Rapid Communications   Editors' Suggestion

When big data fails: Adaptive agents using coarse-grained information have competitive advantage

V. Sasidevan,[1,2] Appilineni Kushal,[3] and Sitabhra Sinha[1,4]

[1]*The Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600113, India*
[2]*Department of Physics, Cochin University of Science and Technology, Cochin 682022, India*
[3]*Indian Institute of Science, C V Raman Road, Bangalore 560012, India*
[4]*Homi Bhabha National Institute, Anushaktinagar, Mumbai 400094, India*

(Received 6 September 2017; revised manuscript received 27 June 2018; published 8 August 2018)

The recent trend for acquiring big data assumes that possessing quantitatively more and qualitatively finer data necessarily provides an advantage that may be critical in competitive situations. Using a model complex adaptive system where agents compete for a limited resource using information coarse grained to different levels, we show that agents having access to more and better data perform worse than others in certain situations. The relation between information asymmetry and individual payoffs is seen to be complex, depending on the composition of the population of competing agents.

DOI: 10.1103/PhysRevE.98.020301

Agents in a population often coordinate their actions with that of their neighbors, resulting in striking forms, such as in swarming and flocking [1,2]. Typically, in such cases, individuals use information obtained from their local environment to adjust their actions in order to achieve some desired objective [3–6]. Emergent coordination is therefore crucially dependent on the information acquired by an agent and its ability to process it appropriately, which determines its future course of action. Often the objectives of different agents in a system may not be compatible with each other, for instance, when they are competing for a limited resource. Examples of such situations are abundant in nature, where individuals vie for food, shelter, and mating opportunities. Even in our more complex social environment, we regularly come across instances of such competition [7], e.g., choosing the least congested route through an urban road network [8,9] or anticipating the relative demand for a financial asset so as to profit by buying or selling it [10,11]. In these settings, individuals may use strategies which project information from past experiences to make decisions about the future course of action [12–14]. Conventional wisdom suggests that the relative success of an agent would increase with the quality and quantity of available data that would form the basis for its decisions. Indeed, the recent excitement about "big data" is partially based on the premise that access to more and better information provides a competitive advantage [15].

In this Rapid Communication, we show that agents using quantitatively more data that are also finely resolved (and hence also qualitatively superior) may not actually do better—and can in fact do worse—in certain situations when they are competing with agents that have access to less, as well as lower-resolution, information. The surprising result arises from emergent coordination in the collective activity of agents who use information at a particular coarse graining (say, level X). This leads to macroscopic patterns of behavior that may be discernible from the data only at a different level of coarse graining (say, level Y). Thus, if there are other agents in the population who have access to information about the system at this latter level Y, they can potentially exploit this predictability to their advantage. We also show that the relation between information asymmetry and the performance of agents is a complex one, depending on the relative fraction in the population of agents of each type. Thus, the utility of "big data" is contingent upon the precise nature of an agent's ecosystem comprising all its competitors. The premise that more and better information will automatically result in better performance, e.g., by improving predictive power, therefore needs to be treated with caution. This is especially true for competitive situations where adaptation through learning occurs, such as in financial markets [16,17].

To investigate how information asymmetry between agents affects their performance, we focus on a complex adaptive system where agents compete for a limited resource. Here, the heterogeneous agents use the different types of information that they have access to for the same purpose, viz., to have preferential access to the resource. In particular, we use the paradigm of the minority game (MG) [18–20] which has all the ingredients to address the above question in a quantitative manner. It also has the advantage that the classical version, in which all agents use the same coarse-grained information, is well understood and can be used as a benchmark for the more complex situation that is investigated here. We consider a population of an odd number N of agents who independently and simultaneously choose between two options (A and B, say) in each round. The option that is chosen by fewer agents is considered the better choice (outcome) in each round and leads to a higher payoff (say, 1), while those who had chosen the alternative receive a lower payoff (say, 0).

We assume that the population consists of different *types* of agents, each type having access only to data coarse grained to a particular level of resolution k (see Fig. 1). For clarity, we focus on the interaction between only two types of agents corresponding to the extremes of coarse graining, viz., the lowest resolution $k = 2$ (which we designate as type 1 agents) and the highest resolution $k = N + 1$ (type 2 agents). The former can only distinguish between $N_A > N/2$ and $N_A < N/2$ (i.e., whether A was chosen by the majority or not) in a

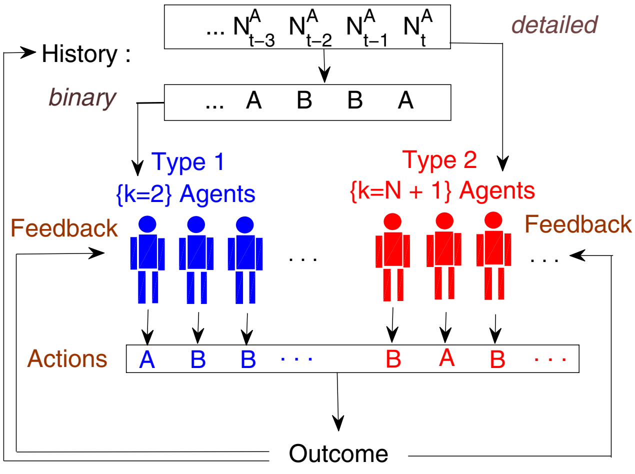

FIG. 1. A schematic representation of a complex adaptive system comprising N agents that are competing for a limited resource. Every agent has to choose between two possible actions (A or B) at each round, with the option chosen by the lesser number of agents being the better choice (outcome) in that round. Agents make decisions using strategies based on information about the collective choice in previous m rounds. The system can be in any of k possible states ($2 \leqslant k \leqslant N + 1$, depending on the level of coarse graining) at each round. Here, agents are distinguished into two classes (types 1 and 2) according to the two extreme levels of coarse-grained information, i.e., $k = 2$ and $k = N + 1$, respectively, that they have access to. After each round t, the information about the total number of agents choosing a specific action A (say), N_t^A, that is accessed only by type 2 agents, as well as binary information, viz., the choice of the minority (A or B) which is accessed only by type 1 agents, are added to the history of outcomes. The information about the outcome is also used as feedback for adaptive selection of strategies by the agents.

particular round [18] while the latter can determine the exact number of agents N_A opting for A in a round [21]. The memory length of each type of agent indicates the number of past rounds whose information they retain, and is denoted by m_1 (m_2) for a type 1 (type 2) agent. Each agent uses *strategies* that map the information about past events [m_1 bits for a type 1 agent, $m_2 \log_2(N + 1)$ bits for a type 2 agent] to the choice of action in the next round (i.e., A or B). Each agent initially chooses at random a small sample of strategies (e.g., of size 2 as here) from the set of all possible strategies, which is of size $2^{2^{m_1}}$ for a type 1 agent and $2^{(N+1)^{m_2}}$ for a type 2 agent. At each round, an agent scores the strategies according to the potential payoffs that would have been obtained by using them in the previous rounds (feedback), and uses the one having the highest score.

We first focus on the simplest case of a *single* type 1 agent with memory length m_1 interacting with a population of $N-1$ type 2 agents with memory length m_2. One may naively expect that agent(s) having more information at their disposal (measured in bits) will have an advantage over the other type of agent(s). Consequently, it would have been expected that when the number of bits m_1 in the information accessible to the type 1 agents is less than $m_2 \log_2(N + 1)$, the corresponding quantity for the type 2 agents, then the latter would have obtained a relatively higher payoff. This would also be in accordance with the intuitive notion that the highly resolved data of type 2 agents is qualitatively better than the low-resolution outcome data of type 1 agents. However, the mean payoffs of the two types of agents shown in Fig. 2 for different memory lengths m_1 and population sizes N reveals that the actual behavior is more complex.

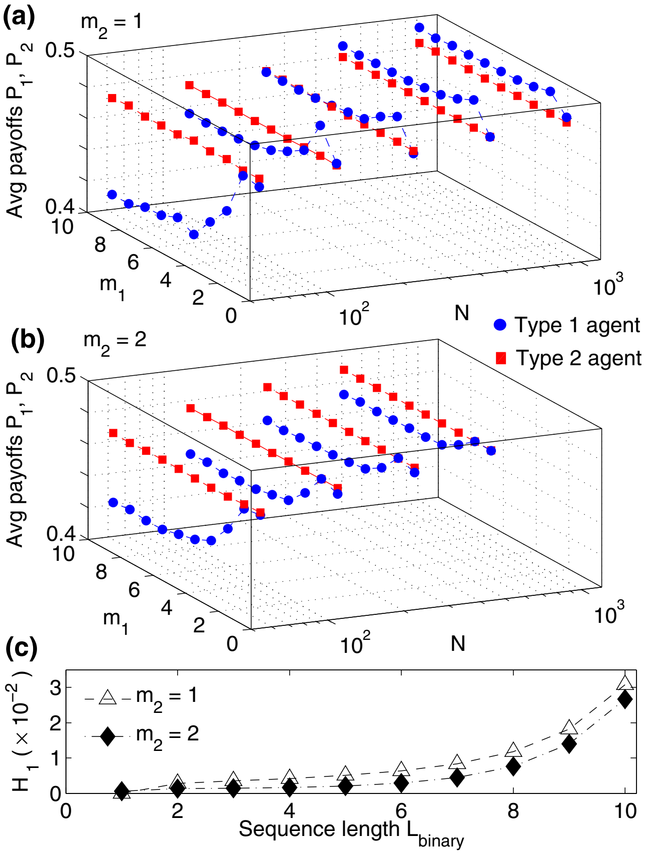

FIG. 2. Average payoffs P_1, P_2 of type 1 and type 2 agents (respectively) shown as a function of the memory length m_1 of a single type 1 agent interacting with $N - 1$ type 2 agents with memory length (a) $m_2 = 1$ and (b) $m_2 = 2$ for different population sizes N. In both cases, the type 1 agent receives the highest payoff when its memory length $m_1 = 2$. Note that only when $m_2 = 1$, the lone type 1 agent has a relative advantage over type 2 agents (viz., when the former has a lower memory m_1). Payoffs are averaged over 10^4 iterations in the steady state and over 100 different realizations. (c) Information content H_1 of the binary sequence containing the history of outcomes for a game involving only type 2 agents shown as a function of the sequence length L_{bin}. Results shown are for $N = 255$ agents and averaged over 100 realizations.

The most surprising outcome for the case when the type 2 agents have memory length $m_2 = 1$ [Fig. 2(a)] is that the type 1 agent is able to acquire a relatively higher payoff at low values of m_1 even though the information accessible to it is highly coarse grained and quantitatively much less compared to the rest of the population. Moreover, the range of m_1 over which the type 1 agent does better than type 2 agents is seen to increase with N. Thus, the success of an agent in a complex adaptive system, where the information accessible by the individual entities differs both in terms of quality and quantity, is not entirely determined by the amount and resolution of the data at its disposal. Instead, as we show below, it depends more on whether discernible patterns in the behavior of the population are present at the level of coarse graining it has access to. When the memory length of the type 2 agents is increased to $m_2 = 2$ [Fig. 2(b)], the type 1 agent is no longer observed to have a higher payoff than the rest of the population, regardless of its memory length m_1. Note that type 2 agents attain the highest degree of emergent coordination among themselves for $m_2 = 2$ independent of N [21]. Thus, it is not surprising that the lone type 1 agent will not be able to outperform the optimally coordinated population of type 2 agents. However, as we shall show below, introducing multiple type 1 agents makes it possible for these mutually competing individuals to develop emergent coordination within themselves by which they can outperform type 2 agents with $m_2 = 2$. If $m_2 > 2$, the behavior of the type 2 agents is indistinguishable from agents randomly choosing between A and B [21]. As there is no predictability in the time series available to the type 1 agent that it can exploit, it will on average receive essentially the same payoff as the rest of the population.

Note that the above results are in stark contrast to the case of a population comprising only type 1 agents, but where one or more agents have a larger memory length than others. When they have access to the same binary history of outcomes, it is known that the agents with smaller memory are always at a disadvantage when playing against a group of larger memory agents [22,23]. Thus, the results reported here cannot be reproduced if we simply replace the type 2 agents with type 1 agents having quantitatively equivalent memory in terms of size (i.e., $m_2 \sim \log_2 N$). Another important point to consider is that if the population is homogeneous, comprising only type 1 agents, it is known that the system behaves essentially identically even if the agents are provided with a random sequence instead of the endogenous history of past outcomes [24,25]. Thus, the only requirement for the agents to display the collective dynamics observed in this system is that all of them possess the same information regardless of whether it is true or false. However, when the population consists of both type 1 and type 2 agents, replacing the actual history by random sequences coarse grained at different levels as appropriate for the different types of agents leads to a different outcome. For example, a single type 1 agent interacting with $N - 1$ type 2 agents no longer enjoys a relative advantage with regard to the rest of the population (see Supplemental Material [26]) unlike what is observed when the agents are provided with the actual history [Fig. 2(a)].

To explain the relative performance of different types of agents having access to information at the two extreme levels of coarse graining, we focus on the information content in the history of outcomes that can be exploited by the agents to their advantage. As we shall see, the collective action of any one type of agents may result in predictable patterns at the other level of coarse graining and hence observable only to these other agents. This "useful" information content above the noise level can be quantified by measuring the predictability of a particular choice (say, A) being the outcome in a particular round, given the history of past outcomes. This history can be either the *binary* sequence of outcomes A, B or the *detailed* time series of the number of agents $\{N_t^A\}$ choosing a particular option A, the former (latter) being accessible only to a type 1 (type 2) agent. We can therefore define two distinct information measures, viz., $H_1 = \sum_{u_{L_{\rm bin}}} P(u_{L_{\rm bin}})[P(A|u_{L_{\rm bin}}) - (1/2)]^2$, and $H_2 = \sum_{u_{L_{\rm det}}} P(u_{L_{\rm det}})[P(A|u_{L_{\rm det}}) - (1/2)]^2$. Here, $u_{L_{\rm bin}}$ is the binary sequence of outcomes for the previous $L_{\rm bin}$ rounds while $u_{L_{\rm det}}$ is the sequence of integers, each lying between 0 and N, representing the number of agents choosing A in the previous $L_{\rm det}$ rounds. The probability with which a particular sequence of L successive outcomes is observed is denoted as $P(u_L)$, while $P(A|u_L)$ represents the conditional probability that the outcome A follows the sequence u_L.

Let us consider a population comprising only type 2 agents having memory length m_2. The collective behavior of such agents generates a history of *binary* outcomes whose information content H_1 is shown in Fig. 2(c) for $m_2 = 1$ and 2. Note that this information cannot be used by the type 2 agents themselves, whose strategies are based on $u_{L_{\rm det}}$ but is accessible in principle to a hypothetical type 1 agent whose strategies use $u_{L_{\rm bin}}$. We observe that H_1 increases with the length of the binary sequence, $L_{\rm bin}$, over the range of sequence lengths considered here, with $H_1 = 0$ when the history is restricted to the immediately preceding round, i.e., $L_{\rm bin} = 1$. Thus, if a type 1 agent with memory length m_1 is introduced into this population, it can make use of the predictability present in the binary sequence accessible to it when $m_1 > 1$. As m_1 increases, the number of possible strategies that can be used by the type 1 agent increases exponentially ($=2^{2^{m_1}}$). It therefore becomes progressively less likely that the agent will randomly pick the strategy that can optimally exploit the predictability present in $u_{L_{\rm bin}}$. This implies that the highest payoff for a type 1 agent is achieved for the lowest value of m_1 having nonzero information content, i.e., $m_1 = 2$, as is indeed confirmed by Fig. 2.

The above arguments explain the performance of a single type 1 agent interacting with a population of agents of the other type [26]. However, in reality, the number of each type of agents having access to data at different levels of coarse graining can be arbitrary. We shall now consider the situation where the relative fraction of the two types of agents present in the population is varied between the two extreme cases considered earlier. Figure 3 shows the average payoffs P_1, P_2 for a population of N agents of types 1 and 2, respectively, for different values of the fraction f_1 and memory length m_1 of type 1 agents, keeping the memory length of type 2 agents fixed [viz., $m_2 = 1$ for Fig. 3(a), and $=2$ for Fig. 3(b)]. As in Fig. 2, here also we see that type 1 agents can outperform type 2 agents even when the quantity of information (m_1 bits) available to the former is much less than that for the latter [$m_2 \log_2(N + 1)$ bits]. This is seen in the low m_1 region in Fig. 3 when f_1 is low, where the few type 1 agents receive a higher payoff than the more numerous type 2 agents. As $f_1 \to 0$, we approach the case of a single type 1 agent (playing $N - 1$ type 2 agents) that achieves a maximum payoff at $m_1^* = 2$, which is indeed observed in Fig. 3. On the other hand, when $f_1 \to 1$, the maximum payoff should occur at $m_1^* \simeq \log_2(0.337N)$, as it is a conventional game between type 1 agents [27]. Indeed, for any fraction $0 < f_1 < 1$ of type 1 agents, their best performance is achieved for a memory size m_1^* that lies between 2 and $\log_2(0.337N)$ (indicated by the dashed curves in Fig. 3). Thus, multiple type 1 agents can achieve a higher individual payoff together than by playing singly against a population of type 2

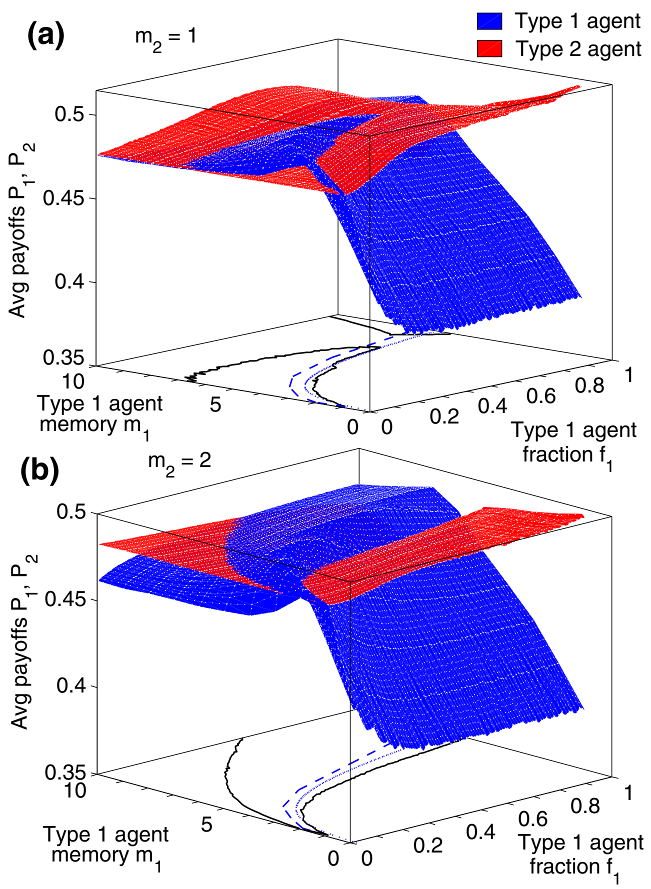

FIG. 3. The average payoffs P_1, P_2 of type 1 (shown in blue) and type 2 agents (red) comprising a population of size N (= 255) for different population fractions f_1 and memory length m_1 of the type 1 agents. The memory lengths of the type 2 agents are fixed at m_2 [=1 for (a) and =2 for (b)]. The contours separate the regions in the (m_1, f_1) parameter space where type 1 agents have a relative advantage over type 2 agents and vice versa. The dashed curve represents the optimal population fraction f_1^* of type 1 agents with a given memory length m_1 at which they receive the highest payoff. The dotted curve is the value of m_1 at which Nf_1 type 1 agents are expected to have maximum payoff in the absence of any type 2 agents. Payoffs are averaged over 10^4 iterations in the steady state and over 100 different realizations.

agents, suggesting an important role of emergent coordination among a group of competing agents who are distinguished by the nature of the information available to them.

Let us now consider the performance of the type 2 agents. When playing against type 1 agents with low memory length m_1, type 2 agents achieve their highest payoff when $f_1 \to 1$ [26]. In other words, it is important to have the size of the group to which detailed data is available as small as possible in order for them to achieve a maximum payoff in this regime of low m_1. When more agents have access to this data (i.e., decreasing f_1), their payoff is eroded until they actually perform worse than the type 1 agents who have coarser-grained data. Thus, access to more and better data is not by itself a determining factor for success in a complex adaptive situation.

As the memory length m_1 of the type 1 agents increases, the optimal population fraction at which type 2 agents achieve the highest payoff decreases from the neighborhood of $f_1 = 1$. In fact, type 2 agents with $m_2 = 2$ (the optimal memory length for a population exclusively composed of such agents) achieve their best performance when $f_1 \to 0$. Thus, in this high m_1 regime ($m_1 \geqslant 6$ for $m_2 = 2$), type 2 agents achieve high payoffs by dominating the population. By contrast, type 1 agents do better for large f_1 as a result of emergent coordination within their group. Indeed, in this regime, for any given m_1, the payoff of type 1 agents increases with f_1. Thus, the outcome is not symmetric for agents having access to information at the two extreme levels of coarse graining [28].

We note here that concerns about the potential pitfalls of big data have been voiced earlier [29]. There have been several critical discussions from different perspectives [30] on how big data projects can fail through improper planning, e.g., as a result of unclear objectives or cost-related issues [31]. Another strand of literature deals with problems related to the analysis of big data in distributed computing frameworks [32]. In this Rapid Communication, however, a more fundamental mechanism is revealed which severely limits the value of big data analytics in competitive environments where agents can adapt their behavior based on information about past outcomes. As most socioeconomic phenomena of interest can be seen to be a result of interactions between agents in a complex adaptive system [33,34], the results reported here may play a key role in explaining the behavior observed in a variety of such systems.

To conclude, we have shown that information asymmetry among agents in a complex adaptive system can have surprising consequences. Specifically, in a system where agents compete for a scarce or limited resource using strategies based on information about the collective behavior in previous interactions, asymmetry arising from individuals having access only to data coarse grained to different levels can result in agents with more and better data performing worse than others under certain circumstances. Such counterintuitive effects arise from predictable patterns emerging in the collective information about the system at a certain level of coarse graining and thus discernible only to agents privy to that level. This provides them a competitive advantage when the population is dominated by agents of a different type who do not have access to the coarse-graining level at which such patterns generated by their own collective activity are apparent. The relation between the relative performance of the different types of agents and the nature of information asymmetry is therefore crucially dependent on the exact composition of the population to which they belong. Our results imply that striving to acquire and process ever increasing quantities of data in the hope of making more accurate predictions in complex adaptive systems, such as financial markets, may sometimes be counterproductive.

We would like to thank Shakti N. Menon for helpful discussions. This work was supported in part by the IMSc Econophysics project (XII Plan) funded by the Department of Atomic Energy, Government of India. We thank IMSc for providing access to the supercomputing cluster "Satpura," which is partially funded by DST.

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

[2] F. Vanni, M. Lukovic, and P. Grigolini, Phys. Rev. Lett. **107**, 078103 (2011).

[3] M. A. Nowak and R. M. May, Nature (London) **359**, 826 (1992).

[4] I. Couzin, Nature (London) **445**, 715 (2007).

[5] D. J. G. Pearce, A. M. Miller, G. Rowlands, and M. S. Turner, Proc. Natl. Acad. Sci. USA **111**, 10422 (2014).

[6] F. L. Pinheiro, F. C. Santos, and J. M. Pacheco, Phys. Rev. Lett. **116**, 128702 (2016).

[7] A. Vespignani, Science **325**, 425 (2009).

[8] H. Youn, M. T. Gastner, and H. Jeong, Phys. Rev. Lett. **101**, 128701 (2008).

[9] S. H. Lee and P. Holme, Phys. Rev. Lett. **108**, 128701 (2012).

[10] D. Challet, A. Chessa, M. Marsili, and Y.-C. Zhang, Quant. Finance **1**, 168 (2001).

[11] S. V. Vikram and S. Sinha, Phys. Rev. E **83**, 016101 (2011).

[12] H. Simon, Q. J. Econ. **69**, 99 (1955).

[13] W. B. Arthur, Am. Econ. Rev. **84**, 406 (1994).

[14] W. B. Arthur, Science **284**, 107 (1999).

[15] C. A. Mattmann, Nature (London) **493**, 473 (2013).

[16] M. Potters, R. Cont, and J. Bouchaud, Europhys. Lett. **41**, 239 (2007).

[17] S. Sinha, A. Chatterjee, A. Chakraborti, and B. K. Chakrabarti, *Econophysics: An Introduction* (Wiley-VCH, Weinheim, 2010).

[18] D. Challet and Y.-C. Zhang, Physica A **246**, 407 (1997).

[19] E. Moro, in *Advances in Condensed Matter and Statistical Mechanics*, edited by E. Korutcheva and R. Cuerno (Nova Science Publishers, New York, 2004).

[20] D. Challet, M. Marsili, and Y.-C. Zhang, *Minority Games: Interacting Agents in Financial Markets* (Oxford University Press, Oxford, UK, 2005).

[21] V. Sasidevan, J. Stat. Mech. (2016) 073405.

[22] N. F. Johnson, P. M. Hui, D. Zheng, and M. Hart, J. Phys. A **32**, L427 (1999).

[23] N. F. Johnson, M. Hart, P. M. Hui, and D. Zheng, Int. J. Theor. Appl. Finance **3**, 443 (2000).

[24] A. Cavagna, Phys. Rev. E **59**, R3783 (1999).

[25] D. Challet and M. Marsili, Phys. Rev. E **62**, 1862 (2000).

[26] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.98.020301 for a discussion of a single type 2 agent interacting with a population of $N-1$ type 1 agents.

[27] D. Challet, M. Marsili, and R. Zecchina, Phys. Rev. Lett. **84**, 1824 (2000).

[28] If m_1 is increased further, the strategy space for type 1 agents becomes too large, with their action effectively reducing to random choice between the two options. If m_2 is also sufficiently large (>2), both types of agents achieve similar payoffs, equal to that obtained by a random choice strategy (see Fig. S5 in Supplemental Material [26]).

[29] N. Silver, *The Signal and the Noise* (Penguin, New York, 2012).

[30] A. Katal, M. Wazid, and R. H. Goudar, in *Sixth International Conference on Contemporary Computing (IC3)* (IEEE, Noida, 2013), p. 404.

[31] B. Marr, Where big data projects fail, retrieved from https://www.forbes.com/sites/bernardmarr/2015/03/17/where-big-data-projects-fail/#35b7cec8239f (2015).

[32] L. Hübschle-Schneider and P. Sanders, arXiv:1710.08255.

[33] J. H. Miller and S. E. Page, *Complex Adaptive Systems: An Introduction to Computational Models of Social Life* (Princeton University Press, Princeton, NJ, 2009).

[34] J. H. Holland and J. H. Miller, Am. Econ. Rev. **81**, 365 (1991).