Non-Thermal Transitions in a Model Inspired by Moral Decisions

Roberto C. Alamino

Non-linearity and Complexity Research Group, Aston University, Birmingham B4 7ET, UK

Abstract. This work introduces a model in which agents of a network act upon one another according to three different kinds of moral decisions. These decisions are based on an increasing level of sophistication in the empathy capacity of the agent, a hierarchy which we name *Piaget's Ladder*. The decision strategy of the agents is non-rational, in the sense they are arbitrarily fixed, and the model presents quenched disorder given by the distribution of its defining parameters. An analytical solution for this model is obtained in the large system limit as well as a leading order correction for finite-size systems which shows that typical realizations of the model develop a phase structure with both continuous and discontinuous non-thermal transitions.

PACS numbers:

1. Introduction

Human societies are inherently complex. While more than a century of social research has generated invaluable understanding of human activities, the limited amount of data available and the difficulty in obtaining more of it to guide and test theories have been a barrier to their efficient application in predicting behaviour to the same level of accuracy as that of the natural sciences [1]. Even though social scientists and psychologists are well aware that human behaviour presents patterns without which their very disciplines would not exist, the amount of data collected in the past was not enough to compensate for the high level of noise present in it.

This noise, one of the main barriers to the mathematical modelling in social sciences, is the result of a variety of sources of disorder affecting individual decisions concerning social interactions [2, 3]. People's reactions can be influenced in uncertain ways by, for instance, the mass media [4, 5], the availability of natural resources [6], the opinions of peers [7, 8, 9] or even instinctive responses which can be, in general, associated with irrational behaviour [10].

When the system is composed by a very large number of individuals though, as is the case for most human activities of interest, deviations are averaged away and typical behaviours can then be analysed with a precision that increases with the number of individual units. Statistical physics has a long history of dealing with systems of this kind which includes a large toolbox of methods to analyse emergent behaviours where fluctuations drive transitions between different phases of the systems [11] resulting in recognisably distinct patterns of macroscopic collective behaviour. In terms of social systems, these transitions might relate to the formation of consensus [12, 13], the manifestations of economic crisis [14] or even the onset of revolutions [15].

It did not take long for statistical physicists to apply their methods to social phenomena and several models inspired by them were studied as early as the 80's [16, 17]. Although most of those initial models were too simple to give precise quantitative predictions in real situations, they were able to provide important qualitative insights in the nature of social modelling and served as guides in identifying the relevant features of those systems.

Based on this pioneering work, posterior refinements allowed the inclusion of more realistic situations [18] and even, in some cases, achieve some success in reproducing and predicting observed behaviour in real social datasets [19, 20, 21, 22]. This gave birth to a whole new branch of physics, still in its infancy, which has been named *Sociophysics* [23]. Still, the importance of the initial, more simplified models cannot be underestimate as their solutions and key insights opened up the way to more sophisticated studies.

One of the most studied problems in sociology concerns moral behaviour and its consequences to the general well-being and survival ability of a certain population [24]. There is evidence that moral behaviour is the result of multi-level selection [25, 26], but the exact mechanism is still not understood. Although the picture is not yet complete, some of its parts are gradually becoming clearer. For instance, the essential role of

intuitive (emotional) judgements in the process of moral formation and decision making is now well accepted and has been recently used in the formulation of the framework known as Moral Foundations Theory (MFT) [27].

This essential role of emotions is, in fact, nothing but expected as motivation from emotional satisfaction, independently of cultural differences, is a factor that has been identified a long time ago as being decisive in the survival of an organism or a group even when their physical needs are fulfilled [28].

In this work, we present a simplified statistical mechanics model inspired by those studies in which interactions of agents in a network are interpreted as moral decisions which affect the overall "emotional satisfaction" of the network defined as a certain function of these decisions.

We will not be interested in the problem of emergence of particular moral beliefs and, therefore, we will assume that the concepts of what is morally acceptable is agreed *a priori* by the whole population. It is obvious that these concepts will vary and even exchange places in different societies, but this is out of the scope of the analysis presented here.

The model analysed here is infinite dimensional, and therefore mean-field, in the sense that each person from the group interacts with every other person. Individuals do that by taking a one-time binary decision to act either *helpfully* or *harmfully* relative to the group's agreed moral rules. We then classify the moral decisions according to the level of empathy embodied in them.

More than often, people tend to take moral decisions based on dogmatic systems like political affiliation or religion, what can be seen as choosing some fixed "moral strategy". Because there is an infinite number of possible strategies, we constrain our functional space in this work to what we call n-th order moral decisions, where the order of the decision identifies the increasing level of empathy necessary to take it based on the work of the well-known psychologist Jean Piaget [29]. Piaget observed that there are distinct cognitive stages in the development of the child intellect before it reaches the adult stage. This evolution occurs in three steps with increasing levels of empathic capacity. It is this sequence of three steps that we call *Piaget's Ladder* and which we use to define the order of a moral decision.

Piaget has proposed, based on empirical observations, that every children first develops a sense of self in which it is capable of understand its own feelings, but is unable to recognize the feelings of others, acting only selfishly. Accordingly, we call this step the Selfish Step and decisions taken selfishly are considered of 0th order. As its development proceeds, the child becomes capable of recognizing that others also have feelings, but it is still unable to see things from others' perspective. We call it the Parental Step (1st order) as it is not uncommon to parents to project their desires in the way they act towards their children. Finally, the cognitive abilities of the child reach a stage in which it can finally understand that others have different needs. This is the Empathetic Step (2nd order). The details of Piaget's original theory of cognitive development have been several times revised to account for further experimental evidence [30], but the details

will not be as important for our purposes, only its key idea of incremental capacity towards empathy.

The agents of our network will act on one another using moral decisions which are combinations of these three steps. These decisions are not planned in order to maximise their well-being (to be defined rigorously later) and simply follow a limited set of pre-determined rules. It is in this sense only that we say that they are not "rational decisions". Notice that this does not mean that some decisions which might be considered non-rational according to this criteria cannot be attributed to the minimisation of other energy functionals defined in different ways (see, for instance, [31, 32]). The average well-being of the group then defines a variable which can be used to characterise the macroscopic phases of the model. We will then be then interested in identifying these phases as the set of disorder parameters of the model is varied.

A detailed explanation of how the model is constructed will be given in Sec. 2 together with its analytical solution. The phase structure of this solution is obtained and analysed in Sec. 3. Finally, we present our conclusions and further discussions in Sec. 4.

2. The Model

Let us consider a population of N agents living on the vertices of a fully connected network. This represents a simplification compared to real situations as there is empirical evidence that human interaction networks tend to be scale-free [33, 34, 35]. A fully connected topology amounts for a mean-field approximation allowing for an exact analytical solution for the model, which provides qualitative insights into the macroscopic properties of the proposed system. Because the present model is still too simplified at this stage to provide quantitative predictions, it is important to explore its qualitative features in order to understand where (and whether) it can be refined or modified to better approximate actual human behaviour.

The interaction between agents will be asymmetric and occurs through the existing links connecting them. The variable J_{ij} represents the strength of the interaction of agent i on j and vice-versa with J_{ij} and J_{ji} not necessarily the same for each fixed pair of vertices (i, j). It is these interactions that we will interpret as *moral decisions* taken by one agent towards the other. These decisions will ultimately define the satisfaction state σ_i of each agent at a vertex $i = 1, ..., N$ of the network according to a rule to be given below.

It is very important at this point to highlight the similarities and differences of this setting with usual spin models. While the lattice and the notation is analogously defined, in our model we will consider the *interactions* J_{ij} as the dynamical variables and not the local spin variables σ_i. The analogue quantities to the local fields and the spin (to be rigorously defined later) will not behave in the usual way. In particular, the fields will influence directly the value of the interactions and not of the spins. The spin variables, which indicate the alignment of the population concerning their satisfaction

towards the assumed interaction configuration, will then be a function of the J_{ij}'s. This is in clear contrast to usual magnetic spin models in which the *spins* are the dynamical variables, the interactions directly multiply pairs of spins and the local fields contribute as an additive term in the Hamiltonian. What remains the same though is that the macroscopic phases will still be characterised by the magnetisation of the spin variables as we define in the following.

For simplicity, we consider only the case of binary interactions with $J_{ij} \in \{\pm 1\}$. We interpret this as modelling the fact that agent i can take either a *harmful* (-1) or *helpful* (+1) decision about how to interact with agent j, where the moral values of the actions are judged according to *a priori* moral concepts agreed by the entire network. Notice that neither choice is directly associated to objective physical or psychological harm, but is based on the subjective classification of these actions by the "community" in which the agent is inserted. Although it might seem a strong simplification, taking the judgement of an action as a binary choice is very common in real life situations and there is empirical evidence that people typically tend to polarise their opinions between two extremes with only a decreasingly small fraction aligning themselves with intermediate positions [36].

To each agent i of the network (i.e., each vertex), we associate a two-dimensional *personality vector* $\pi_i = (u_i, w_i)$ with binary components $u_i, w_i \in \{\pm 1\}$ acting as local fields influencing the alignment of the interactions between agents. In terms of moral interactions, they can be associated to the "emotional preferences" of agent i. The first component, u_i, is related to how the agent i wants (or would be "emotionally inclined") to act towards another agent. When $u_i = -1$, this indicates that the actions which agent i is inclined to choose are judged as harmful ones ($J_{ij} = -1$) according to the consensus in its network, while a value $+1$ would indicate an inclination for helpful actions ($J_{ij} = +1$). The assumed *emotional* nature of these preferences is intended to mean that fulfilling them leads to a subconscious satisfaction of the agent. Conversely, the second component, w_i, represents how the agent i would like to be treated by others and influences the value of J_{ji}.

Assuming binary variables to represent human interactions is usually a strong simplification. This is why we chose to present this model as *inspired by* rather than *modelling* moral decisions. In order to approximate more realistic scenarios, one would have to consider the possibility of continuous distributions. It is interesting to notice, however, that even psychologists make use of discrete personality classifications of human personalities, the best known example being the Myers-Brigg Type Indicator (MBTI) [37]. MBTI classifies human behaviour by considering only 14 archetypes and is largely used by institutions in actual selection processes and also by career advisers.

Although we consider u_i and w_i dependent only on the agent i, it is not difficult to see that the desire of how to act towards another agent might depend on who the other agent actually is due, for instance, to emotional attachments or resentments. Therefore, it would be more appropriate to consider matrices u_{ij} and w_{ij} adding a dependency on the other agent index j.

In realistic interactions, time can be another important factor in the long term as the personality vector might change with it. This is because emotional responses are not only genetically programmed, but are affected by the agent's experiences. The scenario studied in this work does not deal with dynamics and, therefore, this simplification is not relevant in the present case. It is however improbable (although not impossible) that a significant change in behaviour which could affect the whole population occurs in a short time scale. Therefore, we consider stable personalities for a certain macroscopic period of time as a first approximation.

Given the two *personality components* of π_i, whether or not agent i feels fulfilled by the interactions within its network can then be represented by how aligned the actions it takes and receives are with its personality, what we call its (individual) *satisfaction*

$$\sigma_i = \operatorname{sgn} \Omega_i(\pi, J, \gamma), \tag{1}$$

with

$$\Omega_i(\pi, J, \gamma) = \frac{1}{N}[\gamma U_i + (1-\gamma) W_i], \tag{2}$$

and

$$U_i = u_i \sum_{j \neq i} J_{ij}, \qquad W_i = w_i \sum_{j \neq i} J_{ji}. \tag{3}$$

This definition allows three values for the agent's satisfaction. When $\sigma_i = +1$, we say that the agent feels satisfied, when $\sigma_i = -1$ it feels dissatisfied and when $\sigma_i = 0$ the agent is neither.

The scaling $1/N$ in Ω is used to make it an intensive quantity in the number of agents, keeping it finite when $N \to \infty$. Rigorously, due to the sign function, it can be shown that the analytical results will not depend on the chosen scaling in the present case. The large N limit is the one we are interested in as it is only when the number of agents is *exactly* infinite that we can unambiguously identify well separated macroscopic phases and their transitions in the model.

The extensive quantity U_i is the sum of all contributions to the fulfilment of the agent's desire on how to treat other agents, while the also extensive W_i represents the same concerning how the agent feels treated by the others. The parameter γ is taken from the interval $[0, 1]$ for convenience and represents the relative *emotional* importance given by the individual to each of these terms. This parameter will be kept fixed and will be the same for the whole network. More realistically, γ should depend on each agent i and therefore represent another source of disorder in the model. However, we will see that for this simple model the disorder will smooth out the transitions at finite disorder parameters.

For the purpose of the present analysis, we assume that the personality parameters are i.i.d. with distribution

$$\mathcal{P}(u) = (1-s)\delta(u, 1) + s\delta(u, -1), \tag{4}$$

$$\mathcal{P}(w) = (1-m)\delta(w, 1) + m\delta(w, -1), \tag{5}$$

and $s, m \in [0, 1]$ (s and m the same for all agents).

Although there are formal differences in the way the Hamiltonian is constructed in the present model, it shares many conceptual features with previously studied social models in physics. The presence of random local fields influencing the decisions of agents by acting as moral beliefs or simply personal tastes have been used and discussed in classic papers as [12, 31]. In those works, local fields are also the result of a combination of different influences that include personal beliefs, societal norms and the influence of the group within which the agent is interacting. The objective is also the same - to determine how the parameters of the model influences the collective behaviour, reflected in a macroscopic order parameter, of the individual states as a result of the binary decisions taken by the agents.

One key difference though is that we are not interested here in measuring the *moral* alignment of the population, which would relate to the dynamical variables (spins) of the cited models, but a qualitatively different quantity - their satisfaction. In our model, as we explained, this is given a priori in a way that the whole group shares the same moral beliefs. In terms of the models in the above cited papers, this would correspond to a consensus in which the whole population is polarised according to some moral paradigm. The moral actions we study can, in fact, be taken in contradiction with this alignment.

There are two possibilities to proceed from here. One of them is to consider a thermodynamic equilibrium scenario. It can be argued that a the net satisfaction $H(J) = -\sum_i \sigma_i$ of the network is a natural choice for the system's Hamiltonian as, in this case, minimisation of $H(J)$ would lead to the maximisation of the satisfaction for the majority of the population. The dynamical variables would then be the moral decisions J_{ij} (analogous to the spins in a magnetic model), while the personality vectors would play the role of the quenched disorder. As discussed previously, it is important not to be confused by the notation used here though. While the Hamiltonian formally resembles usual magnetic spin models of independent spins σ_i under the influence of a positive constant field, this is not the scenario described by our model. The σ_i's in our Hamiltonian, as we stated before, are not dynamical as would be the case in the usual Ising model, but are functions of the dynamical J_{ij}'s and the local fields. Although the role of the σ_i's are still to indicate a sort of alignment of the population (satisfied/not satisfied), this alignment is not directly prone to choice, but it is the consequence of the agent's actions represented by the J_{ij}'s.

At inverse temperature β, the Gibbs distribution for the system is $\mathcal{P}(J|\pi) \propto e^{-\beta H(J)}$. Although the calculation of the partition function is involved and out of the scope of the present paper, the ground state of this model is easy to find. For $\gamma > 1/2$, it corresponds to $J_{ij} = u_i$, while for $\gamma < 1/2$ it is given by $J_{ij} = w_j$. Depending on the value of the disorder, the ground state has a non-analyticity which reminds (although rigorously being technically different from) quantum phase transitions [38]. Here it is clear that the non-analyticity that appears for some values of the disorder is a consequence of our choice of sign for the satisfaction. Although it might be argued that this would disqualify it as a true phase transition, there is still a sharp change

between two macroscopically different behaviours of the system and, for that reason, we will continue to use the terminology.

In this thermodynamic scenario, the agents work towards a maximisation of the net satisfaction and it seems reasonable to say that they are acting rationally as they change their decisions towards the achievement of that goal. While the game-theoretic definition of rationality is selfish in the sense that each agent will try to maximise its own utility, we are using here a generalisation of this concept where the group plays the role of a super-agent maximising its group utility. However, as discussed in the Introduction, there are several reasons why real persons might not act rationally towards some goal. For instance, they might lack either the will or the resources to follow the strategy based on minimising the energy. In addition, most people tend to avoid the conflict between irrational actions and their conscious justification by adopting dogmatic behaviour systems like religions or similar ideologies.

Inspired by these observations, we will instead analyse the behaviour of our minimal model by assuming that agents act according to some pre-defined fixed strategy concerning the choice of its actions towards others. We will then consider J to be a quenched instead of a dynamical variable, completely eliminating the thermal aspect of the problem. Averages are then going to be taken over quenched realisations of both J and π at the same time. This alternatively can be seen as if the system was at zero temperature, being at the (possibly degenerate) ground state of an appropriately chosen Hamiltonian as in the case of [31].

One feature that however is not captured by the totally quenched scenario we analyse is the characteristic degenerate ground state of random field/interaction spin glasses. This effect, due to the frustration of interactions between neighbouring spins which cannot be satisfied simultaneously, has very interesting interpretations and is the cause of important collective effects when applied to social or political situations [39, 32].

Clearly, the range of possible strategies for choosing J_{ij} are infinite, which requires us to choose some more restricted functional space in order to get meaningful results. The specific distribution of the agent's actions we chose is a key feature of this work. Inspired by the previously cited work of Jean Piaget on the cognitive development of children, we identify what we call n-th order moral decisions where the order n associates the strategy with the corresponding step on Piaget's Ladder. At 0th order, the agent chooses $J_{ij} = u_i$ as it acts selfishly concerning its own desires. Analogous considerations lead us to associate $J_{ij} = w_i$ to the 1st order choice, in which the agent treats another one as it would like to be treated, and $J_{ij} = w_j$ to the 2nd order when the agent treats others as it knows they want to be treated.

More generally, we consider here a convex combination of the three moral strategies

$$\mathcal{P}(J_{ij}|\pi_i, \pi_j) = p_0\delta(J_{ij}, u_i) + p_1\delta(J_{ij}, w_i) + p_2\delta(J_{ij}, w_j), \tag{6}$$

with

$$\sum_i p_i = 1. \tag{7}$$

The probabilities in the above equation can be seen either as the proportion of individuals choosing one of the three levels of moral decision or as the probability of one individual taking each of them in each realisation of a network configuration. Notice that, because the 2nd order decision requires knowledge of the personality vector of others, we will assume that agents possess full noiseless information about the w component of all other agents. An interesting (and more realistic) scenario would be when only partial information is available, adding an extra source of disorder and the possibility of considering learning scenarios [13].

Finally, the *average satisfaction* of the whole network is measured by averaging the individual satisfactions over the decision strategy and over the disorder in the personality vectors as

$$S = \langle \sigma_k \rangle_{\mathbf{u},\mathbf{w},J}, \tag{8}$$

in which the index k inside the average is only for convenience as it disappears due to the averaging over all variables u_i, w_i and J_{ij} represented by the vectors $\mathbf{u}$ and $\mathbf{w}$ and the matrix J.

We will take S as the *order parameter* as it is a convenient measure of the macroscopic phases of the model which we would like to associate with the level of well-being or satisfaction of the whole network. Totally ordered phases would then correspond to situations where $|S| = 1$ and the whole population is either satisfied ($S = +1$) or dissatisfied ($S = -1$) except for a set of zero measure. Other values of S correspond to partially ordered ($0 < |S| < 1$) or totally disordered ($S = 0$) phases. In the next section, we analyse the different behaviour of the system as the parameters of the model are varied.

3. Phase Structure

In the limit $N \to \infty$ of an infinitely sized network, for J_{ij} i.i.d. according to the distribution (6), one can obtain the exact value of the average satisfaction using the Central Limit Theorem as (see appendix Appendix A for a detailed calculation)

$$S = \langle \operatorname{sgn} \mu \rangle_{u,w}, \tag{9}$$

where

$$\mu = p_0[\gamma + (1-\gamma)\bar{u}w] + p_1[\gamma uw + (1-\gamma)\bar{w}w] + p_2[\gamma\bar{w}u + (1-\gamma)], \tag{10}$$

and

$$\bar{u} = \langle u \rangle = (1-2s), \qquad \bar{w} = \langle w \rangle = (1-2m). \tag{11}$$

It is interesting to look at each moral strategy separately to get some insight into the behaviour of the model. In the following, we do this and compare the results with a strategy in which agents choose randomly between the three different orders of moral strategies, i.e., $p_0 = p_1 = p_2 = 1/3$.

3.1. 0th Order Moral

When $p_0 = 1$ then $p_1 = p_2 = 0$ and we interpret this as agents acting selfishly, i.e., they are acting according to their own desires without regard for other's. In other words, agents are taking moral decisions which maximise their own U_i. The expression for the average satisfaction in this case simplifies to

$$S = (1-m)\,\mathrm{sgn}[\gamma + (1-\gamma)(1-2s)] + m\,\mathrm{sgn}[\gamma - (1-\gamma)(1-2s)]. \tag{12}$$

When $\gamma > 1/2$, the system is in the fully ordered phase $S = +1$ for all values of m and s. This is simply a consequence of the fact that the U_i term is dominant with a value $\sim N$ for large N, which can be understood as if agents care much more about what they do, then everyone will be happy by acting selfishly as harmful actions will not have a relevant weight on their overall satisfaction. The case $\gamma = 1/2$ is degenerate, giving

$$S = (1-m)\,\mathrm{sgn}\,(1-s) + m\,\mathrm{sgn}\,s. \tag{13}$$

As $s \in [0,1]$, the above expression gives 1 for all values of m except for $s = 0, 1$, in which case linear relations with m are obtained. For this case, only when all agents are dissatisfied the contributions coming from the W_i's will be large enough to compensate those coming from the U_i's. Fluctuations in the number of satisfied individuals in the population will then tip the delicate balance between these contributions one way or another.

The most interesting cases are when $\gamma < 1/2$, which means that each agent gives more importance to the how it is being treated by others than to the way it treats other agents. Fig. 1 shows the result of the exact expressions and simulations for $\gamma = 0.2$. The simulations are run with a population size of $N = 1000$ agents and averaged over 40 independently generated realizations of J and π configurations. The diagram at the centre is the result of the simulation, the one at the left is the theoretical value for $N = \infty$. The difference between the two diagrams is due to finite size effects.

The diagram to the right represents the leading order theoretical correction for the system's finite size. Because σ^2 in equation (A.21) from appendix Appendix A is identically zero when $p_0 = 0$, we need to consider the variance coming from the average over the u_i's. This leads to the expression

$$S(N) = (1-m)\,\mathrm{erf}\left[\frac{\gamma + (1-\gamma)(1-2s)}{\sqrt{2[1-(1-2s)^2]/N}}\right] + m\,\mathrm{erf}\left[\frac{\gamma - (1-\gamma)(1-2s)}{\sqrt{2[1-(1-2s)^2]/N}}\right]. \tag{14}$$

One can see that three very distinctive bands characterised by their values of S appear in this situation. The central band represents the fully ordered phase $S = 1$. Its width decreases with γ as when more weight is given to the W_i's, a smaller amount of dissatisfaction creates a macroscopic fraction of dissatisfied agents which is more accentuated when the distribution of w_i's is more biased. This width can be calculated in terms of the variable $\epsilon = 1/2 - \gamma$ with $0 < \epsilon \leq 1/2$. The ordered band requires the

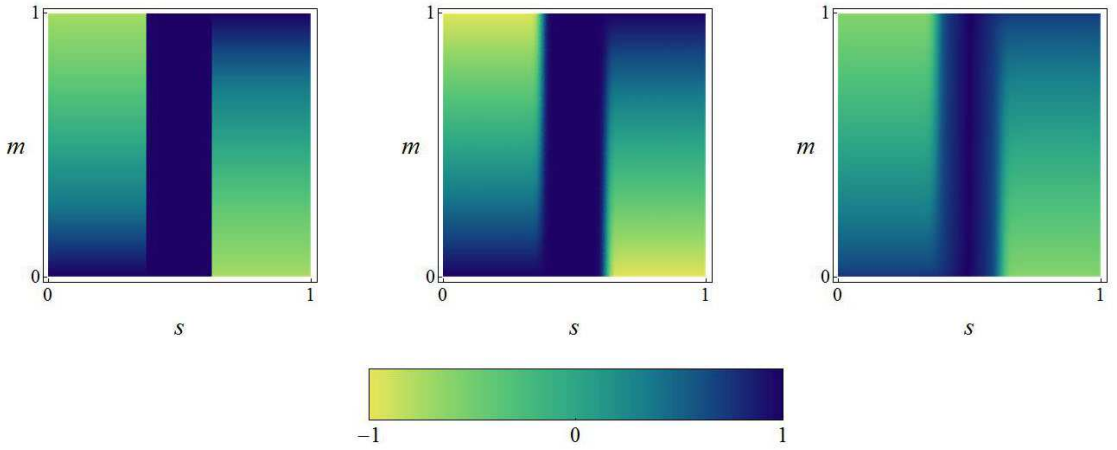

Figure 1. Phase diagrams of selfish decisions for $\gamma = 0.2$ (colour online) represented in the $s \times m$ plane with the real-valued order parameter $S \in [-1, 1]$ given by the colour scheme at the bottom of the picture. Left: exact result for the infinite system; middle: computer simulation with $N = 1000$ averaged over 40 independently generated configurations; right: leading order analytical approximation for $N = 1000$. The central band (online blue) is totally ordered and its width is given by $\delta = (1 - 2\epsilon)/(1 + 2\epsilon)$ with $\epsilon = 1/2 - \gamma$. In the lateral bands, the order parameter S varies linearly from -1 to 1 from top to bottom in the left band and from bottom to top in the right band.

arguments of the two sign functions to be positive independently of the value of m, i.e.

$$\gamma + (1 - \gamma)(1 - 2s) > 0, \qquad \gamma - (1 - \gamma)(1 - 2s) > 0. \tag{15}$$

This implies

$$\frac{2\epsilon}{1 + 2\epsilon} < s < \frac{1}{1 + 2\epsilon}, \tag{16}$$

and the bandwidth is therefore

$$\delta = \frac{1 - 2\epsilon}{1 + 2\epsilon}. \tag{17}$$

Notice that the area of the diagram in which the whole network is fully satisfied with the distribution of moral decisions is larger than or *at most* equal to that in which it is not. This holds for *any* value of γ due to the symmetry of the lateral bands. This is a disheartening result for the network represented in this model as it shows that trying to derive moral decisions from rational principles that objectively satisfy the society as a whole might not be attainable. Such a principle would force one to accept that completely selfish decisions can be justified as morally correct, even when they are meant to do harm to those who do not want it. Of course this result is based on a simplified model and its robustness has to be tested against modifications of it. For instance, while the sharp transitions between bands is softened by disorder in γ, this result still holds, illustrating that tying moral concepts to rational measures of social well-being *in general* does not work.

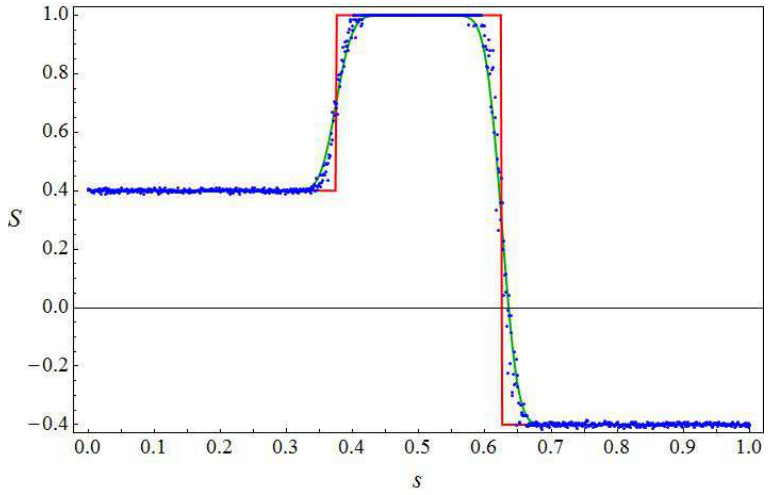

Figure 2. The graph shows the order parameter S along a line of constant $m = 0.3$ in the phase diagrams of fig. 1 (colour online). The dots (online blue) represent the values of the simulations, the smooth line closer to them (online green) is the leading order finite size approximation and the non-smooth distribution (online red) is the infinite system. Notice that in this case $\gamma = 0.2$, giving a width of 0.25 for the ordered central band with edges at $s = 0.375$ and $s = 0.625$.

The structure of the phase changes in the diagrams is also interesting. By keeping m constant and varying s from zero to one, we pass through two discontinuous transitions. We start with a partially ordered phase which becomes fully ordered at the central band and then becomes partially ordered once again, but with an average satisfaction with the opposite sign. These transitions are shown in fig. 2. Inside each of the two partially ordered bands, by keeping s fixed and varying m in the interval $[0, 1]$, we have a continuous linear crossover between the two fully ordered phases with opposite signs of satisfaction which occurs only at the extremes.

Another interesting phase diagram is obtained by plotting the values of S in a $\gamma \times s$ diagram, what is shown if fig. 3 for the fixed value $m = 0.8$. One can clearly see the growth of the central ordered band as the parameter γ increases from zero to 1 separating two lobes representing the partially ordered phases of opposite sign.

Fig. 4 shows a continuous transition in $\gamma = 1/2$ by taking as the order parameter $\bar{S}$, the average of S over the disorder parameters s and m.

3.2. 2nd Order Moral

It seems natural to argue that, from a moral point of view, 2nd order moral decisions in which one takes into consideration the desires of our neighbours are better, or at least more considerate, than 0th order selfish ones. The former is based on the idea of respecting others, while the latter bears no consideration for other's feelings. The simple structure of the minimal model we present was directly inspired by a consideration of how the interactions would affect the satisfaction of each agent. This construction leads

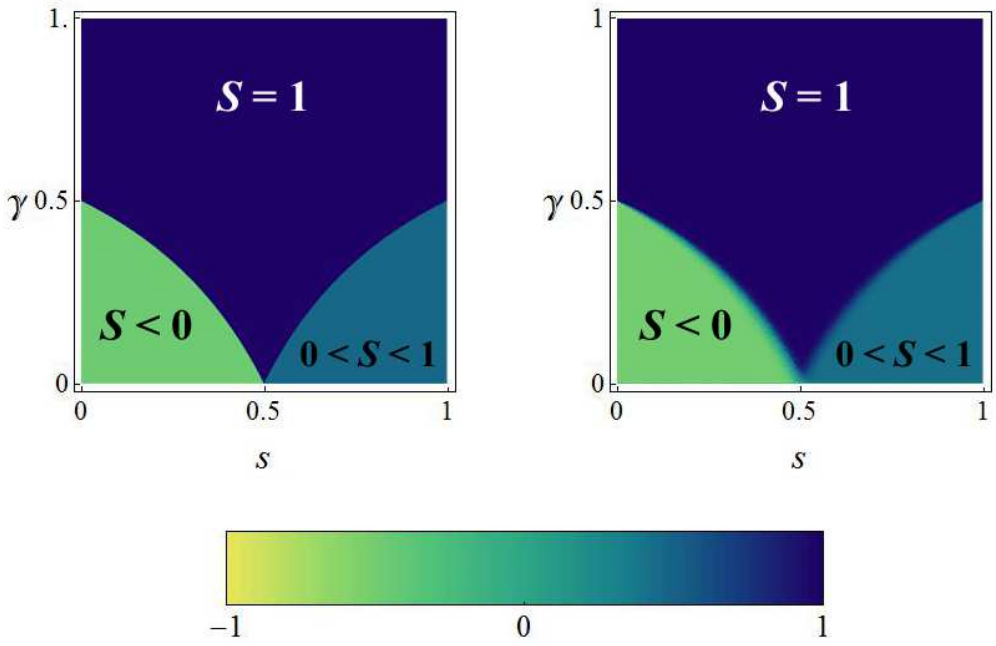

Figure 3. Phase diagram (colour online) showing the discotinuous transitions when γ and s are varied for constant $m = 0.8$. The left graph shows the result for an infinite system, while the right one shows the simulated diagram for $N = 1000$.

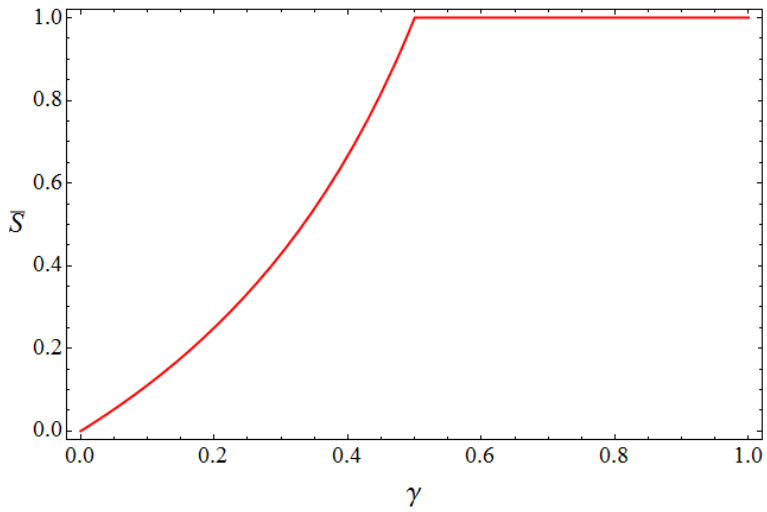

Figure 4. Continuous phase transition from a disordered to an ordered phase at $\gamma = 1/2$ with the order parameter taken as the average over the disorder in the personality vectors of the network satisfaction ($\bar{S}$) for 0th order moral decisions.

to a symmetry connecting 0th and s2nd order moral decisions. The substitution

$$\gamma \to 1-\gamma, \qquad u_i \to w_i, \qquad J_{ij} \to J_{ji}, \tag{18}$$

keeps all σ_i the same and, therefore, also the average satisfaction S. The effect on S is equivalent if we change the last substitution by

$$p_0 = 1 \to p_2 = 1. \tag{19}$$

This means that, within this model, changing from selfish to empathetic moral decisions exchange the vertical and horizontal axes of the diagrams at while $\gamma \to 1-\gamma$. Exactly as in the selfish case, the area of the diagram in which the satisfaction is positive is always larger than the negative one for all values of γ.

Although the present model is admittedly much simpler than a real human society, its core is based on socially reasonable assumptions. Given the agreed concepts of morality, one can then see that two strategies that can be considered as completely opposite morally, when taken by the *whole* population guarantee the well-being of the network in the absolute majority of the parameter space. Here one can then appreciate the ambiguity of attaching the concept of morality to "rational" arguments concerning the overall well-being of a population as both behaviours would be considered equally moral. None would be more harmful to the network than the other.

This result seems paradoxical and one might be tempted to consider the model nonsensical. This behaviour is however not a problem with the model, but arises from associating morality with satisfying the wills of the majority. Satisfying the majority is many times equivalent to ignoring or oppressing the minority, which is (at least in the 21st century) not morally acceptable. The paradox disappears if one recognizes that what can be associated with the definition of morally acceptable behaviour is in fact the value of γ. We tend to consider behaviours to be morally acceptable when everyone is being respected by others, which is equivalent to count only the contributions of the W_i's for the satisfaction. In other words, this is the situation when $\gamma = 0$. For the selfish strategy, a trivial calculation shows that it results in $\bar{S} = 0$ while for the empathetic strategy this gives $\bar{S} = 1$, what would be more sensible as a criteria for moral classification.

3.3. 1st Order Moral

There is no symmetry connecting 0th and 2nd order moral decisions to 1st order ones. Taking only 1st order moral decisions creates a different, although similar in some aspects, phase diagram. The argument of the sign function simplifies in this case to

$$\mu = \gamma u w + (1-\gamma)(1-2m)w, \tag{20}$$

and the average satisfaction is then

$$S = (1-2m)\{(1-s)\,\mathrm{sgn}[\gamma + (1-\gamma)(1-2m)] + s\,\mathrm{sgn}[-\gamma + (1-\gamma)(1-2m)]\}. \tag{21}$$

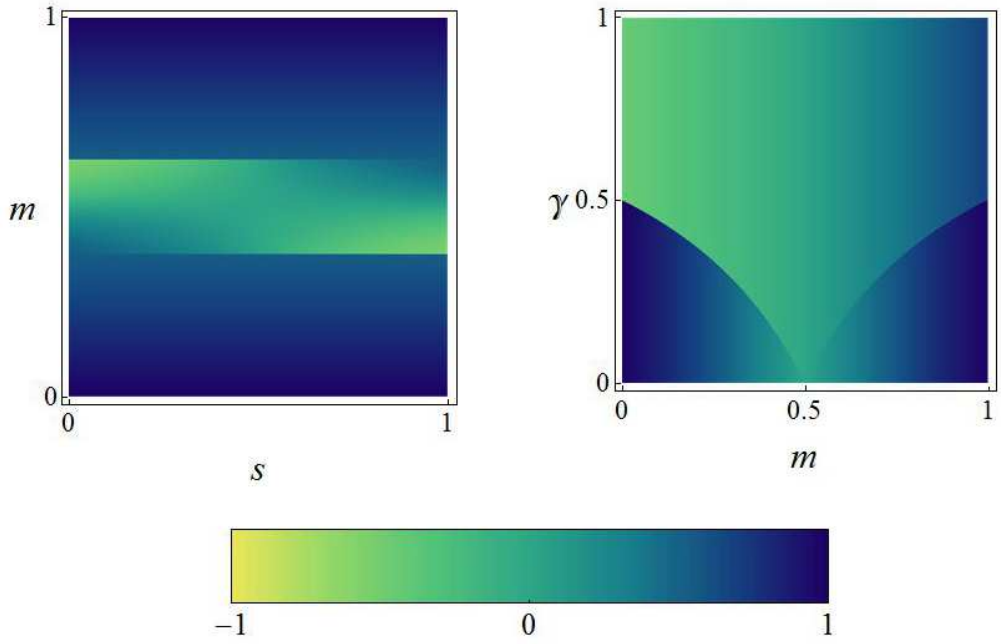

Figure 5. Phase diagrams for the 1st order moral decisions (colour online). Left: $s \times m$ phase diagram for $\gamma = 0.2$ showing the discontinuous transitions on the borders of the central band. Right: $m \times \gamma$ diagram for $s = 0.8$. One can see the continuous transition in which the central band disappears completely at the value $\gamma = 1/2$.

Analogous phase diagrams to the ones for the other strategies are given in fig. 5 which shows a similar structure concerning the discontinuous transitions, but with different details specially on what concerns a clear definition of the phases. For instance, the $s \times m$ diagram still presents a central band, which is however not totally ordered. It is also of opposite sign compared to the 0th and 2nd order strategies, with the central band indicating a partially ordered phase with negative satisfaction. The limits and size of the band are however the same as for the other strategies, which also implies the presence of a continuous transition in $\bar{S}$ when γ is varied. This transition, depicted in fig. 6, is between a partially ordered and a totally disordered phase, differently from those for the other strategies.

This also shows that when $\gamma = 0$ we have $\bar{S} = 1/2$. If we use the criteria suggested before, this is still a behaviour which should not be considered morally acceptable, but would be *more* acceptable than the completely selfish decisions.

3.4. Mixed Strategy

For the sake of completeness, we consider a comparison of the three "pure" strategies analysed above with the mixed strategy for which $p_0 = p_1 = p_2 = 1/3$. In this case, agents simply choose randomly between the three kinds of decision with the same

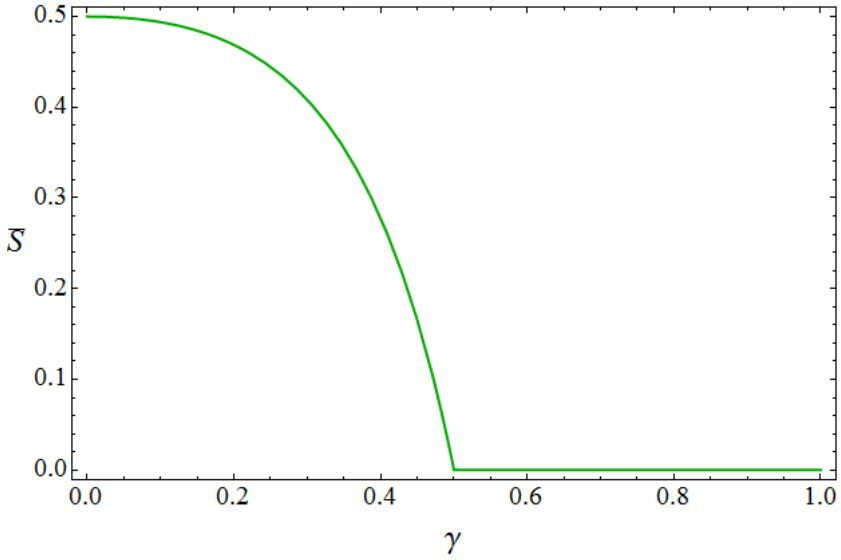

Figure 6. Continuous phase transition from a partially ordered ($\bar{S} = 1/2$) to a disordered phase at $\gamma = 1/2$.

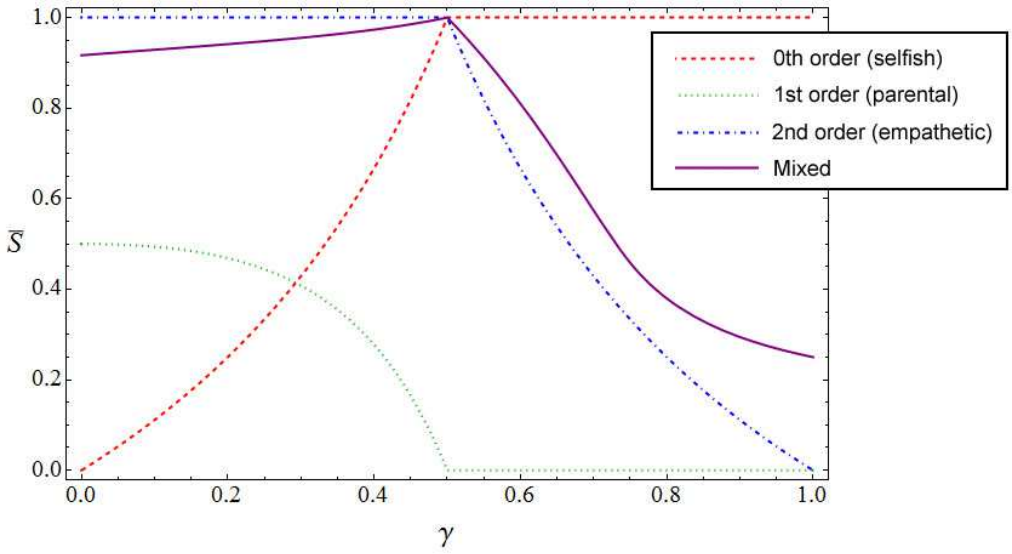

Figure 7. Continuous phase transitions on $\bar{S} = 1/2$ for the four strategies at $\gamma = 1/2$.

probability. Fig. 7 shows a plot comparing the satisfaction averaged over the entire parameter space for three pure moral decisions and the mixed strategy. All strategies present continuous transitions in $\bar{S}$ for $\gamma = 1/2$. The discontinuities in the derivatives at this point characterise the non-analytical behaviour which occurs only for $N = \infty$. Although the mixed strategy remains the second best throughout all range of γ, it only attains $\bar{S} = 1$ at exactly the critical value of γ.

As discussed before, it seems reasonable to associate some sort of *degree of morality* to each strategy by their limiting values at $\gamma = 0$. The reason for this difference can be

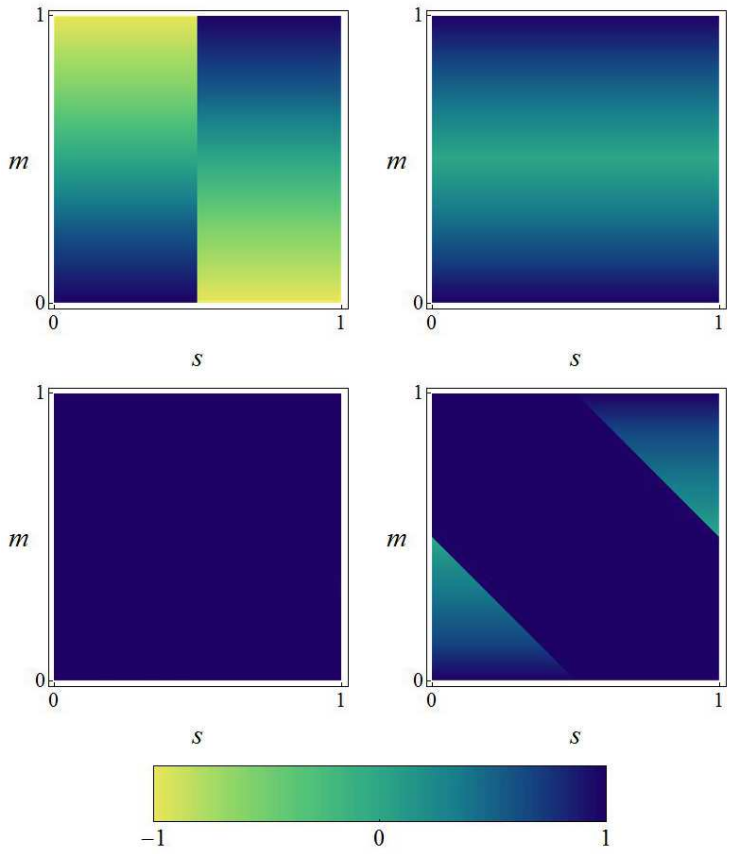

Figure 8. Phase diagrams for the four strategies at $\gamma = 0$ (colour online). From left to right. Top row: 0th order, 1st order. Bottom row: 2nd order, mixed.

better understood by looking at the phase diagrams $s \times m$ for all strategies at $\gamma = 0$ given in fig. 8, where one can clearly see the macroscopic effect of the difference strategies in the whole parameter space.

4. Conclusions

In this work we analysed a model of agents interacting in a fully connected network inspired by moral decisions in human social networks. The agent's decisions were constrained by an increased level of empathy based on the stages of Piaget's theory of cognitive development. Within this model, we defined a measure of (emotional) satisfaction related to the fulfilment of the agents' personal desires to characterise macroscopic phases of the network as some parameters of the model are varied. These desires were modelled by local random binary fields at each site of the network and were called *personality vectors* as they can be interpreted as personality traits particular to each agent. Although sharing many similarities with other sociophysics models [12, 39, 31, 32], there are some key differences (like the role of dynamical variables)

which differentiate it from them.

Being of a mean-field type, the introduced model is exactly solvable. We then presented the analytic expression for the average satisfaction of the network in the limit of an infinite number of agents and also its leading order contribution in powers of $1/N$. Using the average satisfaction as an order parameter, we then were able to find the model's phase diagrams and to identify the existence of both continuous and discontinuous transitions triggered by the variation of the disorder parameters which are linear functions of the means of the personality distributions.

We chose to analyse four different specific strategies of moral decisions in this work. In the first three, agents take decisions according to each stage in Piaget's theory, a three-steps hierarchy which we named *Piaget's Ladder*. Each step in this ladder is associated with an n-th order moral. The 0th order corresponds to completely selfish decisions and the 2nd order to completely empathetic ones. The 1st order corresponds to an intermediate case of partial empathy. We finally compared these three strategies with the case in which each agent takes a decision by randomly choosing one of those strategies at each realisation of the model's configuration, which we called a *Mixed Strategy*.

One of the possible interpretations for the Mixed Strategy is a society in which agents are at different stages in their lives and, therefore, possess different levels of empathy when interacting with others. An alternative scenario, which could possibly be used to obtain empirical data to be compared with the presented model (suggested by one of the anonymous referees), is that of different societies of primates in which the development of the levels of empathy described by Piaget's Ladder may be either limited or non-existent [40, 41].

Interestingly, there is a symmetry between the 0th and 2nd order moral decisions resulting in the fact that the selfish strategy guarantees the *average* satisfaction of the network as well as the empathetic one. Although the model is very simple, its assumptions are reasonable enough to indicate that moral beliefs cannot be simply based on rational minimization of some energy function as two strategies that would clearly be considered morally opposite by most people lead to same size areas of positive satisfaction of the network. One possibility is that the parameters of the disorder are tuned in humans to values for which a species-wide agreement on moral concepts can be derived. A different resolution can be achieved if instead of classifying the moral status of actions based on the overall satisfaction of the network, one considers the particular case in which the satisfaction of *others* is more important than ours, which corresponds to $\gamma < 1/2$. This seems to lead to a sensible classification when all four strategies are compared. A more realistic classification would clearly require a higher level of sophistication, like the one sought by MFT. This would be particularly important in scenarios in which there is a change in the moral classification given by the agents to their actions, introducing a dynamics in the model [21]. The connection of statistical physics models with conceptual frameworks based on empirical data like the MFT is an important aspect of modelling real datasets. Its implementation is however out of

the scope of the present analysis which considers a static situations, but we intend to include it in future works where dynamics play a key role.

From the purely mathematical point of view, the present model has several interesting properties. It is simple enough to be easily interpretable and completely solvable. At the same time, it presents a range of interesting phase diagrams and transitions. There can be some debate on whether these are *real* phase transitions in the sense that the effect of the average over the disorder in the large N limit is equivalent to integrating over a Dirac delta function in their means. However, there are indeed clear regions with different macroscopic behaviours of the system characterisable by values of S ranging from zero (when the system is disordered) to 1 (when it is totally ordered). It is true that disorder in γ, a more realistic situation in human societies, destroys the transitions by smoothing out the non-analytical behaviour of the order parameters, but this is not only a feature specific to this model as there are other statistical physics models in which any amount of disorder destroys any finite temperature transition [42].

The analysis in this paper has focused on the mathematical behaviour of the presented model. It is without doubt that in order to evaluate the actual contribution of it one needs empirical data that might be compared with the simplified scenario given in this work. One possibility already mentioned above was the study of animal groups, in particular of primates. Another possibility would be to collect data, for instance, on the prevalence of corruption in certain societies. Corruption is usually seen as morally wrong, but the degree in which it is ingrained in different societies differ to the point of being endemic in some [43, 44] due to the net reward coming from ineffective punishment. Comparing our model with data collected in this subject could, for instance, identify ranges of parameters that could be influenced, by means of governmental policy, in order to reduce corruption levels still maintaining the overall satisfaction of these groups.

Finally, there are many directions in which this model can be extended in order to include more realistic behaviour. For instance, as several studies on networks have revealed, the topology of human interactions is better modelled by scale-free rather than fully connected networks [33, 34, 35]. One could include dynamics in a way similar to that on [21] in which agents try to infer by a learning algorithm the desires of others when information about it is not freely available. The personality vector also can be extended either to a higher dimensionality to include more realistic personality variations or to continuous instead of binary values. The relevance of theses modifications have however to be considered in the light of how much they would help in modelling real data. We intend to explore this in future works.

Acknowledgments

I would like to thank Dr Juan Neirotti for very useful discussions and comments. I also would like to thank the comments and suggestions from the anonymous referees which greatly contributed for the clarity of this paper as well for its completeness.

References

[1] Tetlock, P. E. and Gardner, D. *Superforecasting: The art and science of prediction.* Signal, (2015).

[2] Reichardt, J., Alamino, R., and Saad, D. *PloS one* **6**(8), e21282 (2011).

[3] Martinez-Vaquero, L. A. and Cuesta, J. A. *Phys. Rev. E* **90**, 022805 Aug (2014).

[4] Peres, L. R. and Fontanari, J. F. *Journal of Physics A: Mathematical and Theoretical* **43**(5), 055003 (2010).

[5] Clementi, N. C., Revelli, J. A., and Sibona, G. J. *Phys. Rev. E* **92**, 012816 Jul (2015).

[6] Sugiarto, H. S., Chung, N. N., Lai, C. H., and Chew, L. Y. *Phys. Rev. E* **91**, 062804 Jun (2015).

[7] Vazquez, F., Krapivsky, P. L., and Redner, S. *Journal of Physics A: Mathematical and General* **36**(3), L61 (2003).

[8] Kozma, B. and Barrat, A. *Journal of Physics A: Mathematical and Theoretical* **41**(22), 224020 (2008).

[9] Roy, P., Biswas, S., and Sen, P. *Journal of Physics A: Mathematical and Theoretical* **47**(49), 495001 (2014).

[10] Ariely, D. *Predictably Irrational.* HarperCollins, New York, USA, (2008).

[11] Solé, R. V. *Phase Transitions.* Princeton University Press, (2011).

[12] Galam, S. and Moscovici, S. *European Journal of Social Psychology* **21**(1), 49–74 (1991).

[13] Vicente, R., Martins, A. C. R., and Caticha, N. *Journal of Statistical Mechanics: Theory and Experiment* **2009**(03), P03015 (2009).

[14] Sornette, D. *Why Stock Markets Crash: Critical Events in Complex Financial Systems.* Princeton University Press, (2003).

[15] Goldstone, J. A. *Annual Review of Political Science* **4**, 139–187 (2001).

[16] Galam, S., Gefen, Y., and Shapir, Y. *Journal of Mathematical Sociology* **9**(1), 1–13 (1982).

[17] Galam, S. *Journal of Mathematical Psychology* **30**(4), 426–434 (1986).

[18] Allard, A., Hbert-Dufresne, L., Nol, P.-A., Marceau, V., and Dub, L. J. *Journal of Physics A: Mathematical and Theoretical* **45**(40), 405005 (2012).

[19] Castellano, C., Fortunato, S., and Loreto, V. *Rev. Mod. Phys.* **81**(2), 591 (2009).

[20] Abrams, D. M., Yaple, H. A., and Wiener, R. J. *Phys. Rev. Lett.* **107**, 088701 (2011).

[21] Vicente, R., Susemihl, A., Jericó, J. P., and Caticha, N. *Physica A: Statistical Mechanics and its Applications* **400**, 124–138 (2014).

[22] Ross, G. J. and Jones, T. *Phys. Rev. E* **91**, 062809 (2015).

[23] Galam, S. *Sociophysics: a physicist's modeling of psycho-political phenomena.* Springer Science & Business Media, (2012).

[24] Wilson, D. S. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society.* University Of Chicago Press, Chicago, (2003).

[25] Okasha, S. *Evolution and the Levels of Selection.* Oxford University Press, Oxford, (2006).

[26] Schonmann, R. H., Vicente, R., and Caticha, N. *arXiv:1208.0863v2 [q-bio.PE]* (2012).

[27] Haidt, J. *Science* **316**, 998–1002 (2007).

[28] Maslow, A. H. *Psychological Review* **50**(4), 370 (1943).

[29] Piaget, J. *The Origin of Intelligence in Children.* International Universities Press, New York, (1965).

[30] Oakley, L. *Cognitive development.* Routledge, Hove, UK, (2004).

[31] Galam, S. *Physica A: Statistical Mechanics and its Applications* **238**(1), 66–80 (1997).

[32] Vinogradova, G. and Galam, S. *Physica A: Statistical Mechanics and its Applications* **392**(23), 6025–6040 (2013).

[33] Milgram, S. *Psychology today* **2**(1), 60–67 (1967).

[34] Albert, R. and Barabási, A.-L. *Rev. Mod. Phys.* **74**, 47–97 Jan (2002).

[35] Barabási, A.-L. and Frangos, J. *Linked: the new science of networks.* Basic Books, (2014).

[36] Nowak, A., Szamrej, J., and Latané, B. *Psychological Review* **97**(3), 362 (1990).

[37] Quenk, N. L. *Essentials of Myers-Briggs type indicator assessment*, volume 66. John Wiley &

Sons, (2009).
[38] Sachdev, S. *Quantum Phase Transitions*. Cambridge University Press, (2011).
[39] Galam, S. *Physica A: Statistical Mechanics and its Applications* **230**(1), 174–188 (1996).
[40] de Waal, F. B. M. *The age of empathy*. Harmony, New York, USA, (2009).
[41] de Waal, F. B. M. *Science* **336**(6083), 874–876 (2012).
[42] Nishimori, H. and Ortiz, G. *Elements of Phase Transitions and Critical Phenomena*. Oxford University Press, New York, USA, (2011).
[43] Weyland, K. G. *Journal of Democracy* **9**(2), 108–121 (1998).
[44] da Silva, M. F. *Revista de Administração de Empresas* **39**(3), 26–41 (1999).

Appendix A. Analytical Calculation of the Average Satisfaction

We can explicitly write the expression for the average satisfaction S as

$$S = \langle \sigma_k \rangle_{\mathbf{u},\mathbf{w},J} = \sum_{\{u_i\}} \sum_{\{w_i\}} \sum_{\{J_{ij}\}} \left[\prod_i \mathcal{P}(u_i)\mathcal{P}(w_i) \right] \left[\prod_{\substack{i,j \\ i \neq j}} \mathcal{P}(J_{ij}|u_i, w_i, u_j, w_j) \right] \times \operatorname{sgn}\left[\frac{1}{N} \left(\gamma u_k \sum_{l \neq k} J_{kl} + (1-\gamma) w_k \sum_{l \neq k} J_{lk} \right) \right]. \tag{A.1}$$

We start by doing the average over J. The argument of the sign contains averages over N variables J_{ij}. Although they are not identically distributed, the fact that they are independent given $\mathbf{u}$ and $\mathbf{w}$ allows us to use an extension of the Central Limit Theorem (CLT). In the following, we explicitly present this extension.

By means of a Dirac delta distribution, one can write the above equation as

$$S = \left\langle \int \frac{dx\, d\hat{x}}{2\pi}\, e^{ix\hat{x}} (\operatorname{sgn} x) \Gamma(\hat{x}, \mathbf{u}, \mathbf{w}, \gamma) \right\rangle_{\mathbf{u},\mathbf{w}}, \tag{A.2}$$

with

$$\Gamma = \sum_{\{J_{ij}\}} \left[\prod_{\substack{i,j \\ i \neq j}} \mathcal{P}(J_{ij}|u_i, w_i, u_j, w_j) \right] \exp\left\{ -\frac{i\hat{x}}{N} \left[\gamma u_k \sum_{l \neq k} J_{kl} + (1-\gamma) w_k \sum_{l \neq k} J_{lk} \right] \right\}. \tag{A.3}$$

The average can now be factorized and written as

$$\prod_{l \neq k} \Lambda^1_{lk} \Lambda^2_{lk} = \exp\left\{ \sum_{l \neq k} \left(\ln \Lambda^1_{lk} + \ln \Lambda^2_{lk} \right) \right\}, \tag{A.4}$$

where

$$\Lambda^1_{lk} \equiv \sum_{J_{kl}} \mathcal{P}(J_{kl}) \exp\left\{ -\frac{i\hat{x}}{N} \gamma u_k J_{kl} \right\}, \tag{A.5}$$

$$\Lambda^2_{lk} \equiv \sum_{J_{lk}} \mathcal{P}(J_{lk}) \exp\left\{ -\frac{i\hat{x}}{N} (1-\gamma) w_k J_{lk} \right\}, \tag{A.6}$$

with k a fixed index.

Given that J is a binary matrix, we can rewrite the probability distributions as

$$\mathcal{P}(J_{ij}|\pi_i,\pi_j) = p_0\left(\frac{1+J_{ij}u_i}{2}\right) + p_1\left(\frac{1+J_{ij}w_i}{2}\right) + p_2\left(\frac{1+J_{ij}w_j}{2}\right), \tag{A.7}$$

which gives

$$\Lambda^1_{lk} = \cos\left[\frac{\gamma\hat{x}}{N}\right] - i(p_0 + u_k w_k p_1 + u_k w_l p_2)\sin\left[\frac{\gamma\hat{x}}{N}\right] \tag{A.8}$$

$$\Lambda^2_{lk} = \cos\left[\frac{(1-\gamma)\hat{x}}{N}\right] - i(u_l w_k p_0 + w_l w_k p_1 + p_2)\sin\left[\frac{(1-\gamma)\hat{x}}{N}\right]. \tag{A.9}$$

Expanding the cosines, sines and logarithms up to order $1/N^2$, we get

$$\ln\Lambda^1_{lk} \approx -\frac{\gamma^2\hat{x}^2}{2N^2}\left[1-\left(\lambda^1_{kl}\right)^2\right] - i\frac{\gamma\hat{x}}{N}\lambda^1_{kl}, \tag{A.10}$$

$$\ln\Lambda^2_{lk} \approx -\frac{(1-\gamma)^2\hat{x}^2}{2N^2}\left[1-\left(\lambda^2_{kl}\right)^2\right] - i\frac{(1-\gamma)\hat{x}}{N}\lambda^2_{kl}, \tag{A.11}$$

where

$$\lambda^1_{kl} = p_0 + u_k w_k p_1 + u_k w_l p_2, \tag{A.12}$$

$$\lambda^2_{kl} = u_l w_k p_0 + w_l w_k p_1 + p_2. \tag{A.13}$$

By defining

$$\mu \equiv \frac{1}{N}\sum_{l\neq k}\left[\gamma\lambda^1_{kl} + (1-\gamma)\lambda^2_{kl}\right], \tag{A.14}$$

$$\sigma^2 \equiv \frac{1}{N^2}\sum_{l\neq k}\left\{\gamma^2\left[1-\left(\lambda^1_{kl}\right)^2\right] + (1-\gamma)^2\left[1-\left(\lambda^2_{kl}\right)^2\right]\right\}, \tag{A.15}$$

we can write

$$\begin{aligned} S &= \left\langle \int \frac{dx\,d\hat{x}}{2\pi}\; e^{-\frac{\hat{x}^2\sigma^2}{2} + i\hat{x}(x-\mu)}(\operatorname{sgn} x)\right\rangle_{\mathbf{u},\mathbf{w}} \\ &= \left\langle \operatorname{erf}\left(\frac{\mu}{\sqrt{2\sigma^2}}\right)\right\rangle_{\mathbf{u},\mathbf{w}}. \end{aligned} \tag{A.16}$$

Notice that this is the extension of the CLT that we alluded to. The average over J became an average over a Gaussian distributed variable x with mean μ and variance σ^2 which are what we would obtain by calculating the mean and variance of each J_{ij} and doing the appropriate linear combination or, using the notation of equation (2),

$$\mu = \frac{1}{N}\sum_{l\neq k}\left[\gamma u_k\langle J_{kl}\rangle + (1-\gamma)w_k\langle J_{lk}\rangle\right], \tag{A.17}$$

$$\sigma^2 = \frac{1}{N^2}\sum_{l\neq k}\left[\gamma^2(1-\langle J_{kl}\rangle^2) + (1-\gamma)^2(1-\langle J_{lk}\rangle^2)\right], \tag{A.18}$$

where

$$\langle J_{ij}\rangle = p_0 u_i + p_1 w_i + p_2 w_j. \tag{A.19}$$

The expressions for μ and σ^2 can be further simplified for $N \to \infty$

$$\begin{aligned}\mu &= \frac{1}{N}\sum_{l\neq k}\left[\gamma(p_0 + u_k w_k p_1 + u_k w_l p_2) + (1-\gamma)(u_l w_k p_0 + w_l w_k p_1 + p_2)\right] \\ &= p_0[\gamma + (1-\gamma)\bar{u}w_k] + p_1[\gamma u_k w_k + (1-\gamma)\bar{w}w_k] + p_2[\gamma\bar{w}u_k + (1-\gamma)],\end{aligned} \tag{A.20}$$

and

$$\begin{aligned}\sigma^2 &= \frac{1}{N^2}\sum_{l\neq k}\left\{\gamma^2\left[1-(p_0 + u_k w_k p_1 + u_k w_l p_2)^2\right] + (1-\gamma)^2\left[1-(u_l w_k p_0 + w_l w_k p_1 + p_2)^2\right]\right\} \\ &= \frac{1}{N}\left\{\left[\gamma^2 + (1-\gamma)^2\right](1 - p_0^2 - p_1^2 - p_2^2) - 2p_0 p_1\left[\gamma^2 u_k w_k + (1-\gamma)^2\bar{C}\right]\right. \\ &\quad \left. -2p_0 p_2\left[\gamma^2 u_k\bar{w} + (1-\gamma)^2\bar{u}w_k\right] - 2p_1 p_2\left[\gamma^2 + (1-\gamma)^2\right]w_k\bar{w}\right\}.\end{aligned} \tag{A.21}$$

where we introduced the definitions

$$\bar{u} = \frac{1}{N}\sum_{l\neq k} u_l, \qquad \bar{w} = \frac{1}{N}\sum_{l\neq k} w_l, \qquad \bar{C} = \frac{1}{N}\sum_{l\neq k} u_l w_l. \tag{A.22}$$

The CLT can now be directly applied in its original form to the above variables, resulting in

$$S = \left\langle \mathrm{erf}\left(\frac{\mu}{\sqrt{2\sigma^2}}\right)\right\rangle_{u_k, w_k, \bar{u}, \bar{w}, \bar{C}}, \tag{A.23}$$

where the hatted variables are distributed according to the following Gaussian distributions

$$\bar{u} \sim \mathcal{N}\big(\bar{u}|\langle u_i\rangle, \sigma_u^2\big), \tag{A.24}$$

$$\bar{w} \sim \mathcal{N}\big(\bar{w}|\langle w_i\rangle, \sigma_w^2\big), \tag{A.25}$$

$$\bar{C} \sim \mathcal{N}\Big(\bar{u}|\langle u_i\rangle\langle w_i\rangle, \sigma_{uw}^2\Big), \tag{A.26}$$

where

$$\mathcal{N}\big(y|\mu_y, \sigma_y^2\big) \equiv \frac{e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}}{\sqrt{2\pi\sigma_y^2}}, \tag{A.27}$$

and

$$\langle u_i\rangle = (1-2s), \qquad \sigma_u^2 = 1 - \langle u_i\rangle^2 = 4s(1-s), \tag{A.28}$$

$$\langle w_i\rangle = (1-2m), \qquad \sigma_w^2 = 1 - \langle w_i\rangle^2 = 4m(1-m), \tag{A.29}$$

$$\sigma_{u_i w_i}^2 = 1 - \langle u_i\rangle^2\langle w_i\rangle^2. \tag{A.30}$$

We do not need to carry the index k anymore and therefore we write u and w instead of u_k and w_k. In the limit $N \to \infty$, the above Gaussians become Dirac deltas

in their means and the error function becomes the sign of its argument, which gives our final expression

$$S = \langle \operatorname{sgn} \mu \rangle_{u,w}, \tag{A.31}$$

with

$$\mu = p_0[\gamma + (1-\gamma)(1-2s)w] + p_1[\gamma uw + (1-\gamma)(1-2m)w] + p_2[\gamma(1-2m)u + (1-\gamma)]. \tag{A.32}$$