [161].

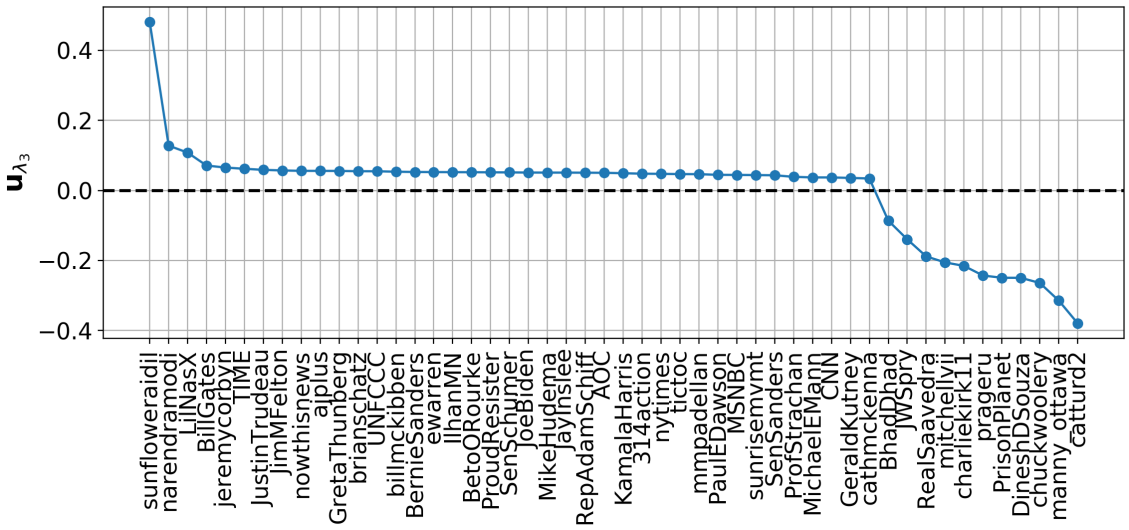


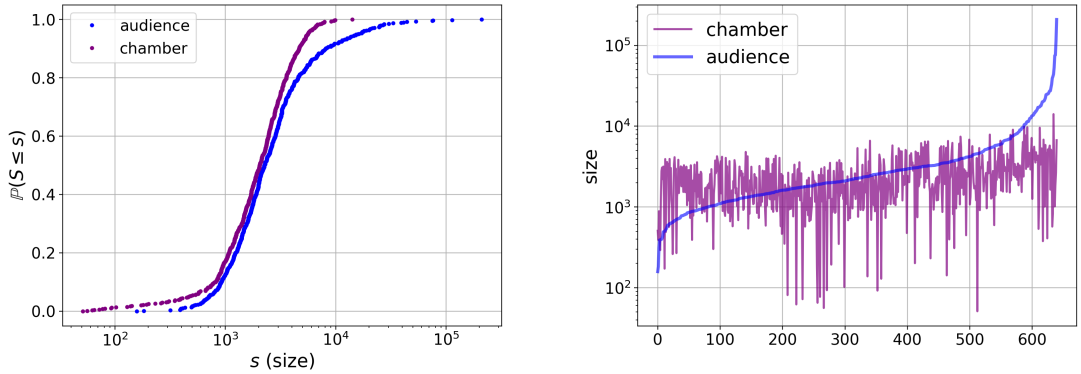
Figure B.1: **Eigenvector associated with λ_3, the third leading eigenvalue, sorted by magnitude.** We identify between climate believers and skeptics with the sign of their corresponding component in $\mathbf{u}_{\lambda_3}$.

In the case of the Laplacian of the aggregate overlap similarity matrix, the second smallest eigenvector separates the satellite users (namely, *BhadDhad*, *narendramodi*, *sunfloweraidi* and *LilNasX*) from the rest of the network. While the separation associated the second eigenvalue separates the network in almost independent connected components, it does not capture the groups seuggested by the bimodality of the overlap distribution. Instead, we take the eigenvector associated with the *third* smallest eigenvalue, λ_3, to partition the leading users into two groups. We identify the entries of $\boldsymbol{u}_3$ that are *greater than zero* as *climate believers* while the entries *smaller than zero* as *climate skeptics*. In Fig. B.1, we show the individual entries of $\boldsymbol{u}_3$, where each entry corresponds to a leading user.

B.3 Comparing chambers and audiences

Many recent studies about polarization on social media have used a first-neighbors approach to characterize the structure of the interaction networks [60, 126, 58, 232]. In Twitter, retweets have been used extensively as a proxy for endorsement [23, 97], so retweet networks have been used to study the assortative structure of several conversations. In this work's terminology, these studies have studied the *audience* of a set of users to quantify polarization, echo chambers, and other nontrivial social structures. We argue that the chamber is, under certain conditions, a more robust tool to quantify such social structures, both conceptually and mathematically. The

difference being that looking at the chamber corresponds to looking at the information sources of the audience, a second-neighbors approach.



(a) Cumulative distribution function

(b) Audience-ordered sizes

Figure B.2: **Chamber and audience sizes.** *a)* Cumulative distribution function. We construct each distribution by concatenating the sizes of the chambers (purple) (audiences (blue)) of the leading users, $\mathcal{I}^{\Delta}(t)$, for every week t on the dataset. The audiences size distribution ($s = 4668 \pm 11472$) has a significantly longer right-tail than that of the chambers ($s = 2425 \pm 1682$). *b)* Audience-sized ordered chamber (purple) and audience (blue) sizes from smallest to largest for the leading users for all weeks. We find a small correlation of $\rho = 0.26$ between audiences and chambers sizes.

From a mathematical perspective, for the null model the *expected overlap* between the *audiences* of two high-impact users is significantly lower than the expected overlap between chambers (see Appendix B.1 for details). Which suggests that typically the chamber overlap should have a higher signal to noise ratio. Moreover, in the case of the climate change retweet network the distribution of chamber sizes has a significantly lower spread than that of the audience, as we show in Fig. B.2a. In other words, many chambers are of roughly the same size while the audiences sizes vary a lot. The overlap similarity is informative whenever the size of the two sets compared are of the same order, so using the chamber gives us more reliable results than the audience. For two arbitrary sets A and B where $|A| \ll |B|$, the Jaccard overlap $|A \cap B|/|A \cup B| \approx |A|/|B|$, which indicates the relative size of A with respect to B and not their overlap.

Besides mathematical arguments, we argue that the audience and the chamber are fundamentally different objects. By construction, one would expect that the chamber is tightly coupled to the audience because the audience determines the chamber. For instance, if a user has a big audience, we expect her to have a big chamber as well.

Surprisingly, the correlation correlation between audience and chamber sizes is very low ($\rho = 0.26$), suggesting that they are loosely coupled (see Fig. B.2b).

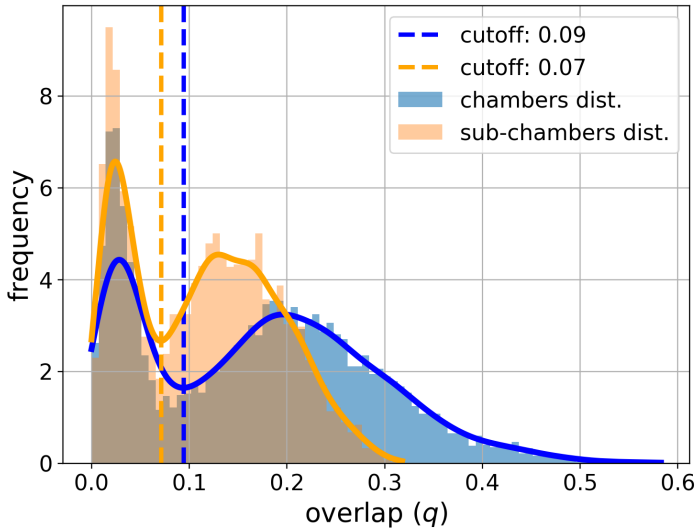


Figure B.3: **Aggregate chamber and subchamber overlap distributions.** We construct the aggregate chamber (blue) (subchamber (orange)) distribution by concatenating the overlap pairs, q_{ij}^t, for every week t on the dataset – i.e., $\boldsymbol{q} = (q_{ij}^t)_{t,i<j}$. A subchamber differs from a chamber in that we remove the overlapping users of the audience before constructing between the subchamber (see main text for details on their construction). While the first peak of the subchamber distribution $q_{off}^s = 0.03 \pm 0.02$ is almost identical to that of the chamber distribution ($q_{off} = 0.04 \pm 0.02$), the second peak, $q_{in} = 0.16 \pm 0.5$ is lower and has a smaller spread ($q_{in} = 0.23 \pm 0.08$). However, both peaks are clearly bimodal and have similar cutoff values.

We further analyze the behavior of the chambers by removing the coupling of their audiences. To do so, when comparing high-impact users i and j, we remove the common audience members of i and j and construct their chambers *without* them. In Fig. B.3, we compare the overlap similarity distributions with and without removing the common audience members. Both are bimodal. We observe that the expected overlap of the second peak decreases but is still significantly high. Moreover, both distributions share the same bimodal structure with a similar cutoff, indicating that we only removed redundant information by removing the coupling caused by the audiences. This suggests that the chamber is more robust to missing information.

Finally, we argue that the audience is a collection of information *consumers*, whereas a chamber is a collection of information *sources*. On the one hand, every member of the audience consumes information from the leading user in a traditional

one-to-many fashion. On the other hand, the audience consumes information from the chamber in a many-to-many fashion, where the ideological coherence comes from considering the chamber as a whole and not its individual members. Thus, the overlap between chambers reflects the similarity between the many-to-many information channels of the audiences, giving us a proxy of the ideological (dis)similarities between leading users/leading users.

B.4 Leading users features

We show in Table B.1 the list of leading users and their main characteristics.

B.5 Descriptions and definitions for polarization and echo chambers

In the main text, we argue that while many authors have studied polarization and echo chambers in online social networks, there is not a consensus yet of what these concepts mean in detail. Intuitively, the polarization between two groups refers to the tendency of each group to interact mostly within its own members, rejecting the interaction with the other group. Similarly, an echo chamber promotes in-group communication while rejecting out-group communication. Moreover, members in an echo chamber tend to systematically reinforce their beliefs because the information posted from a given member of an echo chamber *bounces* back to that member through the flow of communication with other users inside it.

In Tables B.2 and B.3, we describe different notions of polarization and echo chambers, respectively, based on the literature on online social networks.

leading user	Persistence	Total impact	Median impact	Chamber size	Ideology (spectral)	Ideology (manual)
BernieSanders	24	410442	5093	35811	believers	believers
LilNasX	10	320133	10301	6583	believers	other
PaulEDawson	39	282040	7349	88025	believers	believers
AOC	23	243209	7481	40302	believers	believers
CNN	30	158068	3077	38677	believers	other
ewarren	23	116018	3519	29140	believers	believers
MikeHudema	29	112690	3345	49874	believers	believers
sunfloweraidil	7	102276	3951	1989	believers	believers
SenSanders	19	94747	2936	26297	believers	believers
RepAdamSchiff	8	92173	9855	21765	believers	believers
ajplus	21	78273	3179	20646	believers	other
sunrisemvmt	8	77320	1802	11992	believers	believers
GeraldKutney	31	73966	2180	51325	believers	believers
KamalaHarris	12	69900	4365	16269	believers	believers
IlhanMN	7	69250	4806	13716	believers	believers
JimMFelton	9	66397	5102	12995	believers	believers
GretaThunberg	9	64787	5480	20537	believers	believers
BetoORourke	12	59042	2455	16247	believers	believers
nowthisnews	17	58796	3094	18299	believers	other
nytimes	15	49493	2174	19417	believers	other
JayInslee	16	44600	2594	20705	believers	believers
narendramodi	7	43984	4203	3669	believers	believers
JoeBiden	9	40170	3084	8333	believers	believers
ProudResister	9	38873	3048	12921	believers	believers
brianschatz	10	36951	2959	10740	believers	believers
mmpadellan	7	34335	4047	13277	believers	believers
jeremycorbyn	7	33573	2569	11454	believers	believers
UNFCCC	16	32000	2047	24574	believers	believers
SenSchumer	9	31598	3098	10875	believers	believers
JustinTrudeau	8	30146	3582	10281	believers	believers
ProfStrachan	14	28860	1786	26876	believers	believers
billmckibben	8	26331	2398	16595	believers	believers
314action	11	25086	2215	8853	believers	believers
tictoc	7	22500	2588	10495	believers	other
cathmckenna	10	22239	1762	15957	believers	believers
BillGates	8	19071	2400	4872	believers	believers
MichaelEMann	9	17947	1966	18596	believers	believers
TIME	7	16815	2039	7665	believers	other
MSNBC	7	14651	1882	8401	believers	other
PrisonPlanet	15	154928	5180	29046	skeptics	skeptics
DineshDSouza	13	75806	4515	14267	skeptics	skeptics
charliekirk11	7	65267	8427	16081	skeptics	skeptics
BhadDhad	7	60552	2111	2352	skeptics	other
chuckwoolery	9	55050	3691	13988	skeptics	skeptics
RealSaavedra	9	53977	3109	11506	skeptics	skeptics
catturd2	7	53477	3783	14776	skeptics	skeptics
prageru	8	32300	2825	14277	skeptics	skeptics
manny_ottawa	12	28981	2256	8308	skeptics	skeptics
JWSpry	13	21588	1690	16439	skeptics	skeptics
mitchellvii	8	17876	1914	5769	skeptics	skeptics

Table B.1: Description of the $M = 50$ leading users defined in Eq. (4.4) in terms of their persistence, Δ_i, their cumulative impact, $\sum_t w_i^t$, their median impact, $Med(w_i^t)_t$, their aggregate chamber size, $|\cup_t \mathcal{C}_i^t|$, and their ideology based on manual labelling and the spectral clustering algorithm described in Appendix B.2. Users are ordered by ideology according to the spectral clustering algorithm and from largest to smallest impacts.

Concept (Polarization)	**Description**	**Equation**
Polarization index [26]	"A high value indicates that an user systematically interacts with people belonging to her/his same coalition rather than with users from different alliances, but we disregard any information concerning the content of the shared news and we do not take into account how it is amplified by being shared within the same group of people."	$\arg\max_{\alpha} \frac{\lvert\mathcal{N}_i \cap C_\alpha\rvert}{\lvert\mathcal{N}_i\rvert}$, $\mathcal{N}_i$: set of leading users i interacted with $\mathcal{C}_\alpha$ the set of leading users in community α
Adaptive E-I index [58]	"We measure polarization as the relative density of in-group agreement to out-group agreement. We assess these patterns on our endorsement networks, which are a subtype of general communication networks where all ties are publicly conveyed indications of agreement"	$\frac{n^t_{\alpha\alpha} + n^t_{\beta\beta} - (n^t_{\alpha\beta} + n^t_{\beta\alpha})}{n^t_{\alpha\alpha} + n^t_{\beta\beta} + (n^t_{\alpha\beta} + n^t_{\beta\alpha})}$, the notation is the same as in Eq. (4.9).
Random walker controversy measure (RWC) [95]	"Controversial and polarized topics induce graphs with clustered structure, representing different opinions and points of view" "The (random walker) measure captures the intuition of how likely a random user on either side is to be exposed to authoritative content from the opposing side."	$p_{\alpha\alpha}p_{\beta\beta} - p_{\alpha\beta}p_{\beta\alpha}$, $p_{\alpha\beta}$: probability that a random walker starting in community α ends in community beta.

Table B.2: Description of the polarization concept used as in the literature, as well as their associated mathematical equation.

Concept	Description
Echo chamber [43]	"An echo chamber comes into being where a group of participants choose to preferentially connect with each other, to the exclusion of outsiders. The more fully formed this [social] network is, the more isolated from the introduction of outside views is the group, while the views of its members are able to circulate widely within it"
Echo chamber [95]	"Opinions or beliefs stay inside communities created by like-minded people, who reinforce and endorse the opinions of each other"
Echo chamber [20]	"Individuals are exposed only to information from like-minded individuals"
Echo chamber [60]	"Environments in which the opinion, political leaning, or belief of users about a topic gets reinforced due to repeated interactions with peers or sources having similar tendencies and attitudes. Selective exposure and confirmation bias may explain the emergence of echo chambers on social media."
Echo chamber [211]	"One of the consequences of the group polarization phenomenon is to cast new light on an old point, to the effect that social homogeneity can be quite damaging to good deliberation. When people are hearing echoes of their own voices, the consequence may be far more than support and reinforcement. Another consequence is that particular forms of homogeneity can be breeding grounds for unjustified extremism, even fanaticism"

Table B.3: Description of the main notions an echo chamber. In general, echo chambers form as a *consequence* on group polarization, in which, *because* group polarization exists, members inside those groups are only exposed to information of like-minded individuals and *hear echoes* of their own voices.

Appendix C

Estimating initial conditions from incomplete information

C.1 Noiseless non-autonomous linear systems

In this section, we assume that the dynamical system (5.1), the observations (5.2), and the corresponding dynamical mapping (5.3) are given by time-varying linear functions. We further assume that the dynamics are deterministic and the observations are noiseless so that $\boldsymbol{\xi}_k = 0$ and $\epsilon_k = 0$ for all k. Making these assumptions provides us with two advantages over arbitrary nonlinear dynamical systems: 1) linear systems are much easier to handle than nonlinear systems, and 2) nonlinear systems can be approximated by time-varying systems arbitrarily well if they are locally Lipschitz [218].

Additionally, we find it convenient to slightly change the notation introduced in the main text. In the main text, we indexed the observations of the system so that the time stamps describe the observations $\boldsymbol{y}$ in the most natural way. Thus, we defined t_k such that $y_k = y(t_k)$ is the k-th data point of the series. Consequently, we indexed the evolution of the microstates as $\boldsymbol{x}(t_{k+1}) = \boldsymbol{x}(t_k + m\Delta t) = \boldsymbol{\mathcal{M}}(\boldsymbol{x}(t_k))$, i.e., we needed to update the system m times before sampling the next observation. Here, we index the passing of time in the time scale of the microstates, so that $\boldsymbol{x}(t_{k+1}) = \boldsymbol{x}(t_k + \Delta t) = \boldsymbol{\mathcal{M}}(\boldsymbol{x}(t_k))$. Thus, we label the observed time series as $\boldsymbol{y} = (y_0, y_{m-1}, \ldots, y_{mT-1})$ so that y_{mk-1} is the k-th data point of a time series of T observations. This approach emphasizes that $\boldsymbol{y}$ is a coarse-grained sample of the underlying dynamics of the microstates.

Under the above considerations, we can recast Eqs. (5.1-5.2) respectively as follows

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}_k \boldsymbol{x}_k = (\boldsymbol{F}_k \boldsymbol{F}_{k-1} \cdots \boldsymbol{F}_0)\boldsymbol{x}_0 \ , \tag{C.1}$$

$$y_k = \boldsymbol{H}_k \boldsymbol{x}_k \ , \tag{C.2}$$

where, at every time t_k, $\boldsymbol{F}_k$ is an $N_x \times N_x$ matrix representing a linear dynamical process and $\boldsymbol{H}_k$ is an $1 \times N_x$ matrix representing a linear observation operator. We assume that every element of $\boldsymbol{H}_k$ is non zero for every k. Note that under the current notation, the sequence $y_0, y_1, \ldots, y_k, \ldots$ represents the ground-truth dynamics under the (time-varying) observation operator $\boldsymbol{H}_k$, and only those indexes k that are a multiple of m are included in the observations $\boldsymbol{y}$.

From the RHS of Eq. (C.1), we see the explicit dependence of any observation y_k from the initial conditions $\boldsymbol{x}_0$, so we can define an $1 \times N_x$ matrix[1]

$$\boldsymbol{M}_k := \boldsymbol{H}_k \boldsymbol{F}_{k-1} \cdots \boldsymbol{F}_0$$

that takes us from $\boldsymbol{x}_0$ to y_k for any k. Thus, if we possess a time series of T observations such that the last observations happens at time t_{mT-1}, then we can recast the microstate initialization problem as the following linear system of equations of N_x variables and mT equations

$$\begin{aligned} y_0 &= \boldsymbol{M}_0 \boldsymbol{x}_0 \\ y_1 &= \boldsymbol{M}_1 \boldsymbol{x}_0 \\ &\vdots \\ y_{mT-1} &= \boldsymbol{M}_{mT-1} \boldsymbol{x}_0 \ , \end{aligned}$$

or, more compactly,

$$\boldsymbol{y}^* = \boldsymbol{M}^* \boldsymbol{x}_0, \tag{C.3}$$

where $\boldsymbol{y}^* = (y_0, y_1, \ldots, y_{mT-1}) \in \mathbb{R}^{mT}$ is the *extended* sequence of observations (it is extended in that it includes all the observations in $\boldsymbol{y} \in \mathbb{R}^T$ plus its intermediate, unobserved samples) and $\boldsymbol{M}^* = [\boldsymbol{M}_0 | \boldsymbol{M}_1 | \ldots | \boldsymbol{M}_{mT-1}]$, the *extended observation matrix* of size $mT \times N_x$.

We may solve system (C.3) exactly whenever $\boldsymbol{M}$ is invertible. If the matrices $\boldsymbol{M}_k$ are non-singular, then $\boldsymbol{M}^*$ becomes invertible when $mT = N_x$.

We do not possess $\boldsymbol{y}^*$ but the coarse-grained time series $\boldsymbol{y}$ of T observations, where two consecutive samples are m time steps apart. Thus, we may only express a reduced form of system (C.3) with T equations and N_x variables as

$$\boldsymbol{y} = \boldsymbol{M} \boldsymbol{x}_0, \tag{C.4}$$

[1] We define matrix $\boldsymbol{M}$ for time t_0 as $\boldsymbol{M}_0 := \boldsymbol{H}_0$.

with $\boldsymbol{M} = [\boldsymbol{M}_0|\boldsymbol{M}_{m-1}|\dots|\boldsymbol{M}_{mT-1}]$ of size $T \times N_x$. The system (C.4) spans the same time interval as system (C.3), but it has m less equations, so it is underdetermined and might have several solutions.

Under the conditions we describe in what follows, we may obtain a solution $\boldsymbol{x}$ of the reduced system (C.4) that is also a solution of the extended system (C.3). If $\boldsymbol{y}$ consists of $T = N_x/m$ (or more) data points, the solution $\boldsymbol{x}$ is unique and equal to the ground-truth microstate $\boldsymbol{x}_0$. More specifically, if the sampling frequency $(m\Delta t)^{-1}$ is higher than twice the cutoff frequency of the spectrum of the system and the matrices $\boldsymbol{M}_k$ are non singular for $k \in \{0, m-1, \dots, mT-1\}$, then the time series $\boldsymbol{y}$ determines the extended series $\boldsymbol{y}^*$ uniquely, and therefore any solution $\boldsymbol{x}$ that solves (C.4) also solves (C.3). The above conditions establish the necessary and sufficient conditions for the Nyquist-Shannon sampling theorem [198] to be true, so our result is a direct application of the theorem.

Note that the power spectra of the systems considered in this work (and most chaotic systems) exhibit a power-law decay on their power spectrum, so there is not a well-defined cutoff frequency on neither the Mackey-Glass nor the Lorenz systems. Nevertheless, if their power spectrum decays fast enough, we should be able to define an effective cutoff frequency for the previous arguments to be a good approximation. We will investigate the validity of our arguments in future iterations of this work.

In this paper, we chose non-degenerate dynamical systems that are Lipschitz continuous and chose an observation operator (see Eq. (6.15)) that is bijective to the linear operator $\boldsymbol{H} = [1, \dots, 1]^\dagger$. Therefore, the results of this Appendix apply in all our systems whenever the observations are noiseless. In particular, the results of this Appendix explain the critical transition we show in Figs. 6.5 and 6.6 for $T = N_x/m = 25$.

C.2 Comparison of nonlinear observation operators

We assess the robustness of our initialization method using different (nonlinear) observation operators. The way we aggregate the microstates affects how much information we retain about the latent dynamics, so we consider the three following operators each

with different levels of coupling between the microstate components

$$\mathcal{H}_1(\boldsymbol{x}) = \sqrt[3]{S_{\boldsymbol{x}}}, \tag{C.5}$$

$$\mathcal{H}_2(\boldsymbol{x}) = sign\,(P_{\boldsymbol{x}})\ \sqrt[N_x]{|P_{\boldsymbol{x}}|}, \tag{C.6}$$

$$\mathcal{H}_3(\boldsymbol{x}) = sign\,(C_{\boldsymbol{x}})\ \sqrt{|C_{\boldsymbol{x}}|}, \tag{C.7}$$

where

$$S_{\boldsymbol{x}} = \sum_{i=1}^{N_x} \boldsymbol{x}_i^3, \tag{C.8}$$

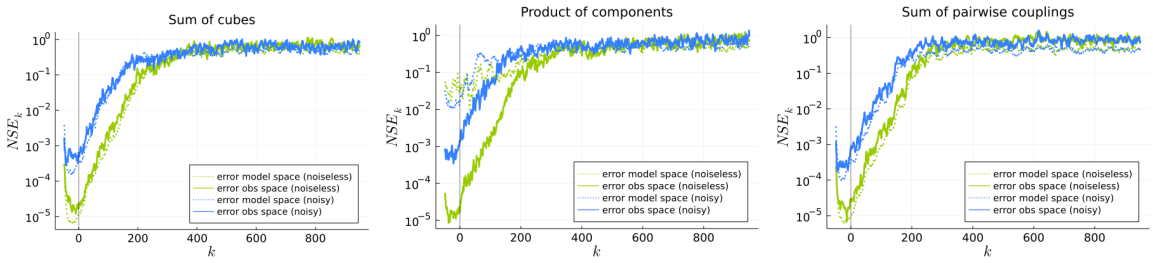
$$P_{\boldsymbol{x}} = \prod_{i=1}^{N_x} \boldsymbol{x}_i, \tag{C.9}$$

$$C_{\boldsymbol{x}} = \sum_{i<j} \boldsymbol{x}_i \boldsymbol{x}_j. \tag{C.10}$$

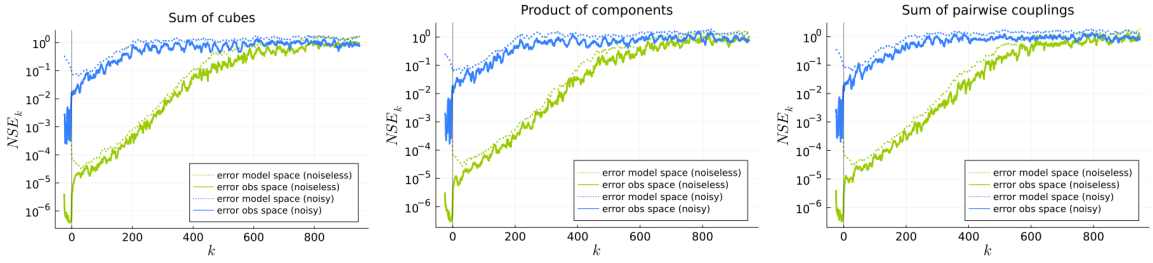
Our rationale for choosing these operators goes as follows. We point out that Eq. (C.5) is the same as Eq. (6.15): it's the cubic root of sum of cubes of the components of the microstates. This operator has no coupling between the microstate variables. As we said before, (C.5) is bijective to any non-degenerate linear operator. For Eq. (C.6), we take an operator that couples all the microstate components by multiplying them. If all the microstate components are positive - i.e. if $\boldsymbol{x}_i > 0$ for all i -, then we can take the logarithm of $\mathcal{H}_2$ and recover a non-degenerate linear operator. If any of the components is non-positive, then we cannot transform $\mathcal{H}_2$ into a linear operator, making the decoupling impossible. Finally, Eq. (C.7) is the sum of pairwise couplings between the microstate components. In this case, there is no smooth transformation between $\mathcal{H}_3$ and a linear operator, so we can make no further simplification of $\mathcal{H}_3$. We consider the cubic-, N_xth-, and square-root of $S_{\boldsymbol{x}}$, $P_{\boldsymbol{x}}$, and $C_{\boldsymbol{x}}$ respectively so that the physical units of the observations are the same as the units of the microstates.

In Fig. C.1, we show the median assimilation ($k < 0$) and prediction ($k \geq 0$) errors over 200 runs for the Lorenz and the Mackey-Glass systems for each of the observation operators described above. We took the same set of 200 initial conditions and noise seeds for each operator to make the results comparable.

In almost every case, we observe that the results for each system are almost identical regardless of the operator we use, both in the observation and the model spaces. This behavior suggests that our method is robust to the observation operator: if the observation convey enough information about the latent dynamics of the system, our method will initialize a microstate with good prediction power.



(a) Lorenz



(b) Mackey-Glass

Figure C.1: **Prediction error for different observation operators.** We show the median normalized squared error over 200 experiments for the observation space (solid lines) and the model space (dotted lines) for the case of noiseless (green) and noisy (blue) observations for *a)* the Lorenz system and *b)* the Mackey-Glass system. From left to right, we show the behaviour for operators representing the sum of cubes (see Eq. (C.5)), the product between microstate components (see Eq. (C.6)), and the sum of pairwise couplings (see Eq. (C.7)). We take all parameters as in Table 6.1.

However, we observe an anomaly on the model space error for the Lorenz system: the model space error corresponding to $\mathcal{H}_2$ –the product of components– is significantly higher than i) its observation space error, and ii) the model error corresponding to $\mathcal{H}_1$ and $\mathcal{H}_3$. Recall that the Lorenz system is symmetric around its x-axis, so the operator $\mathcal{H}_2$ cannot resolve if the product $\prod_i \boldsymbol{x}_i$ corresponds to (x, y, z) or to $(x, -y, -z)$, regardless of the number of measurements we have about the system. Thus, our method sometimes initializes the ground truth microstate and the other times it initializes its reflection about the x-axis, as we show in Fig. C.2.

Summarizing, the feasible set of solutions for the Lorenz system observed through $\mathcal{H}_2$ include both (x, y, z) and $(x, -y, -z)$ which, incidentally, result in the same observation-space dynamics for time windows of arbitrary size. In every other case, the only feasible solution is the ground-truth microstate, hence that all of the other results are almost identical in Fig. C.1. These results suggest that, for future work, we can take an extra step in the initialization procedure, where we consider all the known symmetries of a system when refining our estimate of the initial microstate.

Thus, we could refine the microstate for each symmetry and take the one that gives us either the lowest cost function value or the lowest prediction error.

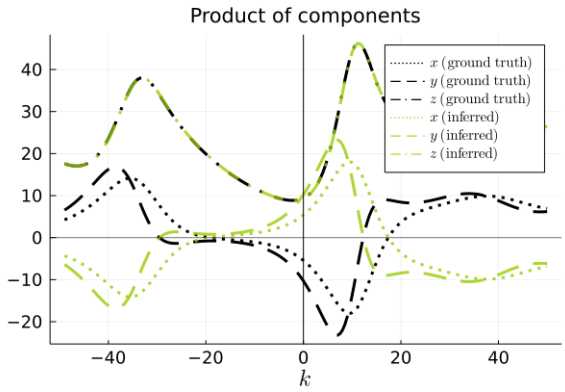


Figure C.2: **Dynamics of the individual microstate components of the Lorenz system** for an example run of the ground truth (black) and the initialized (green) microstate when the observation are taken using the operator of Eq. (C.6) –i.e., the operator that multiplies all the microstate components– in a noiseless scenario. The y and z components are symmetric with respect to the 0 line. Similar to Fig. C.1, $k < 0$ denotes assimilation times and $k \geq 0$ prediction times.

C.3 Initialization method performance & system features

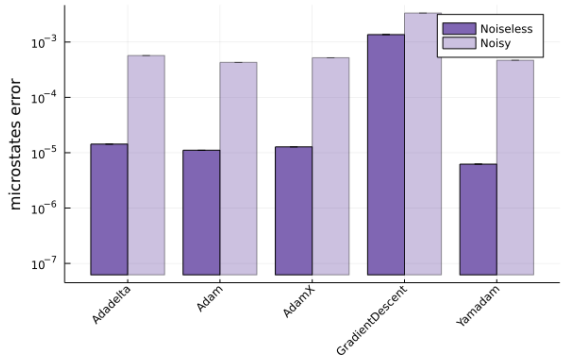


Figure C.3: **Gradient-based optimizers performance on the Lorenz system.** We measure the performance of the gradient-based optimizers (described in Section 6.2.3) with NSE_0^{model}, the average discrepancy between the present-time microstate $\boldsymbol{x}_0$ and the initialized microstate $\hat{\boldsymbol{x}}_0$, for noiseless (dark purple) and the noisy (light purple) time series. We show only the four best performing optimizers –namely Adadelta, Adam, AdamX, and YamAdam– as well as Stochastic Gradient Descent, which serves as our benchmark.

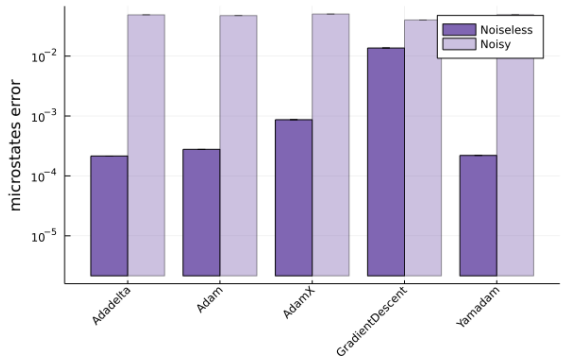
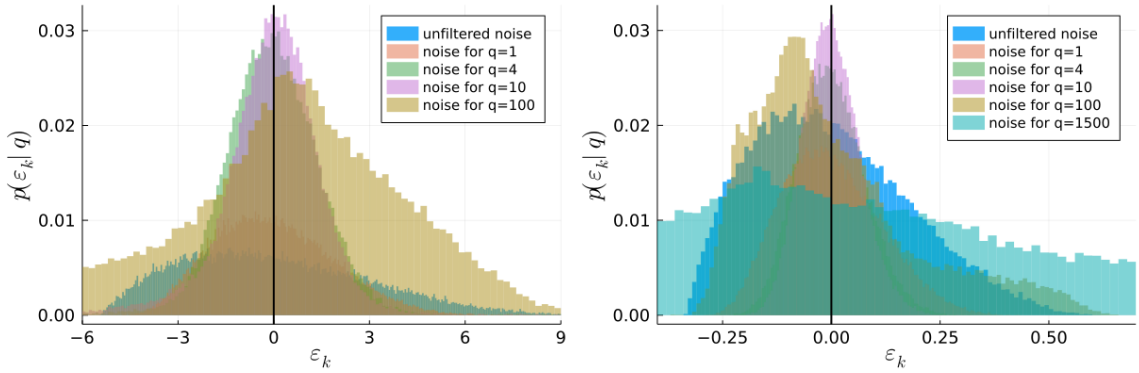
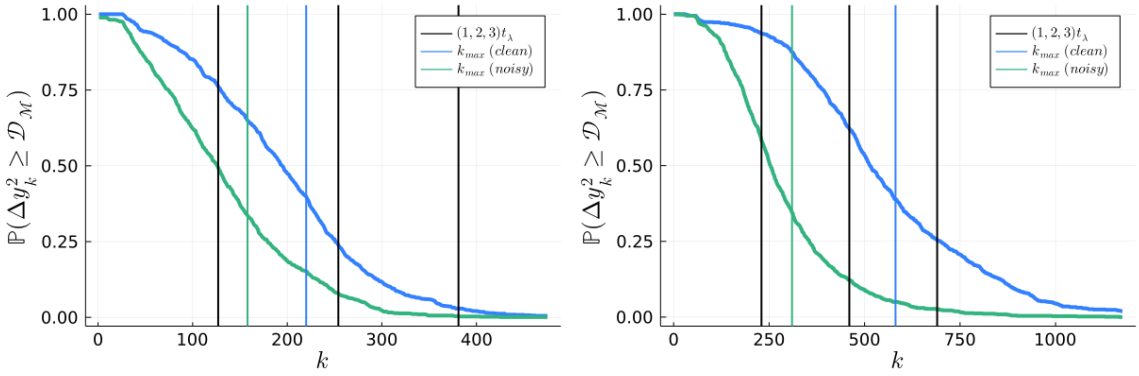


Figure C.4: **Gradient-based optimizers performance on the Mackey-Glass system.** See description in Fig. C.3 for details. The best performing optimizers were the same as with the Lorenz system.



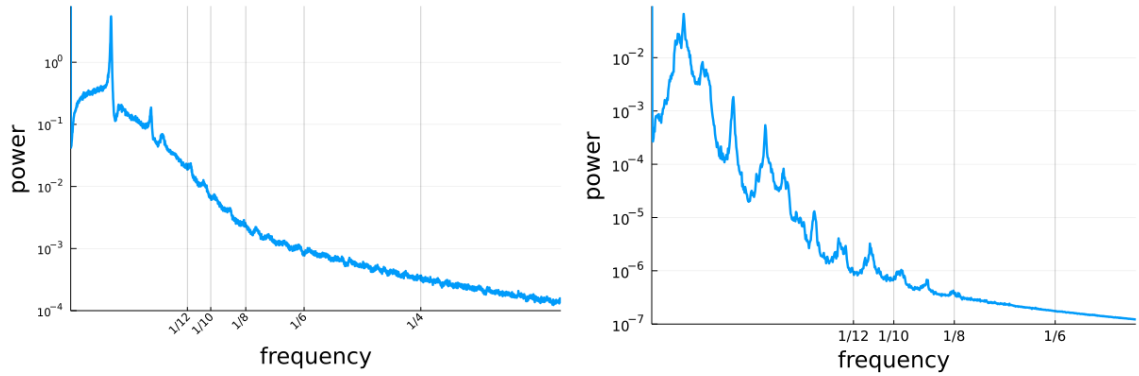
(a) Lorenz (b) Mackey-Glass

Figure C.5: **Filtered noise distributions** Starting from a simulated observational noise described by a left-skewed Beta distribution (see "unfiltered noise" in the legend) of very long time series ($T = 50000$ samples) of the *a)* Lorenz and *b)* Mackey-Glass systems, we plot the noise distribution of the unfiltered noise (solid blue) as well as the noise distribution we obtain after smoothing the corrupted signal with the LMPA filter for filters of increasing strength q. See Section 6.2.1 for details on the filter. For small q, the resulting distribution looks like a Gaussian distribution while for $q \gg 1$, the filter takes a significant of the signal, resulting in exhotic-shaped distributions.



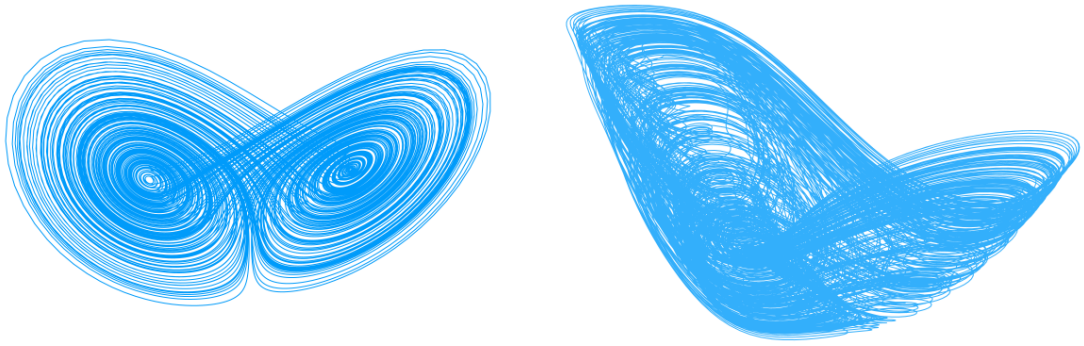
(a) Lorenz (b) Mackey-Glass

Figure C.6: **Cumulative probability of divergence** for *a)* the Lorenz system and *b)* the Mackey-Glass system. We show the fraction of trajectories –out of 1000– for which $NSE_k^{obs} \geq 2$ as a function of the prediction step k. This approach generalizes our definition of prediction horizon of Eq. (6.13) into a distribution-like quantity. In solid vertical lines, we show the Lyapunov 10-fold time of the system as well as the prediction horizons k_{max} for the noiseless and noisy cases.



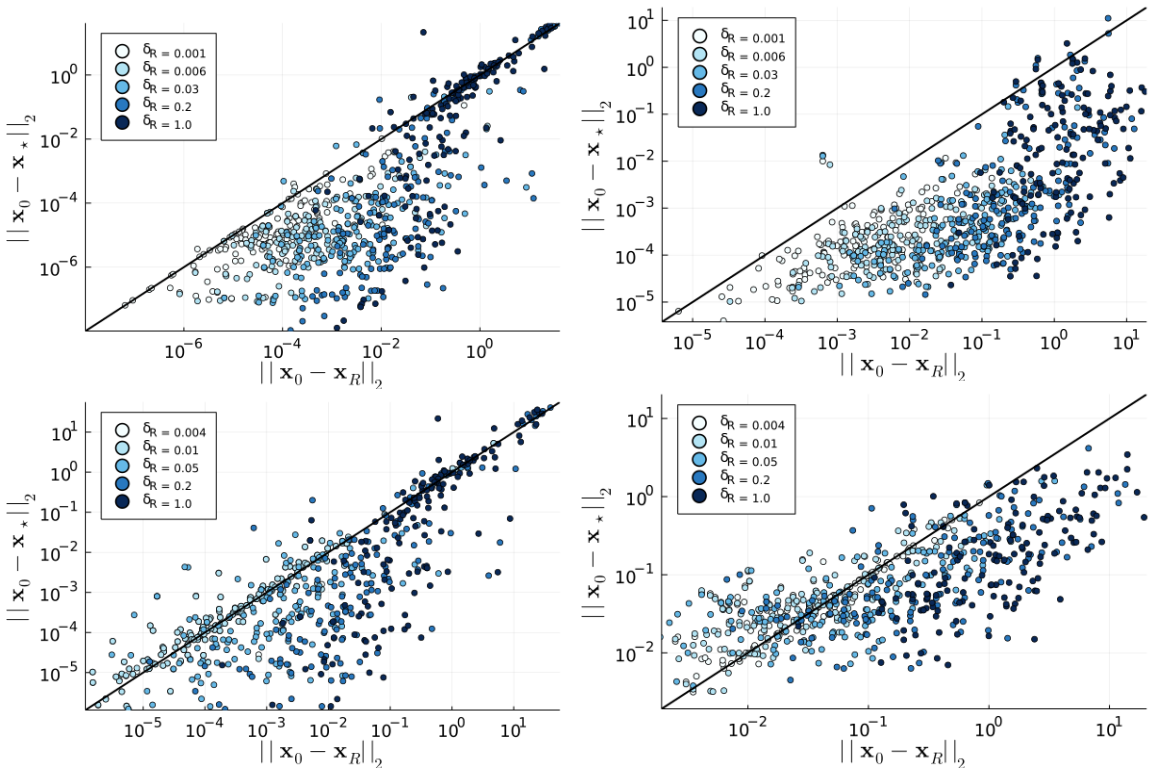
(a) Lorenz (b) Mackey-Glass

Figure C.7: **Normalized power spectra** for *a)* the Lorenz system and *b)* the Mackey-Glass system. The plot shows the average power spectra over 100 random in-attractor initial conditions with trajectories of 2^{12} points. We note that the Lorenz system has a clear power-law frequency decay with only a few low frequency peaks. While the Mackey-Glass system exhibits a non-vanishing spectrum characteristic of chaotic systems, it has more defined frequency peaks and decays much faster than the Lorenz system. Thus, we expect for the Mackey-Glass system to be easier to initialize in general.



(a) Lorenz (b) Mackey-Glass

Figure C.8: **Chaotic attractors** for *a)* the phase space portrait of the Lorenz system, where the axis show x, y and z, respectively, and *b)* the phase space portrait of the Mackey-Glass system, where we show $x(t)$ against $x(t-t_d)$. Each attractor consists of a *long* trajectory of 15000 points coming from a random in-attractor initial condition.



(a) Lorenz (b) Mackey-Glass

Figure C.9: **Initialization performance for different bounding-stage parameters** We present, on the x axis, the discrepancy of $\hat{\boldsymbol{x}}_0^R$ against, on the y axis, the discrepancy of the initialized microstates $\hat{\boldsymbol{x}}_0$ for increasing levels of δ_R (see legend) on *a)* the Lorenz system and *b)* the Mackey-Glass system. The microstate $\hat{\boldsymbol{x}}_0^R$ is the *rough* estimation of $\boldsymbol{x}_0$ after Eq. (6.9) is zatisfied. Values below the identity mean that $\hat{\boldsymbol{x}}_0^R$ improves refining it as described in Section 6.2.3. Values on the diagonal mean that the initialized microstates do not get any better by applying refinement methods.

Appendix D

Latent state estimation of models with network topologies

D.1 The updated ensemble states live near the span of the ensemble forecast

Guided by Vetra-Carvalho et al. [[225], Section V], we show that, under certain conditions, the ensemble analysis $\boldsymbol{X}_k^a$ is a linear transformation of the ensemble forecast $\boldsymbol{X}_k^f$, which implies that $\overline{\boldsymbol{x}_k^a}$ lives in the space spanned by the ensemble forecast members. Even when these conditions are not met exactly, the results we present in what follows are still a good approximation of the real behavior of the ensemble analysis.

For the following argument, we recall the general assumption that the observation noise ϵ_k is uncorrelated with the model dynamics, and thus $\mathbb{E}[\epsilon_k^{(i)}(\boldsymbol{x}_k^{f,(i)})^T] = 0$ for all $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, N_e\}$.

In what follows, we will drop the time index k to sparsify notation. Recall from Eq. (7.22) of the ensemble analysis update that $\boldsymbol{P}_a^e = \tilde{\boldsymbol{X}}_a(\tilde{\boldsymbol{X}}_a)^T$, with $\tilde{\boldsymbol{X}}_a = \boldsymbol{X}_a - \overline{\boldsymbol{x}_a}$. In general, any quantity on a tilde will indicate the deviation with respect to the mean of the original quantity. From Eq. (7.19), the average analysis update is

$$\overline{\boldsymbol{x}_a} = \overline{\boldsymbol{x}_f} + \boldsymbol{K}^e \left(y - \overline{\mathcal{H}(\boldsymbol{x}_f)} \right) \approx \overline{\boldsymbol{x}_f} + \boldsymbol{K}^e \left(y - \mathcal{H}(\overline{\boldsymbol{x}_f}) \right) \ ,$$

where we used that

$$\begin{aligned}\overline{\mathcal{H}(\boldsymbol{X}_f)} =& \frac{1}{N_e}\sum_i \mathcal{H}(\boldsymbol{x}_f^{(i)}) = \frac{1}{N_e}\sum_i \mathcal{H}(\tilde{\boldsymbol{x}}_f^{(i)} + \overline{\boldsymbol{x}_f}) \\ =& \mathcal{H}(\overline{\boldsymbol{x}_f}) + \frac{1}{N_e}\sum_i (\boldsymbol{H}\tilde{\boldsymbol{X}}_f)_i \;\to 0 \; + \mathcal{O}(\|\tilde{\boldsymbol{X}}_f\|^2) \\ \approx& \mathcal{H}(\overline{\boldsymbol{x}_f}).\end{aligned}$$

Here, we expanded the observation operator up to linear terms and used that $\mathbb{E}(\boldsymbol{H}\tilde{\boldsymbol{X}}_f) = \boldsymbol{H}\mathbb{E}(\tilde{\boldsymbol{X}}_f) = \boldsymbol{H}\boldsymbol{0} = \boldsymbol{0}$.

Moreover, we can express the *ensemble analysis* as

$$\begin{aligned}\boldsymbol{X}_a =& \boldsymbol{X}_f + \boldsymbol{K}^e\left(\boldsymbol{Y} - \mathcal{H}(\boldsymbol{X}_f)\right) && \text{(D.1)} \\ \approx& \boldsymbol{X}_f + \boldsymbol{K}^e\left(\boldsymbol{Y} - (\mathcal{H}(\overline{\boldsymbol{x}_f}) + \boldsymbol{H}\tilde{\boldsymbol{X}}_f)\right) , && \text{(D.2)}\end{aligned}$$

where we introduce the observation ensemble $\boldsymbol{Y} = \left[y + \epsilon^{(1)}, \ldots, y + \epsilon^{(N_e)}\right]$ and its corresponding observation ensemble anomaly $\tilde{\boldsymbol{Y}} = \left[\epsilon^{(1)}, \ldots, \epsilon^{(N_e)}\right]$ such that $\mathbb{E}(\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T) = \boldsymbol{R}$, the observation error covariance.

With all these quantities at hand, we may expand the analysis covariance as follows

$$\begin{aligned}\boldsymbol{P}_a^e =& \tilde{\boldsymbol{X}}_a\tilde{\boldsymbol{X}}_a{}^T = (\boldsymbol{X}_a - \overline{\boldsymbol{x}_a})(\boldsymbol{X}_a - \overline{\boldsymbol{x}_a})^T \\ =& \left(\left[\tilde{\boldsymbol{X}}_f + \boldsymbol{K}^e(\boldsymbol{Y} - \mathcal{H}(\boldsymbol{X}_f))\right] - \left[\overline{\boldsymbol{x}_f} + \boldsymbol{K}^e(y - \overline{\mathcal{H}(\boldsymbol{X}_f)}\right]\right)\left(\ldots\right)^T \\ \approx& \left(\tilde{\boldsymbol{X}}_f + \boldsymbol{K}^e\left[\tilde{\boldsymbol{Y}} - \boldsymbol{H}\tilde{\boldsymbol{X}}_f\right]\right)\left(\tilde{\boldsymbol{X}}_f + \boldsymbol{K}^e\left[\tilde{\boldsymbol{Y}} - \boldsymbol{H}\tilde{\boldsymbol{X}}_f\right]\right)^T \\ =& \tilde{\boldsymbol{X}}_f\tilde{\boldsymbol{X}}_f{}^T + \boldsymbol{K}^e\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^T(\boldsymbol{K}^e)^T + \boldsymbol{K}^e\boldsymbol{H}\tilde{\boldsymbol{X}}_f\tilde{\boldsymbol{X}}_f{}^T(\boldsymbol{K}^e\boldsymbol{H})^T \\ & - \tilde{\boldsymbol{X}}_f\tilde{\boldsymbol{X}}_f{}^T(\boldsymbol{K}^e\boldsymbol{H})^T - \boldsymbol{K}^e\boldsymbol{H}\tilde{\boldsymbol{X}}_f\tilde{\boldsymbol{X}}_f{}^T + \mathcal{O}(\underbrace{\mathbb{E}(\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{X}}_f^T)}_{\to 0}) \\ =& (\boldsymbol{I} - \boldsymbol{K}^e\boldsymbol{H})\boldsymbol{P}_f^e - \boldsymbol{P}_f^e(\boldsymbol{K}^e\boldsymbol{H})^T + \boldsymbol{K}^e(\boldsymbol{H}\boldsymbol{P}_f^e\boldsymbol{H}^T + \boldsymbol{R})(\boldsymbol{K}^e)^T .\end{aligned}$$

Here we substituted the approximation (D.2) in the third line of the last equation. Now, using the ensemble Kalman gain of the linearized observation operator (see Eq. (7.17)), the last two terms of the last equality cancel out, resulting in

$$\boldsymbol{P}_a^e \approx (\boldsymbol{I} - \boldsymbol{K}^e\boldsymbol{H})\boldsymbol{P}_f^e , \tag{D.3}$$

just as in the original Kalman filter analysis update.

Finally, if we expand this last result in terms of the anomaly matrices $\tilde{\boldsymbol{X}}_f$ and $\tilde{\boldsymbol{X}}_a$, we obtain the following

$$\begin{aligned}
\tilde{\boldsymbol{X}}_a \tilde{\boldsymbol{X}}_a{}^T =& (\boldsymbol{I} - \boldsymbol{K}^e \boldsymbol{H}) \tilde{\boldsymbol{X}}_f \tilde{\boldsymbol{X}}_f{}^T \\
=& \Big(\boldsymbol{I} - \tilde{\boldsymbol{X}}_f (\underbrace{\boldsymbol{H}\tilde{\boldsymbol{X}}_f}_{\boldsymbol{S}})^T \big[\boldsymbol{H}\tilde{\boldsymbol{X}}_f (\boldsymbol{H}\tilde{\boldsymbol{X}}_f)^T + \boldsymbol{R} \big]^{-1} \boldsymbol{H} \Big) \tilde{\boldsymbol{X}}_f \tilde{\boldsymbol{X}}_f{}^T \\
=& \tilde{\boldsymbol{X}}_f \Big(\underbrace{\boldsymbol{I} - \boldsymbol{S}^T \big[\boldsymbol{S}\boldsymbol{S}^T + \boldsymbol{R} \big]^{-1} \boldsymbol{S}}_{\boldsymbol{T}\boldsymbol{T}^T} \Big) \tilde{\boldsymbol{X}}_f{}^T \\
=& \tilde{\boldsymbol{X}}_f \boldsymbol{T} (\tilde{\boldsymbol{X}}_f \boldsymbol{T})^T \ . \\
\therefore \tilde{\boldsymbol{X}}_a =& \tilde{\boldsymbol{X}}_f \boldsymbol{T} \ , \qquad \text{(D.4)}
\end{aligned}$$

where we introduce the *ensemble observation matrix* $\boldsymbol{S} = \boldsymbol{H}\tilde{\boldsymbol{X}}_f$ and the *transform matrix* $\boldsymbol{T}$ such that $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{I} - \boldsymbol{S}^T \big[\boldsymbol{S}\boldsymbol{S}^T + \boldsymbol{R} \big]^{-1} \boldsymbol{S}$. In the DA literature, the previous approach to obtain the ensemble analysis is called the square-root method, where the main difficulty is to properly find the square root of $\boldsymbol{T}$. Note that we could rewrite $\boldsymbol{T}\boldsymbol{T}^T$ using the nonlinear observation operator in $\boldsymbol{S} \to \mathcal{H}(\tilde{\boldsymbol{X}}_f)$ and the relation (D.4) would still hold. This lets us use either of the adapted Kalman gains (see Eqs. (7.17) and (7.18)) for the current argument.

With Eq. (D.4) we conclude our argument. It states that the ensemble analysis $\tilde{\boldsymbol{X}}_a$ is a linear combination of the ensemble forecast $\tilde{\boldsymbol{X}}_f$ –whenever we use the linearized version of $\mathcal{H}$ at least–, and, therefore, the analysis estimate $\overline{\boldsymbol{x}_a}$ is in the space spanned by the individual members of the forecast ensemble. Mathematically, this says that $\overline{\boldsymbol{x}_a} \in \text{span} \left\{ \boldsymbol{x}_f^{(i)} : i \in \{1, \dots, N_e\} \right\} \subset \mathcal{X}$. Note that this has the desired property of looking for a potential solution in a much lower-dimensional space than the full state space $\mathcal{X}$. If the ensemble members are taken from a low-dimensional attractor of the state space (where we believe that the truth latent states live), then the analysis update will also live in such lower-dimensional attractor as well.

D.2 Square-root form of the localized state error covariance

In this section, following the notation in [144], we show that by taking the modulation product $\tilde{\boldsymbol{Z}}_k = \boldsymbol{W} \triangle \tilde{\boldsymbol{X}}_k^f$ as in (Eq. (7.26)), then $\tilde{\boldsymbol{Z}}_k$ is a square-root matrix of $\boldsymbol{P}_k^{e,loc}$.

We start by noting that the modulation of the n-th ensemble member can be rewritten as

$$\Big[(\tilde{\boldsymbol{X}}_k^f)_n \circ (\boldsymbol{W})_1, \dots, (\tilde{\boldsymbol{X}}_k^f)_n \circ (\boldsymbol{W})_J \Big] = \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\} \boldsymbol{W} \ ,$$

where, if $\boldsymbol{x}$ is an N_x-dimensional vector, then $\text{diag}\{\boldsymbol{x}\}$ is an $N_x \times N_x$ diagonal matrix with the components of $\boldsymbol{x}$ in the diagonal. We can thus rewrite the modulation product as follows

$$\tilde{\boldsymbol{Z}}_k = \boldsymbol{W} \triangle \tilde{\boldsymbol{X}}_k^f = \left[\text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_1\}\boldsymbol{W}, \ldots, \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_{N_e}\}\boldsymbol{W} \right] \ . \tag{D.5}$$

Finally, we substitute expression (D.5) on $\boldsymbol{P}_k^{e,loc}$ to obtain the following expression

$$\begin{aligned}
\left(\tilde{\boldsymbol{Z}}_k \tilde{\boldsymbol{Z}}_k^T\right)_{ij} &= \left(\left[\text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_1\}\boldsymbol{W}, \ldots, \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_{N_e}\}\boldsymbol{W} \right] \left[\cdots \right]^T \right)_{ij} \\
&= \left(\sum_{n=1}^{N_e} \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\}\boldsymbol{W}\boldsymbol{W}^T \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\} \right)_{ij} \\
&= \left(\sum_{n=1}^{N_e} \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\} \ \rho \ \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\} \right)_{ij} \\
&= \sum_{n=1}^{N_e} \left(\sum_{p,q=1}^{N_x} \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\}_{ip} \ \rho_{pq} \ \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\}_{qj} \right) \\
&= \sum_{n=1}^{N_e} \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\}_{ii} \ \rho_{ij} \ \text{diag}\{(\tilde{\boldsymbol{X}}_k^f)_n\}_{jj} \\
&= \rho_{ij} \left(\sum_{n=1}^{N_e} (\tilde{\boldsymbol{X}}_k^f)_{in} (\tilde{\boldsymbol{X}}_k^f)_{jn} \right) \\
&= \rho_{ij} (\boldsymbol{P}_k^{f,e})_{ij} \ ,
\end{aligned}$$

and, therefore

$$\tilde{\boldsymbol{Z}}_k \tilde{\boldsymbol{Z}}_k^T = \rho \circ \boldsymbol{P}_k^{f,e} \ , \tag{D.6}$$

which is what we wanted to show.

There are a lot of indexes going on in the last expression: index k stands for the time index of the variables at time t_k, index n runs through each of the N_e members of the ensemble, and indexes i j, p, and q are dummies that run through the N_x dimensions of the state space. For a given matrix $\boldsymbol{B}$, the expression $(\boldsymbol{B})_j$ denotes the j-th column of $\boldsymbol{B}$, while $(\boldsymbol{B})_{ij}$ denotes the element on the i-th row on the j-th column of $\boldsymbol{B}$.

D.3 Importance sampling bootstrap particle filter

This appendix presents the importance sampling bootstrap particle filter that we use to benchmark our Kalman-based methods. The following steps follow Doucet et al.'s description of the filter [74].

Importance sampling bootstrap particle filter

1. **Initialize the filter**.
   - For $i = 1, \ldots, N_e$, sample particle $\boldsymbol{x}_0^{(i)}$ from the prior $p(\boldsymbol{x}_0)$
   - Set $k \leftarrow 1$
2. **Importance sampling (forecast step)**. For $i = 1, \ldots, N_e$
   - Sample $\boldsymbol{x}_k^{(i)} \sim p(\boldsymbol{x}_k | \boldsymbol{x}_{k-1}^{(i)})$ by driving the system forward.
   - Save the particle trajectory $\boldsymbol{x}_{0:k}^{(i)} = (\boldsymbol{x}_{0:k-1}^{(i)}, x_k^{(i)})$
   - Assign weights by evaluating the particle likelihood $w_k^i = p(y_k | \boldsymbol{x}_k^{(i)})$
   - Normalize weights by setting $w_k^i \leftarrow \frac{w_k^i}{\sum_j w_k^j}$
3. **Particle selection (analysis step)**. For $i = 1, \ldots, N_e$
   - Resample N_e particle trajectories from $\left(\boldsymbol{x}_{0:k}^{(i)}\right)_i$ with replacement according to the weights distribution.
   - Set $k \leftarrow k + 1$ and go back to step 2.

Bibliography

[1] Africa's Top Virus Deaths. https://www.africanews.com/2020/06/25/africa-s-prominent-coronavirus-deaths/. Accessed: 2020-07-07. 135

[2] Argentina entra en cuarentena obligatoria hasta el 31 de marzo. *El País*, Mar. 2020. https://elpais.com/sociedad/2020-03-20/argentina-entra-en-cuarentena-obligatoria-hasta-el-31-de-marzo.html. 135

[3] Coronavirus: 'sombre day' as uk deaths hit 10,000. *BBC*, Apr. 2020. https://www.bbc.com/news/uk-52264145. 135

[4] España registra su primer día sin muertes por covid-19. *El Financiero*, Jun. 2020. https://www.elfinanciero.com.mx/mundo/espana-registra-su-primer-dia-sin-muertes-por-covid-19. 135

[5] 'i can't breathe': Man dies after pleading with officer attempting to detain him in minneapolis. *NBC News*, May 2020. https://www.nbcnews.com/news/us-news/man-dies-after-pleading-i-can-t-breathe-during-arrest-n1214586. 135

[6] Matanza de 23 presos y 83 heridos sacude a una colombia temerosa del covid-19. *Agencia EFE*, Mar. 2020. https://www.efe.com/efe/america/sociedad/matanza-de-23-presos-y-83-heridos-sacude-a-una-colombia-temerosa-del-covid-20000013-4201992. 135

[7] Ministerio de salud decreta cuarentena total para la ciudad de santiago y seis comunas aledañas. *Ministerio de Salud*, May 2020. https://www.minsal.cl/ministerio-de-salud-decreta-cuarentena-total-para-la-ciudad-de-santiago-y-se 135

[8] News consumption in the uk: 2020 report. *Ofcom*, Aug. 2020. https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption. 42

[9] Protests across the globe after george floyd's death. *CNN World*, Jun. 2020. https://edition.cnn.com/2020/06/06/world/gallery/intl-george-floyd-protests/index.html. 135

[10] Sanidad elimina casi 2.000 fallecidos del balance total tras una revisión de los casos. *El Diario*, May 2020. https://www.eldiario.es/sociedad/espana-registra-fallecidos-ultimas-diagnostica_1_5972351.html. 135

[11] IPCC 2022. Summary for policymakers. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, In Press. 8, 45

[12] Sepideh Bazzaz Abkenar, Mostafa Haghi Kashani, Ebrahim Mahdipour, and Seyed Mahdi Jameii. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telematics and Informatics*, 57:101517, 2021. 1

[13] Luca Maria Aiello, Daniele Quercia, Ke Zhou, Marios Constantinides, Sanja Šćepanović, and Sagar Joglekar. How epidemic psychology works on social media: Evolution of responses to the covid-19 pandemic. *arXiv preprint arXiv:2007.13169*, 2020. 16, 19, 20

[14] Victor Amelkin, Francesco Bullo, and Ambuj K. Singh. Polar Opinion Dynamics in Social Networks. *IEEE Transactions on Automatic Control*, 62(11):5650–5665, nov 2017. 117, 123

[15] Hana Anber, Akram Salah, and AA Abd El-Aziz. A literature review on twitter data analysis. *International Journal of Computer and Electrical Engineering*, 8(3):241, 2016. 7

[16] Vadim S Anishchenko, Tatjana E Vadivasova, Andrey S Kopeikin, Jürgen Kurths, and Galina I Strelkova. Peculiarities of the relaxation to an invariant probability measure of nonhyperbolic chaotic attractors in the presence of noise. *Physical Review E*, 65(3):036206, 2002. 80

[17] Salman Aslam. Number of monetizable daily active twitter users (mdau) worldwide from 1st quarter 2017 to 2nd quarter 2022. *Omnicore*, 2022. 7

[18] Salman Aslam. Twitter by the numbers: Stats, demographics & fun facts. *Omnicore*, 2022. 7

[19] Erik Aurell, Guido Boffetta, Andrea Crisanti, Giovanni Paladin, and Angelo Vulpiani. Growth of noninfinitesimal perturbations in turbulence. *Physical review letters*, 77(7):1262, 1996. 84

[20] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015. 147

[21] Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324, 2021. 8, 17, 18

[22] Susan A Banducci and Jeffrey A Karp. How elections change the way citizens view the political system: campaigns, media effects and electoral outcomes in comparative perspective. *British Journal of Political Science*, 33(3):443–467, 2003. 118

[23] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91, 2015. 7, 46, 47, 48, 64, 141

[24] Sylvain Barde. Back to the Future: Economic Self-Organisation and Maximum Entropy Prediction. *Computational Economics*, 45(2):337–358, 2014. 97

[25] John M. Barrios and Yael V. Hochberg. Risk Perception Through the Lens of Politics in the Time of the COVID-19 Pandemic. *NBER Working Paper No. 27008*, 2020. 16

[26] Carolina Becatti, Guido Caldarelli, Renaud Lambiotte, and Fabio Saracco. Extracting significant signal of news consumption from social networks: the case of twitter in italian political elections. *Palgrave Communications*, 5(1):1–16, 2019. 52, 146

[27] Mariano Beguerisse-Díaz, Amy K McLennan, Guillermo Garduno-Hernández, Mauricio Barahona, and Stanley J Ulijaszek. The 'who'and 'what'of# diabetes on twitter. *Digital health*, 3:2055207616688841, 2017. 8, 48, 52

[28] Giancarlo Benettin, Luigi Galgani, and Jean-Marie Strelcyn. Kolmogorov entropy and numerical experiments. *Physical Review A*, 14(6):2338, 1976. 84

[29] Yochai Benkler. The wealth of networks. In *The Wealth of Networks*. Yale university press, 2008. 44

[30] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 u.s. presidential election online discussion. *First Monday*, 21(11), Nov. 2016. https://firstmonday.org/ojs/index.php/fm/article/view/7090. 42

[31] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7), 2016. 66

[32] Gnana K Bharathy and Barry Silverman. Validating agent based social systems models. In *Proceedings of the 2010 Winter Simulation Conference*, pages 441–453. IEEE, 2010. 1

[33] Sudeep Bhatia. Vector Space Semantic Models Predict Subjective Probability Judgments for Real-World Events. In *CogSci*, 2016. 24, 38

[34] Sudeep Bhatia. Predicting risk perception: New insights from data science. *Management Science*, 65(8):3800–3823, 2019. 16, 19

[35] Sudeep Bhatia, Barbara Mellers, and Lukasz Walasek. Affective responses to uncertain real-world outcomes: Sentiment change on Twitter. *PloS one*, 14(2):e0212489, 2019. 7, 16, 20

[36] Craig Bishop and Daniel Hodyss. Ensemble covariances adaptively localized with eco-rap. part 2: A strategy for the atmosphere. *Tellus A: Dynamic Meteorology and Oceanography*, 61(1):97–111, 2009. 104, 107

[37] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 10, 55, 66

[38] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. 28, 137

[39] Alexandre Bovet, Flaviano Morone, and Hernán A Makse. Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump. *Scientific reports*, 8(1):1–16, 2018. 44, 45, 66

[40] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009. 107, 109

[41] Tom Britton, Maria Deijfen, and Fredrik Liljeros. A Weighted Configuration Model and Inhomogeneous Epidemics. *Journal of Statistical Physics*, 145(5):1368–1384, 2011. 30

[42] Matteo Bruno, Renaud Lambiotte, and Fabio Saracco. Brexit and bots: characterizing the behaviour of automated accounts on twitter during the uk election. *EPJ Data Science*, 11(1):17, 2022. 66

[43] Axel Bruns. Echo chamber? what echo chamber? reviewing the evidence. In *6th Biennial Future of Journalism Conference (FOJ17)*, 2017. 7, 45, 56, 58, 65, 147

[44] Jennings Bryant and Dorina Miron. Theory and Research in Mass Communication. *Journal of Communication*, 54(4):662–704, 01 2006. 42

[45] Fred Buckley and Frank Harary. *Distance in graphs*, volume 2. Addison-Wesley Redwood City, CA, 1990. 110

[46] E Bullmore and O Sporns. Complex brain networks: graph theoretical analysis of structural and functional system. *Nature*, 10:186–198, 2009. 74

[47] Gerrit Burgers, Peter Jan Van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126(6):1719–1724, 1998. 103, 105

[48] William J. Burns and Paul Slovic. Risk Perception and Behaviors: Anticipating and Responding to Crises. *Risk Analysis*, 32(4):579–582, 2012. 13

[49] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent systems*, 28(2):15–21, 2013. 11

[50] C. Daryl Cameron and B. Keith Payne. Escaping Affect: How Motivated Emotion Regulation Creates Insensitivity to Mass Suffering. *Journal of Personality and Social Psychology*, 100(1):1–15, 2011. 29

[51] Marine Carrasco and Guy Tchuente. Regularized liml for many instruments. *Journal of Econometrics*, 186(2):427–442, 2015. 107

[52] Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):1–50, 2018. 69, 72, 75, 77, 95

[53] Alberto Carrassi, Marc Bocquet, Jonathan Demaeyer, Colin Grudzien, Patrick Raanes, and Stephane Vannitsem. Data assimilation for chaotic dynamics. *arXiv*, (October), 2020. 113, 123

[54] Martin Casdagli, Stephen Eubank, J. Doyne Farmer, and John Gibson. State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, 51(1-3):52–98, 1991. 79, 84

[55] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009. 105

[56] Roger W Caves. *Encyclopedia of the City*. Routledge, 2004. 45

[57] Andrea Ceron, Luigi Curini, and Stefano M Iacus. First-and second-level agenda setting in the twittersphere: An application to the italian political debate. *Journal of Information Technology & Politics*, 13(2):159–174, 2016. 44

[58] Ted Hsuan Yun Chen, Ali Salloum, Antti Gronow, Tuomas Ylanttila, and Mikko Kivela. Polarization of climate politics results from partisan sorting: Evidence from finnish twittersphere. *Global Environmental Change*, 71:102348, 2021. 7, 45, 46, 47, 48, 56, 64, 65, 141, 146

[59] Matteo Cinelli, Stefano Cresci, Walter Quattrociocchi, Maurizio Tesconi, and Paola Zola. Coordinated inauthentic behavior and information spreading on twitter. *Decision Support Systems*, page 113819, 2022. 62, 66

[60] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), 2021. 7, 44, 47, 58, 65, 66, 141, 147

[61] Robert Clay, Le Minh Kieu, Jonathan A. Ward, Alison Heppenstall, and Nick Malleson. Towards Real-Time Crowd Simulation Under Uncertainty Using an Agent-Based Model and an Unscented Kalman Filter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12092 LNAI(757455):68–79, 2020. 70, 98

[62] Tadeo Javier Cocucci, Manuel Pulido, Juan Pablo Aparicio, Juan Ruíz, Mario Ignacio Simoy, and Santiago Rosa. Inference in epidemiological agent-based models using ensemble-based data assimilation. *Plos one*, 17(3):e0264892, 2022. 70, 98

[63] Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. Climate change sentiment on twitter: An unsolicited public opinion poll. *PloS one*, 10(8):e0136092, 2015. 8, 45, 53

[64] Stepehen Collinson. Trump responds to protests with a strongman act. *CNN Politics*, Jun. 2020. https://edition.cnn.com/2020/06/02/politics/donald-trump-george-floyd-protest-military/index.html. 135

[65] Carmela Comito, Agostino Forestiero, and Clara Pizzuti. Bursty event detection in twitter streams. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(4):1–28, 2019. 63

[66] Wesley Cota and Silvio C Ferreira. Optimized gillespie algorithms for the simulation of markovian epidemic processes on large and heterogeneous networks. *Computer Physics Communications*, 219:303–312, 2017. 98

[67] P Courtier, JN Thépaut, and A Hollingsworth. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly journal of the royal meteorological society*, 120(519):1367–1387, 1994. 100

[68] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9(1):1–20, 2019. 7, 63

[69] Abir De, Sourangshu Bhattacharya, and Niloy Ganguly. Demarcating endogenous and exogenous opinion diffusion process on social networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 549–558, 2018. 12, 118

[70] Wändi Bruine de Bruin and Daniel Bennett. Relationships Between Initial COVID-19 Risk Perceptions and Protective Health Behaviors: A National Survey. *American Journal of Preventive Medicine*, 2020. 16

[71] Donald L DeAngelis and Stephanie G Diaz. Decision-making in agent-based modeling: A current review and future prospectus. *Frontiers in Ecology and Evolution*, 6:237, 2019. 127

[72] R. Maria del Rio-Chanona, Penny Mealy, Mariano Beguerisse-Díaz, François Lafond, and J. Doyne Farmer. Occupational mobility and automation: a data-driven network model. *Journal of The Royal Society Interface*, 18(174):20200898, jan 2021. 74

[73] Michela Del Vicario, Fabiana Zollo, Guido Caldarelli, Antonio Scala, and Walter Quattrociocchi. Mapping social dynamics on facebook: The brexit debate. *Social Networks*, 50:6–16, 2017. 59

[74] Arnaud Doucet, Adam M Johansen, et al. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009. 112, 122, 161

[75] Laura Dovera and Ernesto della Rossa. Multimodal ensemble Kalman filtering using Gaussian mixture models. *Computational Geosciences*, 15(2):307–323, 2011. 122

[76] S. Dryhurst, Claudia R. Schneider, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, David Spiegelhalter, and Sander van der Linden. Risk perceptions of COVID-19 around the world. *Journal of Risk Research*, 2020. 16

[77] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 82

[78] Joel Dyer and Blas Kolic. Public risk perception and emotion on twitter during the covid-19 pandemic. *Applied Network Science*, 5(1):1–32, 2020. 4, 13, 45

[79] Candice LaShara Edrington and Nicole Lee. Tweeting a social movement: Black lives matter and its use of twitter to share information, build community, and promote action. *The Journal of Public Interest Communications*, 2(2):289–289, 2018. 7

[80] Magdalini Eirinaki, Jerry Gao, Iraklis Varlamis, and Konstantinos Tserpes. Recommender systems for large-scale social networks: A review of challenges and solutions, 2018. 45

[81] Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994. 101, 103, 128

[82] Geir Evensen. Advanced data assimilation for strongly nonlinear dynamics. *Monthly weather review*, 125(6):1342–1354, 1997. 3, 75

[83] J Doyne Farmer. Chaotic attractors of an infinite-dimensional dynamical system. *Physica D: Nonlinear Phenomena*, 4(3):366–393, 1982. 89, 113, 114, 122, 128

[84] J Doyne Farmer and John J Sidorowich. Predicting chaotic time series. *Physical review letters*, 59(8):845, 1987. 74, 79, 83

[85] J Doyne Farmer and John J Sidorowich. Optimal shadowing and noise reduction. *Physica D: Nonlinear Phenomena*, 47(3):373–392, 1991. 75, 79

[86] JD Farmer, M Gallegati, C Hommes, A Kirman, P Ormerod, S Cincotti, A Sanchez, and D Helbing. A complex systems approach to constructing better models for managing financial markets and the economy. *European Physical Journal*, 214:295–324, 2012. 74

[87] NT Feather. The effect of differential failure on expectation of success, reported anxiety, and response uncertainty. *Journal of Personality*, 1963. 16

[88] Gustav T Fechner, Davis H Howes, and Edwin G Boring. *Elements of psychophysics*, volume 1. Holt, Rinehart and Winston New York, 1966. 14, 34

[89] Emilio Ferrara. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, 25(6), May 2020. 8, 42

[90] D. Fetherstonhaugh, P. Slovic, S.M. Johnson, and J. Friedrich. Insensitivity to the value of human life: A study of psychophysical numbing. *J Risk Uncertain*, 14:283–300, 1997. 14, 21, 126

[91] David Foster, Jacob Foster, Maya Paczuski, and Peter Grassberger. Communities, clustering phase transitions, and hysteresis: Pitfalls in constructing network ensembles. *Phys. Rev. E*, 81:046115, Apr 2010. 139

[92] J. Friedrich, P. Barnes, K. Chapin, I. Dawson, V. Garst, and D. Kerr. Psychophysical numbing: When lives are valued less as the lives at risk increase. *J Consum Psychol*, 8:277–299, 1999. 14

[93] James Friedrich, Paul Barnes, Kathryn Chapin, Ian Dawson, Valerie Garst, and David Kerr. Psychophysical numbing: When lives are valued less as the lives at risk increase. *Journal of Consumer Psychology*, 8(3):277–299, 1999. 2

[94] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293, 2020. 8

[95] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018. 44, 45, 58, 65, 146, 147

[96] Gregory Gaspari and Stephen E Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757, 1999. 107

[97] Noé Gaumont, Maziyar Panahi, and David Chavalarias. Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 french presidential election. *PloS one*, 13(9):e0201879, 2018. 7, 44, 45, 46, 48, 64, 141

[98] George A Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013. 2, 14

[99] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. Me, my echo chamber, and i: introspection on social media polarization. In *Proceedings of the 2018 World Wide Web Conference*, pages 823–831, 2018. 45, 58, 65

[100] Sarah M Glaser, Michael J Fogarty, Hui Liu, Irit Altman, Chih-Hao Hsieh, Les Kaufman, Alec D MacCall, Andrew A Rosenberg, Hao Ye, and George Sugihara. Complex dynamics may limit prediction in marine fisheries. *Fish and Fisheries*, 15(4):616–633, 2014. 74

[101] Maria Glenski, Tim Weninger, and Svitlana Volkova. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? *IEEE Transactions on Computational Social Systems*, 5(4):1071–1082, 2018. 46, 49, 64, 67, 127

[102] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009. 17

[103] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. The structural virality of online diffusion. *Management Science*, 62(1):180–196, 2016. 47, 51

[104] Peter Grassberger, Rainer Hegger, Holger Kantz, Carsten Schaffrath, and Thomas Schreiber. On noise reduction methods for chaotic data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 3(2):127–141, 1993. 80

[105] Jakob Grazzini, Matteo G Richiardi, and Mike Tsionas. Bayesian estimation of agent-based models. *Journal of Economic Dynamics and Control*, 77:26–47, 2017. 70, 75, 123

[106] Shannon Greenwood, Andrew Perrin, and Maeve Duggan. Social Media Update 2016. *Pew Research Center*, (November), 2016. 42

[107] José Luis Guiñón, Emma Ortega, José García-Antón, and Valentín Pérez-Herranz. Moving average and savitzki-golay smoothing filters using mathcad. *Papers ICEE*, 2007, 2007. 79

[108] Conrad Hackett, Stephanie Kramer, and Anna Schiller. The Age Gap in Religion Around the World. *Pew Research Center*, (June), 2018. 43

[109] Thomas Hale, Sam Webster, Anna Petherick, Toby Phillips, and Beatriz Kira. COVID-19 Government Response Tracker, Blavatnik School of Government. 2020. Data use policy: Creative Commons Attribution CC BY standard. 13, 20

[110] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. 16

[111] Samer Hassan, Juan Pavon, and Nigel Gilbert. Injecting data into simulation: Can agent-based modelling learn from microsimulation. In *World Congress of Social Simulation*, 2008. 70, 97

[112] Samer Hassan Collado. *Towards a Data-driven Approach for Agent-Based Modelling : Simulating Spanish Postmodernisation*. PhD thesis, Universidad Complutense de Madrid, 2009. 97

[113] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002. 3, 113, 116, 122, 128

[114] Jake M Hofman, Amit Sharma, and Duncan J Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017. 1

[115] Remco van der Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016. 139

[116] Matthew J. Hornsey, Emily A. Harris, and Kelly S. Fielding. Relationships among conspiratorial beliefs, conservatism and climate scepticism across nations. *Nature Climate Change*, 8(7):614–620, 2018. 45

[117] P. L. Houtekamer and Fuqing Zhang. Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12):4489–4532, 2016. 69, 107, 121, 123, 128

[118] Peter L Houtekamer and Herschel L Mitchell. A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001. 107

[119] Peter L Houtekamer and Fuqing Zhang. Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12):4489–4532, 2016. 104

[120] Xiaolin Hu and Peisheng Wu. A data assimilation framework for discrete event simulations. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 29(3):1–26, 2019. 70

[121] Mark D Humphries and Kevin Gurney. Network 'Small-World-Ness': A Quantitative Method for Determining Canonical Network Equivalence. *PLOS ONE*, 3(4):1–10, 2008. 24

[122] Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014. 12

[123] Kokil Jaidka, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *PNAS*, 117(19):10165–10171, 2020. 16

[124] S Mo Jang and P Sol Hart. Polarized frames on "climate change" and "global warming" across countries and states: Evidence from twitter big data. *Global Environmental Change*, 32:11–17, 2015. 46, 47, 48, 56

[125] T. Jay and K. Janschowitz. The pragmatics of swearing. *Journel of Politeness Research*, 4:267–288, 2008. 20

[126] Julie Jiang, Xiang Ren, Emilio Ferrara, et al. Social media polarization and echo chambers in the context of covid-19: Case study. *JMIRx med*, 2(3):e29570, 2021. 7, 45, 141

[127] Jeffrey Johnson. The future of the social sciences and humanities in the science of complex systems. *Innovation–The European Journal of Social Science Research*, 23(2):115–134, 2010. 1

[128] Michael N Jones, Jon Willits, Simon Dennis, Jerome R Busemeyer, Zheng Wang, James T Townsend, and Ami Eidels. *Models of Semantic Memory of a single chapter of a title in Oxford Handbooks Online for personal use (for details see Privacy Policy and Legal Notice). Models of Semantic Memory The Oxford Handbook of Computational and Mathematical Psychology.* Number September. 2020. 24

[129] Kevin Judd. Failure of maximum likelihood methods for chaotic dynamical systems. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 75(3):1–7, 2007. 78, 84

[130] Kevin Judd. Forecasting with imperfect models, dynamically constrained inverse problems, and gradient descent algorithms. *Physica D: Nonlinear Phenomena*, 237(2):216–232, 2008. 75

[131] Andreas Jungherr. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics*, 13(1):72–91, 2016. 7

[132] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. 69, 101

[133] Yoed N Kenett, David Anaki, and Miriam Faust. Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, 8:407, 2014. 24

[134] Marc Keuschnigg, Niclas Lovsjö, and Peter Hedström. Analytical sociology and computational social science. *Journal of Computational Social Science*, 1(1):3–14, 2018. 1

[135] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014. 82, 86, 95

[136] Blas Kolic, Fabián Aguirre-López, Sergio Hernández-Williams, and Guillermo Garduño-Hernández. Quantifying the structure of controversial discussions with unsupervised methods: a look into the twitter climate change conversation. *arXiv preprint arXiv:2206.14501*, 2022. 4, 44

[137] Blas Kolic, Juan Sabuco, and J Doyne Farmer. Estimating initial conditions for dynamical systems with incomplete information. *Nonlinear Dynamics*, 108(4):3783–3805, 2022. 4

[138] Joakim Kulin, Ingemar Johansson Sevä, and Riley E Dunlap. Nationalist ideology, rightwing populism, and public views about climate change in europe. *Environmental politics*, 30(7):1111–1134, 2021. 45

[139] Renaud Lambiotte. *Lecture notes in Networks*. Mathematical Institute, Oxford University, 2021. 11

[140] Renaud Lambiotte and Michael T Schaub. *Modularity and Dynamics on Complex Networks*. Cambridge University Press, 2021. 11

[141] Andrew Lamperski and Antonis Papachristodoulou. Stability and consensus for multi-agent systems with Poisson clock noise. *Proceedings of the IEEE Conference on Decision and Control*, 2015-Febru(February):3023–3028, 2014. 118

[142] Francesco Lamperti, Andrea Roventini, and Amir Sani. Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, 90:366–389, 2018. 123

[143] Sandra Laville and Jonathan Watts. Across the globe, millions join biggest climate protest ever. *The Guardian*, 2019. 46, 58, 62

[144] Lili Lei, Jeffrey S. Whitaker, and Craig Bishop. Improving Assimilation of Radiance Observations by Implementing Model Space Localization in an Ensemble Kalman Filter. *Journal of Advances in Modeling Earth Systems*, 10(12):3221–3232, 2018. 108, 112, 160

[145] Hai Liang, Isaac Chun-Hai Fung, Zion Tsz Ho Tse, Jingjing Yin, Chung-Hong Chan, Laura E Pechta, Belinda J Smith, Rossmary D Marquez-Lameda, Martin I Meltzer, Keri M Lubell, et al. How did ebola information spread on twitter: broadcasting or viral spreading? *BMC public health*, 19(1):1–11, 2019. 47, 49, 50, 64

[146] Robert. J. Lifton. Beyond psychic numbing: a call to awareness. *American Journal of Orthopsychiatry*, 52(4), 1982. 21

[147] Simon Lindgren. *Data theory: Interpretive sociology and computational methods*. John Wiley & Sons, 2020. 2, 127

[148] G.F. Loewenstein, E.U. Weber, C.K. Hsee, and E. Welch. Risk as feelings. *Psychological Bulletin*, 127:267–286, 2001. 23

[149] Fabián Aguirre López and Anthony C C Coolen. Transitions in random graphs of fixed degrees with many short cycles. *Journal of Physics: Complexity*, 2(3):035010, may 2021. 139

[150] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963. 77, 85, 86

[151] Jan Lorenz. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838, 2007. 113, 116, 122

[152] Michael C Mackey and Leon Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977. 85, 89, 122, 128

[153] Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. Did you offend me? Classification of Offensive Tweets in Hinglish Language. pages 138–148, 2019. 32, 43

[154] Esteban Ortiz-Ospina Max Roser, Hannah Ritchie and Joe Hasell. Coronavirus Pandemic (COVID-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus. 18

[155] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2):176–187, 1972. 118

[156] Maxwell E. McCombs and Donald L. Shaw. The agenda-setting function of mass media*. *Public Opinion Quarterly*, 36(2):176–187, 01 1972. 42

[157] Aaron M McCright, Riley E Dunlap, and Sandra T Marquart-Pyatt. Political ideology and views about climate change in the european union. *Environmental Politics*, 25(2):338–358, 2016. 8

[158] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. 2, 45

[159] Zhiyong Meng and Fuqing Zhang. Tests of an ensemble kalman filter for mesoscale and regional-scale data assimilation. part ii: Imperfect model experiments. *Monthly Weather Review*, 135(4):1403–1423, 2007. 123

[160] S. M. Mohammad and P. D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, 2010. 15

[161] Bojan Mohar. Some applications of laplace eigenvalues of graphs. In *Graph symmetry*, pages 225–275. Springer, 1997. 10, 55, 65, 140

[162] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165, 2019. 11

[163] Mary Natrella et al. e-handbook of statistical methods. *NIST/SEMATECH*, 49, 2010. 80

[164] Ionel M Navon. Data assimilation for numerical weather prediction: a review. In *Data assimilation for atmospheric, oceanic and hydrologic applications*, pages 21–65. Springer, 2009. 75

[165] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983. 82

[166] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70:056131, Nov 2004. 27

[167] Mark Newman. *Networks*. Oxford university press, 2018. 11, 52

[168] C. Olivola. The Cognitive Psychology of Sensitivity to Human Fatalities: Implications for Life-Saving Policies. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):141–146, 2015. 38

[169] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017. 17

[170] NH Packard, JP Crutchfield, JD Farmer, and RS Shaw. Geometry from a time series. *Physical Review Letters*, 45(9):712–716, 1980. 74

[171] Sergey E. Parsegov, Anton V. Proskurnikov, Roberto Tempo, and Noah E. Friedkin. Novel Multidimensional Models of Opinion Dynamics in Social Networks. *IEEE Transactions on Automatic Control*, 62(5):2270–2285, 2017. 118

[172] Tiago P Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, 2014. 10, 112, 118

[173] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332, 2019. 55, 66

[174] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC2015. 2015. 2, 12, 19, 20

[175] Carlos Pires, Robert Vautard, and Olivier Talagrand. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A*, 48(1):96–121, 1996. 3, 75, 78, 83, 95

[176] Donovan Platt. A comparison of economic agent-based model calibration methods. *Journal of Economic Dynamics and Control*, 113:103859, 2020. 70, 75

[177] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964. 82

[178] Wouter Poortinga, Lorraine Whitmarsh, Linda Steg, Gisela Böhm, and Stephen Fisher. Climate change perceptions and their individual-level determinants: A cross-european analysis. *Global Environmental Change*, 55:25–35, 2019. 45

[179] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect:An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734, 2005. 15

[180] M. J. Prerau, A. C. Smith, U. T. Eden, M. Yanike, W. A. Suzuki, and E. N. Brown. A mixed filter algorithm for cognitive state estimation from simultaneously recorded continuous and binary measures of performance. *Biological Cybernetics*, 99(1):1–14, 2008. 123

[181] Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. Opinion dynamics on interacting networks: media competition and social influence. *Scientific reports*, 4(1):1–7, 2014. 74

[182] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2019. 82, 86, 95

[183] Alpa Reshamwala, Dhirendra Mishra, and Prajakta Pawar. Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ)*, 3(1):113–116, 2013. 11

[184] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663, 2009. 122

[185] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 82

[186] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009. 10

[187] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123, 2008. 55, 66

[188] Camille Roth, Jonathan St-Onge, and Katrin Herms. Quoting is not citing: Disentangling affiliation and interaction on twitter. In *International Conference on Complex Networks and Their Applications*, pages 705–717. Springer, 2021. 45

[189] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 75, 82

[190] Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, 2016. 32, 43

[191] Arif Mohaimin Sadri, Samiul Hasan, Satish V Ukkusuri, and Juan Esteban Suarez Lopez. Analysis of social interaction network properties and growth on twitter. *Social Network Analysis and Mining*, 8(1):1–13, 2018. 7, 48

[192] Peter M Sandman. *Responding to Community Outrage: Strategies for Effective Risk Communication*. American Industrial Hygiene Association, 1993. 23

[193] Hiroki Sayama, Irene Pestov, Jeffrey Schmidt, Benjamin James Bush, Chun Wong, Junichi Yamanoi, and Thilo Gross. Modeling complex systems with adaptive networks. *Computers & Mathematics with Applications*, 65(10):1645–1664, 2013. 1

[194] Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. "Go eat a bat, Chang!": An Early Look on the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. *arXiv e-prints*, page arXiv:2004.04046, April 2020. 16

[195] Jssai Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1911(140):1–28, 1911. 108

[196] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015. 54

[197] Hao Sha, Mohammad Al Hasan, George Mohler, and P. Jeffrey Brantingham. Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives. *arXiv e-prints*, page arXiv:2004.11692, April 2020. 8

[198] C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, Jan 1949. 94, 150

[199] Cynthia S.Q. Siew, Dirk U. Wulff, Nicole M. Beckage, and Yoed N. Kenett. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019, 2019. 24

[200] Orowa Sikder, Robert E Smith, Pierpaolo Vivo, and Giacomo Livan. A minimalistic model of bias, polarization and misinformation in social networks. *Scientific reports*, 10(1):1–11, 2020. 7, 45

[201] Daniel J Simon and D Simon. Kalman Filtering With State Constraints: A Survey of Linear and Nonlinear Algorithms. *Control Theory & Applications*, 4(8):1303–1318, 2010. 3, 106, 107, 121, 128

[202] Luke Sloan. Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media and Society*, 3(1), 2017. 42

[203] Paul Slovic. *"If I look at the mass I will never act": Psychic numbing and genocide*. Springer, Dordrecht, 2010. 14, 21, 33, 39, 41, 126

[204] Akhila Sri Manasa Venigalla, Dheeraj Vagavolu, and Sridhar Chimalakonda. Mood of India During Covid-19 – An Interactive Web Portal Based on Emotion Analysis of Twitter Data. *arXiv e-prints*, page arXiv:2005.02955, May 2020. 15

[205] M. Stella, V. Restocchi, and S. De Deyne. #lockdown: network-enhanced emotional profiling at the times of COVID-19. *Big Data and Cognitive Computing*, 4(2):14, 2020. 15, 24

[206] Massimo Stella. Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media. *PeerJ Computer Science*, 6:e295, 2020. 7

[207] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences of the United States of America*, 115(49):12435–12440, 2018. 8, 42

[208] S. S. Stevens. On the psychophysical law. *Psychological Review*, 64(3):153–181, 1957. 34

[209] S. S. Stevens. *Psychophysics*. New York: Wiley, 1975. 14, 37

[210] C Summers, P Slovic, D Hine, and D Zuliani. 'Psychophysical Numbing': An Empirical Basis for Perceptions of Collective Violence. *Collective Violence: Harmful Behavior in Groups and Governments*, 1994. 14, 37, 40

[211] Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999. 147

[212] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981. 74

[213] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010. 19

[214] Matthew Taylor, Jonathan Watts, and John Bartlett. Climate crisis: 6 million people join latest wave of global protests. *The Guardian*, 2019. 2, 46, 58, 62, 127

[215] Guy Tchuente. Weak identification and estimation of social interaction models. *arXiv preprint arXiv:1902.06143*, 2019. 107, 109, 110

[216] Stefan Thurner, Rudolf Hanel, and Peter Klimek. *Introduction to the theory of complex systems*. Oxford University Press, 2018. 1

[217] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 2012. 82

[218] María Tomás-Rodríguez and Stephen P Banks. *Linear, time-varying approximations to nonlinear dynamical systems: with applications in control and optimization*, volume 400. Springer Science & Business Media, 2010. 148

[219] Kathie M.d.I. Treen, Hywel T.P. Williams, and Saffron J. O'Neill. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):1–20, 2020. 45

[220] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010. 19, 20

[221] Aman Tyagi, Matthew Babcock, Kathleen M Carley, and Douglas C Sicker. Polarizing tweets on climate change. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 107–117. Springer, 2020. 46, 47, 65

[222] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, pages 1–12, 2020. 23

[223] Giuseppe A Veltri and Dimitrinka Atanasova. Climate change on twitter: Content, media ecology and information sharing behaviour. *Public Understanding of Science*, 26(6):721–737, 2017. 8, 48

[224] J Scott Verinis, Jeffrey M Brandsma, and Charles N Cofer. Discrepancy from expectation in relation to affect and motivation: Tests of McClelland's hypothesis. *Journal of personality and social psychology*, 9(1):47, 1968. 16

[225] Sanita Vetra-Carvalho, Peter Jan van Leeuwen, Lars Nerger, Alexander Barth, M. Umer Altaf, Pierre Brasseur, Paul Kirchgessner, and Jean Marie Beckers.

State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems. *Tellus, Series A: Dynamic Meteorology and Oceanography*, 70(1):1–38, 2018. 158

[226] Jonathan A Ward, Andrew J Evans, and Nicolas S Malleson. Dynamic calibration of agent-based models using data assimilation. *Royal Society open science*, 3(4):150703, 2016. 70, 74, 75, 97, 98, 124

[227] Peter C Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140, 1960. 2, 45

[228] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998. 110

[229] Elke U Weber. Perception matters: Psychophysics for economists. *The psychology of economic decisions*, 2:163–176, 2004. 14, 41

[230] Gérard Weisbuch, Guillaume Deffuant, Frédéric Amblard, and Jean-Pierre Nadal. Meet, discuss, and segregate! *Complexity*, 7(3):55–63, 2002. 3

[231] Christopher K Wikle and L Mark Berliner. A bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1-2):1–16, 2007. 102

[232] Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global environmental change*, 32:126–138, 2015. 8, 45, 46, 47, 48, 53, 56, 65, 141

[233] Yan Xia, Ted Hsuan Yun Chen, and Mikko Kivelä. Spread of tweets in climate discussions: A case study of the 2019 nobel peace prize announcement. *Nordic Journal of Media Studies*, 3(1):96–117, 2021. 46, 58, 65

[234] Kazunori D Yamada. Yamadam: a hyperparameter-free gradient descent optimizer that incorporates unit correction and moment estimation. *BioRxiv*, page 348557, 2018. 82, 86, 95

[235] Xin-She Yang. *Nature-inspired Metaheuristic Algorithms*. Luniver Press, 2010. 75

[236] Hao Ye, R. J. Beamish, S. M. Glaser, S. C. H. Grant, Chih-hao Hsieh, L. J. Richards, J. T. Schnute, and G. Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proceedings of the National Academy of Sciences*, 116(41):E1569–E1576, 2015. 74

[237] Dezhi Yin, Samuel D Bond, and Han Zhang. Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS quarterly*, 38(2):539–560, 2014. 19

[238] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 82, 86, 95

[239] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018. 12