Norm violation versus punishment risk in a social model of corruption

Dan Lu,[1,2] F. Bauza,[1,2] D. Soriano-Paños,[1,3] J. Gómez-Gardeñes,[1,3] and L. M. Floría[1,3]

[1]*Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018 Zaragoza, Spain*
[2]*Department of Theoretical Physics, University of Zaragoza, 50009 Zaragoza, Spain*
[3]*GOTHAM Laboratory, Department of Condensed Matter Physics, University of Zaragoza, 50009 Zaragoza, Spain*

(Received 11 September 2019; accepted 22 January 2020; published 13 February 2020)

We analyze the onset of social-norm-violating behaviors when social punishment is present. To this aim, a compartmental model is introduced to illustrate the flows among the three possible states: *honest*, *corrupt*, and *ostracism*. With this simple model we attempt to capture some essential ingredients such as the contagion of corrupt behaviors to honest agents, the delation of corrupt individuals by honest ones, and the warning to wrongdoers (fear like that triggers the conversion of corrupt people into honesty). In nonequilibrium statistical physics terms, the former dynamics can be viewed as a non-Hamiltonian kinetic spin-1 Ising model. After developing in full detail its mean-field theory and comparing its predictions with simulations made on regular networks, we derive the conditions for the emergence of corrupt behaviors and, more importantly, illustrate the key role of the warning-to-wrongdoers mechanism in the latter.

DOI: 10.1103/PhysRevE.101.022306

I. INTRODUCTION

The existence of social norms whose violation is socially agreed to deserve some punishment is perhaps one of the most widespread features across the history of human cultures and societies, to the point that its absence seems a most unexpected observation. Not surprisingly, the conceptual frame of social norm (and its enforcement) is transversal across socioeconomical disciplines, ranging from experimental (e.g., human behavior, experimental economy) to deeply theoretical (e.g., norms ancestry, their evolution and relation to modern social and political institutions) research [1–10].

A ubiquitous kind of norm-violating practice is *corruption*; indeed, corruption is observed in various forms (economical, political, administrative, etc.), at many scales, and in almost any geographical and historical coordinates. Though the tolerance or punishment of a corrupt act is quite relative to sociocultural particularities, we will hereafter simply identify the terms “corrupt” and “deserved-to-be-punished” behavior.

Corruption, explicitly realized as bribery practices in public administration, has received academic attention in social and economical mathematical modeling research [10–17], a field of much recent interest for interdisciplinary physicists [18]. Most of this literature is framed in either classical or evolutionary game theory, a modeling frame for social dynamics which is clearly prevalent in modern theoretical economics, where, in brief, behaviors are formally represented by game’s strategies, each earning a payoff, and economic behavior optimizes benefit. Undoubtedly, greediness is a most clear incentive to corruption. Thus, game theory [19] seems the most suited framework to tackle the analysis of corrupt behaviors, for it combines the calculation of benefits and the posterior decision-making, i.e., strategic adoption, based on the obtained benefits.

At least in the simplest game-theoretical settings, the *honest vs corrupt* behavioral dilemma is somewhat identified with the *cooperator vs defector* strategic dilemma, which has become the standard interpretation of the two-person-two-strategies normal form of games, such as *prisoner’s dilemma*, *stag hunt*, or *hawks and doves*, and group games, such as *public goods*. Nonetheless, the generalization to $n \geqslant 3$ strategies is needed if punishment (the hallmark of norm violation) to defectors has to be introduced in a stronger way than a mere fine to wrongdoers, such as a penalty in their benefit.

From a computational statistical physics perspective, when modeling the social dynamics of corruption in game-theoretical terms, one easily runs into the practical difficulties posed by a large parameter space and/or strategic space that often render near impractical a desirable thorough analysis of model computations. The need of a clear-cut analysis of which ingredients are most relevant and which others are not increases the need to further simplify the modeling assumptions, while simultaneously trying to keep at least some of those that are essential.

Following this line of thought, we will adopt here an almost minimalistic approach that leads us to a simple abstract compartmental flow [20,21] model of corruption (punishable individual norm violation) where individuals can transit among three states: Honest, corrupt, and ostracism (or punished). In this setting, corrupt behavior is not assumed to be a greedy strategy in a population game dynamics but a simpler general formal entity, an infectious state, that nevertheless allows a game-theoretical perspective. This can be seen from the consideration that what makes a behavior spread socially to the point of becoming endemic is the likelihood it is copied, transmitted, imitated, or diffused following any game dynamics perspective [22,23] that might be found more appropriated, e.g. adaptive, best response, evolutionary, etc.

The structure of the work is as follows. In Sec. II the assumptions of the compartmental model are first motivated, and afterward formulated, as a stochastic population dynamics where agents can be, at a given time, in any of three possible states (the compartments), i.e., a socially inspired non-Hamiltonian kinetic spin-1 Ising model [24,25]. The model formulation is made with no restrictive assumptions on the social structure for the population of agents, where the kinetics of agent microstates and the network-dependent Markovian dynamics [26,27] associated to it are defined. Its close connection to compartmental epidemic spreading models, such as the *susceptible-infected-recovered-susceptible* (SIRS), along with some important differences, are also noted at the end of Sec. II. In fact, those readers familiar with epidemic models would easily interpret the proposed model of social norm-violation dynamics as a SIRS model where (a) the transit to recovery of an infected individual needs susceptible agents around and, moreover, (b) a direct flow from the infected to the susceptible compartment is included. This last flow channel is likely unnatural in epidemic scenarios but is germane to the norm-violation dynamics that motivates our model.

In Sec. III we analyze with some degree of generality the mean field, or well-mixed population approximation, dynamics of the model, with explicit predictions on the transitions that occur (Sec. III A) when the model parameters are tuned. Phase diagrams in the three-dimensional (3D) parameter space are explicitly shown and analyzed in full detail in Sec. III B regarding the macrostate transitions. The analysis of the interesting (SIRS, SIR) limit cases of the model is shown in Sec. III C as a benchmark for the assessment on which assumptions determine what specific dynamical effects, inside the mean-field realm. To validate the results of the mean-field formulation, in Sec. IV we compare the results obtained in Sec. III with those from stochastic simulations performed on top of random and nonrandom regular networks. Finally, in Sec. V we summarize the main results and discuss issues concerning its relevance, shortcomings, and potentialities.

II. THE MODEL

When a norm exists, a partition among population individuals (say, observants and law breakers or honest and corrupt people) appears. But if, in addition, the corresponding punishment imposed to those violators that are caught is some type of "ostracism" (e.g., expulsion from society, exile, and prison), one has already three possible states for individuals, say, H (honest), C (corrupt), and O (out of society); it is useful to think of them as "compartments" containing the corresponding fractions, say, $\langle H\rangle$, $\langle C\rangle$, and $\langle O\rangle$, in the population.

To complete the definition of a classical compartmental flow model, one must also formulate sensible hypothesis on the population flows among compartments. That means to postulate a microscopic dynamics. Our modeling assumptions are summarized in the next items, where general structured (in terms of social contacts) populations are being considered:

(i) Our first assumption is that corruption is a socially infective event: Honest individuals become corrupt *only* by infection from their corrupt neighbors, at an infection rate f_α^{HC}, that is a function of their local microstates.

This assumption for the *corruption* flow, $H \to C$, as we already stressed in the introductory Sec. I, formalizes the corrupt behavior as an infective state, a certainly simpler and less elaborate concept than that of a game strategy, without excluding its consideration, because it is the (social) infectious power of a strategy that allows its diffusion. It is this aspect of corrupt behavior that this assumption tries to capture in its simplest form.

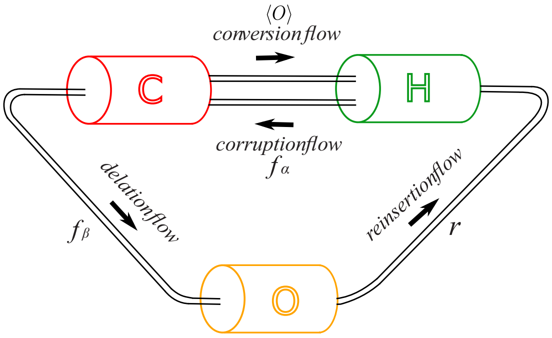

FIG. 1. *Chart flow* representation of the model. Four flows between population compartments are possible. The flow $O \to H$ (reinsertion) occurs at a constant rate r. The flow $C \to H$ (conversion) is fueled by the perception of the delation risk that we simply quantify by $\langle O\rangle$. However, only pairwise social contacts $C - H$ determine the other two flows, say, corruption flow at an infection rate f_α and delation flow at a delation rate f_β.

(ii) We also assume that the flow $C \to O$ is *exclusively* the result of the delation of corrupt individuals by their honest neighbors, at a delation rate f_β^{CO}, also a function of their local microstates.

Let us note that this flow is not the consequence of, e.g., administrative inspection or police investigation; Only interaction with honest agents is the source of this $C \to O$ flow that we call *delation* (or *punishment*) flow. Also note that from the honest agents perspective, delation is not optional. This avoid the need of introducing subtypes of agent states.

(iii) Our third assumption is that, at a given constant rate r, the O individuals are reinserted into social population as H individuals. The flow $O \to H$ is called *reinsertion* flow.

(iv) Finally, we consider a fourth flow, the *conversion* flow $C \to H$, which simply incorporates the *warning-to-wrongdoers* effect of social punishment. The rate at which this flow takes place is controlled by the social perception of risk to be delated, which we simply quantify as the fraction, $\langle O\rangle$, of population in the O compartment.

It is worth emphasizing that corruption and delation flows are the only ones that have their origin in the pairwise interactions among individuals of the socially active population. On the contrary, both the reinsertion and conversion flows do not: The individuals in the O state are socially inactive (i.e., noninteracting), and we only use its fraction (in the fourth assumption above, when implementing the warning-to-wrongdoers effect of social punishment) as the only available information for the estimation of the level of risk that corrupt people perceive. In other words, an O agent does not influence the eventual conversion of its corrupt neighbors more than it does on other far away corrupt agents. However, it is a sort of (temporary) hole in the network of contacts among agents.

See Fig. 1 for a *chart flow* graphical representation of the model, where our assumptions for the flow between any two compartments are simply visualized.

To seek for generality, we assume that the interactions (corruption and delation) among socially active agents define transition probabilities for the corresponding compartmental flows through some functions f_α^{HC} (for corruption of a honest agent) and f_β^{CO} (for the delation of a corrupt one), whose argument is the configuration of agent states in the local neighborhood of the focal agent i, say, $\{\sigma_j(i)\}$, where σ_j $[j = 1, \dots, k(i)]$ denotes the state (H, C, or O) of the neighbor j of i and $k(i)$ is the degree, i.e., the number of neighbors, of the focal agent i.

Due to our assumption on the corruption flow, that it originates exclusively from interaction among individuals in different (H, C) states, the function $f_\alpha^{HC}(i, \{\sigma_j\})$, which gives the transition probability $H \to C$, has to satisfy:

$$f_\alpha^{HC}(i, \{\sigma_j\}) = 0 \ \text{ if } \ \sigma_j \neq C \ \text{ for all } \ j = 1, \dots, k(i). \quad (1)$$

A similar consideration on the delation flow $C \to O$ leads to

$$f_\beta^{CO}(i, \{\sigma_j\}) = 0 \ \text{ if } \ \sigma_j \neq H \ \text{ for all } \ j = 1, \dots, k(i). \quad (2)$$

We will specify later in this section the particular form, see Eqs. (4), (5), (6), and (7) below, that we have used for explicit analytics and computations along the rest of the sections. In addition, for the sake of simplicity, we will use f_α and f_β for the notation of infection rate and delation rate instead of those with superscripts HC and CO in the rest of the paper.

A simple scheme for stochastic (MC) direct simulations of the dynamics is the following: At each time step (t) choose uniformly at random an agent i. Then we have the following:

(i) If $\sigma_i(t) = H$, then $\sigma_i(t+1) = C$ with transition, i.e., conditional, probability f_α, a (yet-unspecified) function of the local configuration around i. The agent remains honest with probability $1 - f_\alpha$.

(ii) If $\sigma_i(t) = C$, then $\sigma_i(t+1) = H$ (warning-to-wrongdoers effect) with probability $\langle O \rangle$, the fraction of population in O state. Then, if not converted (probability $1 - \langle O \rangle$), the corrupt agents will be delated to O state with transition probability f_β, a (yet-unspecified also) function of the local configuration around i. Thus, agent i keeps corrupt at $t+1$ with probability $(1 - \langle O \rangle)(1 - f_\beta)$. (Note that an equally acceptable scheme would try first delation, then conversion, which produces different transition probabilities for $C \to H$ and $C \to O$. We will comment on this later.)

(iii) If $\sigma_i(t) = O$, then $\sigma_i(t+1) = H$ with conditional probability r remaining out with probability $1 - r$.

One can associate to this dynamics on agents' state configurations a nonlinear Markov process in the following way [28,29]. Assign to each agent i, and at time t, a real vector $\vec{\rho}(i;t)$ whose components are the probabilities (at time t) that the agent is in each of the possible states, namely

$$\vec{\rho}(i;t) \equiv (\rho_h(i;t), \rho_c(i;t), \rho_o(i,t)).$$

The transition probabilities (i)–(iii) introduced above define a nonlinear Markov process for the time evolution of these probabilities $\vec{\rho}(i;t+1) = \mathbf{Q}\, \vec{\rho}(i;t)$, where

$$\begin{bmatrix} 1 - f_\alpha & \langle \rho_o \rangle & r \\ f_\alpha & (1 - f_\beta)(1 - \langle \rho_o \rangle) & 0 \\ 0 & f_\beta(1 - \langle \rho_o \rangle) & 1 - r \end{bmatrix}$$

is the matrix representation of $\mathbf{Q}$, and $\langle \rho_o \rangle$ is the fraction of population in O state, i.e.,

$$\langle \rho_o \rangle = N^{-1} \sum_i \rho_o(i). \quad (3)$$

Note our choice of relative order of trial—conversion *before* eventual delation—in the second column of the matrix $\mathbf{Q}$ above written. The alternative choice would correspond to $\mathbf{Q}_{hc} = (1 - f_\beta)\langle \rho_o \rangle$ (instead of $\langle \rho_o \rangle$) and $\mathbf{Q}_{oc} = f_\beta$ [instead of $f_\beta(1 - \langle \rho_o \rangle)$], the rest of the elements being unchanged.

To complete the model formal setting, one has to specify the functions f_α and f_β for the conditional probabilities of corruption and delation, respectively. They define the specific social interactions postulated, and also incorporate the information on the social network, that we assume it is encoded in the *neighborhood* matrix, whose ith row tells us who the $k(i)$ neighbors of the agent i are. The following choice mimics the familiar implementation of infective interactions in Monte Carlo simulations on compartmental epidemic models as SIS, SIR, etc.:

$$f_\alpha(i, \{\sigma_j\}) = 1 - \prod_{j=1}^{k(i)} (1 - \alpha \delta_{\sigma_j, C}), \quad (4)$$

$$f_\beta(i, \{\sigma_j\}) = 1 - \prod_{j=1}^{k(i)} (1 - \beta \delta_{\sigma_j, H}), \quad (5)$$

where $\delta_{x,y}$ is the Kronecker's delta. The rationale for (4) is that an honest focal agent contacts all its corrupt neighbors, and in each one of these contacts, the probability of infection is α. Similarly, for (5) a corrupt focal agent contacts all its honest neighbors and in each contact is delated with probability β.

For the associated Markov process, these expressions translate into

$$f_\alpha(i, \{\vec{\rho}(j)\}) = 1 - \prod_{j=1}^{k(i)} [1 - \alpha \rho_c(j)], \quad (6)$$

$$f_\beta(i, \{\vec{\rho}(j)\}) = 1 - \prod_{j=1}^{k(i)} [1 - \beta \rho_h(j)]. \quad (7)$$

Although in the computations shown below we have used these specific forms for f_α and f_β, other alternative forms, based on some different corruption and delation schemes are, no doubt, of potential interest. The microscopic mechanisms of "becoming corrupt" should inform the appropriate functional form of f_α, as much as those of "delating corrupts" must inform that of f_β.

Despite our emphasis in Sec. I on avoiding a genuine game-theoretic formulation of the model, let us note that game theory could, perhaps only to some extent, be accommodated in the previous modeling frame by appropriately relating f_α and f_β to the payoff earned when playing a formal game where honest and corrupt behavior are considered as strategies. We will come back to this point in Sec. V when discussing model potentialities and current prospects.

For the choice made in Eqs. (4) and (5) or Eqs. (6) and (7), and if the conversion flow channel is suppressed (see Fig. 2 and III C below), then the model should be interpreted (by

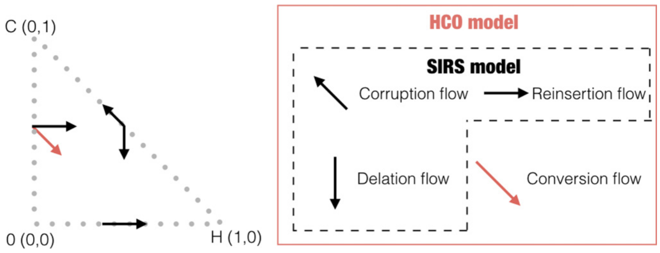

FIG. 2. Schematic visualization of the simplex $\mathcal{S}_2$ [i.e., $0 \leqslant \rho_h, \rho_c \leqslant 1$, $\rho_h + \rho_c \leqslant 1$ (left panel)], and directions of the contributions of each of the four compartmental flows to the (flow vector) $\vec{F}$ field (left panel). The arrows on the face boundaries of the simplex visualize that there is no flow outward, and thus the simplex is an invariant set, as required by consistency. If one excludes the conversion flow, the model (HCO) becomes a (nonstandard) version of the epidemic SIRS model.

the identifications $S \equiv H$, $I \equiv C$, $R \equiv O$) as a kind of SIRS model where the recovery rate is mediated by the interaction of the susceptible neighbors with the infected agent, as if the recovery from infection would crucially depend on the assistance from healthy neighbors [30].

The suppression of the reinsertion flow, $r \to 0$, is another interesting limit case which we will consider below in Appendix B.

III. MEAN-FIELD APPROXIMATION

As a first step in the analysis of a collective phenomenon, a sensible mean-field approximation is a well-known and recommended practice in statistical physics, due to both its simplicity and unbiased character. Often, though not always, it provides a qualitatively correct description of the observed behavior, and, moreover, it reveals basic mechanisms that trigger the collective changes of state for large (macroscopic) systems. In the language of population dynamics, well-mixed population approximation is the usual term employed for the same sensible approach.

A. The mean-field dynamics

Homogeneity of both field (agent state; H, C, or O) and structure of contacts (environment) is the essential assumption of a mean-field approximation; under these circumstances, it seems plausible considering average behavior as a good (least-biased) estimation of agent's behavior. If every agent behaves as the average of all ("average" agent), i.e., $\vec{\rho}(i) = \langle\vec{\rho}\rangle$ (for all i) for the associated Markov process, and the neighborhood of size $k(i) = k$ is "indifferent" regarding i, so that it can be selected at random among the population at each time step (well-mixed population assumption), then one arrives at the following mean-field discrete time evolution equations for the probabilities ρ_h, ρ_c, ρ_o (or, alternatively, for the compartmental fractions ($\langle H\rangle$, $\langle C\rangle$, and $\langle O\rangle$):

$$\vec{\rho}(t+1) = \begin{bmatrix} 1 - f_\alpha & \rho_o & r \\ f_\alpha & (1 - f_\beta)(1 - \rho_o) & 0 \\ 0 & f_\beta(1 - \rho_o) & 1 - r \end{bmatrix} \vec{\rho}(t),$$

where suitable changes [see paragraph just below Eq. (3)] in the second column of the matrix have to be made for a different order of "trial for flow" out from the C compartment.

Due to the normalization constraint, $\rho_h + \rho_c + \rho_o = 1$, the mean-field discrete time dynamics is a nonlinear two-dimensional map of the simplex $\mathcal{S}_2$ (i.e., $0 \leqslant \rho_h, \rho_c \leqslant 1$, $\rho_h + \rho_c \leqslant 1$) onto itself. This simplex is visualized on the left panel of Fig. 2, as the triangle defined by the vertices [$H \equiv (1, 0)$, $C \equiv (0, 1)$, $O \equiv (0, 0)$], in the (ρ_h, ρ_c) plane (say, $\rho_h = 1$, $\rho_c = 1$, $\rho_o = 1$, respectively).

The associated two-dimensional flow (continuous time dynamics) is defined by the velocity (2D vector) field on the simplex $\vec{F}(\vec{\rho})$,

$$\vec{F}(\vec{\rho}) = \dot{\vec{\rho}},$$

whose components are

$$\begin{aligned} F_h(\vec{\rho}) &= -(f_\alpha + r + \rho_c)\rho_h + (r + \rho_c)(1 - \rho_c) \\ F_c(\vec{\rho}) &= [f_\alpha + (1 - f_\beta)\rho_c]\rho_h + [(1 - f_\beta)\rho_c - 1]\rho_c. \end{aligned} \qquad (8)$$

In the right panel of Fig. 2, we indicate the direction of the contribution to the total flow vector field on the plane (ρ_h, ρ_c) of each of the four compartmental flows. The preliminary step of the analysis is to check that the simplex is an invariant set of initial conditions, as obviously required by consistency. Indeed, see left panel of Fig. 2, one easily realizes that

(i) On the hypothenuse of the simplex, where $\rho_h + \rho_c = 1$, both the reinsertion and the conversion flow are null ($\rho_o = 0$); the corruption flow is colinear to this boundary, and the delation flow points vertically inward.

(ii) On the vertical (ρ_c) axis, where $\rho_h = 0$, both the delation and the corruption flows are null; both the nonzero remaining flows point inward.

(iii) On the horizontal axis, where there are no corrupt people, only reinsertion flow is nonzero, which is colinear to this boundary, and points toward the full honesty corner of the simplex, with the proviso that $r > 0$, the generic case.

From now on in this section we will consider the generic case ($r > 0$) where reinsertion flow does not vanish. The interesting $r \to 0^+$ limit is analyzed in Appendix B. Also, we restrict the analysis to one-dimensional functions $f_\alpha(\rho_c)$ and $f_\beta(\rho_h)$. This simplifying restriction amounts to saying that, e.g., the probability that a honest agent becomes corrupt at time t only depends on the agent contacts with corrupt agents, and its contact with others have no influence on its corruption.

It is important to realize that a direct consequence of the model assumptions in Sec. II, namely that infection and delation flows originate *exclusively from* agent interactions, is that the functions f_α and f_β have to be such that $f_\alpha(\rho_c = 0) = 0$ (i.e., no corruption flow without corrupt agents) and $f_\beta(\rho_h = 0) = 0$ (no delation flow without delators). Indeed, this has been implicitly used in the previous simple vector-field analysis, when we considered in item (ii) above that delation flow was null when $\rho_h = 0$ and in item (iii) that corruption flow vanished at $\rho_c = 0$.

Now we look for possible existence of boundary fixed points; from the previous analysis, they can only be located at corners. While $\rho_o = 1$ [i.e., ($\rho_h = 0$, $\rho_c = 0$), the origin] is a fixed (invariant) point only in the limit $r \to 0$, when reinsertion flow vanishes, the two other corners of the simplex, say, *full honesty* ($\rho_h = 1$) and *total corruption* ($\rho_c = 1$), are always fixed points of this dynamics [i.e., zeros of the field $\vec{F}(\vec{\rho})$], irrespective of the parameter values. These are the only fixed points on the simplex boundary.

Let us now analyze, in the linear regime of perturbations, the stability of these corner fixed points by looking at the "restoring forces (flows)" induced by perturbations. For a more formal, albeit fully equivalent, linear stability analysis presentation of these results, see Appendix A, where we show also the irrelevance regarding the stability of them of the order in which the transitions $C \to O$ and $C \to H$ are tried.

$\rho_h = 1$: The full honesty corner H, provided $r > 0$, is clearly stable against a small increase, $\delta\rho_o$, in the population fraction of O compartment, for it just induces a (stabilizing) reinsertion flow. However, a small perturbation $\delta\rho_c$ generates an infection flow $f_\alpha(\delta\rho_c) \simeq f'_\alpha(0)\delta\rho_c$, which unless overcome by the (also induced by perturbation) delation flow $f_\beta(1)\delta\rho_c$, would render unstable the full honesty state. In other words, the full honesty state is a *local attractor* of (nearby) trajectories provided the following stability condition holds:

$$f'_\alpha(0) < f_\beta(1). \tag{9}$$

Note that the rate r of reinsertion has no influence on this stability condition. Only the balance among corruption and delation flows determines the instability of the full honesty state, because inactivity ($\delta\rho_o$) fluctuations induce restoring flow and have no linear effects on the instability driving this *corruption* transition.

$\rho_c = 1$: Regarding the full corruption corner, also for $r > 0$, a small perturbation of component $\delta\rho_h$ generates a restoring corruption flow $f_\alpha(1)\delta\rho_h$, which, to keep this fixed point stable, has to overcompensate the sum of the (destabilizing) delation flow $f_\beta(\delta\rho_h) \simeq f'_\beta(0)\delta\rho_h$, and conversion flow $\delta\rho_o \times 1 \simeq f_\beta(\delta\rho_h)/r$, generated by perturbation. Thus, the linear stability condition for the total corruption is

$$\left(1 + \frac{1}{r}\right) f'_\beta(0) < f_\alpha(1). \tag{10}$$

Note that now the balance corruption-delation is interfered by the influence of ρ_o, which helps small honest fluctuations to further develop. We see that the stability condition of the full corruption state depends on the rate r of reinsertion, via the conversion flow induced by linear perturbation, and then this *honesty* transition is not exclusively driven by agent-agent interactions but also by the (self-consistent, global field) value of the fraction of agents in O state.

In the final stage of our search for attractors (asymptotic, absorbing states) of the dynamics, we pay attention to the $\vec{F}$ field nullclines, i.e., the loci where each of its components vanishes, $F_h(\vec{\rho}) = 0$ and $F_c(\vec{\rho}) = 0$, given explicitly by Eqs. (8). An interior fixed point will exist whenever these loci intersect in the interior of the simplex.

$F_h = 0$: The equation of the F_h nullcline is

$$-(f_\alpha + r + \rho_c)\rho_h + (r + \rho_c)(1 - \rho_c) = 0. \tag{11}$$

Note, in the first place, that this locus is independent of f_β, because the delation flow leaves ρ_h unchanged; next, one quickly convinces oneself that it contains both C and H corners. Finally, one realizes that, provided f_α is independent of ρ_h, Eq. (11) defines, inside the simplex, a unique function $\rho_c(\rho_h)$ whose graph joins those corners.

$F_c = 0$: The F_c nullcline satisfies the following equation:

$$[f_\alpha + (1 - f_\beta)\rho_c]\rho_h + [(1 - f_\beta)\rho_c - 1]\rho_c = 0. \tag{12}$$

One easily realizes that the horizontal axis, $\rho_c = 0$, belongs to this set. This is one of the (curve, in general) branches that are solutions of this nonlinear implicit equation. The rest of them must solve for the equation obtained by dividing (12) by ρ_c:

$$[f_\alpha/\rho_c + (1 - f_\beta)]\rho_h + [(1 - f_\beta)\rho_c - 1] = 0. \tag{13}$$

It is also straightforward to check that the C corner always belong to some of these branches. Another simple general result is the following. There is always one of these branches that crosses the horizontal axis. The argument is simple if one assumes that f_α is an analytic function of ρ_c at 0^+. By keeping second-order terms in the power expansion of $f_\alpha(\rho_c) \simeq f'_\alpha(0)\rho_c + (1/2)f''_\alpha(0)\rho_c^2$, one obtains the following (nonlinear) approximation to the solution of (12) close to the horizontal axis:

$$\rho_c = \frac{1 - [f'_\alpha(0) + 1 - f_\beta]\rho_h}{1 - f_\beta + (1/2)f''_\alpha(0)\rho_h}, \tag{14}$$

which intersects $\rho_c = 0$ at the abscissa value $\rho_h = \rho_h^*$, the solution of the nonlinear equation

$$[f'_\alpha(0) + 1 - f_\beta(\rho_h^*)]\rho_h^* = 1. \tag{15}$$

Whether the curve branch of the F_c nullcline that intersects the horizontal axis at ρ_h^* is the same one that passes through the C corner, or it is a different branch, both are possible situations (conditional to the specific functions f_α and f_β). In fact, we will show numerical examples of both situations below, for a single one-parametric functional form of them [Eqs. (17) and (18)].

The stability of the states of full honesty and full corruption is closely tied to the nullclines' geometrical configuration around them. Indeed, using (9) we easily conclude that the stability condition of the full honesty corner is equivalently expressed as "ρ_h^* is not in the simplex," where ρ_h^* is the intersection of the curve branch of the F_c nullcline with the horizontal axis, defined by (15) above. This implies that the H corner is unstable iff $\rho_h^* < 1$.

In a similar, though geometrically very different, way one can see that if the slope of the curve branch of the F_c nullcline at the C corner is lower than the slope of the F_h nullcline, i.e.,

$$1 + f_\alpha(1) - f'_\beta(0) > 1 + \frac{f_\alpha(1)}{1 + r}, \tag{16}$$

then the C corner is stable, see Eq. (10), and vice versa. From both transitions, we conclude that the difference in relative positions of F_h and F_c nullclines determines the stability of both full honesty and corruption states.

In the next subsection we show numerical and analytical results for the phase portraits and phase diagrams, for the particular choice, inspired by epidemic analogy, that we made above for the flows originated from social interactions $C - H$ [Eqs. (4), (5), (6), and (7)]; the expressions for the conditional

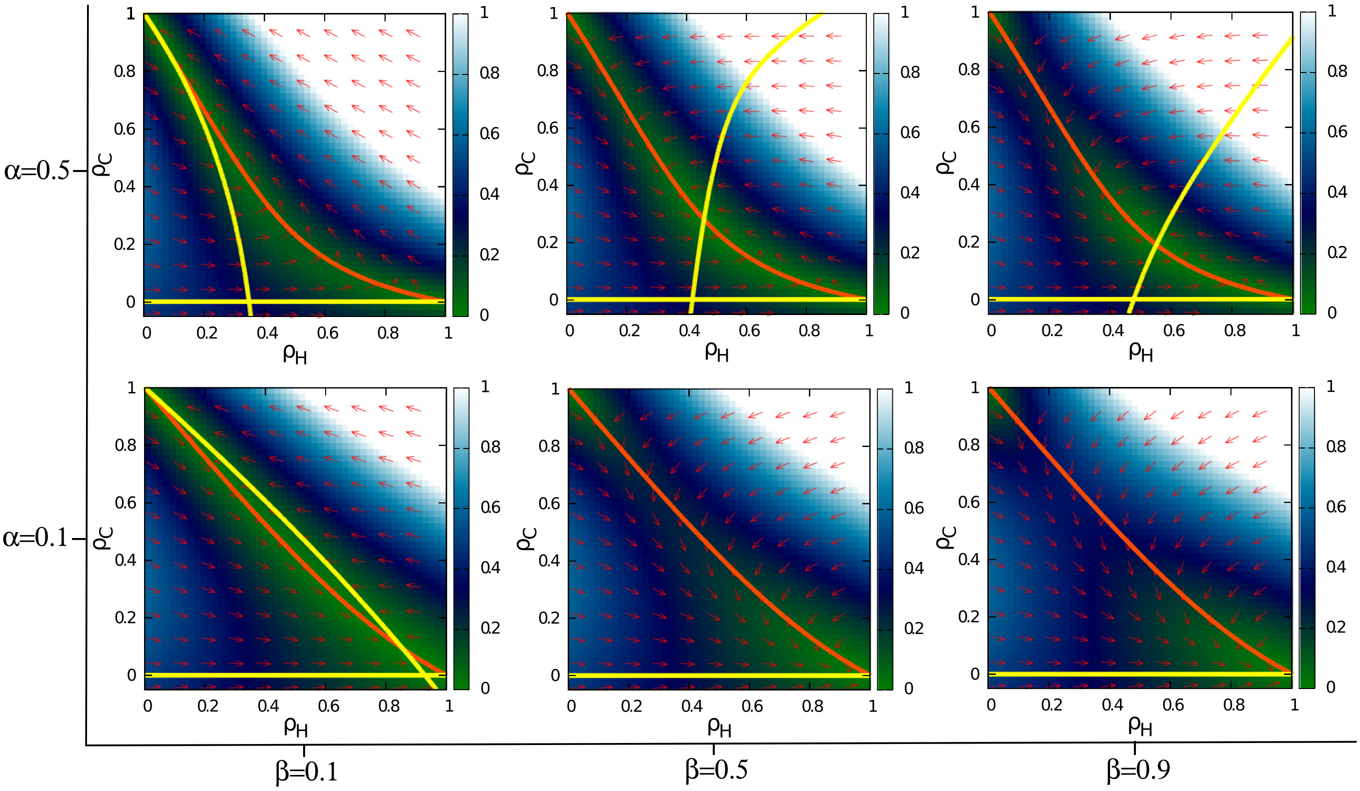

FIG. 3. The nullclines $F_h = 0$ (red line) and $F_c = 0$ (yellow line) with a constant value $r = 0.5$. The infection rate is given as $\alpha = 0.1, 0.5$. The delation rate is set as $\beta = 0.1, 0.5, 0.9$. In each of plane, color level used to represent the numerical value of the flow.

probabilities f_α and f_β in the mean-field approximation are as follows:

$$f_\alpha(k, \vec{\rho}) = 1 - (1 - \alpha\rho_c)^k, \tag{17}$$

$$f_\beta(k, \vec{\rho}) = 1 - (1 - \beta\rho_h)^k. \tag{18}$$

The "epidemic," or "contact interaction," character of this choice for both transition probabilities, corruption and delation, should be kept in mind. On one hand, the knowledge from closely related epidemic models can be capitalized on here, while on the other, the results that we analyze could plausibly be of use in some epidemiology contexts of potential interest, wherever recovery needs assistance from susceptible neighbors.

B. Mean-field phase portraits and diagrams

For the particular *contact interaction* functions (17) and (18), the instability of the full honesty state, from Eq. (9), occurs at the value of the (infection) corruption rate α_c:

$$\alpha_c(\beta) = \frac{1 - (1 - \beta)^k}{k}, \tag{19}$$

for all values of the reinsertion rate r (i.e., it is independent of this parameter value). This value of the infectivity power of corruption is the benchmark for the appearance of observable corruption, under the mean field, well-mixed population, assumptions.

Also, from Eq. (10), the instability of the state of full corruption occurs at a value β_c of the delation rate given by

$$\beta_c(\alpha) = \left(\frac{r}{1 + r}\right)\frac{1 - (1 - \alpha)^k}{k}. \tag{20}$$

From (19) and (20), it is easily seen that the stability regions, in the (α, β) parameter plane, of corner fixed points (H and C) do not overlap, meaning that there is neither bi-stability region in the phase diagram of the model, nor hysthereris behavior. In other words, no discontinuous change full C-full H can occur by tuning a model parameter for our choices (17) and (18) of f_α and f_β. In the region of the parameter plane (α, β) where both fixed points are unstable, an interior (stable) fixed point $\vec{\rho}(r, \alpha, \beta)$ is the unique global attractor of phase-space flow. By no means is this conclusion forcefully valid for more general choices of the corruption, f_α, and delation, f_β, transition probability functions, for multiple interior nullcline's crossing cannot be discarded in general cases.

In Fig. 3 we show the phase portraits for a reinsertion rate $r = 0.5$ and values of $\alpha = 0.1, 0.5$ and $\beta = 0.1, 0.5, 0.9$. Arrows indicate the local direction of the $\vec{F}$ field, the flow, while its modulus is color encoded. The F_h nullclines are plotted in red color; one sees that they are independent of β and that they deviate away from the simplex hypothenuse for increasing values of α.

The branches of the F_c nullclines are plotted as yellow lines. Note that the horizontal axis is always one of them.

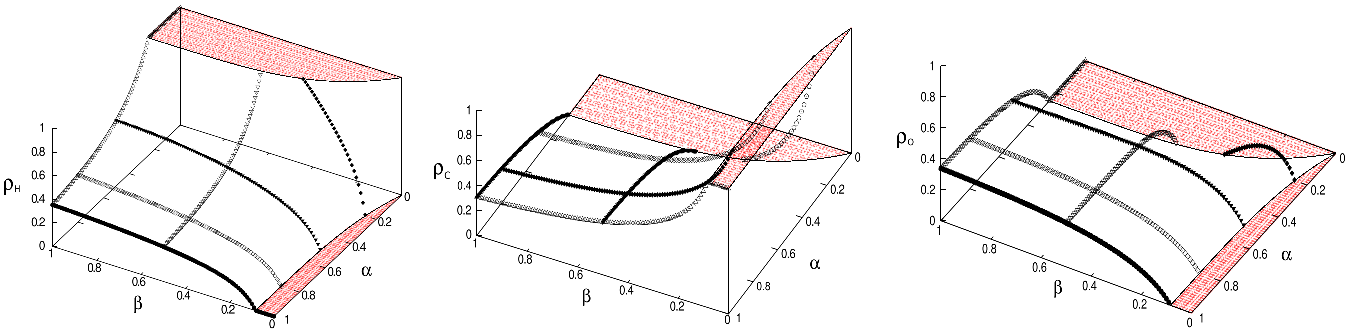

FIG. 4. Visualizations of the equilibrium surfaces $\rho_m(\alpha, \beta, r = 0.5)$, $m = h, c, o$, on the (α, β) parameter plane with a constant value of the reinsertion rate $r = 0.5$. The sections of these surfaces at cutting planes corresponding to values of $\alpha = 0.2, 0.5, 0.8, 1$ and $\beta = 0.5, 1$ are also plotted to help a three-dimensional mental image. The stability regions of full C and full H states are red colored.

When no other branch is visible (as for $\alpha = 0.1$ and $\beta = 0.5, 0.9$), meaning that $\rho_h^* > 1$, the full honesty state H is stable, and it is the global attractor. For the other cases shown in Fig. 3, the trajectories evolve asymptotically to the interior fixed point where the F_h nullcline and a curve branch of the F_c nullcline intersect. While for $\beta = 0.1$ and $\alpha = 0.1, 0.5$ the yellow curve passes through the full corruption corner, and for $\alpha = 0.5$ and $\beta = 0.5, 0.9$ it does not. In these cases, there is a different curve branch of the F_c nullcline, passing through the C corner, from outside the simplex. The transition between these two qualitatively different phase portraits of the entire plane (ρ_h, ρ_c) occurs when the two yellow curve branches "anticross" far outside the simplex; this is a bifurcation on the whole plane phase portrait which has no qualitative effects (no local influence) on the interior of the simplex.

In the three panels of Fig. 4 we try to summarize the effect of parametric variation of α and β (in the mean-field dynamics) on the mixed population absorbing state, through perspective visualizations of the compartmental fractions at the equilibrium (attractor) for a fixed arbitrary value of $r = 0.5$, i.e., of the surfaces $\vec{\rho}(\alpha, \beta, r = 0.5)$ of the asymptotic equilibrium. The regions colored in red in the three panels of this figure correspond to the respective regions of stability of the full honesty $[\alpha \leqslant \alpha_c(\beta)]$ and full corruption $[\beta \leqslant \beta_c(\alpha)]$ absorbing states, where the transition lines are given by (19) and (20). We hope that the simple inspection of this figure is more informative than lengthy and wordy explanations of the general trends of the model behavior.

C. The SIRS limit

Here, leaving aside the fourth of our model assumptions in Sec. II, we will take out from the model the conversion flow (warning to wrongdoers effect). The number of flow channels is thus reduced from four to three (contagion, delation, and reinsertion), and thus the "flow chart" between the three compartments is now that of a SIRS model, with the identifications $S \equiv H$, $I \equiv C$, and $R \equiv O$.

While in the standard SIRS model the rate of recovery $(I \to R)$ is a constant, in this variant of the SIRS model the recovery of an infected individual is only possible through contact interaction with its susceptible neighbors. A plausible epidemic situation leading to it, may be, e.g., one in which the recovery from disease requires, *sine qua non*, imperatively the assistance (care) from relatives [30].

Though the consideration made above may certainly add some interest in the following results by themselves in plausible epidemic contexts, our main purpose in this subsection is to make a precise assessment on the warning to wrongdoers effect in the HCO model by revealing the aspects on which its presence makes a difference and how much this difference amounts to.

The mean-field analysis goes along the same lines as explained in Sec. III A, and one arrives at the following 2D flow on the simplex:

$$\begin{aligned} F_h(\vec{\rho}) &= -(f_\alpha + r)\rho_h + r(1 - \rho_c)F_c(\vec{\rho}) \\ &= f_\alpha \rho_h - f_\beta \rho_c. \end{aligned} \tag{21}$$

Both $\rho_h = 1$ and $\rho_c = 1$ corners are fixed points, whose linear stability analysis we now summarize. The stability condition for the full honesty corner is the same as it was in the presence of "warning to wrongdoers" (conversion flow):

$$f'_\alpha(0) < f_\beta(1). \tag{22}$$

This is an expected result, because we already saw in Sec. III A that the conversion flow has no influence on this transition, which is only determined by the competition of corruption and delation flows generated by linear perturbations.

On the other hand, now the stability condition of the total corruption state no longer depends on the reinsertion rate r:

$$f'_\beta(0) < f_\alpha(1). \tag{23}$$

The stability of both homogeneous population states is unaffected by the value of the reinsertion $(R \to S)$ rate in this SIRS model. In other words, the rate r of reinsertion is an irrelevant parameter regarding both corruption and honesty transitions. However, one should be aware that away from stability of homogeneous populations, this parameter is clearly relevant in determining the mixed equilibrium population fractions, as we will soon discuss below.

Before that, we pay attention to the nullclines (see Fig. 5 for examples). The F_h nullcline connects the corners C and H and is independent of f_β. Indeed, it is qualitatively very similar to that of the HCO model. On the contrary, the F_c nullcline (whose relevant branches are plotted as yellow lines in Fig. 5) shows important differences. The equation of this nullcline is

$$f_\beta \rho_c = f_\alpha \rho_h. \tag{24}$$

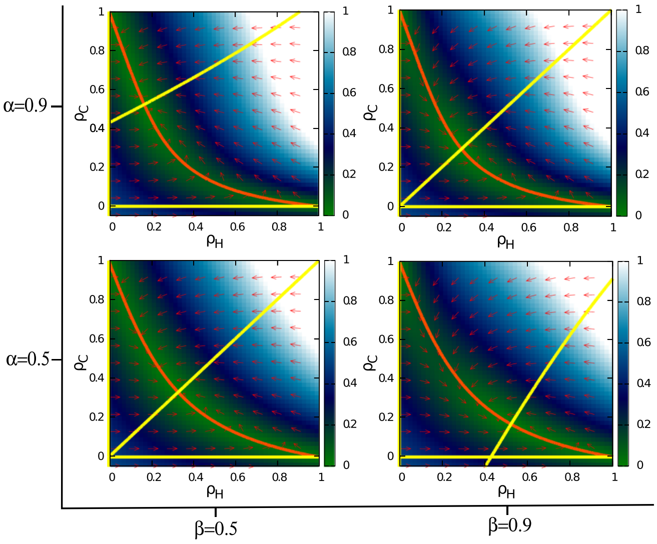

FIG. 5. The nullclines $F_h = 0$ (red line) and $F_c = 0$ (yellow line) in the SIRS model with a constant value $r = 0.5$. The infection rate is given as $\alpha = 0.5, 0.9$, while the delation rate is $\beta = 0.5, 0.9$.

Two branches of this nullcline are easily obtained, namely $\rho_c = 0$ (horizontal axis) and $\rho_h = 0$ (vertical axis). Note that the latter is incompatible with the warning to wrongdoers (or conversion) flow, and thus it is absent in the HCO model, as we know from Sec. III A. The rest of branches of this nullcline must solve for the equation

$$\frac{f_\beta}{\rho_h} = \frac{f_\alpha}{\rho_c}. \tag{25}$$

For the infective type of f_α and f_β functions in Eqs. (17) and (18), this equation has a useful symmetry: It is invariant under the simultaneous interchange $\alpha \leftrightarrow \beta$ and $\rho_c \leftrightarrow \rho_h$. This symmetry of the F_c nullcline is illustrated in Fig. 5. A simple consequence of this symmetry is that if $\alpha = \beta$, then the F_c nullcline is $\rho_c = \rho_h$, the main diagonal. Furthermore, $\rho_h \leqslant \rho_c$ if and only if $\beta \leqslant \alpha$. We will later discuss some other features of the model that are associated to this symmetry.

After (23) and (22), the transition lines, $\beta_c(\alpha)$ and $\alpha_c(\beta)$,

$$\alpha_c(\beta) = \frac{1-(1-\beta)^k}{k}, \tag{26}$$

$$\beta_c(\alpha) = \frac{1-(1-\alpha)^k}{k}, \tag{27}$$

are mirror symmetric around the line $\alpha = \beta$ in the (α, β) plane, and as we already remarked, they do not change with the value of r. However, when the attractor is an interior point of the simplex, and thus the flow through the three channels is, at equilibrium, the same:

$$f_\beta \rho_c = r\rho_o = f_\alpha \rho_h, \tag{28}$$

the reinsertion flow rate r regulates the $S - I$ $(H - C)$ balance. In particular:

For $r = 1$, meaning that the recovery time is just one time step (the shortest possible time scale), we are as closer as the model can be to the limit of zero (instantaneous) recovery time.

In the strict instantaneous recovery limit the $R \equiv O$ state ceases to exist, it just disappears; the feasible region is in this limit case reduced to the hypothenuse ($\rho_o = 0$) of the simplex, and the model becomes a variant of the (kinetic two-states) SIS model, with $I \to S$ rate mediated by S. For our choice of the functions f_α and f_β there is now a strict symmetry of the dynamics (equations of motion) under simultaneous interchange of parameters $\alpha \leftrightarrow \beta$ and labels $h \leftrightarrow c$. Note that though the existence of R (O) state breaks this symmetry, the broken symmetry is still manifest in the F_c nullcline symmetry discussed above.

However, even when recovery takes just one step of time, the instantaneous fraction ρ_o of inactive individuals does not affect infected (do not delate corrupt) neighbors, and the balance infection-recovery is biased toward infection.

For $1 > r > 0$, the larger the characteristic stay time, $1/r$, at the O state the easier the infective state can spread.

For $r = 0$ the model becomes a variant of the SIR model, which will be analyzed in Appendix B.

We conclude that the reinsertion rate r, though being irrelevant regarding the onset of instabilities that operate at both *corruption* and *honesty* transitions, is a determinant factor regarding the stationary values of the compartmental fractions of the SIRS model when the dynamic equilibrium is a mixed population macrostate.

Finally, we pay now a closer attention to the symmetry that a unique choice of the functional form for the corruption and delation transition probabilities, f_α and f_β, induces on this version of the SIRS model: If one assumes that both transition probabilities are given by a unique function $g(x, z)$ in the sense that $f_\alpha(\rho_c) = g(\alpha, \rho_c)$ and $f_\beta(\rho_h) = g(\beta, \rho_h)$, a general simple argument concludes (see Appendix C) that in the mixed population stationary state regime of this model, the fractions (ρ_h, ρ_c) of corrupt and honest people are such that

$$\rho_c(\alpha, \beta) = \rho_h(\beta, \alpha) \;\text{ and }\; \rho_h(\alpha, \beta) = \rho_c(\beta, \alpha), \tag{29}$$

in other words, the *equation of stationary state* $\vec{\rho}(\alpha, \beta)$ is symmetric under the simultaneous interchange $\alpha \leftrightarrow \beta$ and $\rho_c \leftrightarrow \rho_h$. This might be at a first sight unexpected, because the equations of motion, and then the phase portraits (see Fig. 5), are by no means invariant. In the extent that there is no fundamental reason why delation and corruption transition probabilities should be described by the same function, this is an accidental (nonfundamental) symmetry. As already stated, both models, HCO and SIRS, exhibit the same *corruption* transition lines:

$$\alpha_c^{\rm HCO}(\beta) = \alpha_c^{\rm SIRS}(\beta), \tag{30}$$

for all values of β, because the conversion flow has no influence on the onset of corruptive fluctuations. On the contrary, honest instabilities in the full C state are enhanced by the warning to wrongdoers, thus shrinking the full C stability region, see (20) and (27), in the HCO model:

$$\beta_c^{\rm HCO}(\alpha) < \beta_c^{\rm SIRS}(\alpha), \tag{31}$$

for all values of $\alpha > 0$.

As expected, removing the conversion flow closes an input channel of the H compartment and favors higher levels of corruption, thus leading to a decrease of the fraction of honest agents in the SIRS model with respect to the HCO model.

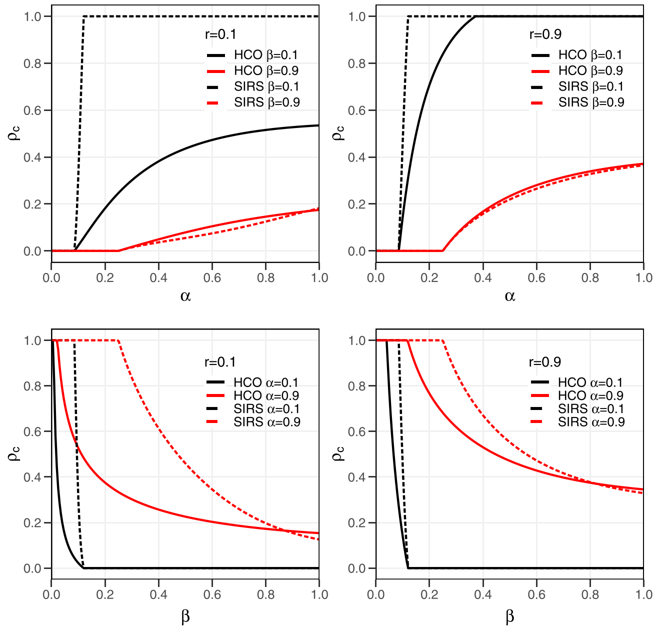

FIG. 6. Top: Fraction of corrupt agents as a function of the corruption rate $\rho_c(\alpha)$ fixing the delation rate to $\beta = 0.1$ and $\beta = 0.9$. Bottom: $\rho_c(\beta)$ for $\alpha = 0.1$ and $\alpha = 0.9$. The mean-field predictions for SIRS model are represented with dashed lines, whereas solid lines correspond to the HCO model. The reinsertion rate is fixed to $r = 0.1$ (left panels) and $r = 0.9$ (right panels).

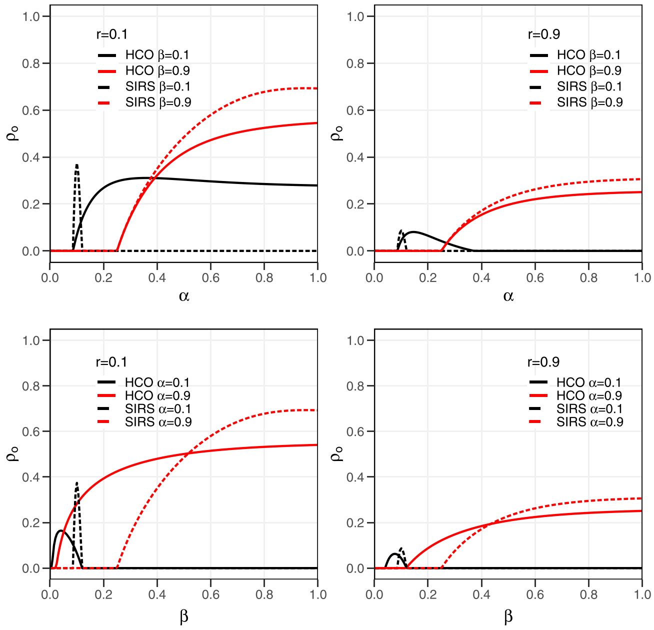

FIG. 7. Top: Fraction of agents in ostracism as a function of the corruption rate $\rho_o(\alpha)$ fixing the delation rate to $\beta = 0.1$ and $\beta = 0.9$. Bottom: $\rho_O(\beta)$ for $\alpha = 0.1$ and $\alpha = 0.9$. The mean-field predictions for SIRS model are represented with dashed lines, whereas solid lines correspond to the HCO model. The reinsertion rate is fixed to $r = 0.1$ (left panels) and $r = 0.9$ (right panels).

For a comparison of both models in their mixed population equilibria regimes, we show in Fig. 6 the mean-field predictions for the fraction of corrupt agents ρ_c. The upper panels on this figure show the graphs of $\rho_c(\alpha)$ at several fixed values of β $(0.1, 0.9)$ and r $(0.1, 0.9)$. Beyond the transition point α_c, one could intuitively expect that the fraction of corrupt agents is always higher for the SIRS model due to the lack of the conversion flow forcing corrupt agents to recover honesty. This holds for low β values since the evolution of corruption is clearly much more boosted in the SIRS model and, as a result, the system undergoes the second transition toward a full corrupt society much before than for the HCO model. Interestingly, for very high values of the delation flow β, this phenomenon is reversed as clearly seen on the curves for $\beta = 0.9$ (upper panels), where $\rho_c^{\mathrm{HCO}}(\alpha) > \rho_c^{\mathrm{SIRS}}(\alpha)$. To heuristically explain this surprising result, we must realize that, for β values very close to 1 (and in the absence of conversion flow), corrupt agents are very likely to be delated and go to ostracism; this is a dynamically inactive state, and thus immune to infection, thus preventing the diffusion of corruption for a typical reinsertion time r^{-1}. In this sense, the existence of the warning to wrongdoers in the HCO model partially prevents the emergence of ostracism, thus facilitating the unfolding of corruption. Obviously, this effect is reinforced as r decreases, for it makes the staying time in the inactive state longer.

The lower panels on Fig. 6 show the graphs of $\rho_c(\beta)$ at several fixed values of α $(0.1, 0.9)$ and r $(0.1, 0.9)$. One sees there how the *honesty* transition occurs at lower delation values for the HCO model and the detrimental effect on corruption of the warning to wrongdoers, provided the delation rate β is not very large. Finally, the *undesired* effect of the warning to wrongdoers is observed for values of β close to maximum, when corruption better spreads for the HCO model.

In Fig. 7 we show, for both models, the mean-field predictions for the fraction of agents out of active population, i.e., in the O compartment. The upper panels in this figure show the graphs of $\rho_o(\alpha)$ at several fixed values of β $(0.1, 0.9)$ and r $(0.1, 0.9)$. We see that close above the *corruption* transition, $\alpha \gtrsim \alpha_c(\beta)$, ostracism increases faster for the SIRS model than for the HCO model, because in the latter converted corrupt agents can no longer be delated. This trend is obviously overcompensated, at very low values of β, before the SIRS transition to the full C state is reached for $\alpha < 1$, because ρ_o should then decrease to zero for the SIRS model, while the HCO model still remains in a mixed population equilibrium. The graphs of $\rho_o(\beta)$ represented on the lower panels of Fig. 7 illustrate further this change of trend in the ρ_o evolution between transitions for $\alpha = 0.1$. On the other hand, the comparison between left (corresponding to $r = 0.1$) and right $(r = 0.9)$ panels of this figure shows the important effect of increasing the reinsertion rate on the ρ_o fraction at equilibrium.

IV. VALIDATION OF THE THEORETICAL RESULTS

The theoretical analysis of the mean-field equations for both HCO and modified SIRS dynamics has shed light onto the interesting phenomena arising from mechanisms which drive the presence of corrupt agents in the society. Some of these phenomena are the existence of two critical transitions

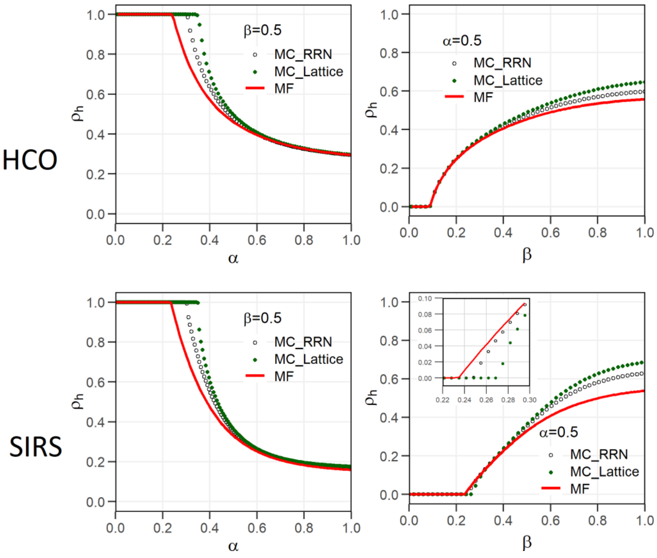

FIG. 8. Fraction of agents in H state, ρ_h, as a function of α and β, for mean-field approximation (red solid lines) and Monte Carlo simulations. Parameter r is fixed to 0.5 for all graphs. Simulations are performed on random regular networks with $\langle k \rangle = 4$ (black empty points) and lattice networks (green filled points), whose size is $N = 10^4$. Top panels correspond to HCO model and bottom ones to SIRS model.

or the crucial role that social interactions like delation or the warning to wrongdoers play in the evolution of corruption. Here we aim at validating these theoretical results by performing extensive Monte Carlo simulations on networked populations. At this point, for the sake of simplicity, we consider homogeneous networks (random regular networks or lattices) as the backbone for corruption and delation processes. A random regular network (RRN) is a random network where all the nodes share the same degree. In what follows, we just include mean-field theoretical predictions since, for regular topologies, both individual-based Markovian and mean-field approaches yield the same predictions for the evolution of corruption. To carry out the simulations, we start with a 10% of corrupt agents and we let the system evolve, following the microscopical rules defined in Sec. II, until the stationary state is reached. In this sense, to reduce stochastic fluctuations, we compute the equation of (stationary) state $\vec{\rho}(\alpha, \beta, r)$ by averaging them over 400 realizations.

Let us first analyze the evolution of the fraction of honest agents as a function of both delation and corruption probabilities. For this purpose, we fix the reinsertion flow to $r = 0.5$ and we represent the curves $\rho_h(\alpha)$ for several β values and its counterpart. Regarding the topologies for the contact networks, we make use of a RRN of $N = 10^4$ agents and $\langle k \rangle = 4$ and a square latice of $N = 10^4$ vertices with periodic boundary conditions. Figure 8 contains the comparison between theoretical predictions obtained via mean-field equations and the results yielded by simulations for both lattices and RRN. There we confirm that the mean-field theory developed above correctly predicts the existence of the two aforementioned transitions: The first one related to the destabilization of a honest population at $\alpha_c(\beta)$ and the second one associated with the irruption of honest agents in a totally corrupted society at $\beta_c(\alpha)$.

Although we are able to reproduce most of the phase diagrams, some relevant differences appear between theory and simulation, especially in the region of the parameters space close to the full honesty corner. In particular, it becomes evident that the value of α_c is underestimated by our formalism. These discrepancies are mainly based on two facts: the so-called echo chamber effect [31,32] and the influence of structural correlations [33–35]. On the one hand, the "echo chamber" effect is caused by the reinforcement of agents corruption from neighboring agents who have been previously corrupted by them. On the other hand, the formalism here presented is constructed by neglecting the possible dynamical correlations existing in the contact network. However, the existence of strongly correlated agents can be of great importance for the evolution of the system, especially close to the transition points. In particular, it has been shown recently that the presence of high-order structures like cycles or motifs tends to make the network more resilient to diffusion processes [35]. In our case, this is reflected in Fig. 8 where it becomes clear that spatially structured topologies (lattices) display a larger value of α_c than uncorrelated ones (RRN).

Interestingly, these structural correlations do not have any impact at the full corrupt corner, since mean-field equations accurately capture the value of β_c for both topologies. To explain this, we must realize that, apart from the delation processes caused by local interactions with honest agents, corrupt agents are also influenced by the warning to wrongdoers. This way, our hint is that having access to information about the global state of the network hinders the role of local interactions, thus giving rise to the same threshold for both topologies.

To confirm this statement, we now remove this effect and perform Monte Carlo simulations using the rules of the modified SIRS model. As observed in Fig. 8, the local nature of delation processes regains its relevance, leading to a splitting of the thresholds $\beta_c(\alpha)$ (see inset). This separation is much smaller than for the former $\alpha_c(\beta)$ splitting, given that the transition from a full corrupt population to the honesty is not affected by any "echo chamber" effect.

V. DISCUSSION AND PERSPECTIVES

In the sections above we have motivated the use of a simple compartmental population-flow model consisting of three states (compartments) and four flow channels connecting them, as a highly stylized model for the social dynamics of corruption, a punishable, and infectious, norm-violating behavior. In familiar terms to nonequilibrium statistical physics studies [36], the model is a non-Hamiltonian kinetic version of the spin-1 Ising model (also referred to as Blume-Emery-Griffiths model [37,38]), where the spin state of null z component represents strict social isolation (ostracism).

The model may also be viewed as an epidemic model, and thus one can capitalize on recent advances in contagion dynamics in complex social nets. However, two major differences respect to usual epidemic models are at work. First, recovery from infectious state (delation) requires the interaction with susceptible people in the local neighborhood, which might, however, be a plausible situation in epidemics. Second, the conversion flow, implementing the warning to wrongdoers effect of punishment, has no obvious counterpart in epidemic contexts.

The mean-field analysis reveals a phase diagram (in the three-dimensional space of model parameters) with three generic absorbing states: (i) full honesty, (ii) full corruption, and (iii) a mixed state with nonzero flow through all the channels. There is no coexistence of stable absorbing states (no multistability). The transition from full honesty to the mixed state is continuous, with a linear increase of the fraction of corrupt population, and is not influenced by the warning to wrongdoers. The transition from full corruption to the mixed state is also continuous, and the fraction of honest people increases linearly, as well; however, the warning to wrongdoers (*wtw*) plays a very important role regarding the onset of honest instabilities. On the one hand, *wtw* reduces the stability region of the full corrupt state. On the other hand, because the rate of conversion flow is assumed to be the fraction of punished population (not a local quantity), the mean-field prediction for the locus (a surface in the parameter space) of this transition becomes exact for random and nonrandom regular (homogeneous) networks. Both features are in contrast with the irrelevance of the *wtw* regarding the transition from full honesty to a mixed state and the (network dependent) shift of the locus of this transition that we observe in homogeneous graphs due to the presence of dynamical correlations induced by the existence of higher-order motifs in the structure of the network.

As for future works it would be interesting to relax some of the assumptions incorporated in our model. For instance, here we have assumed that (for fixed parameters) the rate of corruption is a one-variable function of the local fraction of corrupt agents and that the rate of delation is also a one-variable function of the local fraction of honest agents. A most promising prospective is to build up general and well-informed functions for the corruption and delation flow rates, so that the model in fact allocate a game-theoretic formulation, i.e., that these flow rates correspond to some game dynamics capable of incorporating fewer stylized ingredients than the ones included here. In this regard, the consideration of flow rates based on utility (benefit) functions not only requires many-variable functions but also enlarges the "information horizon" to second neighbors shell, likely expanding the scope of model potential applications.

ACKNOWLEDGMENTS

We acknowledge financial support from Ministerio de Economía y Competitividad (MINECO) through Grants No. FIS2015-71582-C2-1 and No. FIS2017-87519-P, Fondo Europeo de Desarrollo Regional (FEDER), and Gobierno de Aragón/Fondo Social Europeo (Grant No. E36-17R to FENOL group). F.B. and D.S.P. acknowledge financial support from Gobierno de Aragón through doctoral fellowships. D.L. acknowledges the support from the China Scholarship Council through a Ph.D fellowship. We are grateful to Prof. Y. Moreno for discussions on network epidemic models.

APPENDIX A: LINEAR STABILITY ANALYSIS

In Sec. III A, we have discussed the onset of *corruptive* destabilizing fluctuations in the fully honest H state by using a linear approximation to the competing flows that a generic small fluctuation induces, i.e., by analyzing the linear response to generic fluctuations. In the same way, we have also analyzed the onset of instabilities of the fully corrupt C state. In general systems of differential equations this physically appealing approach is rarely doable in such a simple way.

A more formal, and easier to generalize, method of analysis of a fixed point is provided by the spectral analysis (eigenvalues and its associated eigensubspaces) of the Jacobian matrix of the flow at this invariant point. This matrix is the linearized flow in the tangent space of the fixed point. We will use it here to show that both transitions, corruption and honesty, are unaffected if the relative order or priority of channels (delation and conversion) flowing out from the C compartment is reversed.

In the channels' priority scheme used in the main text, the trial for conversion is prior to delation, where from the velocity field, $\vec{F}(\vec{\rho}) = \dot{\vec{\rho}}$, is given by Eqs. (8), while if conversion is conditional on evading delation, then the corresponding equations of motion are slightly different:

$$\begin{aligned} \dot{\rho}_h &= -[f_\alpha + r + (1 - f_\beta)\rho_c]\rho_h \\ &\quad + [r + (1 - f_\beta)\rho_c](1 - \rho_c) \\ \dot{\rho}_c &= [f_\alpha + (1 - f_\beta)\rho_c]\rho_h + [(1 - f_\beta)\rho_c - 1]\rho_c. \end{aligned} \tag{A1}$$

The Jacobian matrix at a point (ρ_h, ρ_c) in phase space is defined as

$$J(\rho_h, \rho_c) = \begin{bmatrix} \frac{\partial \dot{\rho}_h(\rho_h, \rho_c)}{\partial \rho_h} & \frac{\partial \dot{\rho}_h(\rho_h, \rho_c)}{\partial \rho_c} \\ \frac{\partial \dot{\rho}_c(\rho_h, \rho_c)}{\partial \rho_h} & \frac{\partial \dot{\rho}_c(\rho_h, \rho_c)}{\partial \rho_c} \end{bmatrix}. \tag{A2}$$

Though at an arbitrary point of the simplex the Jacobian matrices of the flows (8) and (A1) are generally different, a direct calculation shows that at the full honesty corner ($\rho_h = 1,\ \rho_c = 0$), both are equal:

$$J_H = \begin{bmatrix} -r & -r - f'_\alpha(0) \\ 0 & f'_\alpha(0) - f_\beta(1) \end{bmatrix}.$$

Being this matrix triangular, the eigenvalues are just the diagonal elements, $\lambda_1^H = -r$, and $\lambda_2^H = f'_\alpha(0) - f_\beta(1)$. The H corner is stable whenever both eigenvalues are negative, thus requiring the inequality $f'_\alpha(0) < f_\beta(1)$, as we already know.

At the full corruption corner, the Jacobian matrices of the flows (8) and (A1) are also equal:

$$J_C = \begin{bmatrix} -r - 1 - f_\alpha(1) & -r - 1 \\ f_\alpha + 1 - f'_\beta & 1 \end{bmatrix}.$$

The eigenvalues of J_C are the roots of the characteristic polynomial $\lambda^2 - \lambda T + D$, where $T = \mathrm{Tr}(J_C)$ and $D = \mathrm{Det}(J_C)$ are respectively the trace and the determinant of the Jacobian matrix, explicitly given by

$$T = -r - f_\alpha(1) < 0, \quad D = r f_\alpha(1) - (r + 1) f'_\beta(0). \tag{A3}$$

Thus, the stability of the fully corrupt state requires that $D > 0$, that is,

$$\left(1 + \frac{1}{r}\right) f'_\beta(0) < f_\alpha(1).$$

We should note that the irrelevance of the priority of channels out from C regarding the onset of stability of both H and C corners by no means implies that in the parameter region where the attractor is an interior point of the simplex, this mixed population state is unaffected by the chosen priority; our numerical investigations clearly show that the surfaces of asymptotic equilibrium $\vec{\rho}(\alpha, \beta, r)$ are (in general, slightly) different for different choices.

APPENDIX B: THE $r \to 0^+$ LIMIT

Now we analyze the $r = 0^+$ limit of the model, where the O compartment has no leakages. This is the only locus in parameter space where the O corner, of phase space of feasible solutions, is a fixed point. In fact, as we will see, in the presence of conversion flow (warning to wrongdoers effect), the whole line $\rho_c = 0$ of the simplex (triangle) becomes a line of fixed points, a kind of peculiar, though well-behaved, parametric limit of the model. We will next show that, quite differently, the whole line $\rho_h = 0$ becomes a line of fixed points when the warning to wrongdoers effect is absent (SIRS limit).

Provided the warning to wrongdoers is an ingredient of the model (i.e., our fourth "model assumption" holds), it is easily seen, from the stationarity condition of ρ_o, that in the absence of reinsertion flow:

$$\dot{\rho}_o = -(F_h + F_c) = f_\beta \rho_c (\rho_h + \rho_c) = 0. \tag{B1}$$

First note that the solution that corresponds to $f_\beta = 0$ leads to $\rho_h = 0$ (i.e., the ρ_c axis) where conversion flow is positive unless at the C corner fixed point, where it is null. Also note that the solution $\rho_h + \rho_c = 0$ leads to $\rho_o = 1$, the O corner, which now becomes a fixed point. Finally, the case $\rho_c = 0$, corresponds to an asymptotic extinction of corruption, where the interior trajectories of the simplex should flow somewhere into the $\rho_c = 0$ locus. Also, the ρ_h axis becomes a line of fixed points. We now focus the analysis on this $O - H$ face of the simplex.

One easily realizes that any arbitrary fixed point (in the ρ_h axis) at abscissa ρ_h is marginally stable (zero linear response or induced flow) against a small perturbation $\delta\rho_h$ along the axis. The whole segment [0, 1] in the ρ_h axis is a line of indifferent equilibria regarding honesty perturbations. To inspect its stability against corruptive fluctuations in the linear regime $\delta\rho_c << 1$, note that the sign of the vertical component, $\dot{\rho}_c$, of the interior phase-space flow, close to this axis ($\rho_c << 1$), depends on the relative position respect to the $F_c = 0$ nullcline. As shown in Fig. 9, it is negative on the right side of the nullcline and positive on the left one. The locus of this nullcline is independent of the value of the reinsertion rate r, and when stability condition (9) of full honesty state H holds [say, for values of $\alpha < \alpha_c(\beta)$, in our explicit computations], the sign of the vertical component of the flow in the nearby interior points is negative for the whole unit interval of values of ρ_h. The unit interval is stable against corruptive fluctuations.

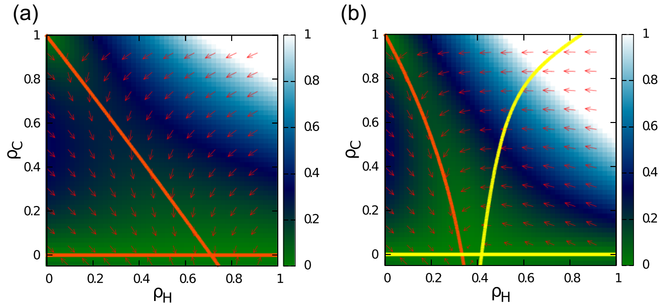

FIG. 9. The $F_h = 0$ nullcline (red line) and the $F_c = 0$ nullcline (yellow line) with a constant value of $\beta = 0.5$, $r = 0$. The infection rate is set to $\alpha = 0.1$ (a) and $\alpha = 0.5$ (b).

When the full honesty state is unstable, i.e., Eq. (9) does not hold [or $\alpha > \alpha_c(\beta)$, for our numerical results], the segment of indifferent equilibria is shortened to $[0, \rho_h^*]$, where ρ_h^* is the solution of the nonlinear equation

$$[1 + f'_\alpha(0) - f_\beta(\rho_h^*)]\rho_h^* = 1, \tag{B2}$$

that is, the point where the nullcline curve $F_c(\vec{\rho}) = 0$ intersects the ρ_h axis.

To the left of this nullcline, the flow (in the interior of the simplex) points vertically down, meaning that the segment $[0, \rho_h^*]$ is now the absorbing segment. To the right, on the contrary, the interior flow points upward and the equilibria with $\rho_h > \rho_h^*$ are linearly unstable.

The nulcline $F_h = 0$ has a peculiar $r \to 0$ limit. For $r = 0$, this nulcline is

$$-(f_\alpha + \rho_c)\rho_h + \rho_c(1 - \rho_c) = 0. \tag{B3}$$

By keeping second-order terms in the power expansion of $f_\alpha(\rho_c)$, we find that for small enough values of ρ_c, one obtains two solutions. The first one is $\rho_c = 0$, and the second solution reads:

$$\rho_c = \frac{\rho_h[1 + f'_\alpha(0)] - 1}{\frac{1}{2}\rho_h(f''_\alpha)(0) - 1}, \tag{B4}$$

This curve, Eq. (B4), is interior to the simplex if $\rho_h < \bar{\rho}_h$, where

$$\bar{\rho}_h = \frac{1}{1 + f'_\alpha(0)}, \tag{B5}$$

while for higher values of ρ_h it takes on negative values, outside the simplex of feasible states, as shown in Fig. 9.

Nevertheless, please note that, as soon as $r > 0$, the $F_h = 0$ nullcline only intersects the $\rho_c = 0$ value at the H corner and thus the reinsertion flow contracts the absorbent segment toward its upper bound, ρ_h^*, whose neighborhood along the nullcline curve $F_c(\vec{\rho}) = 0$ absorbs the whole interior flow. In other words, if we denote by $\rho_h^*(r)$ the asymptotic fraction of honest agents at a value of $r > 0$, then

$$\lim_{r \to 0} \rho_h^*(r) = \rho_h^*, \tag{B6}$$

where ρ_h^* has been just defined above, Eq. (B2), as the upper limit of the absorbent segment for $r = 0$. Note also that, when the H corner is stable [$\alpha < \alpha_c(\beta)$ for computations], and thus

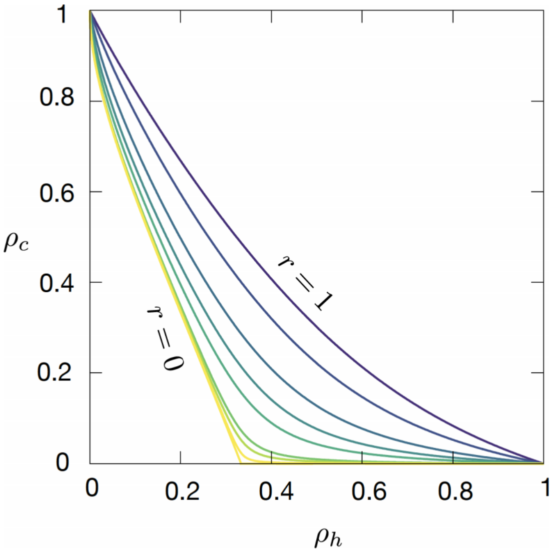

FIG. 10. Nullcline $F_h = 0$ computed according to Eq. (11) for several values of the reinsertion rate r. In particular, the reinsertion values displayed from top to bottom are $r = 1, 0.5, 0.2, 0.1, 0.05, 0.01, 0.005, 0.001, 0$. The corruption rate has been set to $\alpha = 0.5$.

the absorbent segment is the whole unit interval, as soon as $r > 0$, the reinsertion flow contracts it toward the H corner.

In Fig. 10 we show how the $r = 0^+$ limit of the $F_h = 0$ nullcline is approached. For values of $\rho_h < \bar{\rho}_h$ the $r = 0^+$ nullcline is the nonzero $\rho_c(\rho_h)$ solution of Eq. (B3), to which Eq. (B4) is a second-order approximation valid near the horizontal axis. But for $\rho_h > \bar{\rho}_h$, the nullcline limit is the zero solution of Eq. (B3). Thus, the $r = 0^+$ nullcline has a singularity (jump-discontinuous first derivative) at $\bar{\rho}_h$.

Finally, as a check of consistency of the previous analytical conclusions, note that it would be contradictory, and thus impossible, that ρ_h^* be placed to the left of $\bar{\rho}_h$, for that would imply a nullclines crossing leading to an interior fixed point. Some simple algebra on Eq. (B2) shows, with relief, that indeed $\bar{\rho}_h < \rho_h^*$ always holds.

To round up, now we analyze the $r = 0^+$ limit of the SIRS model discussed in Sec. III C. In other words, the situation that we now consider is that when both reinsertion and conversion flows are absent: The SIR limit.

The flow components are now simply

$$F_h = -f_\alpha \rho_h \tag{B7}$$

and

$$F_c = f_\alpha \rho_h - f_\beta \rho_c, \tag{B8}$$

from where the flows, δF_h and δF_c, induced by corruptive $\delta\rho_c$ and honesty $\delta\rho_h$ fluctuations can be easily analyzed in the linear regime.

From the stationarity condition for ρ_o, Eq. (B1), one sees that the $\rho_c = 0$ solution (the ρ_h axis) is a line of equilibria, indifferent regarding honesty fluctuations but unstable against corruptive fluctuations. Also the solution that corresponds to $f_\beta = 0$ is the ρ_c axis, which in the absence of conversion flow becomes an attractive set. In fact, it becomes a line of indifferent equilibria versus corruptive fluctuations, and stable against honesty fluctuations. Let us remark that, from Eq. (42), it becomes clear that the $r = 0^+$ limit of the $F_h = 0$ nullcline is now coincident with both axes: $f_\alpha = 0$ on the ρ_h axis, and $\rho_h = 0$ on the ρ_c axis.

APPENDIX C: ACCIDENTAL SYMMETRY IN THE SIRS MODEL

In Sec. III C we have noted that the kind of SIRS model that results from the removal of the warning to wrongdoers from the HCO model possess an *accidental* symmetry, equation (29), whenever the corruption transition probability $f_\alpha(\rho_c)$ and the delation transition probability $f_\beta(\rho_h)$ have the same functional form. This symmetry is not a fundamental symmetry, because the mechanisms underlying delation of a corrupt individual are surely very different from those driving the corruption of honest people, and one should expect, in general, that those differences translate into different functional forms for their transition probabilities.

We now provide an argument that prove the statement that, under the assumption that both transition probabilities have the same functional form, the *equation of stationary state* $\bar{\rho}(\alpha, \beta, r)$ is symmetric under the simultaneous interchange $\alpha \leftrightarrow \beta$ and $\rho_c \leftrightarrow \rho_h$:

$$\rho_c(\alpha, \beta) = \rho_h(\beta, \alpha) \text{ and } \rho_h(\alpha, \beta) = \rho_c(\beta, \alpha). \tag{C1}$$

In general terms, let us denote $\rho_h = u$, $\rho_c = v$, $x = \alpha$, and $y = \beta$. According to (22), in the stationary state both variables must fulfill

$$u = \frac{r(1 - v)}{r + g(x, v)} \tag{C2}$$

and

$$\frac{g(y, u)}{u} = \frac{g(x, v)}{v}, \tag{C3}$$

where r, x, and y are fixed parameters. By inserting (C3) into (C2), one arrives at

$$v = \frac{r(1 - u)}{r + g(y, u)}, \tag{C4}$$

and thus u and v also solve for Eqs. (C2) and (C4) for the fixed values of the parameters.

Now consider the solution u' and v' of (C2) and (C3) when the values of x and y are interchanged, so that they solve for the equations:

$$u' = \frac{r(1 - v')}{r + g(y, v')} \tag{C5}$$

and

$$\frac{g(x, u')}{u'} = \frac{g(y, v')}{v'}. \tag{C6}$$

Inserting (C6) into (C5), we obtain

$$v' = \frac{r(1 - u')}{r + g(x, u')}, \tag{C7}$$

and thus u' and v' also solve for Eqs. (C5) and (C7) for the fixed values of the parameters. Now note that Eq. (C5) is the same as (C4) and that (C7) is the same as (C2) and then $u = v'$ and $v = u'$, which completes the argument.

[1] T. Yamagishi, J. Pers. Soc. Psychol. **51**, 110 (1986).

[2] E. Fehr and Fischbacher, Evol. Hum. Behav. **25**, 63 (2004).

[3] K. Sigmund, H. De Silva, A. Traulsen, and C. Hauert, Nature (London) **466**, 861 (2010).

[4] H. Peyton Young, in *Genetic and Cultural Evolution of Cooperation*, edited by P. Hammerstein (Dahlem University Press, Cambridge, MA, 2003).

[5] H. Gintis, J. Henrich, S. Bowles, S. Boyd, and E. Fehr, Soc. Justice Res. **21**, 241 (2008).

[6] J. M. Weber and J. K. Murnighan, J. Pers. Soc. Psychol. **95**, 1340 (2008).

[7] W. Güth and H. Kliemt, Metroeconomica **45**, 155 (1994)

[8] C. Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge University Press, Cambridge, 2005).

[9] M. Chudek and J. Henrich, Trends Cogn. Sci. **15**, 218 (2011).

[10] J.-H. Lee, Y. Iwasa, U. Diekmann, and K. Sigmund, Proc. Natl. Acad. Sci USA **116**, 13276 (2019)

[11] V. N. Kolokoltsov, Int. J. Statist. Probab. **1**, 77 (2012).

[12] V. N. Kolokoltsov and O. A. Malafeyev, Dyn. Games Appl. **7**, 34 (2017).

[13] J.-H. Lee, K. Sigmund, U. Dieckmann, and Y. Iwasa, J. Theor. Biol. **367**, 1 (2015).

[14] J.-H. Lee, M. Jusup, and Y. Iwasa, J. Theor. Biol. **428**, 76 (2017).

[15] P. Verma and S. Sengupta, PLoS ONE **10**, e0133441 (2015).

[16] P. Verma, A. K. Nandi, and S. Sengupta, Sci. Rep. **7**, 42735 (2017).

[17] P. Verma, A. K. Nandi, and S. Sengupta, J. Theor. Biol. **450**, 43 (2018).

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

[19] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ, 1944).

[20] N. Boccara, *Modeling Complex Systems* (Springer, Berlin, 2004).

[21] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes in Networks*, (Cambridge University Press, Cambridge, 2008).

[22] H. Gintis, *Game Theory Evolving* (Princeton University Press, Princeton, NJ, 2009).

[23] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, 1998).

[24] Kinetic $s = 1$ Ising models have already a tradition in social dynamics (Schelling model).

[25] L. Gauvin, J.-P. Nadal, and J. Vannimenus, Phys. Rev. E **81**, 066120 (2010).

[26] M. E. J. Newman, The structure and function of complex networks, SIAM Rev. **45**, 167 (2003).

[27] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. **424**, 175 (2006).

[28] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, Europhys. Lett. **89**, 38009 (2010).

[29] S. Gómez, J. Gómez-Gardeñes, Y. Moreno, and A. Arenas, Phys. Rev. E **84**, 036105 (2011).

[30] X. Chen, R. Wang, M. Tang, S. Cai, H. E. Stanley, and L. A. Braunstein, New J. Phys. **20**, 013007 (2018).

[31] W. Wang, M. Tang, H. E. Stanley, and L. A. Braunstein, Rep. Prog. Phys. **80**, 036603 (2017).

[32] B. Karrer and M. E. J. Newman, Phys. Rev. E **82**, 016101 (2010).

[33] J. P. Gleeson, Phys. Rev. X **3**, 021004 (2013).

[34] E. Cator and P. VanMieghem, Phys. Rev. E **85**, 056111 (2012).

[35] S. Chandra, E. Ott, and M. Girvan, arXiv:1905.07433v2.

[36] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattices* (Cambridge University Press, Cambridge, 2005).

[37] In the (Hamiltonian) Blume-Emery-Griffiths model [38], the $s_z = 0$ spin state represents the isotope He-3 in helium 3-4 liquid mixtures, while in the Schelling model of urban segregation, see Ref. [25], it represents an empty flat in the urban neighborhood.

[38] M. Blume, V. J. Emery, and R. B. Griffiths, Phys. Rev. A **4**, 1071 (1971).