ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Trust, law and ideology in a NN agent model of the US Appellate Courts

Nestor Caticha and Felippe Alves *

Instituto de Física, Universidade de São Paulo,
Caixa Postal 66318, 05315-970, São Paulo, SP, Brazil.

Abstract. Interacting NN are used to model US Appellate Court three judge panels. Agents, whose initial states have three contributions derived from common knowledge of the law, political affiliation and personality, learn by exchange of opinions, updating their state and trust about other agents. The model replicates data patterns only if initially the agents trust each other and are certain about their trust independently of party affiliation, showing evidence of ideological voting, dampening and amplification. Absence of law or party contribution destroys the theoretical-empirical agreement. We identify quantitative signatures for different levels of the law, ideological or idiosyncratic contributions.

1 Introduction

In this paper we study a system of agents which exchange opinions about issues to model the interactions of judges in three member panels of US Courts of Appeals. Each agent processes information using a neural network, emitting opinions and learning from the opinion of other agents. The level of distrust to other agents is also dynamically updated. The theory is general and can be applied in many situations. The particular modelled system is chosen because, in addition to its intrinsic interest, data is available in [1] [1]. We propose numerical signatures of behaviors that can be measured in both the theoretical and real systems. Predictions of exact behaviors for particular panels of judges are out of the realm of possibilities of our methodology. The neural networks initial set of weights reflect the fact that, first, judges have been trained in a common set of judicial knowledge base; second, that they have ideological biases associated to the executive in power who made their appointment; and third, have different personalities. The relative weights of these three contributions influence the model statistical signatures leading to conclusions about how judges interact. We can also quantify the influence that the initial attribution of distrust and its uncertainty have on the resulting dynamics.

2 Model of an agent and learning from surprises

An issue is represented by a set of numbers $\boldsymbol{x} = (x_1, x_2..., x_K)$, each representing the relevance of the issue along a given foundation of the law. For simplicity

*Work supported by CNAIPS, the Center for Natural and Artificial Information Processing Systems of the University of São Paulo. FA was supported by a CNPq graduate fellowship.
[1] We thank A. Sawicki for kindly sharing the raw data used in [1].

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

agents are modelled with the simplest possible neural network a perceptron. The state of an agent, call it i, is also given by a set of weights $\boldsymbol{w}_i = (w_{i1}, w_{i2}..., w_{iK})$ and a number C_i. Weight w_{in} is a measure of the importance given to foundation n by agent i and C_i is a measure of the uncertainty of the agent on its weights. Agent-i's private assessment is $h_i = \sum_{n=1}^{K} w_{in} x_n$. A for/against opinion on the issue under consideration is $\sigma_i = \pm 1$. The magnitude $|h_i|$ can be seen as a measure of the confidence on the opinion. Two agents establish a discussion on a common issue. An asymmetric interaction step occurs, first, when the opinion σ_e of the *emitter* agent e becomes available to the *receiver* agent r, and is completed when second, r changes its internal state. Previous to the interaction we can say agents disagree (respect. agree) on the issue if $h_r \sigma_e < 0$ (respect. $h_r \sigma_e > 0$).

We now present an abridged version of the theory. The state of agent i is described by the joint probability density distribution of its vector $\boldsymbol{w}_i$ and its level of distrust $0 \leq \epsilon_{j|i} \leq 1$ for each agent j engaging in the information exchange. We restrict these distributions to a parametric family $Q(\boldsymbol{w}_i, \epsilon_{j|i} \,|\, \boldsymbol{\lambda})$. Bayesian learning leads to a Bayes posterior outside the parametric family but determines the constraints that lead, using Maximum Entropy, to the λ of the max-ent posterior. A reasonable choice is to use a Gaussian parametrization of the distribution of weights, so only the expectation values $\hat{w}_{i,a}$, the components of $\hat{\boldsymbol{w}}_i$ and the covariance $\boldsymbol{C}_i$ (a matrix of components C^i_{ab}) have to be kept. For the noise sector we use a probit transformation of a Gaussian univariate density, which leads to a much simpler results than those obtained using e.g. a beta distribution. Again only mean $\mu_{j|i}$ and variance $s^2_{j|i}$ (or uncertainty) have to be analyzed. Therefore the set of parameters to update is $\boldsymbol{\lambda} = (\hat{\boldsymbol{w}}_i, \boldsymbol{C}_i, \mu_{j|i}, s^2_{j|i})$

For a particular issue $\boldsymbol{x}$, the emitting agent j has an opinion σ_j and the receiving agent i has an assessment $h_i = \boldsymbol{x} \cdot \hat{\boldsymbol{w}}_i$. The dynamics is given by the following equations that update of the mean and covariances that describe i (see [2, 3, 4])

$$\Delta \hat{w}_a^{t+1} = -\sum_b C^t_{ab} \frac{\partial \mathcal{E}_t}{\partial \hat{w}^t_b}, \quad \Delta C^{t+1}_{ab} = -\sum_{cd} C^t_{ac} C^t_{bd} \frac{\partial^2 \mathcal{E}_t}{\partial \hat{w}^t_c \partial \hat{w}^t_d}, \tag{1}$$

and

$$\Delta \mu_{t+1} = -s_t^2 \frac{\partial \mathcal{E}_t}{\partial \mu_t}, \quad \Delta s^2_{t+1} = -s_t^4 \frac{\partial^2 \mathcal{E}_t}{\partial \mu_t^2}, \tag{2}$$

where $\mathcal{E}_t = -\ln\left(\Phi\left(\frac{\mu_t}{\sqrt{1+s_t^2}}\right) + \Phi\left(\frac{\sigma^t h_t}{\gamma_t}\right) - 2\Phi\left(\frac{\mu_t}{\sqrt{1+s_t^2}}\right)\Phi\left(\frac{\sigma^t h_t}{\gamma_t}\right)\right)$ and $\Phi(z)$ is the cdf of a standard Gaussian. The expected value of the mistrust level is then $\hat{\epsilon} = \int_0^1 \epsilon P(\epsilon|\mu s) d\epsilon = \Phi\left(\frac{\mu}{\sqrt{1+s^2}}\right)$,which can be interpreted as the attribution by agent i of a probability that the opinion emitted by agent j is *wrong*. Note that $\hat{\epsilon}_{j|i} > 1/2$ (respect. $< 1/2$) for $\mu > 0$ (respect $\mu < 0$). The changes elicited by the arrival of information are intuitively simple to understand: learning is driven by surprises. We call a surprise the disagreement of a receiver with a trusted emitter or the agreement with a distrusted one. A surprise occurs when $\mu_{e|r} h_r \sigma_e > 0$ and the larger this number the more surprising this exchange

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

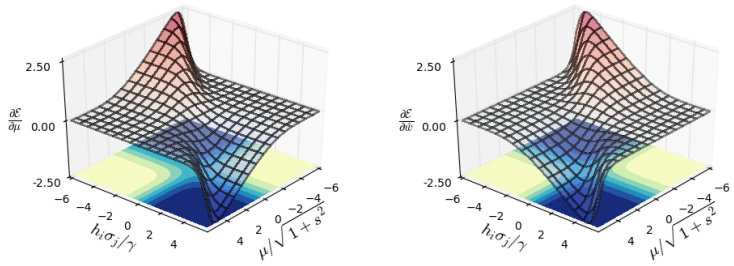

Fig. 1: Complementarity in the blame attribution: Prefactors that set the scales and signs of the changes in μ (left) and $\hat{\boldsymbol{w}}$ (right) respectively. The dark part of the floor are regions of high surprise.

has been. Learning is a process by which surprises are decreased. And this is obtained by blame attribution and then changing mainly the sector - either the weights of the foundations or the distrust - that needs the least change in order to remove the surprise.

The nature and scale of the changes depends whether the blame for the surprise is due to doubts about the assessment h_r being closer to zero or the distrust $\mu_{e|r}$ being closer to zero (neutral). A surprised receiver agent will change. It can either change significantly its internal state $\hat{\boldsymbol{w}}$ leading to a change in assessment. This happens if in doubt about an issue, $|h_r|$ small and $|\mu_{e|r}|$ large. Alternatively. it remains essentially unchanged in assessment of the issue but will change the distrust about the emitter. This happens if $|h_r|$ is large and $|\mu_{e|r}|$ is small. The relative scales are shown in the figures 2. The change in the uncertainties about $\hat{\boldsymbol{w}}$ and μ are not shown. The signs of the changes decrease the surprise.

We study the joint opinion-trust dynamics by simulating the exchange of information numerically. For a panel of N agents we choose the initial weight vectors $\hat{\boldsymbol{w}}_i^{t=0}$ and the initial matrix of distrust and its uncertainty $\mu_{j|i}^{t=0}$ and $s_{j|i}^{t=0}$ for $i \neq j$ ranging from 1 to N, which we restrict to $N = 3$. The the rich and complex behavior (glassy dynamics, phase transitions) that can occur for larger values will be presented elsewhere. The dynamics is as follows. First, an issue $\boldsymbol{x}$ is considered. Two agents are chosen at random, uniformly and independently of anything else. The first acts as the emitter and the second as the receiver on this particular time step of the dynamics. In later time steps, they may interact with any other agent in any of the two possible roles. An exchange of information is performed, the emitting agent sends its opinion σ_e and the receiving agent updates its weights $\hat{\boldsymbol{w}}_r$, the distrust and uncertainty about the emitting agent, $\mu_{e|r}$ and $s_{e|r}^2$. Then another pair is chosen and so on, until a stopping criterion is met. Typically the criterion is related to the fact that interesting changes

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

cease and further exchange of information will not change relevant aspects of the system. Other issues can now be chosen and we can start a new simulation.

3 A model for the Judicial Behavior

The source of the empirical data is [1], whose authors "claim to show both strong conformity effects and group polarization within federal courts of appeals."

The judges can be identified with the party of the appointing president. When looking at the decisions of a republican indicated judge in a panel of two republicans and one democrat, we use the notation $v = Rrd$, capitalizing the initial of the party of the judge under observation. There are six different types of votes v: $Rrr, Rrd, Rdd, Drr, Drd, Ddd$. Specifically, the data is from rulings in 14 areas of the law[2]. The data (e.g. fig 2-2 of [1]) , in the form of a 6×14 matrix containing the percentage of liberal votes, supports their three working hypotheses, that there is (i) Ideological voting: "Republican appointees vote very differently from Democratic appointees"; (ii) Ideological dampening: a judge in the minority party of a panel will be less ideological; and (iii) Ideological amplification: a judge in a pure party panel will be more ideological. Hence they are describing the interactions of the judges in the panel.

We represent their data as a set of 14 dimensional vectors $\boldsymbol{J}_v$, one vector for each v. The angles $\theta(v, v')$ between these vectors give a description of the difference between judges in different panels. For instance $\theta(Rrr, Ddd)$ measures the difference between Republicans and Democrats in pure panels, hence permitting to assess hypothesis (i) about Ideological voting. Comparing $\theta(Ddd, Rdd)$ and $\theta(Ddd, Drr)$ informs about how liberal is a Republican sitting with two democrats and whether it is more so than a Democrat sitting with two republicans, hence probing hypothesis (ii). The angle $\theta(Rdd, Rrr)$ measures the differences of judges in the minority or in a pure panel, relevant for hypothesis (iii). The angles $\theta(Rrr, Rrd)$ or $\theta(Drr, Drd)$ inform about the differences that occur in panels where a companion judge from one party is changed to the other party. The main reason to introduce $\boldsymbol{J}_v$ is that it can be constructed from readily observable quantities. Angles between the vectors that represent the state of the agents $\boldsymbol{w}$ are not empirically available for the judges since $\boldsymbol{w}$ states are only indirectly hinted from voting patterns.

The model considers a two parties (A and B) system. Three adaptive agents interact by exchanging their opinions about a particular issue to be judged. The initial state $\boldsymbol{w}_{i|I}$, where $I = A$ (respect. $I = B$) for agents appointed by party A (respect. by party B), of an agent at the beginning of a discussion reflects three main characteristics the judges ought to have, shouldn't have and simply have. These are, respectively, first a common knowledge of the law; second an ideological bias that depends on the political party I of the executive officer that made the appointment; and third a contribution that is particular to that

[2] Affirmative action, NEPA, 11th Amendment, NLRB, Sex discrimination, ADA, Campaign Finance, Piercing corporate veil, EPA, Obscenity, Title VII, Desegregation, FCC, Contract Clause, Commercial speech.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

agent. The simple mathematical structure of the agents permits a simple way to incorporate these ingredients. This is simply obtained by adding the three contributions $\boldsymbol{w}_{i|I}^{t=0} = \alpha_L \boldsymbol{L} \pm \alpha_P \boldsymbol{P} + \alpha_\eta \boldsymbol{\eta}_i$. The first term $\boldsymbol{L}$ represents *knowledge* of the Law, common to all agents. If this were the only term, agents would have identical opinions on every issue. The second term $\boldsymbol{P}$ represents ideological party lines, perpendicular to $\boldsymbol{L}$. The plus sign indicates an agent appointed by party $I = A$ and the minus, by party $I = B$. The third term is the idiosyncratic component contribution to the agent's position, $\boldsymbol{\eta}_i$ a vector independently chosen at random for each agent. The parameters α_L, α_P and α_η control the relative importance of the Law, the party and personality of the agents. The cases $\alpha_L = 0$ can be called *Lawless* cases and if $\alpha_P = 0$ *partyless* cases. With respect to the initial distrust attribution we consider four different scenarios, arising from relations between agents from different parties being courteous $\mu_{a|b} = -1$ or uncourteous $\mu_{a|b} = 1$, and being certain about it $s^2 = 0.1$ (small) or uncertain $s^2 = 5.0$.

An issue, characterized by its angle ϕ with the Law vector, is chosen and the agents engage in the exchange of opinions. There is a competition between the w dynamics and the μ, s dynamics. We repeat this for $n_{case} = 14$ different ϕ angles in the interval $[0, \pi]$. Then repeat this for a few hundred sets of initial conditions. For each run we record the voting patterns, and the averages are used to construct the 14 dimensional $\boldsymbol{J}_v$ vectors. Then, repeat for all v environments. The angles between two vectors indicate how two agents (or how two judges) are *aligned* in their views. Similar voting patterns will result in small angles. In figure 2 we present our main result, the angles between the vectors $\boldsymbol{J}_v$ obtained from the voting patterns of the judges and of the agents for different conditions. The dark entries represent large angles and different voting patterns, while light colors mean small angles or very aligned voting patterns. The result is that the model behaves similarly to the Appellate Courts if only the agents trust each other and are quite certain about this trust at the beginning of the interactions. But we see that, even in this optimistic scenario, judges appointed by party A behave differently from those appointed by party B, hence we see evidence of ideological voting, a reminiscent behavior of the party dependent initial conditions. We also see evidence of Ideological dampening, for example $\theta(Rrr, Ddd) > \theta(Rrd, Ddd)$ meaning that the difference between a Republican in a pure republican panel and a Democrat in a pure democratic panel is larger that of the same Democrat and a Republican who is interacting with one republican and one democrat. Interestingly a Republican in the presence of two democrats is more liberal that a Democrat in the presence of two republicans. The same results hold if we change R (and D) for A (and B) in the courteous-certain scenario. Also Democrats in the company of one republican and another democrat are more similar to Republicans in the presence of two democrats than to Democrats accompanied by two republicans. This again holds for agents. We can do more, simulating panels of purely ideological judges with $\alpha_L \gg \alpha_P \geq 0$. Also can look at purely non ideological judges, by taking $0 \leq \alpha_L \ll \alpha_P$. Both tests fail to agree with the data. We conclude that to obtain agreement with the empirical

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

signature, the agents have to be (i') quite courteous towards those of another party and certain about this trust; (ii') that there must be a sizable contribution of the common vector $\boldsymbol{L}$ to the initial conditions, (iii') that there is also a sizable contribution of the party bias $\boldsymbol{P}$. Quantitative evaluation of the components of the Law and the Party can be done for individual courts and is currently under study.

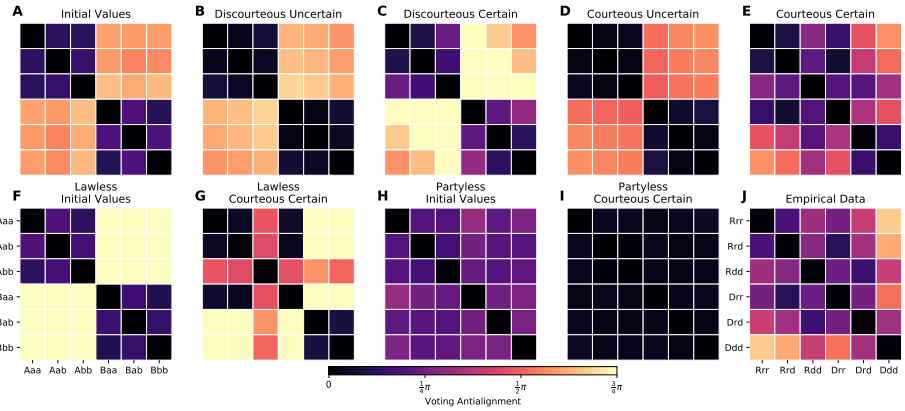

Fig. 2: Alignment angles $\theta_{vv'}$. Top row: all the scenarios have initial values depicted by the leftmost panel (**A**), the other four panels (**B**-**E**) show the asymptotic result for the four $\mu - s$ scenarios with $\alpha_L = \alpha_\eta = 1.0$ and $\alpha_P = 1.25$. Bottom row: (**F**) Initial and (**G**) asymptotic Law-less $\alpha_L = 0, \alpha_\eta = 1.0$ and $\alpha_P = 1.25$. (**H**) Initial and (**I**) asymptotic Ideology-less $\alpha_L = 1, \alpha_\eta = 1.0$ and $\alpha_P = 0$, for the courteous certain scenario. (**J**) empirical data. The reader should concentrate on the similarities of the two right panels.

We have just scratched the surface of the deep set of data that can be amassed from judicial courts. The confrontation of more detailed data with predictions of the model can lead to changes and thus to represent in a more useful manner systems of decision making agents under conditions of incomplete information. More importantly this process should lead to new questions. In particular it can lead to the development of tools to infer the changes in the contributions of law, ideology and personality through time.

References

[1] C R Sunstein, D Schkade, L Ellman, and A Sawicki. *Are judges political? : an empirical analysis of the federal judiciary.* Brookings Institution Press, Washington, D.C., 2006.

[2] M Opper. On-line versus off-line learning from random examples: General results. *Physical Review Letters*, 77:4671 – 4674, 1996.

[3] F Alves and N Caticha. Sympatric multiculturalism in opinion models. *AIP Conference Proceedings*, 1757(1), 2016.

[4] N Caticha, J Cesar, and R Vicente. For whom will the bayesian agents vote? *Frontiers in Physics*, 3(25), 2015.