Universidade de São Paulo
Instituto de Física

Formação de comunidades em modelos de sociedades baseados em agentes

Felippe Alves Pereira

Orientador: Prof.Dr. Nestor Caticha

Tese de doutorado apresentada ao Instituto de Física da Universidade de São Paulo, como requisito parcial para a obtenção do título de Doutor em Ciências.

Banca Examinadora:

Prof. Dr. Nestor Felipe Caticha Alfonso - Supervisor (IFUSP)
Prof. Dr. Renato Vicente - (IME/USP)
Prof. Dr. André Cavalcanti R. Martins - (EACH/USP)
Prof. Dr. Jeferson Jacob Arenzon - (UFRGS)
Prof. Dr. Thadeu J. Pereira Penna - (UFF)

São Paulo – 2020

FICHA CATALOGRÁFICA
Preparada pelo Serviço de Biblioteca e Informação
do Instituto de Física da Universidade de São Paulo

Pereira, Felippe Alves

Formação de comunidades em modelos de sociedades baseados em agentes / Community formation in agent based models of societies. São Paulo, 2020.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Física. Depto. de Física Geral

Orientador: Prof. Dr. Nestor Felipe Caticha Alfonso
Área de Concentração: Física Estatística de Sistemas Complexos

Unitermos: 1.Mecânica estatística; 2. Modelos de aprendizagem; 3. Redes neurais; 4. Sistemas multiagentes; 5. Ciências sociais.

USP/IF/SBI-058/2020

University of São Paulo
Physics Institute

Community formation in agent based models of societies

Felippe Alves Pereira

Supervisor: Prof.Dr. Nestor Caticha

Thesis submitted to the Physics Institute of the University of São Paulo in partial fulfillment of the requirements for the degree of Doctor of Science.

Examining Committee:

Prof. Dr. Nestor Felipe Caticha Alfonso - Supervisor (IFUSP)
Prof. Dr. Renato Vicente - (IME/USP)
Prof. Dr. André Cavalcanti R. Martins - (EACH/USP)
Prof. Dr. Jeferson Jacob Arenzon - (UFRGS)
Prof. Dr. Thadeu J. Pereira Penna - (UFF)

São Paulo – 2020

RESUMO

Neste trabalho apresentamos um modelo para a formação de comunidades em sociedades a partir da dinâmica de trocas de opinião e desconfiança entre agentes. A teoria é desnvolvida com base na Teoria de Probabilidades, Aprendizado de Máquina no princípio de Máxima Entropia (MaxEnt), dos quais deduzimos uma nova forma de Dinâmica Entrópica para Sistemas de Processamento de Informação, em particular para Redes Neurais simples, a Dinâmica Entrópica de Aprendizado.

A teoria e o modelo para a interação de agentes foram analisados em alguns cenários, escolhidos pela natureza intuitiva e de possível associação com circunstâncias reais. Começamos com a análise de sistemas com 2 agentes interagindo em diferentes condições de opinião e desconfiança iniciais, mostrando que a dinâmica deduzina não apresenta apenas fases triviais e não leva a interpretações absurdas. Em seguida, analisamos as propriedades de sociedades com muitos agentes, variando a distribuição de opiniões e desconfianças iniciais, bem como os assuntos que poderiam ser discutidos pelos agentes, mostrando que diferentes condições levam a consenso, polarização ou até mesmo a uma fase frustrada como um video de spin.

Finalizamos com uma aplicação do modelo para o comportamento dos juízes, dada a disponibilidade de dados a respeito da influência ideológico-partidária nos padrões de decisão judicial da Corte de Apelações dos EUA. Nesta aplicação, apesar de se apresentar como uma caricatura com o objetivo de dispor uma ferramenta quantitativa para especialistas na área, tentamos imitar as situações típicas às quais um colégio judicial composto por três juízes estaria submetido, atribuindo aos agentes representantes dos juízes um conhecimento da Lei, um viés do Partido, uma Personalidade e os expondo a diferentes cenários de desconfiança. O único cenárioi capaz de reproduzir o padrão empírico de votações requer que os juízes sejam representados por agentes que atribuem pesos similares à Lei, ao viés Paridário e à Pesonalidade, bem como que extendam a Cortesia Certeira de confiar em juízes com vieses políticos opostos.

Palavras-chaves: Modelos de Agentes, Redes Neurais, Aprendizado de Máquina, Sociologia, Mecânica Estatística

ABSTRACT

In this work we present an agent based model for community formation on societies from the dynamics for opinion exchange and distrust between agents. The development framework relies on Probability Theory, Machine Learning and MaxEnt principles, from which we derive a new form of Entropic Dynamics for Information Processing Systems, in particular for simple Neural Networks, the *Entropic Learning Dynamics*.

The resulting theory and model for agents interactions are analyzed in a few scenarios, chosen due to their intuitive nature and connection with possible real scenarios. We started the analysis with the properties of systems with 2 agents interacting under different trust and opinion initial conditions, and showed that the dynamics is not trivial nor leads to results with absurd interpretations. Then, we analyzed the properties of societies with many agents, varying the distribution of opinions and distrust, as well as the subjects they could discuss, and found different situations leading to consensus, polarization and even frustrated state like a spin glass.

Finally, we applied the model to study the behavior of judges due the availability of data regarding the influence of political party ideology in the voting patterns of judges in the U.S Court of Appeals. In this application, although just a caricature aiming just to provide a quantitative tool for experts in the field, we tried to mimic the typical situations a panel of three judges would be submitted, attributing to agents representing judges a common knowledge of the Law, a Party bias, a Personality and exposing them to different distrust scenarios. The only scenario capable of reproducing the available data had to consider similar contributions of the Law, Party bias and Personality, as well as having Courteous and Certain judges, who extended the courtesy of attributing low distrust to agents of the opposing political party.

Keywords: Agent Models, Neural Networks, Machine Learning, Sociology, Statistical Mechanics

ACKNOWLEDGMENTS

I'm deeply grateful to Prof. Caticha for all the knowledge he taught me, for inspiring passion about research and all the support he provided me.
Thanks to Milton, Marilda, my parents, and Victor, my brother, for all the love and support to achieve my dreams.
To Lilian, for caring and all the patience with my temper.
To all the humans, for being amazing and providing such an interesting subject.
And to CNPq for the financial support.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

We are often faced with many tasks that require interaction among ourselves and most of these tasks will eventually require agreement between the participants in order for something to happen. Still there are countless scenarios where agreement is not achievable, at least not as fast as we would like to, despite our tendencies to cooperate and build common and greater good. Also, there is a lot of scenarios where consensus builds fast, but it is not desirable. This apparent behavioral incongruency is the object of study in many research areas: this persistence of this multiple stances regarding matters in which we would like to agree. The research about this subject is scattered among many of its examples, like the existence of cultural borders, racial segregation, political alignment and international conflict, to cite a few.

In this work we present a mechanism to understand this tendency to form *Communities* we observe in human societies. The strategy we follow is based on the use of *Artificial Neural Networks* and strong use of *Probabilistic* and *Entropic* inference, which, as we hope to show, favor a clearer interpretation at the cost of a slightly more complex theoretical model. More precisely, we present a model for agents that emulate conversations, within a fixed abstract broad subject, learn to form opinions and assign trust to each other. More then the model results, we try to show the idea behind the tools used here to build models for other phenomena related to human behavior in society, hoping this helps to establish a solid framework for this type of queries. We seek a framework expressive enough to investigate many scenarios of social interaction while still being simple enough to enable analysis. Also, it should be rooted in cutting edge inference methods and with a clear connection among the entities representing the studied phenomenon and the results of the analysis. Fulfilling these requirements are the techniques employed to establish learning algorithms by *Entropic Dynamics*. We apply it for sequential inference on parametric distributions.

Although we focus on the model development and its results, the motivation comes from the growing amount of data and experimental results in many fields related to individual and social human behavior. Through the last few decades, the study of human behavior and social phenomena was developed in many directions, bringing new insights, perspectives and hope for the design of a theory on this complex subject. These results lie in many different research and application fields, like the availability of social network data and

computational methods to analyze it, studies in behavior and moral psychology, evolutionary game theory and, of particular interest to us, advances in information theoretical methods, statistical mechanics and machine learning. From these multiple points of view about the nature of behavior and social phenomena, we are beginning to understand some key aspects, or at least what seems to be the key aspects, to achieve such theory. As will be discussed ahead, we believe these keys aspects to be the ability to process information, to communicate and to learn.

Given the variety of research fields interested in the matters of social behavior, it is only natural to face many different questions regarding the subject. For instance, it is possible for a political scientist to be interested in understanding how a particular nation deals with international conflict, while social psychologists may be interested in how citizens would behave in such situation, a neuroscientist may want to identify the patterns of brain activity relevant for that behavior and a physicist could be interested in the symmetries exhibited in this kind of events among nations or different social settings. The rest of this chapter illustrates some of the results that serve as motivation for our work and a brief explanation of the history behind the models used here and other interesting results from it.[1]

1.1 PSYCHOLOGICAL AND SOCIOLOGICAL BASIS

1.1.1 *Reinforcement and Social Learning*

Adaptability is one of the perks of human behavior that seems to be of most importance when one tries to model some phenomenon of social nature. We believe this adaptability is related to the ability to learn from from examples and there is some evidence on how the social interaction can be a source of learning.

Studies show[2] how the region of the *Anterior Cingulate Cortex* (ACC) together with the *Dopaminergic Mesencephalic System* work as an error detection and rewarding mechanism while learning how to execute a task. This conclusion is based on Electroencephalogram (EEG) analysis for Error Related Negativity (ERN).

The ability to detect errors by the ACC may be related to sensation of social exclusion, suggesting that social awkwardness may be considered a type of error subject to learning social behavior. Parallel findings[3] point to an increase in the activity of the ACC when a person experiences social exclusion or is negatively rated by others.

1 See 13, 14, 44–46, for previous works with a similar approach.
2 See 25, for ACC as rewarding mechanicsm.
3 See 17, 40, for ACC as activity under social exclusion.

Another interesting result[4] is the correlation between the ERN amplitude and one's self declared political attitude, observed to be different in self declared conservatives when compared to self declared liberals on several justice issues[5]. The study finds that the former group present smaller ERN than the later group, indicating that surprising information is processed differently by each group.

These results point to the possibility of linking social interaction with the dopamine system for reinforcement learning in the brain, and provides a way to capture human interaction through communication. Studies by Sherif [37] and Asch [5] give support to the idea of the influence of social pressure on a person seeking an answer to a hard task, in which errors are easy to identify.

The setup for Sherif's experiment was the following: a group of three participants, two of which are confederates for the experiment, are sitting in a completely dark room with the exception of one light dot projected within a fixed distance from them and they are asked to estimate the displacement of the light dot[6]. The confederates are in previous agreement to estimate the displacement within a determined range of values, while the 'naive' participant will give his/her own estimate. After a few seconds staring to the light dot, which is completely still, the participants state their estimates for its displacement, followed by a new round of observation and estimation. The experiment is repeated in the next day only with the naive participant, with no confederates.

The experiment's results show the convergence of the naive participant estimates to the predetermined ranges of values given by the confederates, depending on the degree to which this range is credible. Also, the estimated ranges persist at least until the naive participant is tested alone in the next day, indicating a kind of learning of the social influence from the previous day.

The objective of Asch's experiment was to determine the effect of group size and range precision on the influence exerted by the group on the naive participant. To test this, groups with 6 to 9 participants, always with all confederates except one naive, were presented with two white cards, one of which had three black bars of different sizes and the other one with a single black bar. The participants were asked to choose which of the three bars matched the size of the bar on the single bar card, always with a single correct answer. The confederates were instructed to choose a wrong answer a fixed number of times and the number of times the naive participant chose the wrong card was measured in two situations, one when alone and the other within the group. This procedure pointed a clear influence of the group, affecting the typical error rate of 1% when the naive participant was

4 See 4, for the ERN relation with self-declared political attitude.

5 The terms "conservative" and "liberal" do not apply, in the context of Amodio's experiment, to economic issues.

6 This phenomenon is known as *Auto-Kinetic Illusion*

alone up to 30% when he/she was under the influence of the confederate group.

On a different perspective, a recent study[7] shows, through Functional Magnetic Resonance Imaging (FMRI), the effect of social influence on tasks of value attribution at a brain activity level. The task of attributing value is related to activity in the *Ventral Striatum* region in the brain, which presents a varied activation under social influence.

1.1.2 *Morality and Intuition Primacy*

An important aspect of communication is the subject of the information being communicated. When looking for a good representation of information and how it is processed by the communicants, it would be useful to be able to translate a piece of real world of opinionated information into an object in the theory.

As an example, consider the *Moral Foundation Theory* (MFT),[8] which contributes to the understanding of human morality through four principles. The first principle, the *Intuition Primacy* states that moral judgments happens as an automatic response to incoming moral dilemmas. Notice that Primacy doesn't mean domination over reasoning but a first path to the conclusion, possibly being later replaced by higher level information processing mechanisms. The second principle states the universality of a fixed set of foundations, the most primitive ideals to human morality. This implies that, despite moral values changing from place to place and time to time, the moral reasoning happens in a common stage. Using questionnaires applied to people in many countries, Haidt and colleagues could determine the five most representative foundations out of a set of plausible candidates. The determined foundations, they say, are the concerns with *Violence/-Care*, with *Justice/Reciprocity*, with *Group Inclusion/Loyalty*, with *Authority/Respect* and with *Purity/Sanctity*. Using a set of questionnaires with moral dilemmas, the MFT proponents were able to capture to which extent each of the foundations was relevant to different groups of different self-declared political attitudes. Of immediate interest to us is the possibility of consistently recovering an internal/hidden characteristic of behavior through a set of questions that can be represented by a small set of numbers, in this case 5.

1.1.3 *Opinion Dynamics, Cultural Dissemination and Segregation*

One ambitious branch of the study of human behavior goes by the name of *Opinion Dynamics* and gather the endeavor of many researchers on the task of finding suitable models for how human communication affects the society. A large collections of models with different

7 See 10, for social influence over value attribution.

8 See 23, 24, for MFT and MFQ.

degrees of success have been introduced in the last decades, based on a even larger set of techniques and empirical results.[9]

One of the most influential works in the field of "Opinion Dynamics" is the model for culture dissemination presented by Axelrod [6] in the early 1990's. The model has a simple dynamics for agents in square lattice with interactions based on cultural similarity and exhibits a counter intuitive behavior qualitatively similar to the coexistence of multiple cultures even when cultural convergence is locally enforced. In more detail, Axelrod's model consists of agents represented by a set of cultural features, each feature with a fixed number of possible traits, and each agent in a site in a square lattice, possibly with interactions reaching further than the first neighbors. The interaction between two agents occur with probability proportional to the number of equal traits they have at that instant and, when two agents interact, one of them copies one trait from the the other.

Varying the number of agents, the distance of interactions and the numbers of features and traits, the model shows that, even when the interaction enforces local convergence, global polarization can be sustained in some situations. The main results are the reduction in number of stable distinct cultural regions when the number of neighbors of each agents or the size of the lattices increase. Also the number of stable cultural domains increases with the number of traits. These results provide an argument for the possibility of cultural diversity arising from local interactions favoring similarity, when the there is enough room to diversity at the feature or spatial levels.

Other interesting, although not exactly an opinion model, result is the model for spatial segregation by Schelling [36]. As with the Axelrod's results, this study defines a set of rules for local interaction of agents with a definite character, a label representing the skin color, in a square lattice. The color of each agent is of public knowledge the agents of the same color share the same demands for same color density of neighbors. When an agent's demands are not met, it is considered to be dissatisfied and moves to other site where it can be satisfied, if possible. The results show that, even for tolerant or integrative demands, segregation can be sustained and large domains of like agents become stable. Although these results are affected by varying the demands, distribution of color and size of neighborhood, as long as likeness is demanded, segregation is a stable equilibrium. One interesting argument in Schelling's work is that the definition of agents' neighborhood can reduce the dissatisfaction of agents and, consequently, reduce the possibility and intensity of segregation effects in the society.

Both Axelrod's and Schelling's work brought a new perspective on the study of social influence and behavior. However, both of them can be considered oversimplified to actually represent what they pro-

9 See 6, 19, 21, 26, 35, 41, 44, for an broad class of approaches to opinion dynamics.

pose to understand. More precisely, both models successfully show the possibility of community formation from homophilic interactions, however each suffer at least one severe drawback, for instance Schelling's model relies on global knowledge of the the agents positions and features and Axelrod's requires the narrow information bottleneck from a square lattice. We consider both to take a very "kinematic", or algorithmic, approach to represent the social behavior, which present several difficulties when one tries to compare it to real data. Instead of setting an algorithmic behavior to agents at the action level, we will try to understand the dynamics behind the processing of information, which can be easier to reason about. We will show that this "dynamic" way of thinking about human behavior leads to promising analytical tools to start considering theoretical and data analysis for real social phenomena.

1.1.4 *Information Processing Systems based models for social behavior*

The deep connection between *Information Theory* and *Statistical Mechanics* makes possible the theoretical analysis of information processing systems using methods designed to study collective phenomena. For instance, many important results in *Machine Learning* on *Artificial Neural Networks* were obtained using *replica methods* or *MaxEnt*.[10] The combined advances in these areas also allowed the deployment of methods of entropic inference, statistical mechanics, neural network and machine learning to model certain aspects of societies, where its members are represented by adaptive agents such as artificial neural networks.[11]

In [3] we introduced a reasonable, but unprincipled, mechanism for the appearance of effective communications barrier. Such barriers were not the result of blocking information exchanges but rather from a dynamics of distrust based "cognitive rewards" agents expected from interactions. In the present work, we develop a principled dynamics of distrust, within a framework general enough to enable the modeling of other types of behaviors.

One of the main differences between the way we tackle this kind of phenomena and the more traditional way is our usage of *Statistical Mechanics*, *Bayesian* or *Entropic Inference* and *Machine Learning*. The reason for this choice lies in the simplicity and expressive power of such theoretical tools when compared to others, like typical techniques of dynamics systems or classical statistics. Although the mathematics behind it are a little bit more demanding the rewards are much greater. Also, we believe our approach is easier to explain to people with no mathematical training, despite the greater mathematical complexity.

10 See 20, 28, 29, 34, 39, for Statistical Mechanics applications to Neural Network and Machine Learning.

11 See 3, 13, 14, 30, 38, 44–46, for other NN agent based models of social behavior.

Finally, the spontaneous natural interpretations it leads us may help to break the "language barrier" between the social sciences and mathematical theories for social behavior.

1.2 CONTRIBUTIONS AND THESIS STRUCTURE

This thesis contribution is threefold: we present a simplifying assumption to deduce optimal learning algorithms from MaxEnt; an *Agent Model*, our main contribution, to study community formation based on opinion exchanges and distrust; and an application of the model to study data from the U.S. Federal Court of Appeals.

In Chapter 2 we develop the *Entropic Learning Algorithm*, which serves a the basis for the *Agent Model* presented in Chapter 3, where we develop and discuss how agents behave, the interpretation of the model entities and how to reason about community phenomena within this model. Chapter 4 presents an application to model a data set and analysis[12] regarding the political party ideological influence over judicial decisions in the U.S. Appellate Court.

12 See 42, for an expert analysis of the judicial decisions of the U.S. Court of Appeals.

2

ENTROPIC LEARNING DYNAMICS

As pointed out in Chapter 1, we want a theory for understanding social behavior in terms that can be used to reason about whatever phenomenon we might be interested in. It should be clear that the necessary mathematical tools must be the most general and universal available and still be as simple as possible. Probability Theory is one strong candidate to fit these criteria. More specifically, we refer to the logical view of Probability Theory, as regarded by P. Laplace, Cox [15], Jaynes and Bretthorst [27] and other names in the most modern studies of inference methods.[1] This, somehow generalizing, view of Probability lead us to a strong method of inference using *Entropy Maximization*. The method not only enables us to make inference based on a extremum principle, which allows subject specific knowledge to be incorporated as constraints, but also provides us with simple procedures to derive mechanisms based on simple hypotheses. This approach is no stranger to physicists, used to apply Variational methods and think in terms of Hamiltonians. Given the simplicity of the method, it is very obvious which ingredients in our model are responsible for each resulting aspect of the theory and how to introduce, modify or exclude hypothesis about whatever we are trying to understand, social behavior being no more than an example.

An inherent companion to the subject of inference is the problem of representing information. We could think about many ways to mathematically represent ideas like "opinion" and "trust", some reasonable and some not so much. For instance, ontologies in first-order logic, a set of Ising spins or real numbers, a multiplayer game, or an Artificial Neural Network (ANN), is a non-extensive list of mathematical objects used to this end. Any choice has its pros and cons, and we chose ANNs the later for the following reasons: First, its is easy to do inference for simple ANNs, like the *Perceptron*; second, the theory of Machine Learning is very well integrated with ANNs, and learning is a very reasonable way to introduce adaptability to information processing; third, we have some experimental indication that it may not be a far fetched representation of how our brain deal with information, at least for some of its processes; and finally, its relatively simple to interpret small (shallow) ANNs,[2] and finally, it permits addressing more complex problems than those modeled with Ising variables.

1 See 12, 22, and references therein.

2 In opposition to the problems of interpretation in Deep Neural Networks, shallow networks carry the semantic interpretation of each unit.

We begin with a brief explanation of how to use *Probability Theory* as an inference tool rooted on predicate logic. The principles behind this take on Probability has been explored by *E. T. Jaynes* and others together with the extended interpretation of *Entropy* as a measure of ignorance regarding the subject in focus. The aspect of generality in Probability Theory resulting from this construction as an extension to predicate logic provides a strong method to start building models and theories about phenomena typically regarded by some as "impossible to mathematize", like social and economic behavior or the functioning of the brain to name a few. We then proceed with a review of Entropy as a Probability updating method and how we use it to build a Dynamics for the inference of parameters in a given model. Finally, we apply the results for the particular case of *Exponential Family* of probability distributions, with *Gaussian Distributions* as a special case, leading to a simple dynamics for updating the expected values of a Gaussian model on incoming information. This method, combined with clever tricks to build Gaussian models, is very useful to tackle a broad category of subjects and we call it *Entropic Learning Dynamics* (ED).[3]

2.1 PROBABILISTIC INFERENCE

Classical Probability Theory is based on concepts of Set Theory with the Kolmogorov's Axioms and nested with concepts of Measure Theory and Combinatorics. This approach to probability clouds the its power to deal with more general concepts, ideas hard to express in such terms. A more inviting approach is to realize that probability is an extension to Propositional Logic, which enable us to deal directly with assertions about the subject and perform inference in a much clearer way. Bear in mind, however, that the axioms and theorems in Measure and Classical Probability theories are still valid and constitute base of Probability Theory from an operational perspective. The Logical approach to Probability theory can be seen as the establishment of a map between Logic and Probability, building a bridge between logical deduction and statistical analysis, which is the main route of science to understand anything.[4]

Consider assertions, or propositions, $a, b, c, \ldots$, which could be any phrase evaluating to TRUE or FALSE. For instance, a could be "Its raining outside" and b could be "There are clouds outside". In Propositional Logic we are interested to deduce the truth about an assertion given other propositions. The most natural way to achieve this is to consider the propositional algebra for the operators AND OR

3 See 11, for an introduction to *Entropic Dynamics*.

4 See 12, 15, 22, 27, for the mathematical development and applications of Probability as an extension of Logic.

and NOT [5], which will be denoted as $\wedge, \vee$ and $\neg$. The AND is defined so $a \wedge b$ is TRUE if and only if both a and b are TRUE. For the operator OR, $a \vee b$ is FALSE if and only if both a and b are FALSE. And the NOT operator is defined so $\neg a$ is TRUE if a is FALSE and vice-versa.

For assertions a and b we can use these operations to define the implication operator $a \Rightarrow b = \neg(a \wedge \neg b) = \neg a \vee b$, the equivalence operator $a \Leftrightarrow b = (a \Rightarrow b) \wedge (a \Rightarrow b)$, and with the two extra quantifying operators $\forall$ and $\exists$ we would be able to establish all the classical Deductive Inference, which is base for Mathematics[6]. As an example, we could say $R =$"It is raining outside", $C =$"There are clouds in the sky" and $I =$"If it is raining there is clouds in the sky"$= (R \Rightarrow C)$ and ask ourselves if C is TRUE, which would follow by *Modus Ponens*.

One weakness of Propositional Logic, from the point of view of dealing with incomplete information which is the case for most real phenomena, is its complete inability to do the reverse of such deductive inference. For instance, with the assertions R, C and I given above, if know in advance that C is TRUE, i. e. "There are clouds in the sky", but don't know if R is TRUE or FALSE, i.e. whether "It is raining outside" or not, C and I alone could tell us nothing about R. And yet, we somehow develop the intuition to prepare ourselves for the rain before leaving our houses in a cloudy day.

To understand how we develop such intuition, and many others in a similar informational context, we can follow the steps of *R. Cox*, and others after him, and see that the propositional algebra for $\wedge, \vee, \neg$ can be extended to support this kind of inference, leading to Probability Theory.[7] Call the probability of an assertion a being TRUE conditioned on information I by $p(a|I)$. To say the probability of an assertion a to be conditioned on some other assertion I is equivalent to say how much should we one *rationally believe* a to be TRUE given the certainty about the truth of I. In this notation, the propositional algebra translates into probability algebra as follows

$$p(a \wedge b|I) = p(b|I)p(a|b \wedge I) = p(a|I)p(b|a \wedge I) \tag{1}$$

$$p(a \vee b|I) = p(a|I) + p(b|I) - p(a \wedge b|I) \tag{2}$$

$$p(a|\neg a) = 0 \tag{3}$$

$$p(a|a) = 1 \tag{4}$$

From this rules follows the *Bayes'* Theorem

$$p(a|b \wedge I) = \frac{p(a|I)}{p(b|I)} p(b|a \wedge I) \tag{5}$$

5 There are alternative operators defining the same Propositional Logic, but we wont use them here

6 By "base for Mathematics" we mean that Logic is used to deduce theorems from axioms, and to reason about them in general, regardless of which field of math is being studied. Not to be confused with the usage of Propositional Logic to construct every mathematical structure.

7 See 15, for a deduction of Probability Theory as an extension of Propositional Logic.

which is the basis for inference in face of incomplete information. For instance, we can use the Bayes' Theorem to analyze the "It is raining outside" example above

$$P(R|C \wedge I) = P(C|R \wedge I)\frac{P(R|I)}{P(C|I)}$$

since $R \Rightarrow C$, we can say $P(C|R \wedge I) = 1$ and because $0 \leq P(\cdot|\cdots) \leq 1$ we have

$$P(R|C \wedge I) = \frac{P(R|I)}{P(C|I)} \geq P(R|I)$$
$$\Rightarrow \frac{P(R|C \wedge I)}{P(R|I)} \geq 1$$

which, with the information we have about rain and clouds, allows us to assign a greater probability to face rain when there is clouds in the sky compared to when the sky is clear, in accordance with our intuition.

For a more interesting example consider the propositional descriptions H for the hypothesis about some physical phenomenon, D a data set of values about it and I as all the prior Physics knowledge we have up to this point. We can ask what is the probability of H being TRUE once we observe the data D given our current knowledge I, which is readily answered by the Bayes' Theorem:

$$p(H|D \wedge I) = \frac{p(H|I)}{p(D|I)}p(D|H \wedge I) \tag{6}$$

This result tells us how to think about the process of inference, regarding the posterior distribution $p(H|D \wedge I)$ to be obtained from the prior $p(H|I)$ by weighting it with the evidence $p(D|I)$ and the likelihood of observing the data if the theory is true $p(D|H \wedge I)$. We call Bayes' Rule the iterated application of the Bayes' Theorem for sequentially incoming data. The Bayes' Rule can be illustrated following the Data-Hypothesis example by considering a prior $Q_0(H|I)$ and an experiment with results D_1, which leads to posterior $P_1(H|D_1 \wedge I)$ and, once we perform another experiment with result D_2, we treat our previous posterior as the prior for the next inference, i. e.$Q_1(H|I) = P_1(H|D_1 \wedge I)$, leading to $P_2(H|D_2 \wedge D_1 \wedge I)$ and so on. By not following the Bayes' Rule we would be required to either forget P_1 or to have full access D_1 to improve our knowledge.

In this work we make extensive use of the Bayes Rule, as it is the template for sequential probability encoded knowledge updating.

The main advantage of thinking Probability Theory as an extension of Propositional Logic is that probabilities distributions are the representatives of our knowledge and can be consistently updated when we get new information. This procedure of updating through Bayes' rule can be generalized to a method capable of incorporating information directly into probability distributions, called *Entropy Maximization* or *MaxEnt* and is the subject of the next section.

2.2 MAXENT AND THE EXPONENTIAL FAMILY

Entropy is one of those terms with a long history in the scientific literature. It came from Thermodynamics as a system's property related to its "disorder", a natural counterweight to the order induced by energy minimization. By the time of the Second World War, the concept of Entropy was finding its way to the field of Information Theory, as a measure of lost information through noisy communication channels. This coincidence was further extended to the point we find ourselves in, using Entropy as a measure of the lack of knowledge about a given object, in the most general meaning of the word "object". It turns out that this general concept of Entropy is very useful to update the knowledge about a given object on the base of incoming related information, having as the Bayes' Theorem as a special case. In what follows, we consider the Relative Entropy as given and use it as a "metric"[8] and describe how to use it as a tool to update probability distributions.[9]

The idea behind the method of *Entropy Maximization*, or *MaxEnt* for short, is to rank probability distributions according to their relative information content from some prior distribution and subject to some constraints and select the one which attains the maximum value in this ranking. For probability distributions q and p the *Relative Entropy* from p to q is defined as

$$S(q,p) = -\int \mathrm{d}x\, q(x) \ln \frac{q(x)}{p(x)} \tag{7}$$

and can be regarded as a measure of difference in the information required to leave p and get to q. Note that $S(q,p) \geq 0$ for all distributions, with equality only when $q = p$, and its not symmetric. The relative Entropy provides a functional method to find probability distributions satisfying given constraints, or in other word to assign a probability distribution based only on knowledge we may have about it. Consider constraints of the type $\kappa(q,a) = 0$ for some index $a \in \Lambda$, then we can form the objective function to be extremized as

$$\mathcal{L}(q,p,\kappa) = S(q,p) - \int \mathrm{d}a\, \theta(a)\kappa(q,a) \tag{8}$$

for continuous a or

$$\mathcal{L}(q,p,\kappa) = S(q,p) - \theta^a \kappa_a(P) \tag{9}$$

for discrete a[10].

8 It not a metric for it is not symmetric, but symmetry is the only lacking property for it to be a metric

9 See 11, 15, 27, for an inductive approach to Entropy.

10 We are using Einstein's convention for implicit summation of repeated indices:

$$u^a v_a \equiv \Sigma_a u^a v_a$$

The simplest form of constraints are linear in q, *i.e.* the distributions is constrained to expected values η_a of functions F_a, or explictly $\kappa_a(q) = \mathbb{E}(F_a|q) - \eta_a = 0$, where $\mathbb{E}(F|q) = \int \mathrm{d}x\, q(x)F(x)$. The normalization constraint falls in this type with $F_0(x) = 1$ and $\eta_0 = 1$ and will always be imposed. The Entropy with a finite number of linear constraints and a prior $p(x) = \mathrm{e}^{C(x)}$ gives us the objective functional to be maximized

$$\mathcal{L}(q) = S(q, \mathrm{e}^{C(x)}) - \theta^a[\mathbb{E}(F_a|q) - \eta_a] - \theta_0\left[\int \mathrm{d}x\, q(x) - 1\right] \tag{10}$$

which is maximized when the functional derivative of $\mathcal{L}$ relative to q and is zero:

$$0 = \frac{\delta\mathcal{L}}{\delta q} = -1 - \theta_0 - \theta^a F_a(x) - \ln\frac{q(x)}{\mathrm{e}^{C(x)}} \tag{11}$$

$$\Rightarrow q(x) = \mathrm{e}^{C(x)+\psi(\theta)-\theta^a F_a(x)} = \frac{1}{\zeta(\theta)}\mathrm{e}^{C(x)-\theta^a F_a(x)} \tag{12}$$

where $\psi(\theta) = -\ln\zeta(\theta)$ and $\zeta(\theta) = \int \mathrm{d}x\, \mathrm{e}^{C(x)-\theta^a F_a(x)}$. Note how the set function F_a define a family of distributions where each member of the family is a distribution with this same set of *Generating Functions*[11] but with different values of for their expectations η_a, or equivalently different values of the Lagrange multipliers θ^a. This family of probability distributions is called *Exponential Family*[12]. We identify a member of the Exponential Family by the values of the Lagrange multipliers, i. e. $Q(z|\theta)$ or Q_θ. Alternatively, we could identify a member the their expectation values, $Q(z|\eta)$ or Q_η, although its not always clear how the these values are related with generating functions and. There is, however, at least one subset of the Exponential Family that can be easily identified by their expected values: *Gaussian Distributions*, for which a simple relation between the Lagrange multipliers θ and the expected values η can exploited.

The Exponential Family has two interesting analytical properties strongly related to Maxwell's relations in Thermodynamics. Let us show these properties, since they are crucial for the development of the Entropic Dynamics in the next section. First, we can recover the

11 also called *Sufficient Statistics* in fields outside *Information Geometry*

12 Since any distribution satisfying finite linear constraints will be of this form, the Exponential Family includes, actually, all the distributions that maximize the entropy with any set of linear constraints. This can lead to ambiguities when we want to refer to member with the same set of generating functions, so a better name could be F-Exponential, where F are the chosen generating functions. However, we will try to make clear from the context what we mean to avoid cluttering the notation

expected values of a given Exponential Family member if we know its "*Partition Function*" ζ

$$\begin{aligned}
\frac{\partial}{\partial\theta^a}\psi = -\frac{\partial}{\partial\theta^a}\ln\zeta &= \frac{1}{\zeta}\frac{\partial\zeta}{\partial\theta^a} \\
&= -\frac{1}{\zeta}\frac{\partial}{\partial\theta^a}\int \mathrm{d}x\, \mathrm{e}^{-\theta^b F_b(x)} \\
&= -\frac{1}{\zeta}\int \mathrm{d}x\, \frac{\partial}{\partial\theta^a}\mathrm{e}^{-\theta^b F_b(x)} \\
&= -\frac{1}{\zeta}\int \mathrm{d}x\, [-F_a(x)]\mathrm{e}^{-\theta^b F_b(x)} \\
&= \int \mathrm{d}x\, F_a(x)\frac{1}{\zeta}\mathrm{e}^{-\theta^b F_b(x)} \\
&= \int \mathrm{d}x\, F_a(x) Q(x|\theta) \\
&= \eta_a \qquad (13)
\end{aligned}$$

The other property is the ability to recover the constraints by taking derivatives of the distribution

$$\begin{aligned}
\frac{\partial}{\partial\theta^a}Q(x|\theta) &= \frac{\partial}{\partial\theta^a}\mathrm{e}^{\psi(\theta)-\theta^b F_b(x)} && (14)\\
&= \mathrm{e}^{\psi(\theta)-\theta^b F_b(x)}\frac{\partial}{\partial\theta^a}\left[\psi(\theta)-\theta^b F_b(x)\right] && (15)\\
&= Q(x|\theta)\left[\frac{\partial\psi}{\partial\theta^a} - \frac{\partial}{\partial\theta^a}\theta^b F_b(x)\right] && (16)\\
&= Q(x|\theta)[\eta_a - F_a(x)] && (17)
\end{aligned}$$

where we made use of equation (13). Using these properties, we derive a method for sequential inference in the Exponential Family using MaxEnt and a clever choice of constraints. Not only the results form the theoretical core of our models for social behavior, its construction serves as an illustration of the use of MaxEnt as an inference method.

2.3 ENTROPIC DYNAMICS WITHIN THE EXPONENTIAL FAMILY

An interesting development from the MaxEnt method is the possibility to implement many kinds of *Dynamics from Sequential Inference*, meaning a systematic way of performing inference from incoming of (recurrent structurally consistent) data. Many results that inspired our adherence to this inference scheme come from the efforts of A. Caticha[13] to establish *Quantum Mechanics* on more probabilistic grounds aiming for easier interpretation and reasoning, under the name of *Entropic Dynamics*. Here we present an special case of Entropic Dynamics for inference constrained to remain in the Exponential Family for processes with iterated structured data. The results

13 See 11.

were heavily inspired in the theory of *Machine Learning*, specially in the results for the Optimized and Bayesian Perceptron Learning.[14]

Lets start by considering x as the variable representing the state of some object of interest for which we want to assign a proper value but can't be directly observed. Instead, the we can get information from the observation of quantities y, related to x by some encoded by a *Likelihood* $L(y|x)$. Suppose the meaningful questions we can ask about the quantity x can be encoded into a set of generating functions $F_a(x)$ and, as it seems natural, our prior distribution for the values of x is a member of the Exponential Family[15] $Q(x|\theta) = \mathrm{e}^{\psi(\theta)-\theta^a F_a(x)}$, where θ encodes our current knowledge about x.

Now, the Bayesian posterior for x after observing y is given by the Bayes' Theorem

$$P(x|y\theta) = \frac{1}{Z} Q(x|\theta) L(y|x) \tag{18}$$

where $Z = \int \mathrm{d}x\, Q(x|\theta)L(y|x)$. The Bayesian posterior is the best inference we can do from our prior with the observed value for y, however it not guaranteed that P will be a member of the Exponential Family[16]. If we want to remain in the parametric family defined by F, we need to constrain the inference to prevent the generating functions $F_a(x)$ from being continuously displaced by data as the process goes on. As we will see, the distribution we want is a member of the Exponential Family that preserves the information content about the Generating Functions of the Bayesian posterior. This amounts to say that we want Q' that simultaneously maximizes the Entropy from $Q(x|\theta)$ with constraints of the form $\mathbb{E}\left(F_a|Q'\right) = \mathbb{E}\left(F_a|P\right)$ and the Entropy from Q' to P.

More precisely, we want Q' that maximizes the Entropy

$$\mathcal{L}(Q', Q) = S(Q', Q) - \Delta\theta^a \left[\mathbb{E}\left(F_a\middle|Q'\right) - \mathbb{E}\left(F_a|P\right)\right] - \Delta\theta^0 \left[\int \mathrm{d}x\, Q'(x) - 1\right] \tag{19}$$

which, using the same procedure as in equations (11) and (12), can be readily found

$$0 = \frac{\delta\mathcal{L}}{\delta Q'} = -1 - \Delta\theta^0 - \Delta\theta^a F_a(x) - \ln\frac{Q'}{Q(x|\theta)} \tag{20}$$

$$\Rightarrow Q' = Q(x|\theta') = Q(x|\theta)\mathrm{e}^{\Delta\psi - \Delta\theta^a F_a(x)} \tag{21}$$

$$= \mathrm{e}^{\psi(\theta)+\Delta\psi-[\theta^a+\Delta\theta^a]F_a(x)} \tag{22}$$

$$= \mathrm{e}^{\psi(\theta')-\theta'^a F_a(x)} = \frac{1}{\zeta(\theta')}\mathrm{e}^{-\theta'^a F_a(x)} \tag{23}$$

14 See 28, 29, 34, 39, for the development of optimal algorithms for the *Perceptron*.
15 The choice $C(x) = 0$ corresponds to an uniform prior.
16 Its true when L is a conjugate distribution to Q, but not in general

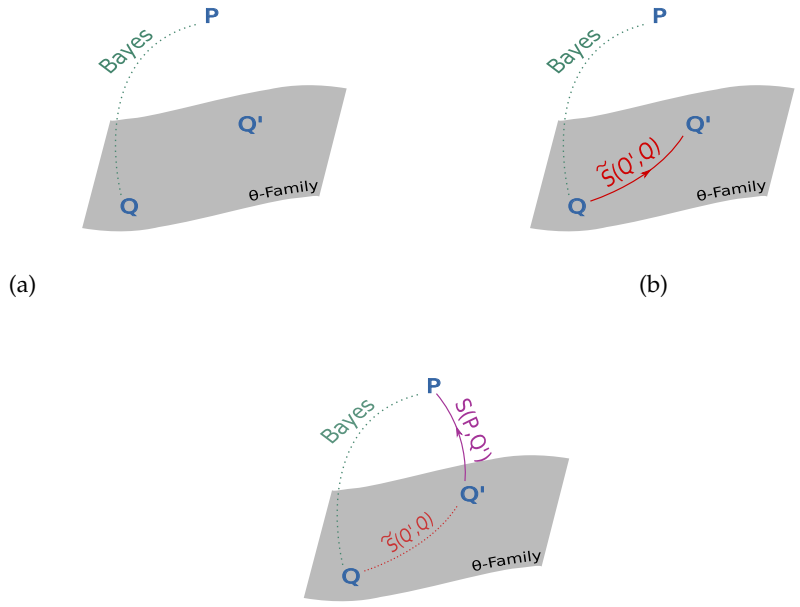


Figure 1: Illustration of the three possible ways to perform this inference: (a) show inference through Bayes' Theorem, (b) shows inference through maximization of $\tilde{S}(Q', Q) = S(Q', Q) + \Delta\theta[\mathbb{E}(F|Q) - \mathbb{E}(F|P)]$ and (c) shows inference by maximization of $S(P, Q')$ over Q', i. e.with reversed direction. Inference (a) gives P while Q' from (b) are (c) are the same and only match P on the expected values of the generating functions

where θ' is a new point in the parametric space and $\psi(\theta') = -\ln \zeta(\theta')$, with $\zeta(\theta') = \int \mathrm{d}x\, \mathrm{e}^{-\theta' a F_a(x)}$. Notice how the fact that both Q and Q' are in the same Exponential Family comes from imposing constraints on the same set of generating functions F.

Now, for our prior we had the constraint $\mathbb{E}(F_a|Q) = \eta_a$ and for the posterior we must have

$$\mathbb{E}\left(F_a\middle|Q'\right) = \eta'_a = \mathbb{E}\left(F_a|P\right) = \int \mathrm{d}x\, P(x|y\theta)F_a(x) \tag{24}$$

$$= \int \mathrm{d}x\, \frac{1}{Z} L(y|x)Q(x|\theta)F_a(x) \tag{25}$$

$$= \frac{1}{Z}\int \mathrm{d}x\, L(y|x)Q(x|\theta)[F_a(x) + \eta_a - \eta_a] \tag{26}$$

$$= \eta_a + \frac{1}{Z}\int \mathrm{d}x\, L(y|x)Q(x|\theta)[F_a(x) - \eta_a] \tag{27}$$

$$= \eta_a + \frac{1}{Z}\int \mathrm{d}x\, L(y|x)\left[-\frac{\partial Q(x|\theta)}{\partial\theta^a}\right] \tag{28}$$

$$= \eta_a - \frac{1}{Z}\frac{\partial}{\partial\theta^a}\int \mathrm{d}x\, L(y|x)Q(x|\theta) \tag{29}$$

$$= \eta_a - \frac{1}{Z}\frac{\partial Z}{\partial\theta^a} \tag{30}$$

$$\Rightarrow \eta'_a = \eta_a - \frac{\partial \ln Z}{\partial\theta^a} \tag{31}$$

where we used equation (17) to write equation (28). This is the Entropic Dynamics for the Exponential Family. Notice how we have a simple mechanism to perform sequential updating on the expected values by following the gradient of the evidence Z[17]. Also, notice that the derivative is relative to θ but the update is explicit on η. This can generate difficulties to implement the dynamics to most cases of exponential families, as a method of inversion must be used to properly find whatever is the representation, θ or η, we use to refer to a distribution. This is the case, for example, with the Beta Distribution but can be entirely avoided in the case of Gaussian Distributions, as we will see next.

2.3.1 *Entropic Dynamics for Gaussian Distributions*

Working out the Entropic Dynamics for Gaussian Distributions we will arrive at one of the simplest forms it can take. Gaussian distributions are those that maximize the Entropy from a uniform distribution with constraints on the expected values of the first and second moments. Lets begin with the one dimensional case. Consider $x \in \mathbb{R}$, calling $F_1(x) = x$ and $F_{11}(x) = x^2$, we can impose the constraints $\mathbb{E}\left(F_1|Q\right) = \mathbb{E}\left(x|Q\right) = \mu$ and $\mathbb{E}\left(F_{11}|Q\right) = \mathbb{E}\left(x^2\middle|Q\right) = V + \mu^2$. As we saw above, the MaxEnt distribution with this constraints from a uniform distribution is

$$Q(x|\theta) = \mathrm{e}^{\psi(\theta) - \theta^1 x - \theta^{11}x^2} \tag{32}$$

17 Or, more precisely, the log evidence $\ln Z$

But the canonical notation for Gaussian distribution with mean μ and variance V is

$$Q(x|\mu V) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{V}} \tag{33}$$

and expanding the exponent

$$Q(x|\mu V) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{V}} \tag{34}$$

$$= e^{-\frac{1}{2}\ln 2\pi V - \frac{\mu^2}{2V} + \frac{\mu}{V}x - \frac{1}{2V}x^2} \tag{35}$$

we can identify with (33) if

$$\theta^1 = -\frac{\mu}{V} \tag{36}$$

$$\theta^{11} = \frac{1}{2V} \tag{37}$$

$$\psi(\theta) = -\frac{1}{2}\ln 2\pi V - \frac{\mu^2}{2V} \tag{38}$$

This identification is enough to guaranty the general Entropic Dynamics, equation (31), to be valid for the Gaussian distribution. However, Gaussians have additional properties that can be used to remove the crossed dependency between the θs and μ and V. First, notice that the equations (36) and (37) can be easily inverted to give μ and V in terms of θs

$$V = \frac{1}{2\theta^{11}} \tag{39}$$

$$\mu = -\frac{\theta^1}{2\theta^{11}} \tag{40}$$

and we can use this to replace the derivatives relative to θs

$$\frac{\partial}{\partial\theta^1} = \frac{\partial\mu}{\partial\theta^1}\frac{\partial}{\partial\mu} + \frac{\partial V}{\partial\theta^1}\frac{\partial}{\partial V}$$

$$= -V\frac{\partial}{\partial\mu} \tag{41}$$

$$\frac{\partial}{\partial\theta^{11}} = \frac{\partial\mu}{\partial\theta^{11}}\frac{\partial}{\partial\mu} + \frac{\partial V}{\partial\theta^{11}}\frac{\partial}{\partial V}$$

$$= -2\mu V\frac{\partial}{\partial\mu} - 2V^2\frac{\partial}{\partial V} \tag{42}$$

This replacements completely remove all the references to the θs and allow us to work only with the expected values, and the Entropic dynamics now have the form

$$\mu' = \mu + V\frac{\partial}{\partial\mu}\ln Z \tag{43}$$

$$V' + \mu'^2 = V + \mu^2 + \left[2\mu V\frac{\partial}{\partial\mu} + 2V^2\frac{\partial}{\partial V}\right]\ln Z \tag{44}$$

$$\Rightarrow V' = V - \left[V\frac{\partial}{\partial\mu}\ln Z\right]^2 + 2V^2\frac{\partial}{\partial V}\ln Z \tag{45}$$

The other property of Gaussians will replace the derivative relative to V for a second derivative relative to μ, by simply showing their equivalence, except by a constant factor, when applied to the distribution. The first two derivatives of $Q(x|\mu V)$ relative to μ are

$$\frac{\partial}{\partial\mu}Q(x|\mu V) = \left[\frac{x-\mu}{V}\right] Q(x|\mu V) \tag{46}$$

$$\frac{\partial^2}{\partial\mu^2}Q(x|\mu V) = \left[\left[\frac{x-\mu}{V}\right]^2 - \frac{1}{V}\right] Q(x|\mu V) \tag{47}$$

and its derivative relative to V

$$\frac{\partial}{\partial V}Q(x|\mu V) = \frac{1}{2}\left[\left[\frac{x-\mu}{V}\right]^2 - \frac{1}{V}\right] Q(x|\mu V) \tag{48}$$

$$= \frac{1}{2}\frac{\partial^2}{\partial\mu^2}Q(x|\mu V) \tag{49}$$

Now, this result applies only to derivatives of Q, so we need to be careful when taking the derivative of $\ln Z$

$$\frac{\partial}{\partial V}\ln Z = \frac{1}{Z}\frac{\partial Z}{\partial V} \tag{50}$$

$$= \frac{1}{Z}\frac{\partial}{\partial V}\int \mathrm{d}x\, L(y|x)Q(x|\mu V) \tag{51}$$

$$= \frac{1}{Z}\int \mathrm{d}x\, L(y|x)\,\frac{\partial}{\partial V}Q(x|\mu V) \tag{52}$$

$$= \frac{1}{Z}\int \mathrm{d}x\, L(y|x)\frac{1}{2}\frac{\partial^2}{\partial\mu^2}Q(x|\mu V) \tag{53}$$

$$= \frac{1}{2}\frac{1}{Z}\frac{\partial^2}{\partial\mu^2}\int \mathrm{d}x\, L(y|x)Q(x|\mu V) \tag{54}$$

$$= \frac{1}{2}\frac{1}{Z}\frac{\partial^2 Z}{\partial\mu^2} \tag{55}$$

With this results we can rewrite equations (43) and (45) in their final form

$$\mu' = \mu + V\frac{\partial}{\partial\mu}\ln Z \tag{56}$$

$$V' = V - \left[V\frac{1}{Z}\frac{\partial Z}{\partial\mu}\right]^2 + V^2\frac{1}{Z}\frac{\partial^2 Z}{\partial\mu^2} \tag{57}$$

$$= V + V^2\frac{\partial^2}{\partial\mu^2}\ln Z \tag{58}$$

The equations (56) and (58) are the Entropic Dynamics of inference for one-dimensional Gaussian distributions under a very weak hypothesis, the independence of the likelihood on distribution parameters, i. e. $L(y|x\theta) = L(y|x)$.

For the multivariate Gaussian distribution, consider the generating functions $F_i(\mathbf{w}^\star) = w_i^\star$ and $F_{ij}(\mathbf{w}^\star) = w_i^\star w_j^\star$ with constraints

$\mathbb{E}(F_i|Q) = \mathbb{E}(w_i^\star|Q) = w_i$ and $\mathbb{E}\left(F_{ij}\middle|Q\right) = \mathbb{E}\left(w_i^\star w_j^\star\middle|Q\right) = C_{ij} + w_i w_j$. The MaxEnt distribution under this constraint is

$$Q(\mathbf{w}^\star|\mathbf{w}C) = \mathrm{e}^{\psi(\theta) - \theta^i x_i - \theta^{ij} w_i^\star w_j^\star}$$

which can be put in the canonical notation

$$Q(\mathbf{w}^\star|\mathbf{w}C) = \det(2\pi C)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{w}^\star - \mathbf{w})^T C^{-1} (\mathbf{w}^\star - \mathbf{w})}$$

if we identify the constraints as $\theta^i = -C^{-1}_{ij} w_j$, $\theta^{ij} = -\frac{1}{2} C^{-1}_{ij}$ and $\psi(\theta) = -\frac{1}{2}\ln\det(2\pi C) - \frac{1}{2} C^{-1}_{ij} w_i w_j$. The remaining steps to reach the Entropic Dynamics equations are analogous to the 1-dimensional case, leading us to

$$\mathrm{w}_i' = \mathrm{w}_i + C_{ij} \frac{\partial}{\partial \mathrm{w}_j} \ln Z \tag{59}$$

$$C_{ij}' = C_{ij} + C_{ik} C_{lj} \frac{\partial^2}{\partial \mathrm{w}_k \partial \mathrm{w}_l} \ln Z \tag{60}$$

or in vector notation

$$\mathbf{w}' = \mathbf{w} + C \frac{\partial}{\partial \mathbf{w}} \ln Z \tag{61}$$

$$C' = C + C \frac{\partial^2 \ln Z}{\partial \mathbf{w} \partial \mathbf{w}^T} C \tag{62}$$

This set of equations form the *Entropic Learning Dynamics* (ELD) within the Gaussian Family, where parameters of the distribution representing the accumulated knowledge in evolve as a gradient ascend in the log Evidence. The Evidence Z, or more precisely $\mathcal{E} = -\ln Z$, acts as a kind of energy driving the parameters towards the regions in the parametric space where inference is the best possible given for the available data and constraints.

One useful consequence of the ELD equations (61) and (62), or its 1-dimensional analogues, is that we can plug as many of them as we want in a model and as long as they are pair wise independent we get the same set of equations. The only difference being in the Likelihood and Evidence. Suppose, for instance, we are modeling a system with a vector parameter $\mathbf{w}^\star$ and a scalar parameter x with a certain Likelihood $L(y|\mathbf{w}^\star x)$ we find appropriate. Choosing Gaussian distributions $G(\mathbf{w}^\star|\mathbf{w}C)$ and $G(x|\mu V)$ where $\mathbf{w}, C, \mu$ and V are unrelated leads us to a distribution $Q(\mathbf{w}^\star, x|\mathbf{w}CxV) = G(\mathbf{w}^\star|\mathbf{w}C)G(x|\mu V)$, an Evidence $Z(y|\mathbf{w}C\mu V) = \mathbb{E}(L|Q)$ and the set of ELD equations

$$\mathbf{w}' = \mathbf{w} + C \frac{\partial \ln Z}{\partial \mathbf{w}} \tag{63}$$

$$C' = C + C \frac{\partial^2 \ln Z}{\partial \mathbf{w} \partial \mathbf{w}^T} C \tag{64}$$

$$\mu' = \mu + V \frac{\partial \ln Z}{\partial \mu} \tag{65}$$

$$V' = V + V^2 \frac{\partial^2 \ln Z}{\partial \mu^2} \tag{66}$$

To completely specify the ELD for a model all we need is to fix the functional form of the Likelihood L, or the Evidence Z or even the energy $\mathcal{E} = -\ln Z$, allowing us to think of inference problems in terms of energy with familiar interpretations. This result is general for any model within the Gaussian Family, as long as L only depends on the underlying variable $\mathbf{w}^\star$ and data y but no on the parameters (μ and V or $\mathbf{w}$ and $\mathbf{C}$). The specification of the Likelihood is where we capture the Physics of whatever phenomenon we are studying and, in our case will be done in Chapter 3 for the agents behavior.

3

AN AGENT MODEL OF OPINION AND DISTRUST

In this chapter we introduce our main contribution, an agent model to study community formation, based on *opinion exchange* and *trust-distrust relations* between agents. We extend the approach of previous works from Nestor Caticha and other[3, 13, 14, 44, 46], in particular from [2], with the addition of a distrust drive, giving the agents the ability to sustain disagreement without cutting relationships with others. Agents are modeled as simple *Neural Networks* representing their behaviors and evolve through exchanging information among them, through a learning dynamics derived from *probabilistic* and *entropic* inference principles. They interact in a way that resembles a casual conversation: an agent exposes its opinion about a given issue to another agent, who learns from the combination of the received opinion and the trust or distrust it attributes to the emitting agent.

This chapter is focused on the interpretation and results of the model as a given consequence of Entropic Learning Dynamics, please refer to chapter 2 to see the development of the dynamics.

We start by motivating our approach for the representation of behavior drives as simple neural networks based on studies in Social Psychology, namely the Moral Foundation Theory (MFT). Then, we proceed to establish the mathematical description and physical interpretation of opinion and distrust drives, how they are related within an agent cognition and how they evolve when the agent is supplied with new information.

Once the mathematical description is introduced, we'll study the interaction of agents in many situations within the model's freedom, starting by analyzing the interaction of few agents and going to bigger groups.

3.1 THE AGENTS BEHAVIORS AND ARCHITECTURE

We are interested in studying typical human interaction situations, where a person exposes his or her opinions about particular issues and how they change how they think based on the exposed opinions of others. A very important character of human interaction is the fact that we need to deal with *incomplete information* about most of the relevant aspects involved. In a discussion, for instance, where people talk about an issue and "state their mind" about it, they don't really expose the whole process to reach that opinion, the communication of the subject and ideas involved require language constructs that may limit one or other important aspect of the issue or opinion and

so many other details are hidden behind a huge complexity reduction mechanism we have builtin in our biology and culture to enable efficient communication. Nonetheless, communication is, indeed, efficient and effective in enough situations for the viability of society.

Thus, in order to develop a physical model for the process of communication, opinion exchanges and it consequences we need to find a mathematical structure with the ability to adapt from limited information, easily extensible to incorporate the interaction of different subjects, information sources and situations, or in other words of universal applicability, and, hopefully, simple enough to still be under our grasp to analyze.

Most of these requirements are met by a combination of *Probability Theory* and *Machine Learning* (ML) on *Neural Networks* (ANNs)[1], and applied together with some insights from Moral and Social Psychology,[2] this challenging quest can be amended.

The *Moral Foundation Theory* (MFT) is an effort to understand what humans consider *Moral* in a extensive way, meaning regardless of cultural and historical backgrounds. The *Moral Foundations Questionnaire* was surveyed across the world and answered by all sorts of people, which enabled the researcher to pin 5 or 6 *Foundations* humans base their moral judgments on. Their conclusion entails that any situation that purports a moral problem, say the Trolley Problem for instance, is *"parsed"* within the human mind as set of projections on each of this Foundations. Also from MFT, we have a prescription for the role of Morality as mainly a mechanism for social binding and that it precedes Reason in the task of making judgments, what they called *Intuition Primacy*.

3.1.1 *A mathematical representation for Opinion*

Morality is an example of subject that every human being capable of understanding it has an opinion about, and the studies described above help us to visualize a description of how we can model *opinions*. We start with the assumption that *issues* within the category of subjects discussed by the agents, can be represented as a vector $\mathbf{x}$ in a K–dimensional vector space $\mathbb{R}^K$. Here, K reflects the number of "Foundations" in this subject category, in analogy with MFT.

Looking again to our example kinds of subject, one's stance about Morality or Politics have a binary aspect, meaning one "is pro/against a certain Government action", "think is right/wrong an certain action in a moral conundrum", for instance. Clearly, this is a very blunt simplification of such topics and since we are not modeling how experts deal with these subjects but rather how people deal with this topics in conversations, we can say that this binary nature

1 See Chapter 2 for the theoretical development

2 See 23, 24, for the developements of the Moral Foundation Theory.

hinges on *Intuition Primacy*. An alternative way is to claim that *Binary Opinions* are the simplest way to deal with any subject that allows binary stances. Its rather difficult to be "pro/against $1+1=2$" once the axioms are set, but its rather easy to have this kind of opinion on matters of human life where there are no axioms to guide us.

To capture this binary nature of *Opinions*, we represent the agent's opinion machinery by the simplest neural network for classification, the *binary classifier* or *Perceptron*. A perceptron for issues in K–dimensional spaces is a hyperplane of $K-1$ dimensions, which can be easily represented by a vector $\mathbf{w} \in \mathbb{R}^K$ orthogonal to the hyperplane. The perceptron, then, is a bisection of the issue space, classifying issues on each side according to their projection on the perceptron weight vector $\sigma \equiv \text{sign}(h\sigma) = \text{sign}(\mathbf{w} \cdot \mathbf{x})$, much in the spirit of how Moral issues resonate with people according the its projections on the Moral Foundations.

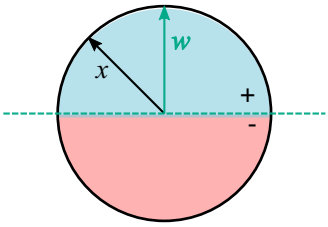


Figure 2: Illustration of a Perceptron with weight vector $\mathbf{w}$ classifying an issue $\mathbf{x}$.

We give each agent i in a society a perceptron weight vector $\mathbf{w}_i$ and call $\sigma_i = \text{sign}(h_i\sigma) = \text{sign}(\mathbf{w}_i \cdot \mathbf{x})$ the opinion of agent i about $\mathbf{x}$ and $\mathbf{w}_i$ is the opinion weight vector of agent i representing the weight it gives to each foundation in the subject category. We call $h_i = \mathbf{w}_i \cdot \mathbf{x}$ agent's i *internal response*[3] for $\mathbf{x}$.

This assumptions allow us to represent agents as simple *Artifitial Neural Networks* (ANN) that classify issues and make inference about agents behavior using *Probability Theory*. At this point we are ready to use ML give the our agents the ability to evolve from received information, except that in a typical learning scenario there is a special "agent", the *Teacher*, who provides *reliable*, examples so the *Student* agent can learn. It should be clear that we are not learning a classification task, but modeling human interaction, for which there is no teacher but ourselves. With that in mind, each agent would take each role, sometimes exposing its opinion about an issue, or teaching, sometimes learning from others opinions, or studying.[4]

3 Also called *stability* or *internal repesentation* in Machine Learning

4 See 7–9, 18, 28, 29, 31, 32, 34, 39, 45, 47, for the theory and development of ML algorithms.

To use the Entropic Learning Dynamics (ELD) equations (63) and (64), we need to model the *Likelihood* L of a receiver agent r getting an opinion σ about an issue $\mathbf{x}$ from an emitting agent e, if we knew its weight vector $\mathbf{w}_e$ and a model distribution Q for $\mathbf{w}_e$. Intuitively, a paired issue $\mathbf{x}$ and opinion σ should be more likely if $\sigma = \text{sign}(\mathbf{w}_e \cdot \mathbf{x})$, so consider the Likelihood:

$$\begin{aligned} L_w(\mathbf{x}, \sigma|\mathbf{w}_e) &= L(\sigma|\mathbf{w}_e\mathbf{x})L(\mathbf{x}|\mathbf{w}_e) \\ &= \int \mathrm{d}\mathbf{x}_e \, G(\mathbf{x}_e|\mathbf{x}, v_e)\Theta(\mathbf{w}_e \cdot \mathbf{x}_e\sigma) \\ &= \Phi\left(\frac{\mathbf{w}_e \cdot \mathbf{x}\sigma}{\sqrt{v_e\|\mathbf{w}_e\|^2}}\right) \\ &= \Phi(\mathbf{w}_e \cdot \mathbf{x}\sigma) \end{aligned}$$

where we accounted for "parsing errors" on issue $\mathbf{x}$ through additive white noise is variance $v_e = \|\mathbf{w}_e\|^{-2}$, which can be interpreted as errors in communicating the issue duo to inexperience on this subject category. If we choose a Gaussian distribution $Q = G(\mathbf{w}_e|\mathbf{w}_r C_r)$, where $\mathbf{w}_r$ is the receiver agent weight vector and C_r its covariance matrix, we get the Evidence

$$\begin{aligned} Z_w(\mathbf{x}, \sigma_e|\mathbf{w}_r C_r) &= \mathbb{E}\,(L|G) \\ &= \int \mathrm{d}\mathbf{w}_e \, G(\mathbf{w}_e|\mathbf{w}_r C_r)L(\mathbf{x}, \sigma_e|\mathbf{w}_e) \\ &= \Phi\left(\gamma_r^{-1}\mathbf{w}_r \cdot \mathbf{x}\sigma_e\right) = \Phi(\tilde{h}_r\sigma_e) \end{aligned} \tag{67}$$

The variable C_r has the role of uncertainty for agent's i opinion weights, while $\gamma \equiv \sqrt{1 + \mathbf{x} \cdot C_r\mathbf{x}}$ is an uncertainty scale for the agent's r internal response h_r to the received opinion on an issue. We call $\tilde{h}_r \equiv \mathbf{w}_r \cdot \tilde{\mathbf{x}} = \gamma^{-1}h_r$ the *effective internal response* of agent r to $\mathbf{x}$.

Plugging this evidence Z_w on the ELD equation with pairs of issue and opinions generated by other agents in the society would already gives us a model for opinion dynamics. This first implementations of neural network agents for opinion dynamics was published by Caticha and Vicente[13, 14, 44], showing that this agent model can *only* generate social disorder or, under certain circumstances, consensus, with the exception of [14] which relies on the communication topology to enable community formation.

Before we put the agents to interact we focus on the *reliable* part we mentioned above, as it hooks the motivation for our approach on modeling community formation.

3.1.2 *A mathematical representation for Distrust*

A major struggle for the scientific community in the field of Social and Opinion Dynamics is to find models of community formation.

The challenge arises from the incompatibility between agents abilities to learn by imitation and to choose those how are similar to learn from. So far, most of the models for social and opinion dynamics either relied on topology, typically regular lattices with very narrow information bottlenecks and with no resemblance with social networks topologies observed in real world, or on specific class tags to enable community formation and thus starting the model with communities already.[5]

Our approach to the community formation problem is to introduce a behavior that runs "in parallel" with the opinion, the ability to distrust other agents and to adapt their distrust based on their interactions.

It is not hard to imagine people exposing non reliable opinions, duo to honest mistakes or intentional lies, neither is hard to imagine people protecting themselves from this possibility. To introduce *Distrust* in our model, we give agents the ability to doubt the opinions they receive based on who exposed it. Consider a receiver agent r attributes a distrust value of $z_{e|r} \in \mathbb{R}$ for an emitting agent e from which it computes the probability $\varepsilon_{e|r} = \Phi(z_{e|r})$ for the opinion σ_e about an issue $\mathbf{x}$ being wrong, i.e. having its sign flipped. For the sake of readability, we will drop the indices on the receiver agent variables, but it should be clear that $\mathbf{w}, C, \mu$ and V (the last two yet to be introduced) refers to the receiver agent.

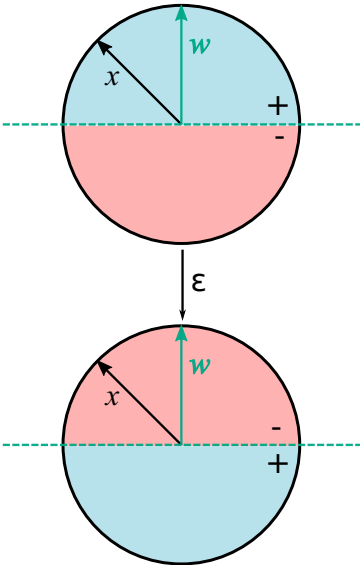


Figure 3: Illustration of a Perceptron with weight vector $\mathbf{w}$ classifying an issue $\mathbf{x}$ with distrust ε

The complete ELD requires a Likelihhod of r receiving the pair $(\mathbf{x}, \sigma_e)$ and a distribution $Q(\mathbf{w}_e, \varepsilon)$ to use ELD equations (63), (64), (65) and (66). Consider the Likelihood to be:

5 See 6, 19, 37, 41, for other approaches to Opinion Dynamics and Community formation.

$$\begin{aligned} L(\mathbf{x}, \sigma_e|\mathbf{w}_e\varepsilon_{e|r}) &= \varepsilon_{e|r} L_w(\mathbf{x}, -\sigma_e|\mathbf{w}_e) + (1-\varepsilon_{e|r}) L_w(\mathbf{x}, \sigma_e|\mathbf{w}_e) \\ &= \varepsilon_{e|r}\Phi(\mathbf{w}_e \cdot \mathbf{x}(-\sigma_e)) + (1-\varepsilon_{e|r})\Phi(\mathbf{w}_e \cdot \mathbf{x}\sigma_e) \\ &= \Phi(z_{e|r})(1-\Phi(\mathbf{w}_e \cdot \mathbf{x}\sigma_e)) + (1-\Phi(z_{e|r}))\Phi(\mathbf{w}_e \cdot \mathbf{x}\sigma_e) \\ &= \Phi(z_{e|r}) + \Phi(h_e\sigma_e) - 2\Phi(z_{e|r})\Phi(h_e\sigma_e) \end{aligned}$$

To conclude, we choose Q to be a product of independent Gaussian distributions for $\mathbf{w}_e$ and $z_{e|r}$, as $Q = Q_w Q_z = G(\mathbf{w}_e|\mathbf{w}C)G(z_{e|r}|\mu V)$, leading us to the Evidence

$$\begin{aligned} Z(\mathbf{x}, \sigma|\mathbf{w}C\mu V) &= \mathbb{E}\left(L|Q_w Q_z\right) \\ &= \Phi(\tilde{h}\sigma_e) + \Phi(\tilde{\mu}) - 2\Phi(\tilde{h}\sigma_e)\Phi(\tilde{\mu}) \end{aligned} \tag{68}$$

where we called $\tilde{\mu} \equiv \lambda^{-1}\mu$ the *effective distrust* agent r attributes to agent e and $\lambda = \sqrt{1+V}$ is the uncertainty scale for agent's r internal response to agent e.

This choice of L, Q and the resulting Evidence gives us a set of equations for the evolution of an agent r state given once it receives an opinion from equations. We proceed to analyze how this evolution occurs, first by looking at many details of the equations, then looking at the patterns that appear when we let agents interact.

3.1.3 *The Agent Interaction Dynamics*

With the Evidence Z from (68) in hands, we can explicitly state the evolution equations for an agent's state when it receives the opinion from other:

$$\mathbf{w}' = \mathbf{w} + C\frac{\partial \ln Z}{\partial \mathbf{w}} = \mathbf{w} + F_w C\frac{\mathbf{x}\sigma}{\gamma} = \mathbf{w} + \Delta\mathbf{w} \tag{69}$$

$$C' = C + C\frac{\partial^2 \ln Z}{\partial \mathbf{w}\partial \mathbf{w}^T}C = C + F_C C\frac{\mathbf{x}\mathbf{x}^T}{\gamma^2}C = C + \Delta C \tag{70}$$

$$\mu' = \mu + V\frac{\partial \ln Z}{\partial \mu} = \mu + F_\mu\frac{V}{\lambda} = \mu + \Delta\mu \tag{71}$$

$$V' = V + V^2\frac{\partial^2 \ln Z}{\partial \mu^2} = V + F_V\frac{V^2}{\lambda^2} = V + \Delta V \tag{72}$$

where we define

$$F_w = \frac{\partial \ln Z}{\partial(\tilde{h}\sigma)} = \frac{1-\Phi(\tilde{\mu})}{Z}G(\tilde{h}\sigma) \tag{73}$$

$$F_C = \frac{\partial^2 \ln Z}{\partial(\tilde{h}\sigma)^2} = -F_w[F_w + \tilde{h}\sigma] \tag{74}$$

$$F_\mu = \frac{\partial \ln Z}{\partial \tilde{\mu}} = \frac{1-2\Phi(\tilde{h}\sigma)}{Z}G(\tilde{\mu}) \tag{75}$$

$$F_V = \frac{\partial^2 \ln Z}{\partial \tilde{\mu}^2} = -F_\mu[F_\mu + \tilde{\mu}] \tag{76}$$

These equations are the result of plugging our Evidence (68) into the ELD equations for distributions in the Gaussian Family (63),(64),(65),(66) and they tells us that the evolution of an agent's state occurs in the natural gradient of the Evidence under the state variables.

We call the functions F, resulting from the derivatives of $\ln Z$, *Modulation Functions*, a rather cryptic name meaning they modulate the amplitude of changes in a neural network for a given example. A much more interesting name for the functions F would *Surprise Functions* of *Blaming Functions* because, as will become clear, they serve the purpose of indicating surprising situations and blaming which behavior, Opinion or Distrust, is responsible for such surprise.

In order to understand the Modulation functions, lets take a moment to understand what the effective internal reactions tells us about the situation faced by the receiver agent. Since the opinion $\sigma_i = \text{sign}(h_i)$ of an agent about an issue is the sign of its internal reaction $h_i = \mathbf{w}_i \cdot \mathbf{x}$, and $\tilde{h}_i$ is just a scaling transformation that preserves the opinion, the product of $\tilde{h}_r \sigma_e = \tilde{h}\sigma > 0$ when agents e and r agree on the given issue, and $\tilde{h}\sigma < 0$ when they disagree. On the other hand, agent r attributes a probability $\varepsilon_{e|r} = \Phi(\tilde{\mu}_{e|r}) = \Phi(\tilde{\mu})$ of agent's e being wrong, either by mistake of lie, therefore when $\tilde{\mu} > 0$ agent r believes that agent e is wrong with probability higher than 0.5 so r distrusts e, and when $\tilde{\mu} < 0$ the situation is reversed and r trusts e.

The Modulation functions *depend on the agreement and distrust scales*, and only have non negligible values in surprising situations: when r trusts e and they disagree or when r distrusts e and they agree, as we can see in Figure 4.

Also, the modulation functions control how much its associated behavior should change according to the magnitude of effective internal reactions. In Figure 5, we can see how the modulation function grows in each agreement situation as we change the distrust. Its clear in each scenario that the $|F_\mu|$ grows until it reaches the maximum around $|\tilde{h}\sigma| \approx |\tilde{\mu}|$ and then falls again. Since F_μ gives the size of chance in μ, we can attribute blame by region: when there is agreement and trust ($\tilde{h}\sigma > 0$ and $\tilde{\mu} < 0$) there is no surprises and therefore the $F_\mu \approx 0$; when there is disagreement and trust ($\tilde{h}\sigma < 0$ and $\tilde{\mu} < 0$) F_μ is higher for most of the values of $|\tilde{\mu}| < |\tilde{h}\sigma|$ and smaller for most of the values of $|\tilde{\mu}| > |\tilde{h}\sigma|$. This situation is analogous for distrust scenarios and for the behavior of F_w.

In other words, an agents changes its opinion when it disagrees with an agent it trusts if it is more sure about the attributed trust than about its opinion, or vice-versa. What this aspects of the Modulation functions tells us may seem a bit obvious, but that exactly what we want. It gives some confidence that the assumptions we made to develop the ELD equations did not lead us to absurd consequences and that reasoning about how we deal with information still holds, so far, as good approach to understand how we form groups.

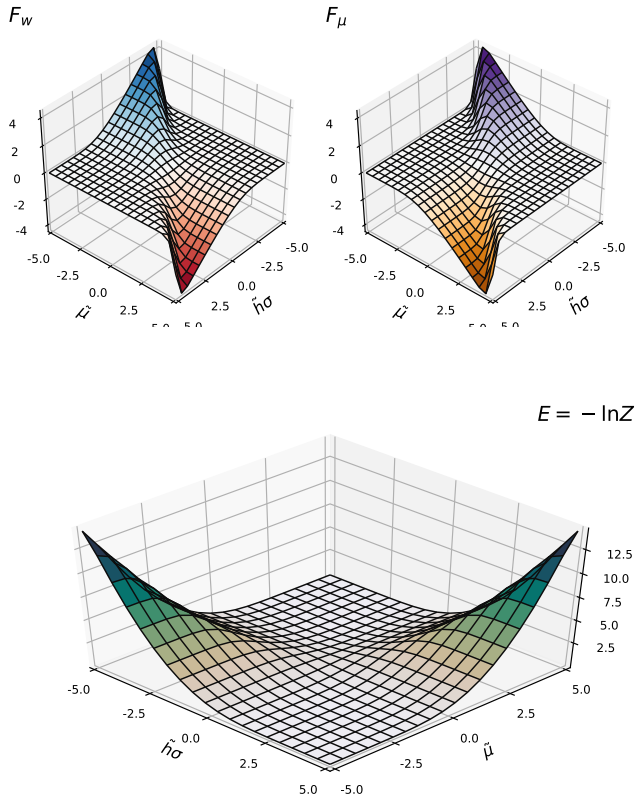


Figure 4: Surface plot for F_w, F_μ (top) and $E = \ln Z$ (bottom) as functions of the effective internal reactions $\tilde{h}\sigma$ and $\tilde{\mu}$. Notice how the modulation is only relevant on surprising situations of agreement with distrust or disagreement with trust and the complementary behavior on the modulation functions, where the biggest surprise depends on which behavior is more certain about the situation.

For the roles of the uncertainties in the Modulation functions and in the Δs for the agent state, remember they depend on $\gamma = \sqrt{1 + \mathbf{x} \cdot C\mathbf{x}}$ and $\lambda = \sqrt{1 + V}$, which satisfy $\gamma \geq 1$ and $\lambda \geq 1$ and only appear in the denominator so they either scale down the internal reactions and Δs or leave them untouched. This implies that for higher values of [6] C or V the internal reactions, $h\sigma$ and μ, can be brought back to the surprise range of the Modulation functions, while smaller values will not help in that[7] The interpretation for this effect is simple: behaviors with low uncertainties are harder to change in comparison to behaviors with high uncertainties and, therefore, agents who are very sure of their assessments are harder to convince than agents with doubts.

From the Modulation functions perspective, the uncertainties regulate the relevance of *Corroborative Information*, when there is agree-

6 By a "large value" matrix variable we mean a matrix with large norm.

7 Although changing the Likelihood and Evidence may change this aspect. For instance, had we not accounted for parsing errors in the Z_w we would have $\gamma = \sqrt{x \cdot Cx}$, which take the opinion internal reaction far away from the surprise range when C is close to zero.

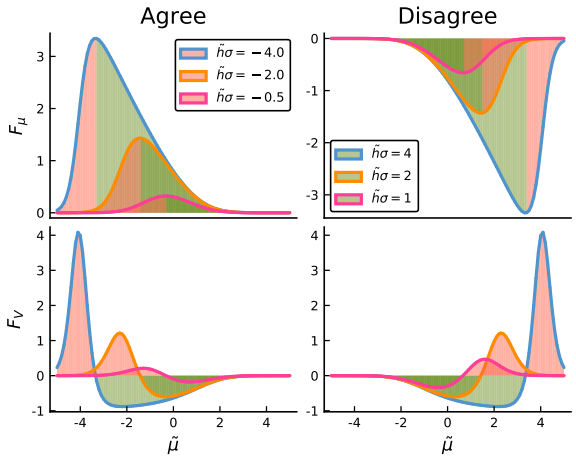


Figure 5: Modulation functions F_μ (top) and F_V (bottom) for agent's distrust for a few different values of agreement (left) and disagreement (right). The shaded areas indicate the blame attribution for the surprise: green areas blame $\tilde{\mu}$, red areas are the transition in blame attribution from $\tilde{\mu}$ to $\tilde{h}\sigma$ and the areas with no shading correspond either to corroboration, when there is no surprise, or to blame $\tilde{h}\sigma$.

ment and trust or disagreement and distrust, relative to the relevance of *Novelty Information*, when there is disagreement and trust or agreement and distrust, an agent receives. Higher values of uncertainties lead to similar relevance between Corroborative and Novelty information, while lower values of uncertainty lead to Novelty predominance. This aspect of the Modulation functions is illustrated in Figure 6.

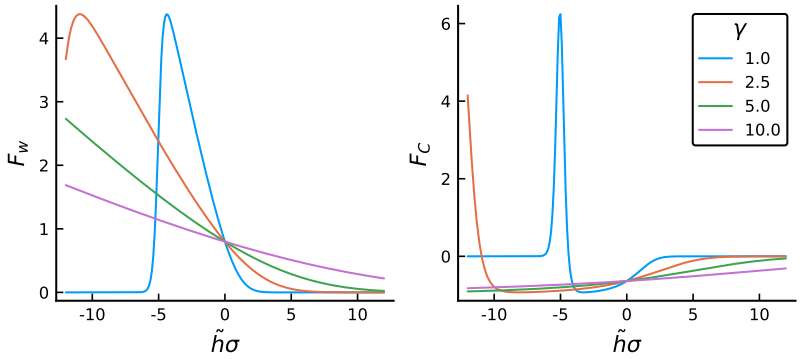


Figure 6: Modulation Functions for $\mathbf{w}$ and C for fixed $\tilde{\mu} = -5$ and some values of γ

As a last remark, the norm of an issue $\mathbf{x}$ also plays a role in the Modulation functions, somewhat related to the role of C. Because $\gamma \geq 1$, the norm of the effective issue is always smaller than or equal to the norm of issue, i. e. $\|\tilde{\mathbf{x}}\| = \gamma^{-1}\|\mathbf{x}\| \leq \|\mathbf{x}\|$. Let $C = c\mathbb{1}$ and rewrite $\tilde{\mathbf{x}}$ as

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\sqrt{1+\mathbf{x}\cdot C\mathbf{x}}} = \frac{\mathbf{x}}{\sqrt{1+c\mathbf{x}\cdot\mathbf{x}}} = \frac{\mathbf{x}}{\sqrt{1+c\|\mathbf{x}\|}} = \frac{\hat{\mathbf{x}}}{\sqrt{\|\mathbf{x}\|^{-2}+c}}$$

and we see that $\|\tilde{\mathbf{x}}\| \to \|\mathbf{x}\|$ when $c \to 0$ and $\|\tilde{\mathbf{x}}\| \to 0$ when $c \to \infty$, while $\|\tilde{\mathbf{x}}\| \to c^{-1/2}$ when $\|\mathbf{x}\| \to \infty$ and $\|\tilde{\mathbf{x}}\| \to 0$ when $\|\mathbf{x}\| \to 0$. This shows us that $\|\mathbf{x}\|^{-2}$ plays the same role as C, and since the issues are not under the ELD we need to be careful when choosing the issue vector $\mathbf{x}$. There are a few possible interpretations for this, the most appealing one is considering the norm of the issue as an *Emphatic weight* given by the emitter agent when expressing its opinion which may take the issue out of the surprise range for the receiving agent, triggering a corroborative rather than novelty information. In our simulations, unless explicitly stated otherwise, we use choose $\|\mathbf{x}\| = 1$ for all issues, which fix the scale for h and C to trigger the surprise range for opinion.

In the next section we will start putting agents to interact and see how what are the social consequences of the agents dynamics.

3.2 A SOCIETY OF AGENTS

In this section we will look at the evolution of the agents state variables when they interact in a variety of conditions. We begin by reviewing our notation, defining the *Social Network Topology*, the *Subject Category Universe* and some macroscopic observable variables, then we analyze the evolution of only two agents interacting and conclude by analyzing bigger societies.

3.2.1 *Agents states, Social Network and Subject Category*

We consider a *Society* with N agents a_i, $i \in \{1\ldots,N\}$, and give each agent the ability to form opinions about issues and to distrust other agents. We call $a_i = (\mathbf{w}_i, C_i, \mu_{j|i}, V_{j|i})$ the state of agent i, where $\mathbf{w}_i \in \mathbb{R}^K$ is its opinion weight vector, $C_i \in \mathbb{R}^{K\times K}$ its opinion uncertainty, $\mu_{j|i} \in \mathbb{R}$ and $V_{j|i} > 0$ for $j \in \{1,\ldots,N\}$ are the distrust and the corresponding uncertainty agents i attributes for each other agent j. The Society dynamics is a sequence of "conversations" between pairs of agents, as one agent j states its opinion about an $\mathbf{x}$ to agent i, who updates its state according to the ELD equations. Two questions to ask about the Society dynamics are "who interacts with whom?" and "what issues do they talk about?", both outside our theoretical framework and deserving some attention now.

To decide "who interacts with whom?" is equivalent to choose a *Social Network Topology*, an interesting and difficult subject and un-

der heavy study in the recent years.[8] These studies indicate a few common patterns of social interaction, but some important details depend on the particular to the context where the interactions take place. Also, such topologies hinders, if not make impossible, any analytical progress for models on top of it. A way out of these complications is to use a regular lattice for the Social Network, an approach followed by many of the classical agent based models of social behavior[9] and carrying its own drawback: the introduction of artificial information bottlenecks in social interactions. As a consequence, the interesting macroscopic behaviors of many models based on regular lattices are dependent on a narrow information bottleneck, i.e. they are not *robust* on the Social Network topology and loose the phenomena when the number of social neighbors increase to values closer to the ones suggested by real topologies. That said, in [13] they studied a Neural Network agent based model for opinion on top of *Barabasi-Albert* topology with great success.

Our way out of this conundrum is to choose the simplest regular lattice as Social Network topology, the *fully connected graph* with N agents, since it introduces no bottlenecks and the macroscopic effects it exhibits, if any, can only intensify or delayed by approaching a real topology but won't vanish. We already followed this path in [3], where we introduced a mechanism to implement the appearance of effective communications barriers, using a different concept of distrust (actually more like a expected cognitive reward). In practical terms, this choice of topology implies that at any given moment any two agents in the society can interact.

For the "what issues do they talk about?" we introduce the concept of *Subject Category Universe*, which is a set of issues belonging to the same subject category, say Morality or Politics for instance. The properties of this set will have an effect on the social states and we will take a look at some of these effects. Particularly interesting is the possibility of spin-glass like states to appear in a society when a the subject category universe is large. However, further investigation on the relation between how society is affect by what people talk about and how people choose what to talk is, to the authors knowledge, lacking in the research community and beyond our current scope. That said, a quote from Abraham Lincoln gives some inspiration

> The process is this: Three, four or half a dozen questions are prominent at a given time, the party selects its candidate, and he takes his position on each of these questions.
>
> — Abraham Lincoln,
> Speech in the U.S. House of Representatives, July 27, 1848.

8 See 1, 33, for an overview of Social Network Topologies and Graph Theory.
9 See 6, 19, 35, 41, for other approach to opinion dynamics.

With that in mind, by the principle of insufficient reason, we fix a set of P issues $\mathbf{x}_p$, such that $\|\mathbf{x}_p\| = 1$ for all $p \in \{1, \dots, P\}$, chosen uniformly on the unit sphere on $\mathbb{R}^K$. This gives us control over the number of issues the agents can discuss and allow us to choose what "questions are prominent at a given time" for that Subject Category.

3.2.2 *Social state observables*

Our objective to study how community formation in society demands us to define what is a community and how we recognize one in our model. Agents have the ability to express their opinions and to attribute distrust to other agents through their weight vectors $\mathbf{w}_i$ and $\mu_{j|i}$, so the simplest way to identify communities is to identify agents how share the same opinions and trust each other but distrust agents who differ in opinion. That is easy to accomplish in our model by looking at *Overlaps* and *Disbelief Matrices*.

The *overlap* between agents i and j is defined as $\rho_{ij} = \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|}$, and the Overlap Matrix R is the $N \times N$ matrix with entries ρ_{ij}, while the disbelief agent i has on agent j is given by $\varepsilon_{i|j} = \Phi(\mu_{j|i})$ and the Disbelief Matrix D is the matrix with entries $\varepsilon_{j|i}$. Notice that R is symmetric by definition while D is not.

Since $\varepsilon_{j|i}$ is a monotonic transformation of $\mu_{j|i}$, we can also use the distrusts directly and the *Distrust Matrix* M is the matrix with entries $\mu_{i|j}$.

Other observables of interest are calculated from overlaps and either the distrusts or disbelief values. An interesting one is the *Balance of Relations*, defined as all the agent triangles of transitive behavior. For instance, we could check if the distrust is balanced for a given social state by computing $\mu_{j|i}\mu_{k|j}\mu_{k|i}$ for all social neighbors i, j and k. This quantity measure if distrust relations are balanced by the sign of the product and how much by its magnitude: if i distrusts j and distrusts k, but j trusts k than the oriented triangle ijk is balanced, on the other hand if j distrusts k it is unbalanced. The magnitude of the product can be used to identify how much this relation is consolidated, since values of μ closer to zero carry less information about that pair of agents. The same can be done for the overlaps or the product of overlaps and distrusts, so we could check if the agents behaviors are "consistent" within that state.

The uncertainty variables C_i and $V_{j|i}$ also provide dynamical information about the system, as well as the norms of $\mathbf{w}_i$ and $\boldsymbol{\mu}_i$, if we think of $\mu_{j|i}$ as a vector in $\mathbb{R}^{N-1}$.

Although the model observables are not directly measurable in real societies, we could estimate them by performing well designed questionnaires for whatever the subject category one could want to compute the appropriate weight vectors. This idea is not different from what was done with the MFT or from "asking" each particle

in a mechanical system what its position at a given moment and computing its momentum to get the system energy.

...its just a slightly more difficult

3.2.3 *Studying a couple of agents*

When we look at the case of just a couple of agents interacting, $N = 2$, there is a lot of information of about the dynamics that we would not be able to handle in bigger societies. In this section we will take a look at two agents interacting under the following simplifying circumstances: we consider all the uncertainties $C_i = c_i\mathbb{1}$, where $c_i \in \mathbb{R}_+$, and $V_{j|i} = v_i \in \mathbb{R}_+$, i.e. we give *scalar* uncertainties for each agent, with initial equal initial values, $c1 = c2 = c(0) = 1$ and $V_{1|2} = V_{i|j} = v(0)$; we fix[10] $K = 5$ and $P = 100$; we choose the initial weight vectors $\mathbf{w}_i$ so their norm is 1 and their overlap is $\rho_{12}(0) = \rho_{21}(0) = \rho(0) = 0.5$ and $\mu_{1|2}(0) = -\mu_{2|1}(0) = \mu(0) = 0.25$; we count the number of interactions as t and $\alpha = \frac{t}{K(N-1)N} = \frac{t}{10}$ as the number of interaction per adjustable weight in $\mathbf{w}_i$ and $\mu_{j|i}$.

At a given moment t this 2 agents systems evolves following:

1. randomly choose $j \in \{1, 2\}$ as the emitter agent
2. choose $i \in \{1, 2\} \setminus \{j\}$ as the receiver agent
3. randomly choose $\mathbf{x}(t) \in \{\mathbf{x}_1, \ldots, \mathbf{x}_P\}$
4. compute $\sigma_j(t) = \text{sign}(\mathbf{w}_j \cdot \mathbf{x})$
5. update $a_i(t+1) = a_i(t) + \Delta a_i(\mathbf{x}(t), \sigma_j(t))$
6. increase $t \leftarrow t + 1$ and repeat from step 1

where $a_i(t+1) = a_i + \Delta a_i(\mathbf{x}, \sigma)$ is just a shorthand for the ELD equations (69), (70), (71) and (72)

In Figure 7 we can see how changing only the initial values of the distrust uncertainty affects the evolution of the society. The figure show that any result is possible for within a fixed window of time, or number of interactions if $v(0)$ is small enough and becomes more deterministic in on the initial conditions if its is big enough. The sizes are relative to other parameters in the model, like the initial overlap $\rho(0)$, initial distrusts $\mu(0)$, $c(0)$ and the properties of the Modulation functions F. In all figures, $\alpha = \frac{t}{KN(N-1)}$ is the *effective number of interactions*, i.e. the number of interactions per adjustable weight in the society.

If we repeat this simulation several times and look at the first sample moments of some observable, like we do for the overlap in Figure 8, we can see what to expect from a this two agents dynamics by varying the initial distrust uncertainty $v(0)$.

10 This choice was made based on the number of Foundations proposed by MFT.

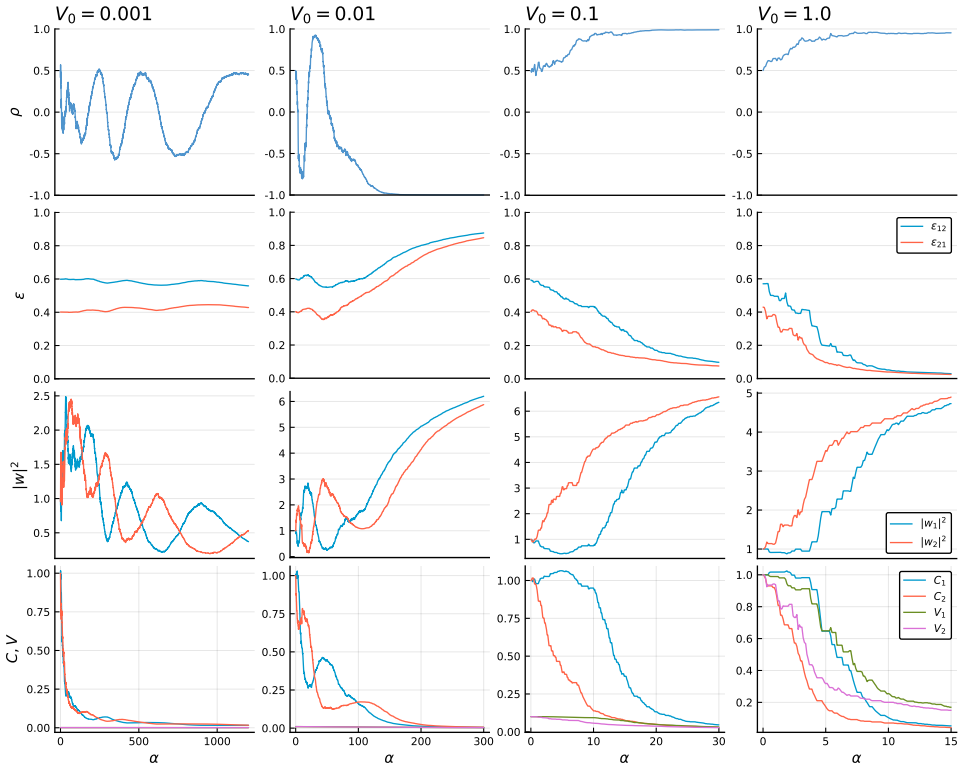


Figure 7: Evolution of the observables in 1 realization of the dynamics of two agents for different values of $v(0)$. Here, $\alpha = \frac{t}{KN(N-1)} = \frac{t}{10}$ is the number of interactions per adjustable weight. Notice the how even that as $v(0)$ grows, the transient period and the fluctuations get smaller.

What figures 7 and 8 shows is that agents starting with similar opinion ($\rho(0) > 0$) and unbalanced distrust relations ($\mu_{1|2}(0) = -\mu_{2|1}(0)$) may end up agreeing on everything and trusting each other or disagreeing on everything and distrusting each other, as long as $v(0)$ is small enough, i.e. they are quite sure of their initial attribution of distrust. We can see that, under these circumstances, for $v(0) < 0.01$ the initial values of ρ are not enough to predict the equilibrium state, but higher values of $v(0)$ lead to states that emphasize the initial conditions. The sample moments for the repeated realizations show that the transients extend for a longer period for smaller $v(0)$ and that the odd moments are the only asymptotically non zero moments when the initial conditions are enough to predict the equilibrium, and the even moments are the only asymptotically non zero when they are not. Also, all the moments fluctuate around zero during the transient period.

We can run simulations like this for other situations, many of which will present a similar statistical profile.

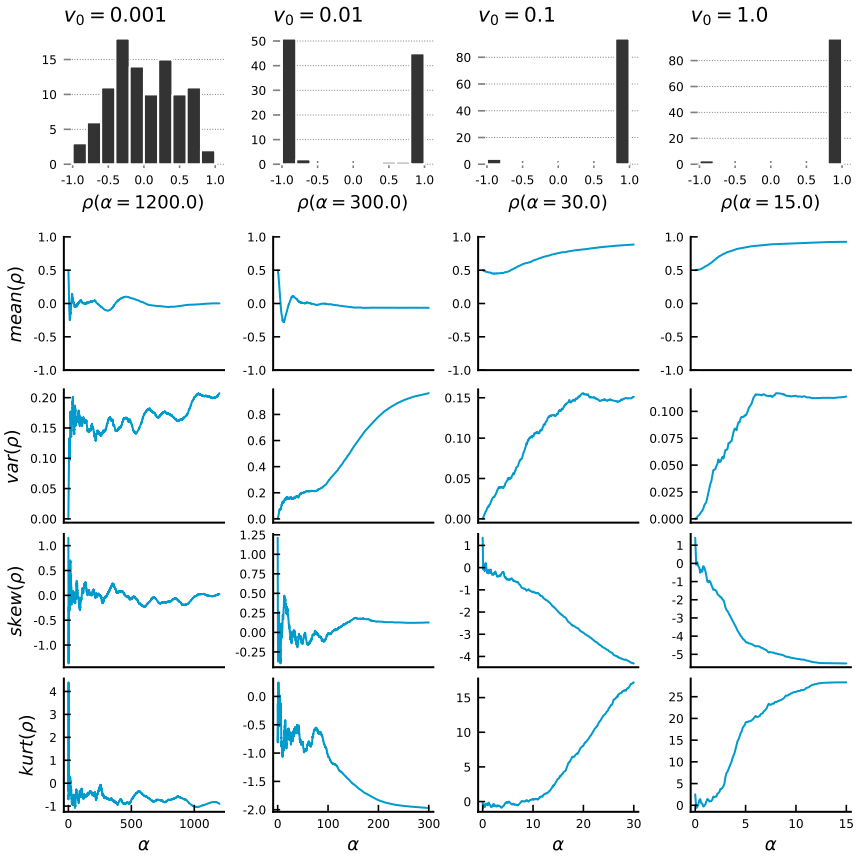


Figure 8: Histogram of ρ for the last interaction recorded and the evolution of the first four sample moments over a 100 realizations of simulated two agents dynamics varying only the initial values of $v(0)$.

3.2.3.1 *A note about other small N scenarios*

The simplest $N = 2$ agents scenario is the typical *Supervised Machine Learning* scenario where one agent always plays the role of emitter and other of receiver, or in the traditional jargon *teacher* and *student*. This scenario has been extensively studied for many architectures, specially for the binary classifier.[11] and, although not for the architecture we introduced in this work, its profiles are similar to the ones we showed above.

The next simple scenario with few agents that can be studied is the case of $N = 3$, which we wont do in this theoretical discussion but is present in an application of the model for in Chapter 4.

11 See 7–9, 18, 28, 29, 31, 32, 34, 39, 45, 47, for the many approaches to the Perceptron Teacher-Student scenario.

3.3 AGENT COMMUNITIES

When we increase the number of agents in the society new forms of organization start to appear. Consider a society where each agent $i \in \{1, \ldots, N\}$, with $N > 2$, can interact with any other agent and has a weight vector $\mathbf{w}_i \in \mathbb{R}^K$.

We are looking for what situations can lead this society of agents to the macroscopic states of *consensus*, *communities* or *disorder*. This investigation requires us to choose the initial conditions for agents states and the size of the subject category, i.e. the number of issues they can discuss.

3.3.1 *Effects of distrust in a society of agents*

Lets start with the investigation of the most intuitive case, a society of trustful agents. For this scenario, consider $\mu_{j|i}(t=0) < 0$ for all $i, j \in \{1, \ldots, N\}$, or in other words, agents initially trust each other, and $\mathbf{w}_i(t=0)$ are uniformly distributed in a K–dimensional sphere. This choice for the initial weight vectors gives a disordered society, with all sorts of opinions about any possible issue. Lets restrict the number of issues to $P = 1$ and choose $\mathbf{x} \in \mathbb{R}^K$ so $\|\mathbf{x}\| = 1$. We choose the initial magnitudes, $|\mu_{j|i}(t=0)|$ and $\|\mathbf{w}_i(t=0)\|$, and initial uncertainties, $C_i(t=0)$ and $V_{j|i}(t=0)$, in a way that $\tilde{w}_i^0 = \frac{\|\mathbf{w}_i\|}{\sqrt{1+\|C_i\|}}$ and $\tilde{\mu}_{j|i}^0 = \frac{|\mu_{j|i}|}{\sqrt{1+V_{j|I}}}$ can be easily compared. This is done by setting $C_i(t=0) = c^0 \mathbb{1}$, $V_{j|i} = v^0$ and choosing $\|\mathbf{w}_i(t=0)\| = \sqrt{1+c^0}$ and $|\mu_{j|i}(t=0)| = \beta\sqrt{1+v^0}$, for $\beta > 0$. The reason for this choices is the fact that the Modulation functions F_w and F_μ depend only on the internal reactions, for which $|\tilde{h}\sigma| \propto \frac{\|\mathbf{w}\|}{\sqrt{1+C}}$ and $|\tilde{\mu}| = \frac{|\mu|}{\sqrt{1+V}}$.

Figures 17 and 9 show what happens when we choose large values of $\tilde{\mu}_{j|i}$, with $\mu_{j|i} < 0$, a society full of agents who trust each other with high certainty, but with homogeneous opinions, $\mathbf{w}_i$ is uniformly distributed. Both use figures show statistics for the overlaps $\rho_{j|i} = \frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\|\|\mathbf{w}_j\|}$ and disbelief $\varepsilon_{j|i} = \Phi\left(\frac{\mu_{j|i}}{\sqrt{1+V_{j|i}}}\right)$.

In Figure 9 show the evolution of the first four moments of the distributions of overlaps and distrust in the society as a function of the number of interactions per adjustable weight. Figure 17 shows a snapshot of the initial and final matrices with entries $\rho_{j|i}$ and $\varepsilon_{j|i}$, as well as the histograms. That when trust is high, agents opinions align and society achieves *Consensus*, as every one agrees on most issues and trust is pervasive.

If we change just the initial $\tilde{\mu}$ to a smaller value we get a different scenario, as shown in figures 10 and 18[12]. When initial trust is not high enough, society breaks into two *Communities*, and we say its

12 The matrices were sorted using the algorithm in [43].

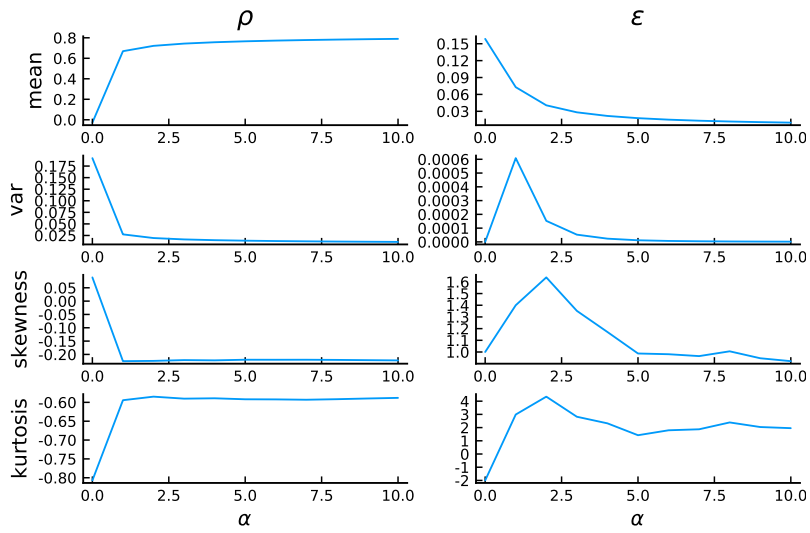


Figure 9: Evolution of the first four moments of the distrust and overlaps of an agent society with $K = 5$, $N = 30$, $\tilde{\mu}^0_{j|i} = \tilde{w}^0_i = 1$ and $\varepsilon_{j|i} < \frac{1}{2}$ for all agents.

Polarized. Notice that a *community* is defined by the antagonist groups regarding opinions and distrust.

From figures 9 and 10 we see that the first moments of the overlap and distrust distributions can be used to identify the different social states. In Figure 11 we can see how the equilibrium value of each distribution moment vary with the initial value $\tilde{\mu}$.

If instead of starting with $\tilde{\mu}_{j|i} < 0$ we start with a distrustful situation, where $\tilde{\mu}_{j|i}(t = 0) > 0$ for all agents, or let distrust be normally distributed, i. e. $_{j|i}(t = 0) \sim G(\mu^0, 1)$ for some μ^0, we can only get polarized, or disordered when distrust is to high, social states.

3.3.2 *Effects of opinion in a society of agents*

We can reverse the situations, starting from a society in consensus, with $\rho_{j|i} > 0$ for all agents, and start from a normally distributed distrust. The results are analogous to the trustful situation, but the relative values of $|\tilde{\mu}_{j|i}|$ and $|\tilde{w}_i|$ are very different. While trust requires a $1 : 1$ relation with opinion to percolate the system, opinion need a $20 : 1$ relation with distrust to percolate for our choices of $N = 30$ and $K = 5$ in figures 12 and 13. This can be interpreted as if trust was harder to swing than opinion, although this is an intrinsic distinction between entropic dynamics in 1–dimensional spaces versus dynamics in $(K > 1)$–dimensional spaces. Had we choose $Q(\mathbf{w}^\star|\mathbf{w}C)$ with $C \in \mathbb{R}$ instead of $C \in \mathbb{R}^{K\times K}$ our model would lead to hardly swinging opinions. Also, the snapshots of the effects of initial consensus can be seen in figures 19 and 20.

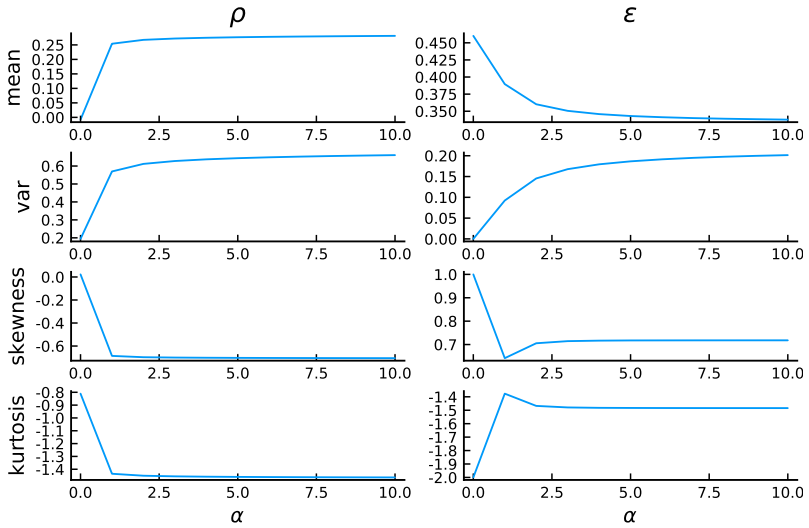


Figure 10: Evolution of the first four moments of the distrust and overlaps of an agent society with $K = 5$, $N = 30$, $\tilde{\mu}^0_{j|i} = 0.1$, $\tilde{w}^0_i = 1$ and $\varepsilon_{j|i} < \frac{1}{2}$ for all agents.

3.3.3 *An Spin-Glass states, or the effect of the number of issues*

One interesting aspect of the model revolves around the number of issues agent have at disposal to talk about, the size of category subject universe. So far we used $P = 1$ for the analysis, but what happens when we increase this number? Consider a society starting with randomly attributed $\mu_{j|i}$ and $\mathbf{w}_i$ and high initial uncertainties, $C_i \gg 1$ and $V_{j|i} \gg 1$. If we give $P = 1$ issue and let them discuss we get a polarized society shown in figures 21 and 15.

If we only change to a huge number of issues, say $P \propto K^4$, we get a weird stagnation in social states, as shown in figures 22 and 16.

As a consequence, for a fixed window of interactions the transitivity of distrust relations breaks as the number of issues grow. This show as slowdown in polarization speed due to the diversity of subjects in agents communication. We say the relations of distrust are balanced for agent i regarding agents j and k when i distrust k if i trusts j and j distrusts k. We can create a quantitative index for this relations by computing $b_{k\to j\to i} = \tau_{j|i}\tau_{k|j}\tau_{k|i}$, where $\tau_{b|a} = 1 - 2\varepsilon_{b|a}$. The relation is balanced or unbalanced when $b_{k\to j\to i} > 0$ or $b_{k\to j\to i} < 0$, respectively.

Figure 14 show the average of the distrust relations balance distribution for all triads of agents after $\alpha = 25$ for different numbers of issues.

This scenario can be used as an abstract description of a public debate, where society focuses on a small set of issues for a period of time and turns back on discussing "trivialities" until something, like

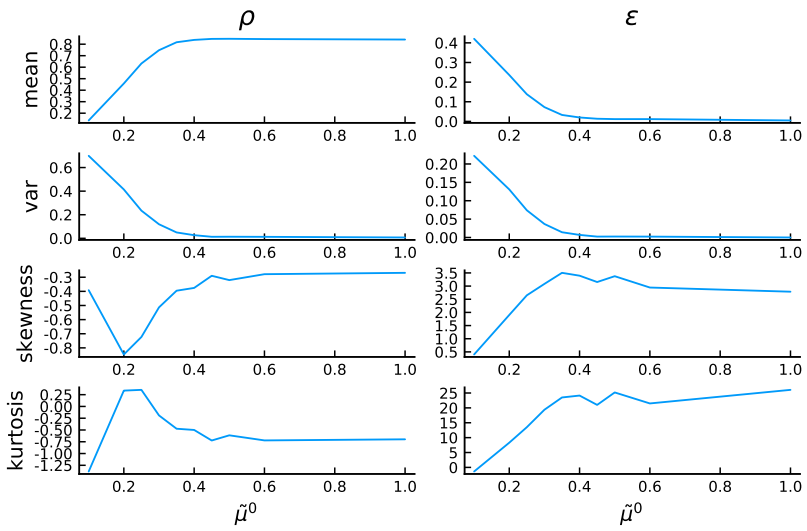


Figure 11: Average over a 100 simulations of the equilibrium values for the first four moments of the overlaps and distrust distributions as a functions of $\tilde{\mu}^0_{j|i}$ for an agent society with $K = 5$, $N = 30$, $\tilde{w}^0_i = 1$ and $\varepsilon_{j|i} < \frac{1}{2}$ for all agents.

the media or a pandemic spread of virus, pulls their attention back to just a few issues.

This slow dynamic is similar to spin-glass states in typical Statistical Mechanics, where frustration can extend to mesoscopic domains preventing the system from getting to its minimum energy configurations and depending on the "communication" conditions, like the couplings J_{ij} distribution in a spin system.

3.3.4 *Discussion*

The results presented in this section show many situations where this model can generate consensus and polarization. Polarized structures can be identified by antagonist communities, different from complete disorder, where no structure can be found, or consensus (which also can be seen as mono-polarized). Also, there is a slow spin-glass like scenario, where the distrust relations are unbalanced and neither community, consensus or complete disorder dominate the society.

Although we did not exhaust the possible scenarios and variations, these results, and the ones in Chapter 4, show the descriptive power of this model achieves without compromising quantitative analysis.

One point to emphasize is that the macroscopic observables used in this analysis (the overlaps ρ and distrust ε) are measurable, although not directly. All entities in the model are directly related to observable behavior in human societies and their interpretations allow us to properly reason about real social scenarios.

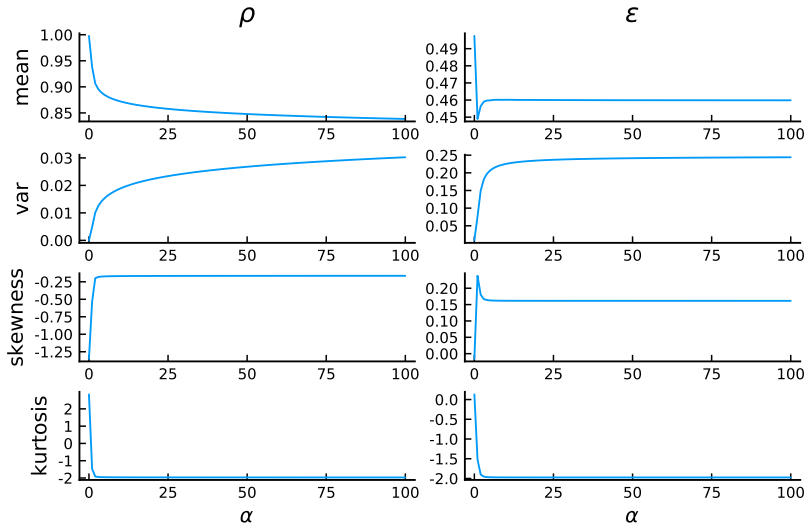


Figure 12: Evolution of the first four moments of the distrust and overlaps of an agent society with $K = 5$, $N = 30$, $\tilde{\mu}^0_{j|i} = \tilde{w}^0_i = 1$ and $\rho_{j|i} > 0$ for all agents.

Just for fun, and let it be clear that the following argument is just for fun, lets play around with our model. Consider the "not-at-all obvious" spin-glass state for the agents, from a reversed perspective and ask "how did Brazil ended up in its current political state?". If we look back a few year, starting at 2013,[13] we will see that most of what was reported in the news was about political corruption and the impeachment of the president and the presence of alternative source of subjects to traditional media was not as prominent as today. Such situation shaped our daily interactions with those we care, putting us in a scenario of 1 or 2 issues in our subject category, and we lived through a fast change of paradigm accordingly. Now, in 2020, we still have a high perceived polarization and we have to deal with several other issues, like the Covid-19 pandemic, the constant misbehavior of our government and institutions and the disbelief in traditional media, removing its monopoly on the subject category, leading us in a "many issues" scenario and slowing down the rate of change in the political points. Of course, this isn't the whole story (actually it may just be a delirious oversimplification of the situation), but its neat that our model can get to the point of allowing us to give a (somewhat informed) shot at such complex real world event[14].

13 See 2, for an application of a previous version of this model to a political analysis previous to 2013 in Brazil (in Portuguese).

14 Also, notice that we can test some parts of this argument, like the number of different news headlines at a given moment, the adhesion to communication apps as source of news, the approval rates of politicians and parties, all of which could be used to test this playful argument.

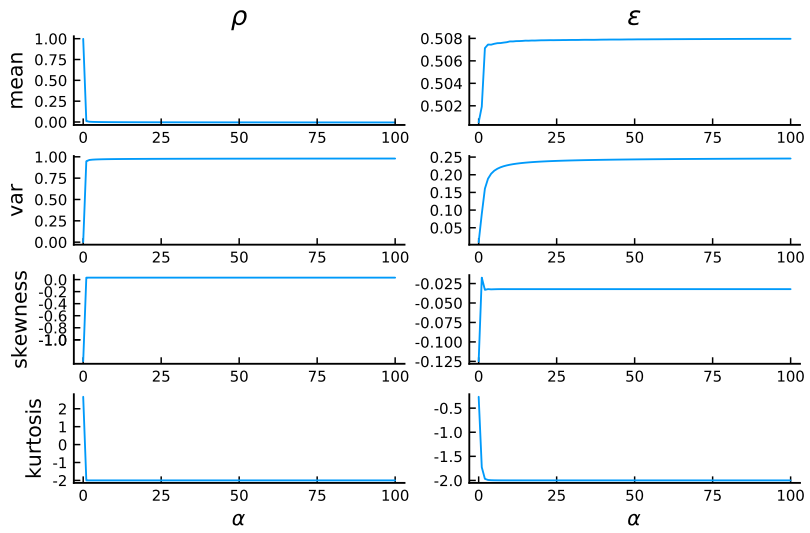


Figure 13: Evolution of the first four moments of the distrust and overlaps of an agent society with $K = 5$, $N = 30$, $\tilde{\mu}^0_{j|i} = 0.1$, $\tilde{w}^0_i = 1$ and $\rho_{j|i} > 0$ for all agents.

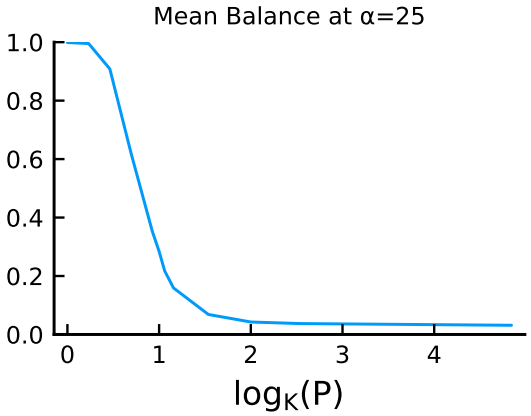


Figure 14: Average over a 100 simulations of the for mean of the distrust relations balance distribution over a society of agents with $N = K = 20$, initial $C_i = V_{j|i} = 10$ and uniformly distributed initial $\mathbf{w}_i$ and $\mu_{j|i}$.

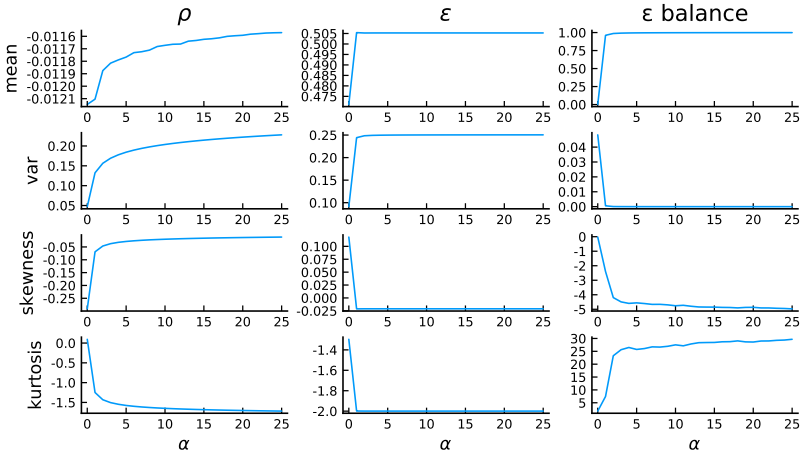


Figure 15: Evolution of the first four moments of the distrust and overlaps of an agent society with $K = 20$, $N = 20$, uniformly distributed $\mathbf{w}_i(t = 0)$ and $\mu_{j|i}(t = 0)$, $C_i(t = 0) = V_{j|i}(t = 0) = 10$ and $P = 1$.

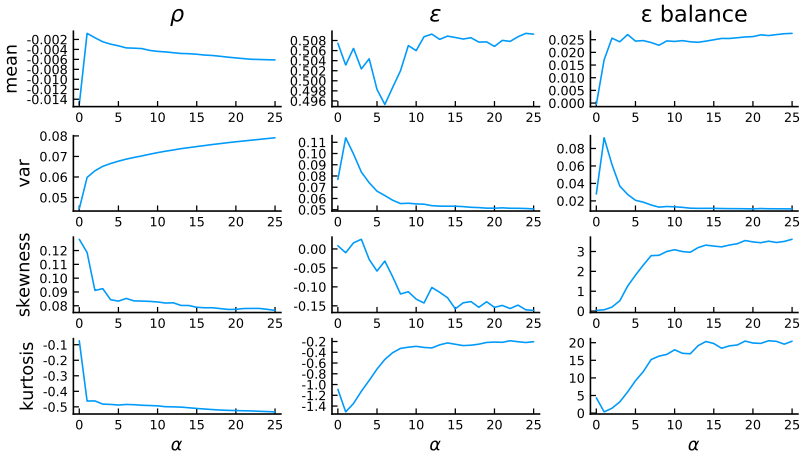


Figure 16: Evolution of the first four moments of the distrust and overlaps of an agent society with $K = 20$, $N = 20$, uniformly distributed $\mathbf{w}_i(t = 0)$ and $\mu_{j|i}(t = 0)$, $C_i(t = 0) = V_{j|i}(t = 0) = 10$ and $P = 2000000$.

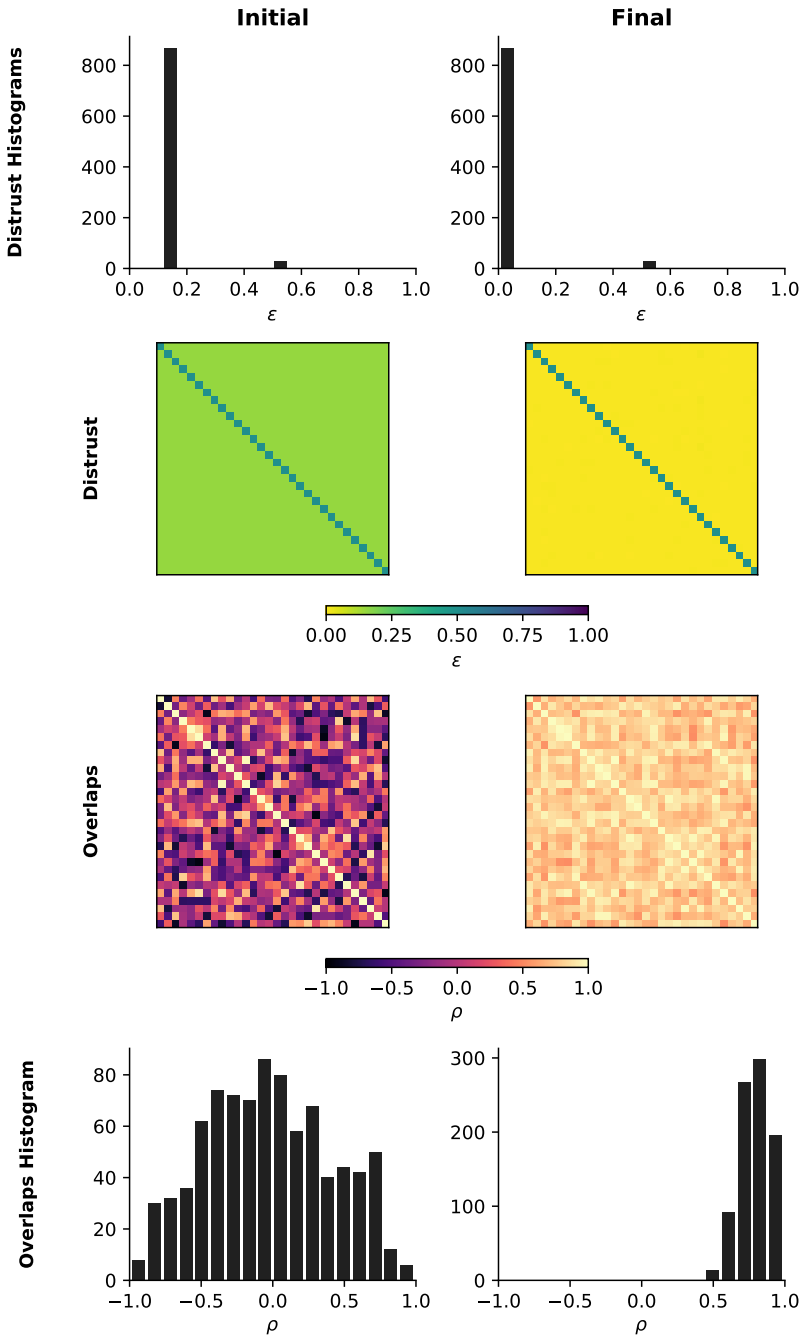


Figure 17: Histograms and heatmaps of the distrust and overlaps for the initial and final states of an agent society with $K = 5$, $N = 30$ and $\tilde{\mu}^0_{j|i} = \tilde{w}^0_i = 1$ and $\varepsilon_{j|i} < \frac{1}{2}$ for all agents. The histograms are computed from all the entries in the matrix, so the values sum to N^2.

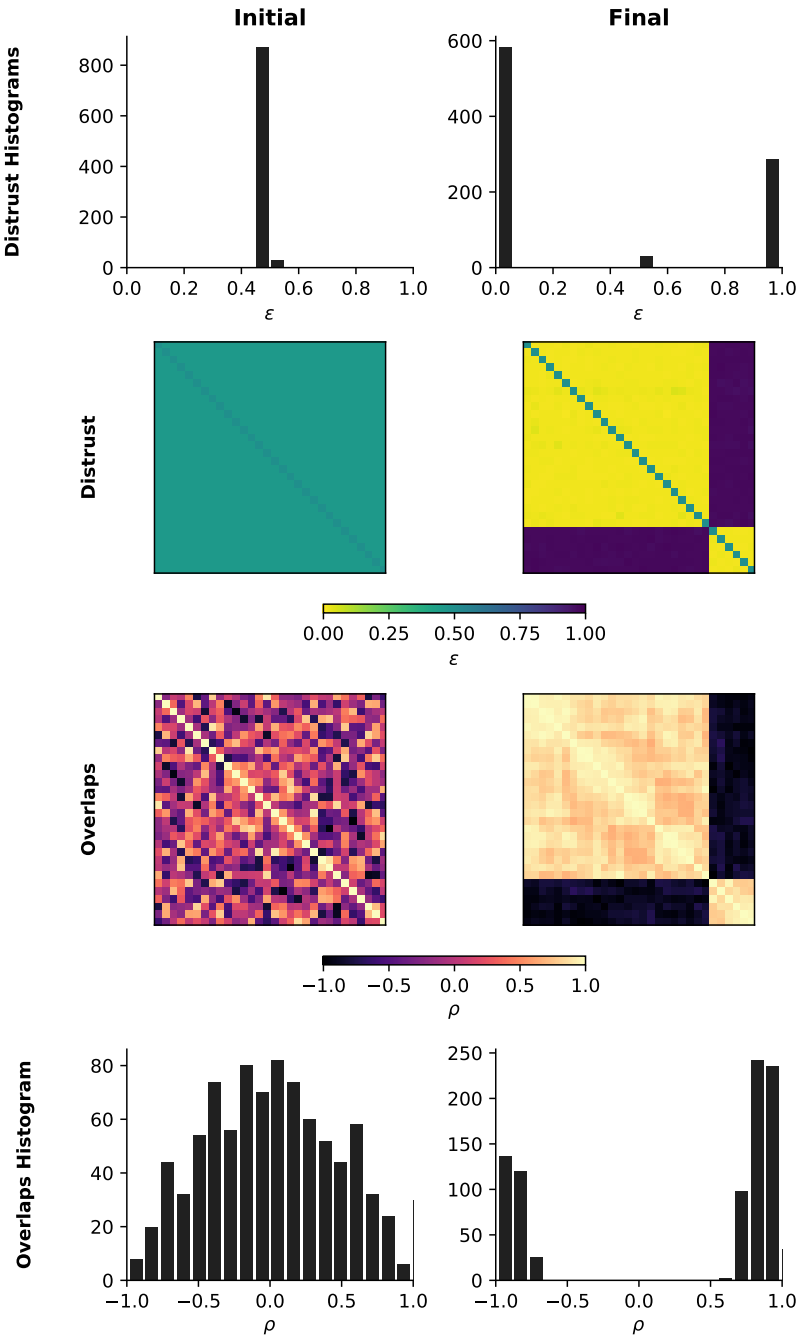


Figure 18: Histograms and heatmaps of the distrust and overlaps for the initial and final states of an agent society with $K = 5$, $N = 30$ and $\tilde{\mu}^0_{j|i} = 0.1$ and $\tilde{w}^0_i = 1$ and $\varepsilon_{j|i} < \frac{1}{2}$ for all agents. The histograms are computed from all the entries in the matrix, so the values sum to N^2.

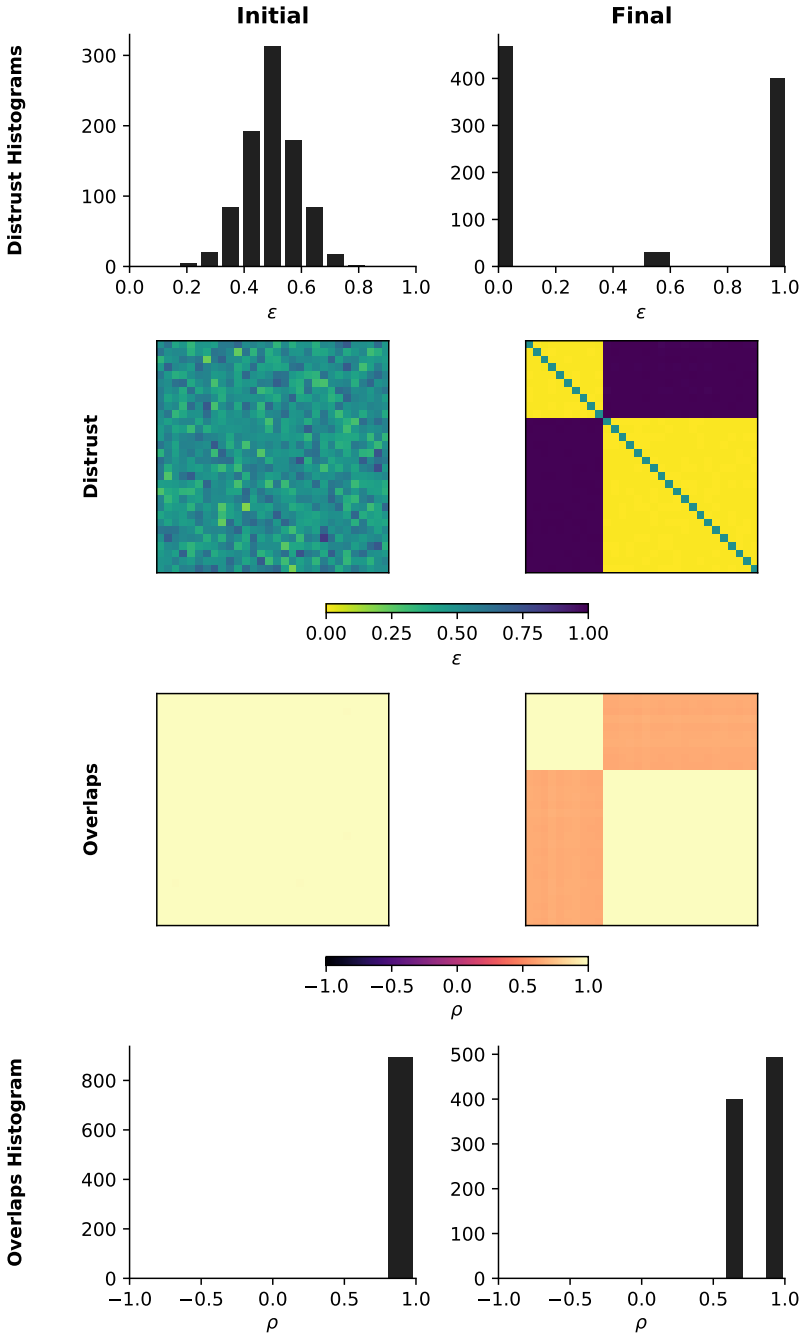


Figure 19: Histograms and heatmaps of the distrust and overlaps for the initial and final states of an agent society with $K = 5$, $N = 30$ and $\tilde{\mu}^0_{j|i} = 0.25$, $\tilde{w}^0_i = 5$ and $\rho_{j|i} > 0$ for all agents. The histograms are computed from all the entries in the matrix, so the values sum to N^2.

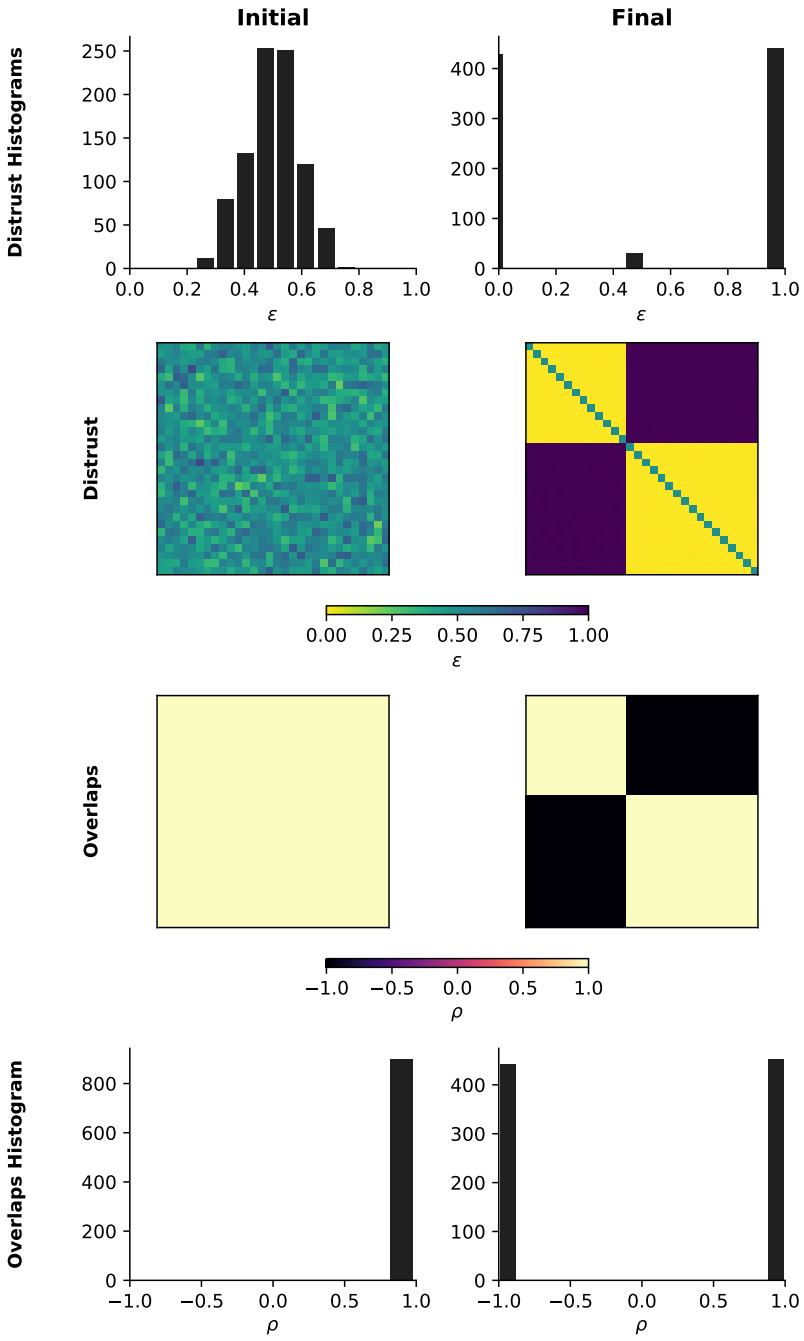


Figure 20: Histograms and heatmaps of the distrust and overlaps for the initial and final states of an agent society with $K = 5$, $N = 30$, $\tilde{w}_i^0 = \tilde{\mu}_{j|i}^0 = 0.25$ and $\rho_{j|i} > 0$ for all agents. The histograms are computed from all the entries in the matrix, so the values sum to N^2.

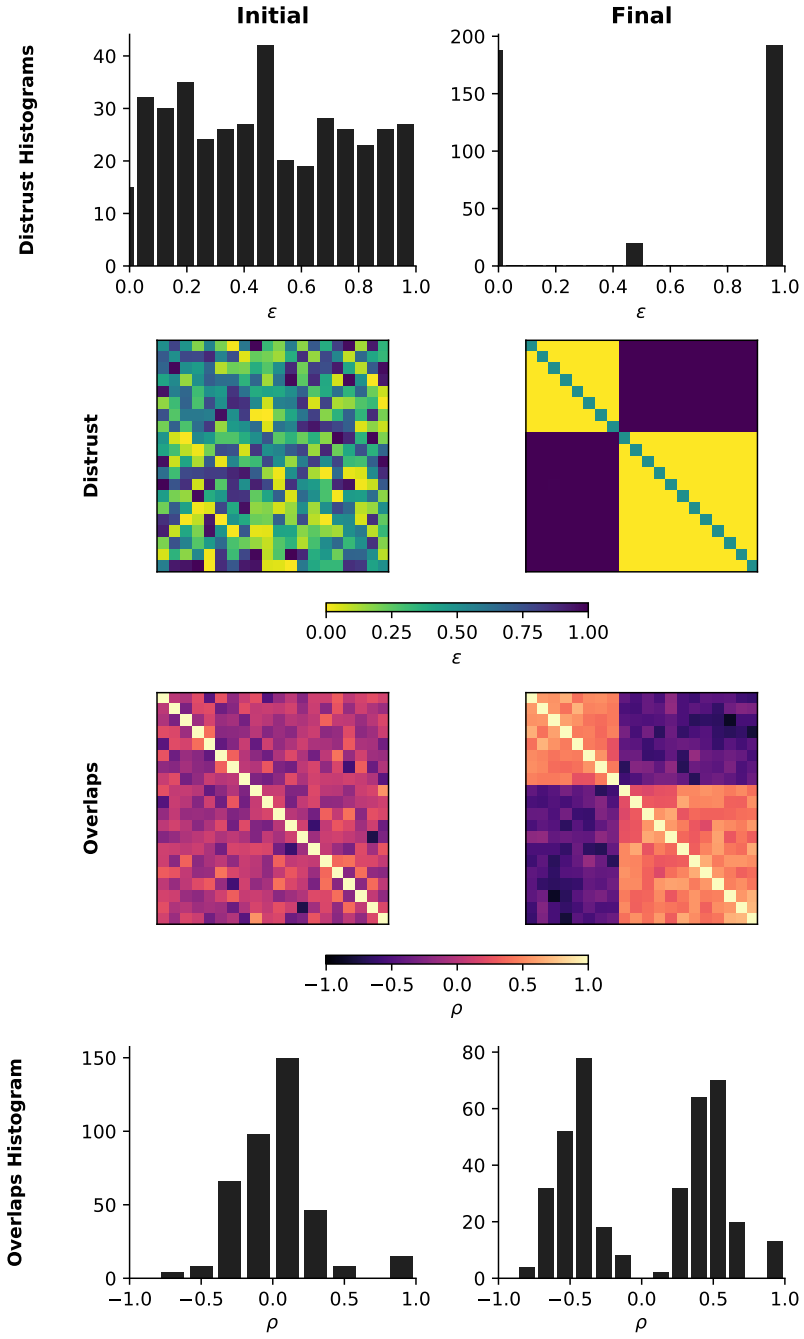


Figure 21: Histograms and heatmaps of the distrust and overlaps for the initial and final states of an agent society with $K = 20$, $N = 20$, uniformly distributed $\mathbf{w}_i(t = 0)$ and $\mu_{j|i}(t = 0)$, $C_i(t = 0) = V_{j|i}(t = 0) = 10$ and $P = 1$. The histograms are computed from all the entries in the matrix, so the values sum to N^2.

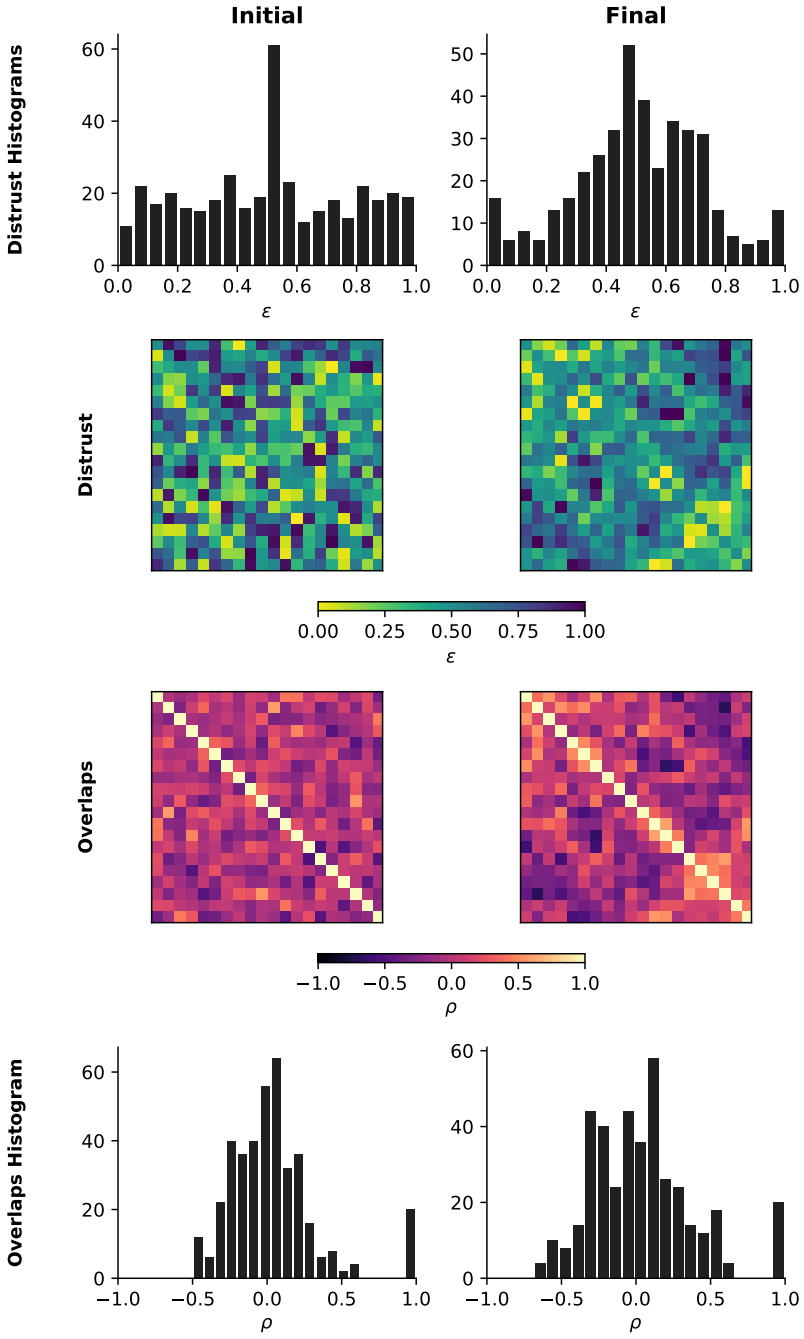


Figure 22: Histograms and heatmaps of the distrust and overlaps for the initial and final states of an agent society with $K = 20$, $N = 20$, uniformly distributed $\mathbf{w}_i(t = 0)$ and $\mu_{j|i}(t = 0)$, $C_i(t = 0) = V_{j|i}(t = 0) = 10$ and $P = 2000000$. The histograms are computed from all the entries in the matrix, so the values sum to N^2.

4

IDEOLOGY AND VOTING PATTERNS IN THE U.S. COURT OF APPEALS

In this section we apply the theory developed in chapters 2 and 3 to model the data and analysis in Sunstein et al. [42] about judicial decisions made by panels of judges in the U.S. Court of Appeals. Our interest in this topic, besides its obvious intrinsic interesting aspect, is the availability of data regarding the political party allegiances influence on the decisions of members in the panel. We must disclaim that we are, by no means, specialists in Political Sciences or U.S. Law, so our only objective is to provide a mathematical model as a tool of numerical analysis that could be of use for those who are. As we will show, however, our approach serves not only to identify statistical signatures in the judges behavior but also as a candidate mechanism to explain how these signatures may appear. This is accomplished by setting up the agents sharing a common knowledge of the Law, having opposite party influences and a personality component on the opinion weight vectors, choosing if and how much they trust (or distrust) agents with an opposite political allegiances and choosing set of issues (representing the cases) they discuss with different projections on the Law. We find that changing the relative weights betweeen Law, Party influence and Personality lead to different statistical signatures, with only one case where it matches the data signature.

4.1 ARE JUDGES POLITICAL?

The data available in Sunstein et al. [42] comprises the fraction of "Liberal Decisions" taken by different compositions of three judges in fourteen types of cases[1]. By "Liberal Decision" they mean a decision that would be aligned with values of self declared liberals. Judges are nominated by a president, which in the U.S. case can be either a Republican or a Democrat, so there are six types of panels regarding the political allegiance of the nominating president: *Rrr, Rrd, Rdd, Ddd, Ddr and Drr*, where *D or d* stands for Democrat, *R or r* stands for Republican and we capitalize the letter representing who the vote we are looking belongs to, i.e. if we are looking at the decision of a Democrat judge[2] in a panel with another Democrat and a Republican

1 Specifically, the data is from rulings in 15 areas of the law: Affirmative action, NEPA, 11th Amendment, NLRB, Sex discrimination, ADA, Campaign Finance, Piercing corporate veil, EPA, Obscenity, Title VII, Desegregation, FCC, Contract Clause, Commercial speech

2 To simplify the discussion we will extend to the judge the political party of his/her appointing president

we would refer to the panel as *Drd* or *Ddr*. The data published in the book is reproduced in figure 23

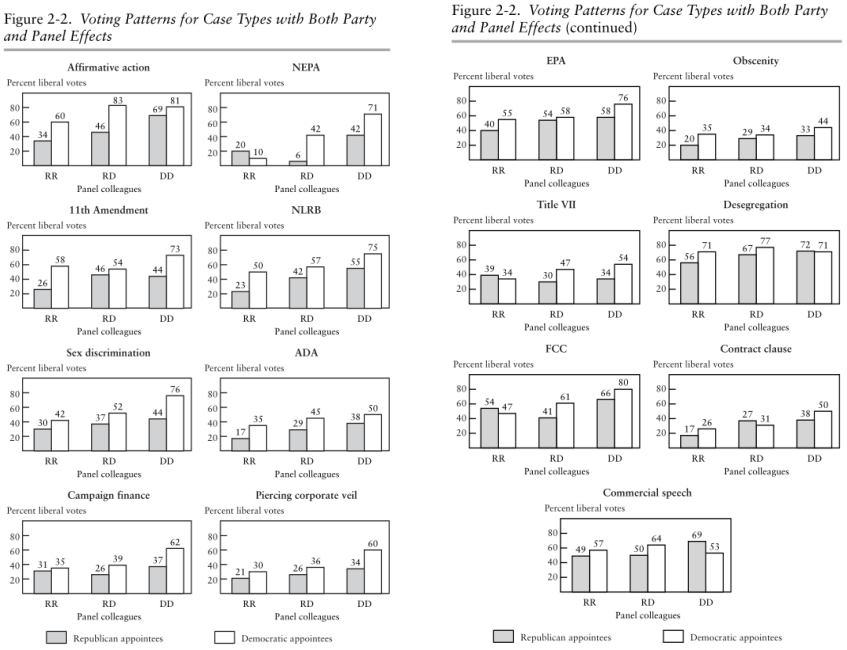


Figure 23: Party and Panel ideological effects per Case Type, extracted from [42]

They claim to have identified three ideological effects in the data:

1. Ideological Voting: Republican appointees vote differently than Democrat ones
2. Ideological Dampening: A judge in a minority party in the panel will be less ideological
3. Ideological Amplification: Judges in a pure party panel will be more ideological.

These effects were found through statistical regularities in the data, presented as the percentage above 50% of liberal decisions by a given panel over the cases for each of the 15 areas of the Law. We represent their data as 15–dimensional vectors $\mathbf{J}_g$, one for each panel compositions $g \in \{Rrr, Rrd, Rdd, Ddd, Ddr, Drr\}$, and define the angles $\theta(g, g') = \arccos\left(\frac{\mathbf{J}_g \cdot \mathbf{J}_{g'}}{\|\mathbf{J}_g\|\|\mathbf{J}_{g'}\|}\right)$ as a distance to measure the difference between rulings on each area of the Law. For example, if we had $\theta(Ddd, Rrr) = 0$ than a pure Democrat panel and a pure Republican panel would be no different in how they vote, but $\theta(Ddd, Rrr) = \pi$ means these panel take opposite decisions in each area of the Law. The matrix with the empirical angles for each panel composition can be seen in Figure 24.

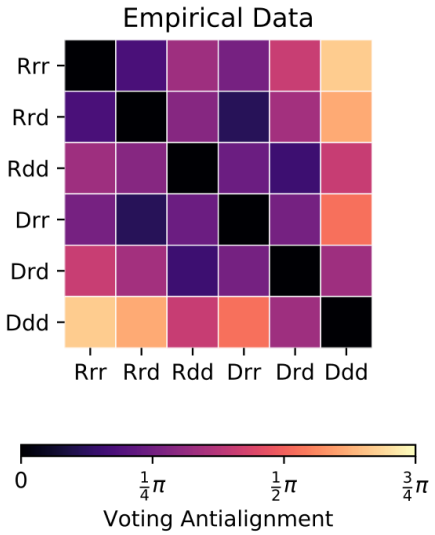


Figure 24: Alignment angles for different panel compositions computed from data in Figure 23

The advantages of defining $\mathbf{J}_g$ and $\theta(g, g')$ are the possibility of constructing them readily from data and the ease in probing the working hypothesis. Each hypothesis translate into $\theta(g, g')$ as indicated in table 1.

Also, the angles $\theta(Rrr, Rrd)$, $\theta(Ddd, Drd)$, $\theta(Rdd, Rrd)$ and $\theta(Drr, Drd)$ could be used to see how much influence a single judges brings to a panel composition, when majority is preserved or reversed.

4.2 A MODEL FOR THE JUDICIAL BEHAVIOR

To model the behavior of three-judges panels representing the U.S. Federal Appellate Court we will make use of our agent model developed in Chapter 3, fixing $N = 3$, $K = 5$ and choosing appropriate initial conditions[3]. We consider a two-party system, with parties A and B, and each agent i in a panel of $N = 3$ agents has a set of parameters $a_i = (\mathbf{w}_{i|p(i)}, C_{p(i)}, \mu_{p(j)|p(i)}, V_{p(i)})$, where $p(i) \in A, B$ is the agent's appointing executive officer political party. The subscripts refer to parties so we can control agents similarities and differences through ideological influence, except for $\mathbf{w}_{i|p(i)}$ where we admit a personality component as well. To each panel of agents we give a *case*, an issue that falls in the issue space formed by the *Law* and *Party* components, and the agents interact by exchanging their opinion about the case.

The initial state for the agents weight vectors $\mathbf{w}_{i|p(i)}$ are chosen to reflect the aspects of judges opinions regarding this subject category, namely their knowledge of the Law, the Party bias of executive offi-

3 This choice of K just reflects the number of Moral Foundations in MFT[24] and is, otherwise, arbitrary.

	Hypothesis	Example	Interpretation
1.	Ideological Voting	$\theta(Ddd, Rrr)$ is the largest angle	Largest differences between *Rrr* and *Ddd*
2.	Ideological Dampening	$\theta(Drr, Rrr) < \theta(Rdd, Rrr)$	*D* in *Drr* is more conservative than *R* in *Rdd*
		$\theta(Rdd, Ddd) < \theta(Drr, Ddd)$	*R* in *Rdd* is more liberal than *D* in *Drr*
3.	Ideological Amplification	$\theta(Ddd, Drd) < \theta(Ddd, Drr)$	*D* in *Drd* is more liberal than *D* in *Drr*
		$\theta(Rrr, Rrd) < \theta(Rrr, Rdd)$	*R* in *Rrd* is more conservative than *R* in *Rdd*

Table 1: Translating the working hypothesis in [42] to angles between voting patterns.

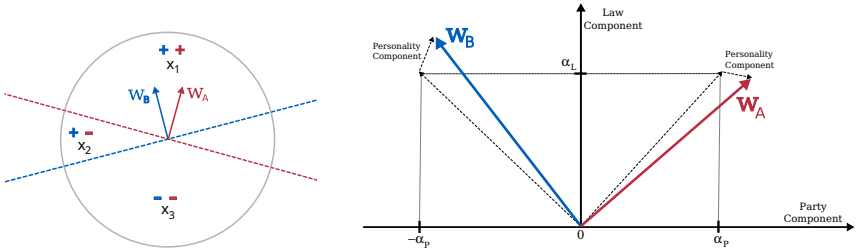
cer political party and a *Personality* components unique to that agent. More precisely, $\mathbf{w}_{i|p(i)}(t=0) = \beta_L \mathbf{L} + \beta_{p(i)} \mathbf{I}_{p(i)} + \beta_R \mathbf{R}$, where $\mathbf{L}$ is the Law component, $\mathbf{I}_{p(i)}$ is the Party bias and $\mathbf{R}$ the agents Personality. We assume all agents have the same knowledge of the Law, so β_L is the same for all agents, and $\mathbf{I}_A = \mathbf{P} = -\mathbf{I}_B$, so if assume parties to exert influence over subscribed agents then $\beta_A = -\beta_B$. Also, we assume Party ideology to be orthogonal to the Law, so $\mathbf{L} \cdot \mathbf{P} = 0$ and they form an independent base in the issue space for the cases. The personality component $\mathbf{R}$ is chosen randomly in $\mathbb{R}^K$. An illustration of the initial vectors of two agents can be seen in Figure 25b.

The parameters $\beta_L, \beta_p, \beta_R$ control the relative importance agents give initially to each component in the issue space and their idiosyncratic component. We consider three *Opinion Scenarios*, a *Lawless scenario* when $\beta_L = 0$, a *Partyless scenario* when $\beta_A = \beta_B = 0$ and a *Fair scenario* when $\beta_L \approx \beta_p \approx \beta_R$. In this model, we choose the *Fair* scenario to have $\beta_L = \beta_R = 1$ and $\beta_p = 1.25$.

The uncertainty C_p is related to how agents ponder corroborative or novelty information and lower values represent novelty predominance. In this model we used $C_A = 0.25$ and $C_B = 0.5$, which can be interpreted as judges from party A being more difficult to convince than judges of party B.

For the initial conditions for $\mu_{p'|p}$ and V_p, for $p', p \in \{A, B\}$, have we consider four possible scenarios, given in Table 2.

The distrusts scenarios are divided by the initial values of $\mu_{p'|p}$, which we called *Courteous or Distrusts*, and V_p, which we called *Certain or Uncertain*. We say agents are *Courteous* they, regardless of party,



(a) (b)

Figure 25: Two dimensional representation of the state vectors and issues. (a) Two agents (arrows), $\mathbf{w}_A$ and $\mathbf{w}_B$ and their hyperplanes of class separation (dot-dashed lines). Both classify agreen on issues $\mathbf{x}_1$, with $\sigma_A(\mathbf{x}_1) = \sigma_B(\mathbf{x}_1) = 1$, and $\mathbf{x}_3$, with $\sigma_A(\mathbf{x}_3) = \sigma_B(\mathbf{x}_3) = -1$. They disagree on issue , as $\sigma_A(\mathbf{x}_2) = 1 \neq -1 = \sigma_B(\mathbf{x}_2)$. (b)The initial states of the agents and issues. Continuous arrows: Initial state of agents $\mathbf{w}_A$ and $\mathbf{w}_B$ are made up by three contributions. The *Law Component* is common to all agents with weight α_L. The direction of the *Party Component* contribution depends on the party being A or B with weights α_P and $-\alpha_P$ respectively. A random contribution (*Personality Component*) that is unique to each agent models the idiosyncratic part of a judge (which has a weight α_η not shown in the figure).

Courteous Certain	Courteous Uncertain	Discourteous Certain	Discourteous Uncertain
$\mu_{p'\|p} = -0.5$	$\mu_{p'\|p} = -0.5$	$\mu_{p'\|p} = 0.5$	$\mu_{p'\|p} = 0.5$
$V_p = 0.1$	$V_p = 5$	$V_p = 0.1$	$V_p = 5$

Table 2: Four different distrust initial conditions scenarios were considered. $\mu_{p'|p}$ refers to agents of different parties. For all pairs of agents of the same party $\mu = -0.5$.

extend the courtesy of attributing low distrust to other agents, i. e.$\mu_{A|A} = \mu_{B|B} = \mu_{A|B} = \mu_{B|A} = \mu < 0$, and *Discourteous* when distrust attribution depends on party so member of the same party trust each other and distrust members of the opposing party, i. e.$\mu_{A|A} = \mu_{B|B} = \mu < 0$ and $\mu_{A|B} = \mu_{B|A} = -\mu > 0$. We say agents are *Certain or Uncertain* when they are, respectively, more or less confident about their initial distrust attributions, i. e.$V_p \approx 0$ or $V_p \gg 0$. In this model we choose $\mu = -0.5$ and either $V_p = 0.1$ or $V_p = 5$ for the Certain or Uncertain scenarios, respectively.

The *Cases* are represented by issues in the space formed by *Law and Party bias*. Each issue $\mathbf{x}_c$ represents an area of the Law among the 15 areas and have an angle ϕ_c with the Law vector $\mathbf{L}$, i. e.$\mathbf{x}_c = \cos(\phi_c)\mathbf{L} + \sin(\phi_c)\mathbf{P}$, for $c \in \{1, \ldots, 15\}$ and $\phi_c \in [0, \pi]$.

4.2.1 *Voting pattern in panels of agents*

Having established how we initialize the agents and how to represent the cases, we need to probe the decision of each panel composition, $g \in \{Aaa, Aab, Abb, Bbb, Bab, Baa\}$, over each case type, $\{\mathbf{x}_c | c \in \{1, \ldots, 15\}\}$, under each *Distrust* scenario, *Courteous-Certain, Courteous-Uncertain, Discourteous-Certain, Discourteous-Uncertain*, and 3 *Opinion* scnearios, *Lawless, Partyless and Fair*. The voting pattern of panel $\mathbf{J}_g$ is the average of many panels with compositions g over each case and the angles $\theta(g, g')$ tells us how much panel compositions differ on how the rule the cases with similar voting patterns having small angles.

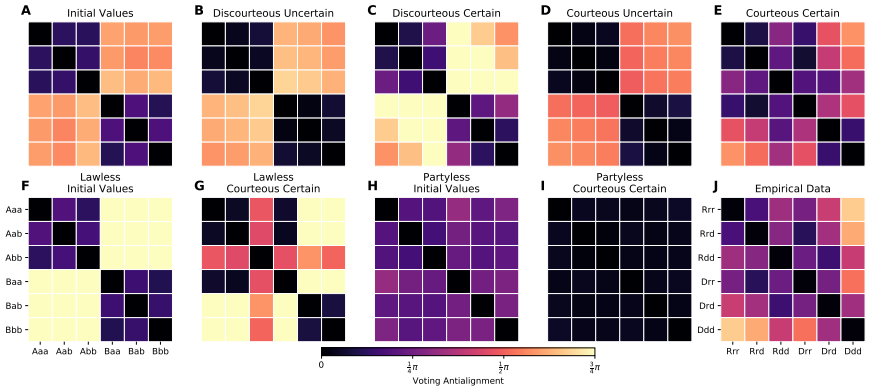


Figure 26: Voting alignment as represented by the matrix of angles $\theta(g, g')$ between vectors $\mathbf{J}_g$ and $\mathbf{J}_{g'}$. All the scenarios have initival values depicted by the rightmost panel, the four middle panels show the asymptotic state of the evolution under the different $\mu - V$ scenarios and with $\beta_L = \beta_R = 1.0$ and $\beta_p = 1.25$ (figures **A-E**) or best case scenarios for the $\beta_L = 0$ (figures **F** and **G**) and $\beta_p = 0$ (figures **H** and **I**). Empirical angles from Figure 24 are reproduced in figure **J**.

In Figure 26 we can see the initial and final angles for each of these scenarios and the empirical angles. We can easily see that the only scenario resembling the empirical patterns is the *Courteous Certain* distrust under *Fair* opinions. This means that, under our assumptions, the only way we would see the statistical voting patterns presented in [42] is if the judges do value Ideology as much as the Law and trust each other despite of party influence. All other scenarios lead to different statistical signatures.

There are two main points to take away from this application of our agent model. First, we are not predicting the behavior of single judges, but we could, predict the behavior of a judicial system if we could pin the scenario defining the relevance of the law, personality and the influence of party ideology over opinions and distrust. Alternatively, we if we were to design a judicial system, we could use the model to inquire the necessary features to avoid ideological vot-

ing. A major advantage over other approaches is that no parameter in the model lacks an interpretation or is intrinsically impossible to measure, although the experimental design to measure some of them may be hard. Second, the reason we are able to get away with at least this amount of empirical validation is due to the deep connection between *Information Processing Systems* and *Statistical Mechanics*, allowing us to use a *Universal Inference* framework to deal with any kind of situation where only partial information is available with the least amount of additional hypothesis we can impose.

5 CONCLUSION

In this work we developed a model for community formation in systems of based on the dynamics of opinion and distrust. The development was done with the framework provided by Probability Theory, Machine Learning and MaxEnt principles, from which we derived a new form of Entropic Dynamics for Information Processing Systems, in particular for simple Neural Networks, the Entropic Learning Algorithm, as we call it in Chapter 2.

Once in possession of this theoretical tool, in Chapter 3 we invested in describing the concepts of opinion and distrust within the Neural Network formalism and modeled the behavior dynamics under the ELD prescription. The resulting theory and model for agents interactions were analyzed in a few scenarios, chosen due to their relationship with realistic counterparts. We started the analysis with the properties of systems with 2 agents interacting under different trust and opinion initial conditions, and showed that the dynamics is not trivial nor leads to results with absurd interpretations. Then, we analyzed the properties of societies with many agents, varying the distribution of opinions and distrust, as well as the subjects they could discuss, and found different situations leading to consensus, polarization and even stagnation.

Finally, in Chapter 4 we applied the model to study the behavior of judges due the availability of data[1] regarding the influence of political party ideology in the voting patterns of judges in the U.S Court of Appeals. In this application, although just a caricature aiming just to provide a quantitative tool for experts in the field, we tried to mimic the typical situations a panel of three judges would be submitted, attributing to agents representing judges a common knowledge of the Law, a Party bias, a Personality and exposing them to different distrust scenarios. The only scenario capable of reproducing the available data had to consider similar contributions of the Law, Party bias and Personality, as well as having Courteous and Certain judges, who extended the courtesy of attributing low distrust to agents of the opposing political party.

The model present in this work has some loose ends, but we can possibly cut them with further investigation on each and they don't seem to be intrinsic flaws of the approach. For instance, we do not know the actual number of dimensions of the opinion weight vectors, if any would make sense, or the actual number of people one can directly attribute a distrust, although we can imagine it cannot go

1 42, See.

beyond 250 based on Dunbar [16], or the its behavior on real social network topologies, although we just extend it and simulate or seek guidance in other works.[2] A more difficult limitation is related to the issues under discussion in a society, called Subject Category above, since our model predicts different states depending on it.

That said, some of our prospects for future work, from the theoretical perspective, are to apply this model to more complex networks topologies, include the choice of issues into the framework and develop a variation for competing opinion-like behaviors, and from an experimental perspective, to identify the typical number of dimensions in the public debate of political issues, measure to idea spreading, typical times to reach polarization, consensus or to fall in disorder and to identify the relation, if any, between what people talk about has to with the social state.

2 See 13.

BIBLIOGRAPHY

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks." In: *Reviews of modern physics* 74.1 (2002), p. 47.

[2] Felippe Alves. "Quebra de simetria espontânea, limites cognitivos e complexidade de sociedades." DOI: 10.11606/D.43.2015.tde-27042015-101234. text. Universidade de São Paulo, Mar. 27, 2015. URL: http://www.teses.usp.br/teses/disponiveis/43/43134/tde-27042015-101234/.

[3] Felippe Alves and Nestor Caticha. "Sympatric multiculturalism in opinion models." In: *AIP Conference Proceedings* 1757.1 (July 26, 2016), p. 060005. ISSN: 0094-243X. DOI: 10.1063/1.4959064. URL: http://aip.scitation.org/doi/abs/10.1063/1.4959064 (visited on 08/09/2017).

[4] David M. Amodio, John T. Jost, Sarah L. Master, and Cindy M. Yee. "Neurocognitive correlates of liberalism and conservatism." In: *Nature Neuroscience* 10.10 (Oct. 2007), pp. 1246–1247. ISSN: 1097-6256. DOI: 10.1038/nn1979. URL: http://www.nature.com/neuro/journal/v10/n10/full/nn1979.html?foxtrotcallback=true.

[5] Solomon E. Asch. "Opinions and Social Pressure." In: *Scientific American* 193.5 (Nov. 1, 1955), pp. 31–35. ISSN: 0036-8733. DOI: 10.1038/scientificamerican1155-31. URL: https://www.nature.com/scientificamerican/journal/v193/n5/pdf/scientificamerican1155-31.pdf.

[6] Robert Axelrod. "The Dissemination of Culture: A Model with Local Convergence and Global Polarization." In: *Journal of Conflict Resolution* 41.2 (Apr. 1, 1997), pp. 203–226. ISSN: 0022-0027. DOI: 10.1177/0022002797041002001. URL: http://dx.doi.org/10.1177/0022002797041002001 (visited on 08/05/2017).

[7] M. Biehl and P. Riegler. "On-Line Learning with a Perceptron." In: *EPL (Europhysics Letters)* 28.7 (1994), p. 525. ISSN: 0295-5075. DOI: 10.1209/0295-5075/28/7/012. URL: http://stacks.iop.org/0295-5075/28/i=7/a=012 (visited on 08/11/2017).

[8] M. Biehl and H. Schwarze. "Learning by on-line gradient descent." In: *Journal of Physics A: Mathematical and General* 28.3 (1995), p. 643. ISSN: 0305-4470. DOI: 10.1088/0305-4470/28/3/018. URL: http://stacks.iop.org/0305-4470/28/i=3/a=018 (visited on 08/11/2017).

[9] Michael Biehl, Peter Riegler, and Martin Stechert. "Learning from noisy data: An exactly solvable model." In: *Physical Review E* 52.5 (Nov. 1, 1995), R4624–R4627. DOI: 10.1103/PhysRevE.52.R4624. URL: https://link.aps.org/doi/10.1103/PhysRevE.52.R4624 (visited on 08/11/2017).

[10] Daniel K. Campbell-Meiklejohn, Dominik R. Bach, Andreas Roepstorff, Raymond J. Dolan, and Chris D. Frith. "How the Opinion of Others Affects Our Valuation of Objects." In: *Current Biology* 20.13 (July 2010), pp. 1165–1170. ISSN: 09609822. DOI: 10.1016/j.cub.2010.04.055. URL: http://linkinghub.elsevier.com/retrieve/pii/S0960982210005956 (visited on 08/06/2017).

[11] Ariel Caticha. "Entropic dynamics." In: vol. 617. AIP, 2002, pp. 302–313. DOI: 10.1063/1.1477054. URL: http://aip.scitation.org/doi/abs/10.1063/1.1477054 (visited on 08/11/2017).

[12] Ariel Caticha, Ali Mohammad-Djafari, Jean-François Bercher, and Pierre Bessiére. "Entropic Inference." In: 2011, pp. 20–29. DOI: 10.1063/1.3573619. URL: http://aip.scitation.org/doi/abs/10.1063/1.3573619 (visited on 08/11/2017).

[13] Nestor Caticha, Jonatas Cesar, and Renato Vicente. "For whom will the Bayesian agents vote?" In: *Frontiers in Physics* 3 (2015). ISSN: 2296-424X. DOI: 10.3389/fphy.2015.00025. URL: http://journal.frontiersin.org/article/10.3389/fphy.2015.00025/full.

[14] Nestor Caticha and Renato Vicente. "Agent-based social psychology: from neurocognitive processes to social data." In: *Advances in Complex Systems* 14.5 (Oct. 1, 2011), pp. 711–731. ISSN: 0219-5259. DOI: 10.1142/S0219525911003190. URL: http://www.worldscientific.com/doi/abs/10.1142/S0219525911003190 (visited on 08/05/2017).

[15] Richard Threlkeld Cox. *The algebra of probable inference*. OCLC: 749469235. Baltimore: Johns Hopkins University Press, 2001. ISBN: 978-0-8018-6982-2.

[16] R.I.M. Dunbar. "Neocortex size as a constraint on group size in primates." In: *Journal of Human Evolution* 22.6 (1992), pp. 469–493. ISSN: 0047-2484. DOI: https://doi.org/10.1016/0047-2484(92)90081-J. URL: http://www.sciencedirect.com/science/article/pii/004724849290081J.

[17] Naomi I. Eisenberger, Matthew D. Lieberman, and Kipling D. Williams. "Does Rejection Hurt? An fMRI Study of Social Exclusion." In: *Science* 302.5643 (Oct. 10, 2003), pp. 290–292. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1089134. URL: http://science.sciencemag.org/content/302/5643/290.

[18] A. Engel and C. van den Broeck. *Statistical mechanics of learning*. Cambridge, UK ; New York, NY: Cambridge University Press, 2001. 329 pp. ISBN: 978-0-521-77307-2 978-0-521-77479-6.

[19] Serge Galam. "Sociophysics: a review of Galam models." In: *International Journal of Modern Physics C* 19.03 (2008), pp. 409–440.

[20] Elizabeth Gardner. "Maximum storage capacity in neural networks." In: *EPL (Europhysics Letters)* 4.4 (1987), p. 481.

[21] Francesca Giardini, Daniele Vilone, and Rosaria Conte. "Consensus emerging from the bottom-up: the role of cognitive variables in opinion dynamics." In: *Frontiers in Physics* 3 (2015), p. 64. ISSN: 2296-424X. DOI: 10.3389/fphy.2015.00064. URL: https://www.frontiersin.org/article/10.3389/fphy.2015.00064.

[22] Adom Giffin, Ariel Caticha, Kevin H. Knuth, Ariel Caticha, Julian L. Center, Adom Giffin, and Carlos C. Rodríguez. "Updating Probabilities with Data and Moments." In: vol. 954. AIP, 2007, pp. 74–84. DOI: 10.1063/1.2821302. URL: http://aip.scitation.org/doi/abs/10.1063/1.2821302 (visited on 08/11/2017).

[23] Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. "Mapping the moral domain." In: *Journal of Personality and Social Psychology* 101.2 (2011), pp. 366–385. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/a0021847. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0021847 (visited on 08/04/2017).

[24] J. Haidt. "The New Synthesis in Moral Psychology." In: *Science* 316.5827 (May 18, 2007), pp. 998–1002. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1137651. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1137651 (visited on 08/04/2017).

[25] Clay B. Holroyd and Michael G. H. Coles. "The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity." In: *Psychological Review* 109.4 (Oct. 2002), pp. 679–709. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.109.4.679. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.109.4.679 (visited on 08/06/2017).

[26] Wander Jager. "Enhancing the Realism of Simulation (EROS): On Implementing and Developing Psychological Theory in Social Simulation." In: *Journal of Artificial Societies and Social Simulation* 20.3 (2017), p. 14. ISSN: 1460-7425. DOI: 10.18564/jasss.3522. URL: http://jasss.soc.surrey.ac.uk/20/3/14.html.

[27] E. T. Jaynes and G. Larry Bretthorst. *Probability theory: the logic of science*. Cambridge, UK ; New York, NY: Cambridge University Press, 2003. 727 pp. ISBN: 978-0-521-59271-0.

[28] O. Kinouchi and N. Caticha. "Optimal generalization in perceptions." In: *Journal of Physics A: Mathematical and General* 25.23 (1992), p. 6243. ISSN: 0305-4470. DOI: 10.1088/0305-4470/25/23/020. URL: http://stacks.iop.org/0305-4470/25/i=23/a=020 (visited on 08/11/2017).

[29] Osame Kinouchi and Nestor Caticha. "Learning algorithm that gives the Bayes generalization limit for perceptrons." In: *Physical Review E* 54.1 (July 1, 1996), R54–R57. DOI: 10.1103/PhysRevE.54.R54. URL: https://link.aps.org/doi/10.1103/PhysRevE.54.R54 (visited on 08/11/2017).

[30] Juan Neirotti. "Anisotropic opinion dynamics." In: *Physical Review E* 94.1 (2016), p. 012309.

[31] Juan P. Neirotti and Nestor Caticha. "Statistical mechanics of program systems." In: *Journal of Physics A: Mathematical and General* 39.33 (2006), p. 10355. ISSN: 0305-4470. DOI: 10.1088/0305-4470/39/33/006. URL: http://stacks.iop.org/0305-4470/39/i=33/a=006 (visited on 08/05/2017).

[32] Juan Pablo Neirotti and Nestor Caticha. "Dynamics of the evolution of learning algorithms by selection." In: *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 67.4 (Apr. 2003), p. 041912. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.67.041912.

[33] Mark Newman. *Networks*. Oxford university press, 2018.

[34] Manfred Opper. "On-line versus Off-line Learning from Random Examples: General Results." In: *Physical Review Letters* 77.22 (Nov. 25, 1996), pp. 4671–4674. DOI: 10.1103/PhysRevLett.77.4671. URL: https://link.aps.org/doi/10.1103/PhysRevLett.77.4671 (visited on 08/05/2017).

[35] Thomas C. Schelling. "Models of Segregation." In: *The American Economic Review* 59.2 (1969), pp. 488–493. ISSN: 0002-8282. URL: http://www.jstor.org/stable/1823701 (visited on 08/06/2017).

[36] Thomas C. Schelling. "Dynamic models of segregation." In: *The Journal of Mathematical Sociology* 1.2 (July 1, 1971), pp. 143–186. ISSN: 0022-250X. DOI: 10.1080/0022250X.1971.9989794. URL: http://dx.doi.org/10.1080/0022250X.1971.9989794 (visited on 08/06/2017).

[37] Muzafer Sherif. "An Experimental Approach to the Study of Attitudes." In: *Sociometry* 1.1 (1937), pp. 90–98. ISSN: 0038-0431. DOI: 10.2307/2785261. URL: http://www.jstor.org/stable/2785261 (visited on 08/06/2017).

[38] Lucas Silva Simões and Nestor Caticha. "Mean Field Studies of a Society of Interacting Agents." In: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. Ed. by Adriano Polpo, Julio Stern, Francisco Louzada, Rafael Izbicki, and Hellinton Takada. Cham: Springer International Publishing, 2018, pp. 131–140. ISBN: 978-3-319-91143-4.

[39] Sara Solla and Ole Winther. "On-line Learning in Neural Networks." In: ed. by David Saad. New York, NY, USA: Cambridge University Press, 1998, pp. 379–398. ISBN: 978-0-521-65263-6. URL: http://dl.acm.org/citation.cfm?id=304710.304758 (visited on 08/05/2017).

[40] Leah H. Somerville, Todd F. Heatherton, and William M. Kelley. "Anterior cingulate cortex responds differentially to expectancy violation and social rejection." In: *Nature Neuroscience* 9.8 (Aug. 2006), pp. 1007–1008. ISSN: 1097-6256. DOI: 10.1038/nn1728. URL: https://www.nature.com/neuro/journal/v9/n8/full/nn1728.html.

[41] Dietrich Stauffer. "Opinion dynamics and sociophysics." In: *arXiv preprint arXiv:0705.0891* (2007).

[42] Cass R. Sunstein, David Schkade, Lisa M. Ellman, and Andres Sawicki. *Are Judges Political?: An Empirical Analysis of the Federal Judiciary*. Google-Books-ID: G2DRDAAAQBAJ. Brookings Institution Press, Feb. 1, 2007. 194 pp. ISBN: 978-0-8157-8235-3.

[43] D. Tsafrir, I. Tsafrir, L. Ein-Dor, O. Zuk, D. A. Notterman, and E. Domany. "Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices." In: *Bioinformatics* 21.10 (May 15, 2005), pp. 2301–2308. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti329. URL: https://academic.oup.com/bioinformatics/article/21/10/2301/206463/Sorting-points-into-neighborhoods-SPIN-data.

[44] R. Vicente, A. Susemihl, J. P. Jericó, and N. Caticha. "Moral foundations in an interacting neural networks society: A statistical mechanics analysis." In: *Physica A: Statistical Mechanics and its Applications* 400 (Apr. 15, 2014), pp. 124–138. ISSN: 0378-4371. DOI: 10.1016/j.physa.2014.01.013. URL: http://www.sciencedirect.com/science/article/pii/S037843711400017X (visited on 08/05/2017).

[45] Renato Vicente, Osame Kinouchi, and Nestor Caticha. "Statistical Mechanics of Online Learning of Drifting Concepts: A Variational Approach." In: *Machine Learning* 32.2 (Aug. 1, 1998), pp. 179–201. ISSN: 0885-6125, 1573-0565. DOI: 10.1023/A:1007428731714. URL: https://link.springer.com/article/10.1023/A:1007428731714.

[46] Renato Vicente, André C. R. Martins, and Nestor Caticha. "Opinion dynamics of learning agents: does seeking consensus lead to disagreement?" In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.3 (2009), P03015. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2009/03/P03015. URL: http://stacks.iop.org/1742-5468/2009/i=03/a=P03015 (visited on 08/09/2017).

[47] Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. "The statistical mechanics of learning a rule." In: *Reviews of Modern Physics* 65.2 (Apr. 1, 1993), pp. 499–556. DOI: 10.1103/RevModPhys.65.499. URL: https://link.aps.org/doi/10.1103/RevModPhys.65.499 (visited on 08/11/2017).