Impact of committed minorities: Unveiling critical mass of cooperation in the iterated prisoner's dilemma game

Zhixue He,[1,2] Chen Shen,[3,*] Lei Shi,[1,4,†] and Jun Tanimoto[3,2]

[1]*School of Statistics and Mathematics, Yunnan University of Finance and Economics, 650221 Kunming, China*
[2]*Interdisciplinary Graduate School of Engineering Sciences, Kyushu University, Fukuoka 816-8580, Japan*
[3]*Faculty of Engineering Sciences, Kyushu University, Fukuoka 816-8580, Japan*
[4]*Interdisciplinary Research Institute of Data Science, Shanghai Lixin University of Accounting and Finance, 201209 Shanghai, China*

(Received 17 July 2023; accepted 2 January 2024; published 16 January 2024)

The critical mass effect is a prevailing topic in the study of complex systems. Recent research indicates that a committed minority of cooperators, unwavering in their beliefs and consistently maintaining cooperation, can effectively foster widespread cooperation in social dilemma games. However, achieving a critical mass of cooperation in the one-shot prisoner's dilemma requires stricter conditions. The underlying mechanism behind this effect remains unclear, particularly in the context of repeated interactions. This work aims to investigate the influence of a committed minority on cooperation in the iterated prisoner's dilemma game, a widely studied model of repeated interactions between individuals confronting a social dilemma. In contrast to previous findings, we identify tipping points for both well-mixed and structured populations. Our findings demonstrate that a committed minority of unconditional cooperators can induce full cooperation under weak imitation conditions. Conversely, a committed minority of conditional cooperators, who employ extortion strategy, can promote widespread cooperation under strong imitation conditions. These results are consistent across various network topologies and imitation rules, suggesting that critical mass effects may be a universal principle in social dilemma games. Moreover, we discovered that an excessive density of committed extortioners can hinder cooperation in structured populations. This research advances our understanding of the role of committed minorities in shaping social behavior and provides valuable insights into cooperation dynamics.

DOI: 10.1103/PhysRevResearch.6.013062

I. INTRODUCTION

The study of the *critical mass effect* explores significant changes in state and properties that occur within complex nonlinear systems [1–5]. In natural systems, even minor perturbations near tipping points can lead to sudden and drastic transformations, often referred to as "explosive changes." For instance, the intricate interplay between system structure and dynamic characteristics can result in explosive synchronization [6]. Similarly, establishing a specific threshold for quarantine probability has been observed to effectively mitigate the spread of epidemics by isolating infected individuals [7], while a minority of dissenting particles can disrupt the flocking state in active Brownian motion [8]. Remarkably, these critical phenomena also appear in social systems such as voting and social segregation. Whether it involves the convergence of social opinion in voter models [9–11] or the emergence of distinct segregation patterns characterized by similar preferences, such as observed in Schelling's segregation model [12], these social phenomena are initiated by a few individuals and propagate from individual behavior to collective behavior [13,14]. Such critical phenomena provide valuable insights into system dynamics, particularly emphasizing the crucial role of a minority of individuals in driving overall evolution of social systems.

Studying the evolution of human behavior in social dilemmas provides profound insights into the development of social systems, as it involves the inherent conflict between individual interests and collective interests [15–22]. Pairwise social dilemma games, such as the harmony game, stag-hunt game, snowdrift game, and prisoner's dilemma (PD) game, capture various forms of conflicting interests and exhibit different equilibrium properties within these game types [15,16]. In the realm of social dilemma games, the influence of committed individuals who steadfastly adhere to their beliefs and consistently maintain their behavior on the emergence of cooperation has been explored using evolutionary game theory [9,23–31]. Specifically, research indicates that committed cooperators, known as "zealots," who always cooperate with their opponents, even as a minority, can effectively trigger widespread cooperation in stag-hunt dilemma games, which represent coordination problems [27]. Stag-hunt games have two pure strategy Nash equilibria, and the initial fraction of cooperators in the population heavily influences the outcomes, enabling committed cooperators to more easily induce large-scale cooperation [27]. Conversely, in snowdrift games, which represent situations where

*steven_shen91@hotmail.com
†shi_lei65@hotmail.com

individuals benefit from being different from others (anticoordination problems), the presence of committed defectors as a minority can facilitate large-scale cooperation in well-mixed populations, but not in structured populations [31]. As for prisoner's dilemma games, which capture the fundamental conflict between individual and collective interests and are frequently used to address the cooperation conundrum, the emergence of large-scale cooperation depends on factors such as population structure, imitation intensity, or update rules, rather than relying solely on the presence of committed cooperators [26,27,30].

However, previous research has primarily focused on analyzing one-shot game scenarios, where players do not have repeated encounters with the same opponents [25–31]. In contrast, in realistic scenarios with repeated interactions, the dynamic nature of the process allows for the emergence of a broader range of strategies that can be adopted by players over time [17–19]. One well-known strategy in iterated games is tit-for-tat (TFT), where individuals start by cooperating and then mirror their opponent's previous actions in subsequent interactions. Despite its simplicity, TFT has proven highly successful in sustaining cooperation within populations in iterated prisoner's dilemma (IPD) games [17,18]. Expanding on the concept of iterated games, Press and Dyson introduced a class of strategies known as zero-determined (ZD) strategies [19], which establish a linear relationship between an individual's payoff and their opponent's payoff. Extortion strategies, a subset of ZD strategies, enable players to achieve higher payoffs than their opponents by a certain percentage [32] and can serve as a catalyst for promoting cooperation under specific conditions or in certain population scenarios [33–37].

Building upon this, the following question arises: Can committed individuals initiate a universal critical mass effect in the PD game regardless of factors such as imitation intensity, network topologies, or update rules when repeated interactions are considered? To investigate this, we concentrate on the PD game, which poses a significant challenge for the emergence of cooperation, instead of examining the critical mass of cooperation in games such as stag hunt and snowdrift. Our objective is to understand the mechanisms behind how a minority of committed individuals can trigger widespread cooperation within the context of the IPD game. We focus on three typical strategies in the IPD game, namely *unconditional cooperation*, *unconditional defection*, and *extortion*, as well as taking into account the presence of committed individuals who adopt one of these strategies. Our findings reveal that even a minority of committed individuals can effectively trigger widespread collective cooperation in both well-mixed and structured populations, surpassing the scope of a one-shot PD game. Interestingly, we have discovered that committed extortioners, under a strong imitation scenario where a more successful individual is always imitated and a less successful one is never imitated [38], exhibit a significant critical mass effect that enhances cooperation. Similarly, committed cooperators demonstrate a similar critical mass effect in a weak imitation scenario where the strategy imitation is basically random. Additionally, we evaluate the robustness of this phenomenon by analyzing different network topologies and alternative strategy-updating rules, thereby confirming the impact of committed individuals on the critical mass effect of cooperation. In conclusion, these results imply that the critical mass of cooperation could serve as a universal principle in realistic social dilemma games, as it remains resilient given different network topologies, network structures, imitation intensities, and update rules.

II. MODEL

Game model and strategies setting. In a typical IPD game, a player continuously interacts with the same coplayer in pairs. In each interaction, both players either choose to cooperate (C), incurring a cost c to yield a benefit b for the coplayer, or defect (D) and do nothing. Players can receive the rewards $R = b - c$ through mutual cooperation, whereas mutual defection leads to the punishment $P = 0$ for them. A defector can gain the temptation to defect $T = b$ from unilateral defection, while a cooperator only receives the sucker's payoff, $S = -c$. The memory-1 strategy is a type of strategy frequently discussed in IPD games, in which players take actions based on the results of the previous interaction. A memory-1 strategy can be represented by a four-element tuple $[p_{CC}, p_{CD}, p_{DC}, p_{DD}]$, where p_{ij} $(i, j \in \{C, D\})$ denotes the probability of choosing cooperation when the player chose action i and the coplayer chose action j in the previous interaction. Along this line, the typical strategies can be written as follows: *unconditional cooperation* C [1, 1, 1, 1], *unconditional defection* D [0, 0, 0, 0], and tit-for-tat TFT [1, 0, 1, 0]. The extortion strategy can be denoted as $E\chi$ $[p, 0, q, 0]$, where the elements satisfy the condition $0 < p, q < 1$. Compared with the TFT strategy, the extortion strategy exhibits a lower probability of cooperation in response to a cooperative opponent in the previous interaction. This characteristic of reducing cooperative tendencies enables an extortioner x to guarantee that his or her own "surplus" exceeds that of the coplayer y by χ-fold, represented by $r_x - P = \chi(r_y - P)$, where r_i is the long-term payoff of player i [19,32]. The TFT strategy can be regarded as a limiting case of the extortion strategy with $\chi = 1$ [32], where the player and coplayer receive the same payoff. This connection enables us to solely consider the extortion strategy in the mode and also catch a glimpse of the results associated with the TFT strategy. Moreover, although the values of p and q can impact the χ in the extortion strategy [19,32], our interest lies in understanding the impact of introducing the extortion strategy, rather than delving into the specific variations resulting from different values of p and q. Therefore we concentrate solely on the extortion factor χ to explore the extortion strategy, without considering the specific values of p and q.

Investigating evolutionary results necessitates building upon the outcomes of strategy pair interactions. Adhering to the methodology outlined in Ref. [32], the mean payoffs per interaction between the aforementioned strategies can be expressed using the following matrix:

$$\begin{array}{c} \\ C \\ D \\ E_\chi \end{array}\!\!\begin{array}{c} \begin{array}{ccc} C & D & E_\chi \end{array} \\ \begin{pmatrix} b-c & -c & \frac{b^2-c^2}{b\chi+c} \\ b & 0 & 0 \\ \frac{(b^2-c^2)\chi}{b\chi+c} & 0 & 0 \end{pmatrix} \end{array}. \tag{1}$$

TABLE I. List of parameter symbols in our model.

Symbol	Description	Range of parameter values
N	Population size	3000–40 000
ρ	Proportion of committed individuals in a population	[0, 0.5]
$\langle K\rangle$	The average number of neighbors of a player in a structured population	4
b	Temptation to defection	[1.0, 2.0]
χ	Extortion factor, a multiple of the extortioner's surplus relative to the coplayer's surplus	[0, 10]
α	Imitation intensity, measures how strongly players base their decisions on payoff comparisons	[0,1]

In the given payoff matrix, it is preferable for a player to choose strategy C in response to an extortioner, rather than selecting strategy $E\chi$. When facing a cooperator, choosing $E\chi$ is also better than choosing C. This suggests a snowdriftlike relationship between C and $E\chi$, as equilibrium arises when both parties choose different strategies, similar to a snowdrift game. We set $b - c = 1$ to simplify the payoff elements; so we only need to focus on the extortion factor $\chi \geqslant 1$ and the temptation b.

Population setting. Consider a population of size N which consists of a proportion ρ $(0 < \rho < 0.5)$ of committed extortioners who always adopt the $E\chi$ strategy and a proportion of normal players equal to $1 - \rho$. To examine the robustness of the model, we explore the results based on two typical populations: well-mixed and structured populations. In a well-mixed population (equivalent to a fully connected network), each player interacts in a pairwise game with all $N - 1$ other players. In structured populations, players are placed on network nodes and can only interact with immediate neighbors, with both regular and heterogeneous networks being taken into account. For the structured regular network, we utilize both grid lattice and regular small-world (SW) network structures, where each node in these networks has the same degree. The construction of a regular SW network follows the method described in Ref. [39]; in this method we start with a regular ring network and the connected edges of some nodes are randomly exchanged with a rewiring probability of 0.1. Compared with a lattice network, a regular SW network with the same average degree exhibits a higher clustering coefficient and a smaller average shortest path length [16,40]. For heterogeneous networks, we utilize the random network generated by the Erdős-Rényi (ER) method in Ref. [41] and the scale-free networks generated by the Barabási-Albert (BA) algorithm in Ref. [40]. These networks exhibit a certain level of randomness where each node has a distinct degree; degree distributions in these networks are Poisson and power-law distributions [16,40,41], respectively. All of the mentioned networks are consistently configured with an average degree $\langle K\rangle$ of 4.

Game dynamic. We utilize an asynchronous Monte Carlo simulation (MCS) approach. Initially, each normal player holds one of three strategies with equal probability. In each MCS time step, players acquire payoffs through playing pairwise games with their direct neighbors. Then, they update their strategies through social imitation. Following these works [20–22], we implement the finite population analog of replicator dynamics in strategy updating, allowing for the inclusion of network heterogeneity in game dynamics. Specifically, a focal player, denoted as i, imitates the strategy of a randomly chosen neighbor, denoted as j, with a probability $w_{i\leftarrow j}$. This probability is determined by the difference in payoffs:

$$w_{i\leftarrow j} = \begin{cases} \frac{r_j - r_i}{H\cdot \max(K_j, K_i)} & \text{if } r_j > r_i \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where r_x is the player's payoff obtained from interactions with all neighbors at the current MCS time step, K_x is the number of neighbors of the player x, and H is the maximum possible payoff difference in a pairwise game ($H = b + c$ since $\chi \geqslant 1$). Equation (2) represents the deterministic strong imitation scenario where only successful players (who have the high payoffs in the game) can be imitated, and the imitation depends on payoff comparison. The denominator ensures that this difference in payoff can be transformed into a probability form. We introduce the imitation intensity parameter α to investigate weak imitation scenarios in which player imitation is less dependent on payoff comparisons, thereby testing the model's robustness. The modification to the player's strategy imitation probability is as follows:

$$\tilde{w}_{i\leftarrow j} = \frac{1-\alpha}{2} + \alpha \cdot w_{i\leftarrow j}. \tag{3}$$

As $\alpha \to 1$, the game dynamics revert to Eq. (2). Conversely, as $\alpha \to 0$, the imitation dynamics tend to be random and independent of payoff comparisons (i.e., the weak imitation case). Otherwise, we also explore the results based on the so-called Fermi rule [38,42] to assess the robustness of the results. Compared with Eq. (2), the Fermi update rule introduces some level of noise into the player's strategy imitation, allowing even unsuccessful individuals to potentially be imitated (refer to Appendix A 3 for additional details).

The MCS is conducted on networks of size $N =$ 3000–40 000. We primarily focused on the scenario of strong imitation (i.e., $\alpha = 1$). We average the results from more than 50 simulations, where each simulation is obtained by averaging the last 3000 time steps over a total of 5×10^4 MCS time steps. This ensures that the fraction of each strategy in the population remains stable. The parameter symbols used in our model are summarized in Table I.

III. RESULTS

A. Results for well-mixed populations

Before delving into the impact of committed individuals in the IPD game, let us revisit the influence of committed cooperators in a one-shot prisoner's dilemma game. In the absence of any reciprocal mechanism, defection undeniably

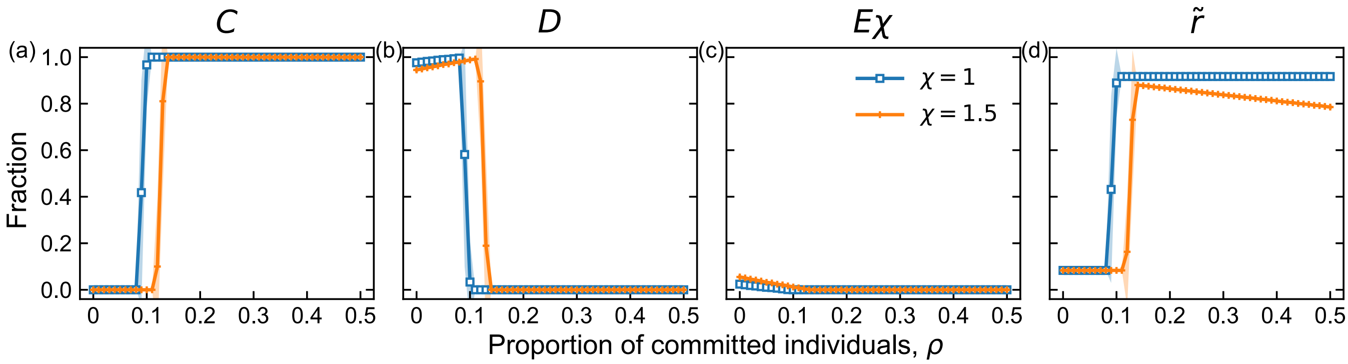

FIG. 1. Committed extortioners exert a critical mass effect on the emergence of cooperation in the iterated prisoner's dilemma game within a well-mixed population. The figure illustrates the average fraction of (a) *cooperation* C, (b) *defection* D, (c) *extortion* $E\chi$, and (d) normalized average payoff $\tilde{r}$ for normal players, as a function of the proportion of committed individuals (ρ). Shown are the results of introducing the committed extortioners at $\chi = 1$ and $\chi = 1.5$, with the parameter b set to 1.1. The shaded areas represent the standard deviation of the results.

emerges as the only Nash equilibrium strategy in such games. Cardillo and Masuda [27] found that introducing committed cooperators into a well-mixed population does not lead to the occurrence of critical mass effects in a one-shot PD game. The research suggests that committed cooperators exhibit a critical mass effect promoting cooperation only under weak imitation conditions. These findings suggested that the occurrence of critical mass effects in a one-shot PD game is dependent on the imitation intensity parameter.

In this paper, we broaden our scope beyond the committed cooperators and explore whether a minority of committed individuals, encompassing both committed unconditional cooperators and committed conditional cooperators, can demonstrate a universal critical mass effect. When taking into account the more common scenario of repeated interactions between players, interestingly, we observe the emergence of crucial phenomena triggered by committed individuals in the IPD game, as illustrated in Fig. 1. The introduction of a minority of committed conditional cooperators who adopt the $E\chi$ strategy can have a significant impact on the outcome of cooperation. In particular, when $\rho \approx 0.125$, the presence of committed extortioners can significantly alter the outcome from the extinction of cooperation (i.e., $F_C = 0$) to its dominance (i.e., $F_C = 1$). Even for $\chi = 1$, where the extortioners are unable to gain an advantage over their opponent, this phenomenon exists, and the corresponding critical value of ρ is approximately 0.09. On the one hand, the snowdrift-like relationship between C and $E\chi$ can sustain reciprocity between committed extortionists and normal cooperators. On the other hand, defectors are unable to gain benefits by exploiting committed extortioners, reducing the advantage of defection over cooperation. Therefore committed extortioners can support the emergence of cooperation.

Then, we explore the impact of committed individuals on the payoffs of regular players. Figure 1(d) shows the normalized average payoff (i.e., the payoffs scaled to the range 0–1) for normal players. When defectors are entirely eliminated, individuals adopting the $E\chi$ do not affect the payoffs of normal cooperators when $\chi = 1$, their behavior being equivalent to that of unconditional cooperators. Nevertheless, for $\chi > 1$, extortioners can achieve a greater surplus compared with cooperators, as indicated by Eq. (1). As the proportion of committed extortioners rises, the anticipated result is a decrease in payoffs for regular players.

We further explore whether committed individuals can still induce critical effects under varying levels of imitation intensity. In strong imitation scenarios, the presence of committed extortioners facilitates the evolution of cooperation. The existence of committed extortioners in the population diminishes the profit advantage of defectors since they are unable to exploit extortioners as stated in the payoff matrix. Conversely, in weak imitation scenarios, committed individuals act more as broadcasters of their own strategies to the entire population. As a result, the critical mass effect of promoting cooperation, triggered by committed extortioners, diminishes as the imitation intensity decreases (see Fig. 6 in Appendix A 1). However, committed cooperators can still induce the critical mass effect and achieve pure cooperation under weak imitation. These findings suggest that committed individuals possess a broad capability to elicit the critical mass effect in repeated games, regardless of the imitation intensity. However, as the extortion factors increase, the committed extortioners further reduce the profits of normal cooperators, which weakens the facilitating role of committed extortioners. This attenuation effect eventually leads to the disappearance of critical mass effects (as depicted in Fig. 5 in Appendix A 1).

B. Results for structured populations

The critical mass effects induced by committed cooperators in a one-shot PD game within a structured population depend on the specific network structures. This effect is primarily observed in heterogeneous networks, such as scale-free networks [27]. In lattice networks, the role of committed cooperators is counterproductive as it inhibits the evolution of cooperation [26]. Firstly, we focus on the scenario of a lattice network. Figure 2 suggests that the presence of a minority of committed extortioners can greatly enhance the overall level of cooperation among normal players. In particular, under the condition of high temptation (i.e., $b > 1.06$) where cooperation cannot emerge traditionally, the presence of even a small fraction of committed extortioners can shift the outcome from complete extinction of cooperation to its existence, similar to

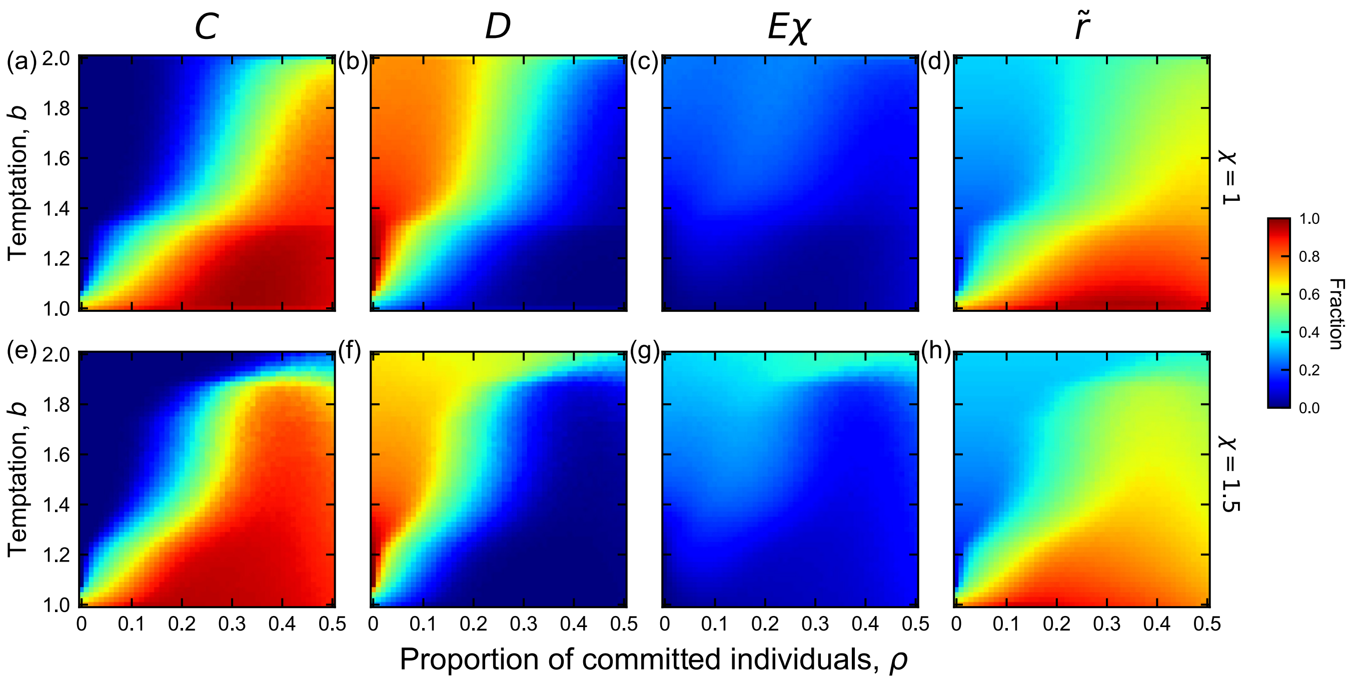

FIG. 2. On a regular lattice, a minority of committed extortioners can exert a critical mass effect on the emergence of cooperation in the iterated prisoner's dilemma game under the strong imitation scenario, which goes beyond the scope of the one-shot prisoner's dilemma game. The figure illustrates the average fraction (color-coded) of C, D, and $E\chi$ strategies and the normalized average payoff $\tilde{r}$ for normal players, plotted against the proportion of committed individuals ρ and the temptation parameter b. (a)–(d) Results obtained at $\chi = 1$. (e)–(h) Results obtained at $\chi = 1.5$.

observations made in well-mixed populations. For $\chi = 1$ and $\chi = 1.5$, the introduction of committed extortioners results in similar outcomes. Consistent with results from the well-mixed population, committed extortioners have been found to be effective under strong imitation conditions, whereas committed unconditional cooperators can also trigger critical effects in weak imitation scenarios (see Fig. 7 in Appendix A 2). Furthermore, a high extortion factor negatively impacts their ability to promote cooperation (as illustrated in Fig. 11 in Appendix A 2).

It is notable that while a minority of committed individuals can trigger critical mass effects, introducing more committed individuals may not always efficiently promote cooperation. In particular, under low-temptation conditions (i.e., $b \lesssim 1.06$), cooperators are capable of resisting defector invasion even in the absence of committed individuals. Introducing a minority of committed individuals can further enhance cooperation. However, when more committed individuals are introduced (i.e., $\rho \gtrsim 0.38$), regardless of whether they are able to obtain a surplus payoff (i.e., $\chi = 1$ and $\chi = 1.5$), it can be observed that cooperation declines [see Figs. 15(a) and 15(e) in Appendix A 3]. Otherwise, the introduction of more committed individuals results in decreased payoff levels for normal players [see Figs. 15(d) and 15(h) in Appendix A 3]. This reveals the existence of an optimal ρ that enables the committed individuals to effectively promote cooperation. In subsequent research, we will delve deeper into investigating the underlying reasons behind this phenomenon.

To further investigate the generality of critical effects induced by committed individuals in structured populations, we explore the outcomes based on regular small-world networks, Erdős-Rényi (ER) random networks, and Barabási-Albert (BA) scale-free networks [4,16,40]. In Fig. 3, results suggest that the outcomes obtained on SW and ER networks are similar to those observed on a lattice network. Specifically, the presence of committed extortioners effectively triggers the critical mass effect and promotes cooperation under strong imitation (see Fig. 3). The critical mass effect of committed unconditional cooperators can also be observed under weak imitation (see Figs. 7–9 in Appendix A 2). Remarkably, the fraction of cooperation increases significantly when $\rho = 0.06$ in the BA network, which implies that committed extortioners exhibit a critical mass effect on promoting cooperation. We also explore scenarios involving varying levels of imitation intensity (see Fig. 10 in Appendix A 2), where we observe that committed extortioners are effective exclusively under conditions of strong imitation. In contrast, committed unconditional cooperators are capable of promoting cooperation regardless of the imitation intensity within the scale-free (SF) network.

In order to validate the robustness of our findings, we examine a scenario where individual strategy imitation follows the Fermi rule [43]. This rule introduces a certain degree of noise during the strategy-updating process. In contrast to what is described in Eq. (2), this rule allows for the possibility of imitating an unsuccessful strategy. Significantly, the model consistently produces similar results, even when the strategy update rule is altered (as shown in Figs. 12–14 in Appendix A 3). These results demonstrate the universality of critical mass effects, irrespective of the particular strategy update rules employed.

We note that an excessive proportion of committed extortioners can lead to a lower fraction of C in Figs. 2 and 3.

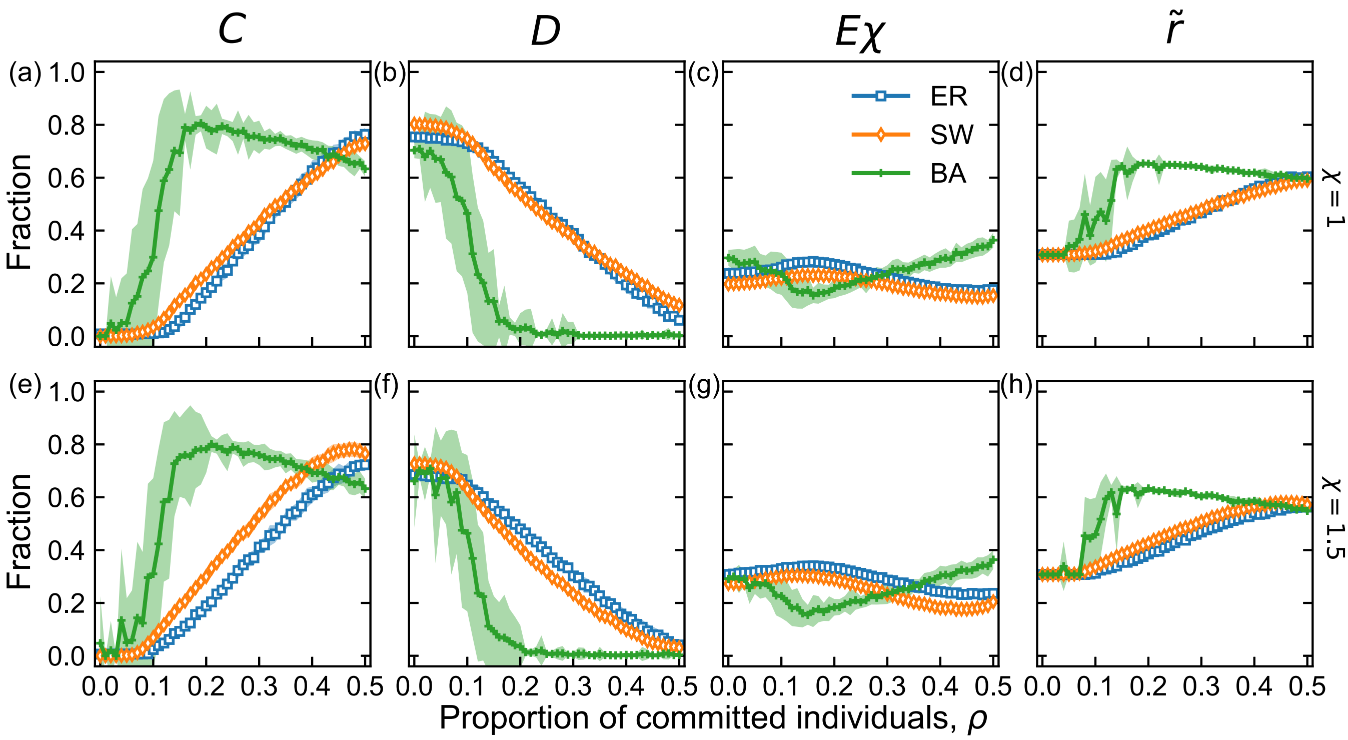

FIG. 3. Committed extortioners can promote collective cooperation in regular and heterogeneous networks. Depicted are the average fractions of C, D, and $E\chi$ strategies and the normalized average payoff $\tilde{r}$ for normal players as a function of the proportion of committed extortioners ρ for (a)–(d) $\chi = 1$ and (e)–(h) $\chi = 1.5$. The parameters are set to $N = 4000$ and $b = 1.8$. We consider three representative network structures: regular small-world (SW) networks, Erdős-Rényi (ER) networks, and Barabási-Albert (BA) scale-free networks. The average degree $\langle K \rangle$ of these networks is fixed at 4. When generating the network structure, the rewiring probability of the regular SW network is set to 0.1. The generation of the BA network starts with five nodes and adds four edges per step. The shaded areas represent the standard deviation of the outcome.

In order to explore the underlying mechanisms governing the impact of committed extortioners on cooperation, we analyze the temporal evolution and representative evolutionary snapshots across various proportions of committed extortioners in a lattice network. We focus on the lattice network as it provides more intuitive insights. During the initial stages of evolution, certain cooperators form clusters and manage to survive within the population, as depicted in the third-from-left panel in Fig. 4. In the absence of committed individuals, the formation of cooperation-extortion alliances becomes unfeasible due to a low extortion factor (i.e., $\chi = 1.5$) [35]. Consequently, the fraction of C experiences a significant decline, ultimately leading to the extinction of cooperation, as illustrated in Fig. 4(a). Conversely, even a minority of committed extortioners (i.e., $\rho = 0.05$) can sustain the existence of cooperative clusters through cooperation-extortion alliances, as shown in Fig. 4(b). When a moderate proportion (i.e., $\rho = 0.3$) of committed extortioners is present, the likelihood of normal cooperators encountering them increases, thereby promoting the extensive formation of cooperation-extortion alliances, which contributes to the expansion of the cooperative cluster, as indicated in the rightmost panel of Fig. 4(c). As a result, the fraction of C increases, leading to the eventual elimination of defectors. However, an excessive proportion of committed extortionists can lead to a segregation of normal noncooperative individuals from cooperators, as depicted in the second-from-right panel in Fig. 4(d). This segregation impedes the continued expansion of the cooperative cluster, resulting in an overall lower level of cooperation compared with the scenario with $\rho = 0.3$.

IV. CONCLUSIONS

In conclusion, our study expands the theory of the critical mass of cooperation from the one-shot prisoner's dilemma game to the iterated prisoner's dilemma (IPD) game. Unlike the findings in the one-shot prisoner's dilemma game, where the critical mass of cooperation relies on weak imitation intensity or scale-free networks under strong imitation intensity [27], our extensive Monte Carlo simulations provide compelling evidence for the significant role of committed individuals in driving widespread cooperation in the context of the IPD game. We consistently observed that a minority of committed unconditional cooperators play a pivotal role in generating critical mass effects of promoting cooperation, especially in the case of weak imitation intensity, which is in line with previous research [27,30,31]. Our findings suggest that when the imitation intensity is strong, indicating that individuals place significant value on the differences in payoffs when updating their strategies, even a minority of committed conditional cooperators, such as committed extortioners, can effectively initiate critical mass effects of cooperation. This holds true irrespective of the extortioners' ability to gain surplus payoffs.

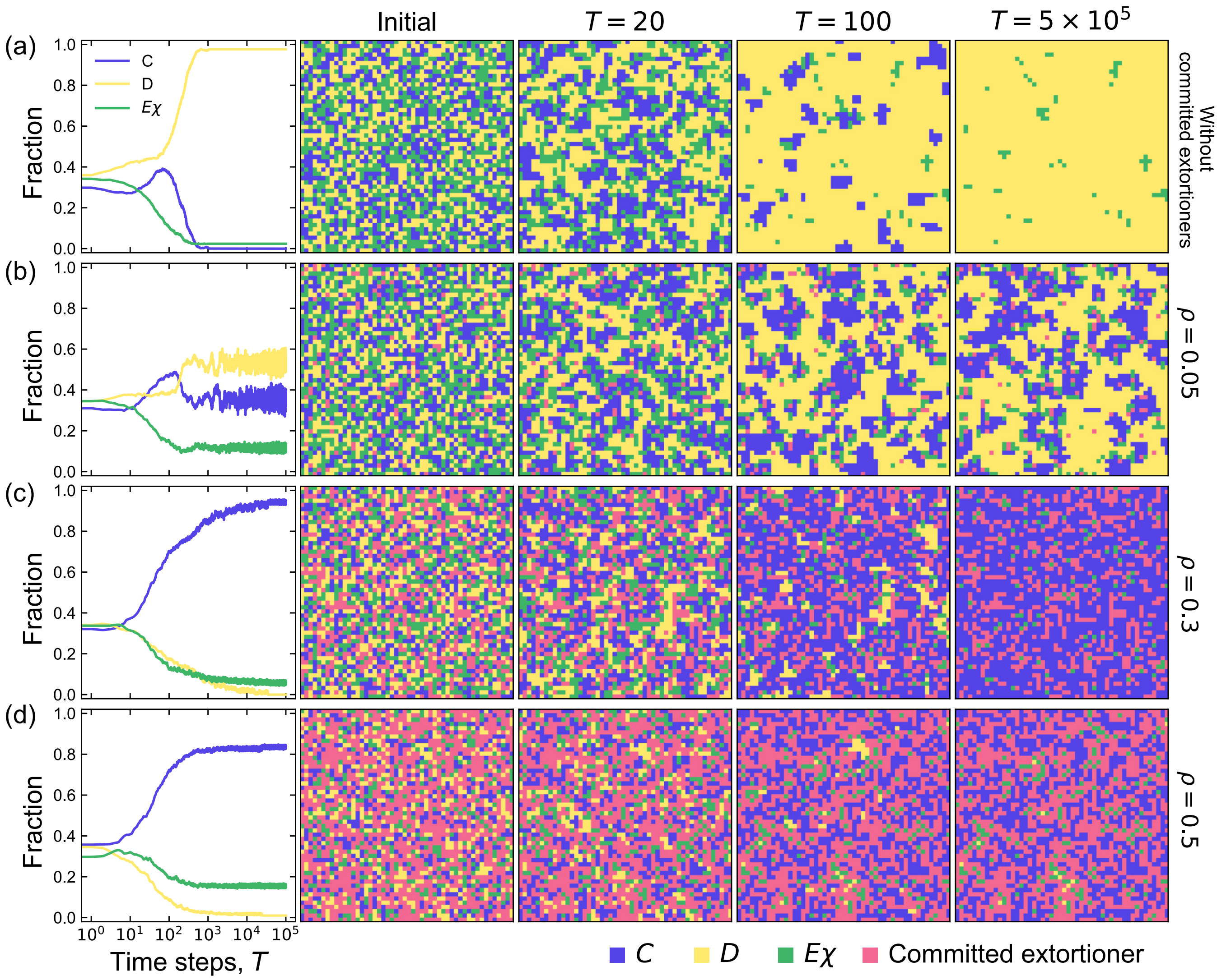

FIG. 4. An excessive proportion of committed extortioners hinders the evolution of cooperation. Depicted are the fractions of the three strategies' payoffs as a function of the number of time steps T (leftmost column) and typical evolutionary snapshots. Blue, yellow, green, and pink colors in the snapshots represent the cooperator, defector, extortioner, and committed extortioner, respectively. The other parameters are set to $b = 1.2$ and $\chi = 1.5$.

Our thorough testing across various network structures and strategy update mechanisms reinforces the reliability of these findings. The results demonstrate that the presence of committed individuals significantly influences cooperation, regardless of network structure or update rules. This indicates that the critical mass effect, driven by committed participants, might be a widespread phenomenon in social systems with self-interested individuals. Additionally, we found that a small number of committed extortioners can significantly boost cooperation in structured populations. However, an excessively high number of such extortioners can create barriers between cooperators and defectors, hindering cooperation spread and diminishing the overall benefits for regular players. Therefore there appears to be an optimal proportion of committed individuals that can maximize cooperation within a network.

In contrast to the one-shot and anonymous PD game, where participants must choose between unconditional cooperation or defection without knowledge of their opponents' information, this scenario presents significant challenges for sustaining cooperation. Such a simplistic approach confines individuals to one of two rigid roles: either committed cooperator or committed defector. However, this limitation is alleviated when we move beyond the one-shot game framework and consider scenarios involving repeated interactions. Repeated games introduce a spectrum of strategies influenced by various memory lengths, as elaborated in Ref. [44]. In these contexts, players can adopt more sophisticated strategies, such as tit-for-tat strategies, zero-determined strategies, and others. Transitioning to repeated PD games unveils a universal principle: Committed individuals can drive the critical mass effect for cooperation in various scenarios, whether with strong or weak imitation strength scenarios. This contrasts with the traditional one-shot PD game where defection dominates, even considering committed cooperators.

To examine the robustness of the critical mass of committed individuals given different player update rules, we utilized two specific rules as examples, based on the disparity in players' payoffs. The first is a replicator-like dynamics rule, a deterministic update rule where players imitate only the best-performing strategies. The second is the Fermi update rule, functioning stochastically and allowing for decision errors and imitation of less successful players. These rules exemplify stochastic and deterministic processes, respectively. Our findings show that the critical mass of cooperation is robust given both stochastic and deterministic rules. Given the shared underlying principle of these rules, we posit that our results are likely to extend to other update rules incorporating payoff differences. This suggests that committed conditional cooperators can trigger a critical mass of cooperation in scenarios with strong imitation.

In the study of cooperation evolution within repeated games, a notable variant is the finitely repeated game. In these games, players interact with their opponents across multiple rounds, with the knowledge of when these interactions will conclude. As the end of the game approaches, end-game effects often emerge, leading to a decline in cooperative behavior [45]. This phenomenon is characterized by players opting to defect early, aiming to maximize their short-term gains. Intriguingly, Mao *et al.* [46] conducted behavioral experiments on finite repeated games and observed the presence of a minority of individuals who engage in conditional cooperation in real-world scenarios. This minority plays a crucial role in preventing premature defections, thereby stabilizing the erosion of cooperation. This finding underscores the significance of a minority in maintaining cooperation in finite repeated games. In contrast, our model is based on another type of repeated game—the infinitely repeated game, where players participate in an unlimited number of interactions without knowledge of their termination. Given the fundamental nature of both game types in the study of cooperation evolution, our work aims to furnish a theoretical framework. This framework seeks to delve into the impact of committed minorities on cooperation, enriching our understanding in a more holistic manner.

The study of strategies is an essential aspect of comprehending the dynamics of cooperation evolution in repeated games. Previous research has shown that strategies such as TFT and win-stay, lose-shift (WSLS) effectively promote cooperation [17,18,34]. When considering structured populations, the intermediate "crossover" strategy that lies between TFT and WSLS emerges as the winner in multiple strategy evolutions and exhibits robustness [47]. In this paper, we focus primarily on investigating the extortion strategy. When compared with the TFT strategy, individuals who adopt extortion tend to cooperate less with their opponents over long-term interactions [32,34], thereby weakening the effect of direct reciprocity within our model, allowing us to focus on analyzing the role of committed individuals in repeated games. Despite stringent conditions, committed individuals can still demonstrate the ability to sustain cooperation. To gain a deeper understanding of the impact of committed individuals' strategies, it is essential for future research to encompass the comprehensive strategy space with memory-1 strategies. Furthermore, it is important to consider the influence of irrational factors and external environmental disturbances, such as decision noise and mutation, on the decision-making process of real individuals. Integrating these factors can contribute to a deeper understanding of how committed individuals influence collective behavior in more realistic situations.

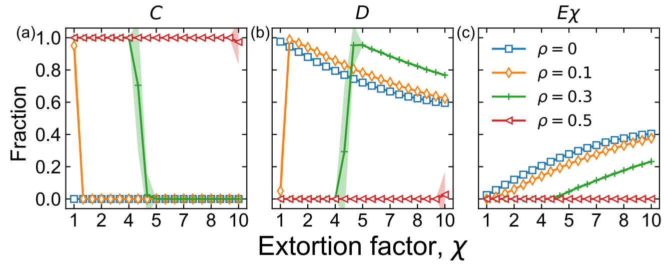

FIG. 5. As the extortion factor increases, the committed extortioner's role in promoting cooperation diminishes in a well-mixed population. Depicted are the average fractions of (a) C, (b) D, and (c) $E\chi$ among normal players as a function of the extortion factor χ for committed extortioners. The parameter b is set to 1.1. The shaded areas represent the standard deviation of the outcome.

Although one-shot games with some prior information available are more common in realistic scenarios than repeated games or one-shot prisoner's dilemma games without any information, we chose to disregard this situation in our model. Our primary focus was to answer how and why committed individuals can elicit a universal critical mass effect of cooperation, independent of imitation intensity, network topologies, or other updating rules. However, we acknowledge that future studies could explore the critical mass of cooperation within such a framework, especially in the context of indirect reciprocity. By expanding our understanding of the critical mass of cooperation within the framework of indirect reciprocity and reputation dynamics, we can gain a more comprehensive understanding of the factors that influence cooperation in real-world social systems. This could shed light on the mechanisms that facilitate or hinder the formation of critical mass effects of cooperation in complex social systems.

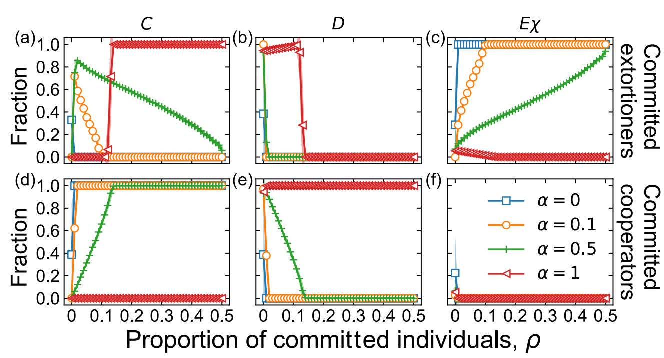

FIG. 6. In a well-mixed population, committed extortioners only promote cooperation under strong imitation intensity, while committed cooperators only promote cooperation under weak imitation. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners individuals and cooperators. Strong imitation $\alpha \to 1$ means that the player's strategy imitation depends on the difference in payoffs, while the individual performs random imitation under weak imitation $\alpha \to 0$. The parameters are set to $b = 1.1$ and $\chi = 1.5$. The shaded areas represent the standard deviation of the outcome.

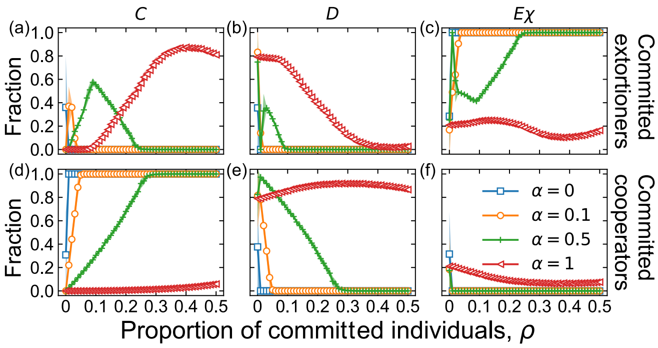

FIG. 7. The impact of committed individuals on cooperation at various imitation intensities in a lattice network. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. The parameters are set to $b = 1.5$ and $\chi = 1.5$. The shaded areas represent the standard deviation of the outcome.

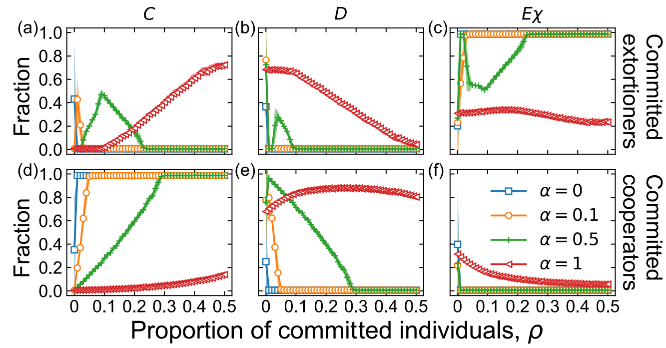

FIG. 9. The impact of committed individuals on cooperation at various imitation intensities in an Erdős-Rényi network. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. The parameters are set to $b = 1.5$, $\chi = 1.5$, $N = 4000$, and $\langle K \rangle = 4$. The shaded areas represent the standard deviation of the outcome.

Overall, our model uncovers extensive critical effects initiated by committed individuals and underscores the pivotal role played by a minority of committed individuals in promoting large-scale cooperation. Our results thus enrich the study of the critical mass of cooperation within social systems, especially in systems that contain conflicts between personal interest and collective interest.

ACKNOWLEDGMENTS

We acknowledge the support provided by (i) Major Program of National Fund of Philosophy and Social Science of China (Grants No. 22&ZD158 and No. 22VRCO49)) to L.S.; (ii) JSPS Postdoctoral Fellowship Program for Foreign Researchers (Grant No. P21374) and an accompanying grant-in-aid for scientific research from JSPS KAKENHI (Grant No. JP 22F31374) to C.S.; (iii) a grant-in-aid for scientific research from JSPS KAKENHI (Grant No. JP 20H02314) to J.T.; and (iv) China Scholarship Council (Grant No. 202308530309) and Yunnan Provincial Department of Education Science Research Fund Project (Project No. 2023Y0619) to Z.H.

Z.H. and C.S. conceptualized and designed the study; Z.H. and C.S. performed simulations and wrote the initial draft of the manuscript; L.S. and J.T. provided overall project supervision; all authors read and approved the final manuscript.

The authors declare no conflict of interests.

APPENDIX

1. The influence of imitation intensity and extortion factor on committed individuals in a mixed population

Figure 5 reports the fractions of three strategies as a function of extortion factor χ when committed extortioners are introduced into a well-mixed population. While the existence of committed extortioners can weaken the advantage of defectors, an increase in the extortion factor allows extortioners to obtain even greater surplus benefits from their interactions

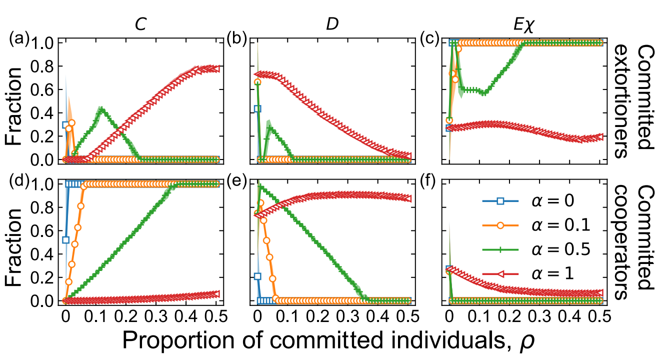

FIG. 8. The impact of committed individuals on cooperation at various imitation intensities in a regular small-world network. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. The parameters are set to $b = 1.8$, $\chi = 1.5$, $N = 4000$, and $\langle K \rangle = 4$. The shaded areas represent the standard deviation of the outcome.

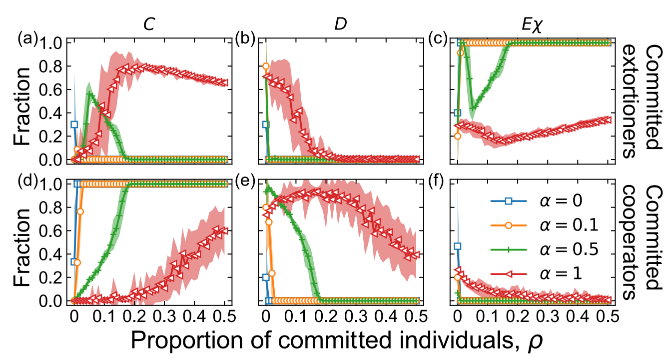

FIG. 10. The impact of committed individuals on cooperation at various imitation intensities in a Barabási-Albert scale-free network. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. The parameters are set to $b = 1.8$, $\chi = 1.5$, $N = 4000$, and $\langle K \rangle = 4$. The shaded areas represent the standard deviation of the outcome.

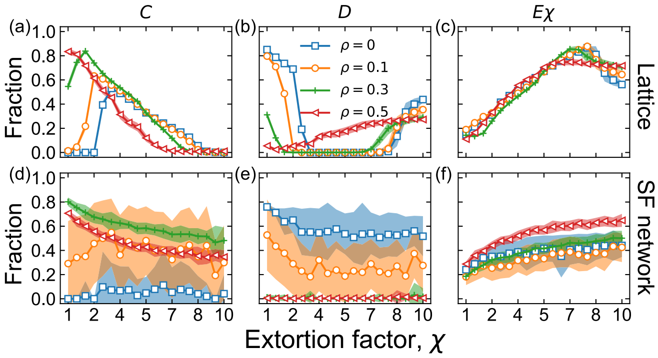

FIG. 11. In structured populations, the impact of committed extortioners under different extortion factors. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the extortion factor χ for committed extortioners at $\alpha = 1$. In lattice and BA networks, b is set to 1.5 and 1.8, respectively. The shaded areas represent the standard deviation of the outcome.

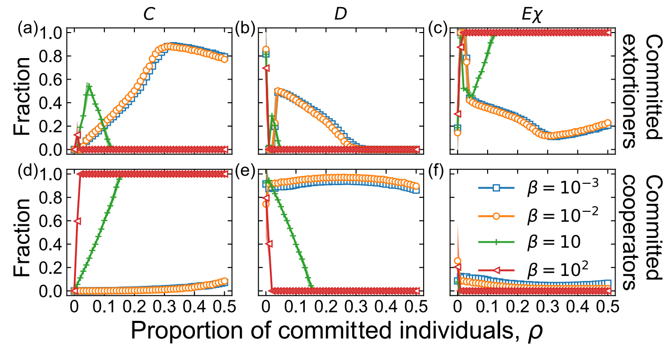

FIG. 13. The impact of committed individuals on cooperation using the Fermi rule in a lattice network. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. The parameter χ is set to 1.5.

with cooperators, which reduces the payoff of cooperators. As a result, normal players become more inclined towards extortion strategies. With high extortion factors, the critical mass effect triggered by committed extortioners becomes unsustainable, see Fig. 5(a).

Here, we explore the influence of committed individuals under varying imitation intensities, where the updating is governed by Eq. (3). Figure 6 presents compelling evidence of the critical mass effect exhibited by committed extortioners under strong imitation. Even under moderate imitation intensity ($\alpha = 0.5$), they still demonstrate a critical mass effect, albeit with diminishing impact as ρ increases. Conversely, under weak imitation cases ($\alpha = 0.1$ and $\alpha = 0$), the presence of committed extortioners facilitates the prevalence of extortion rather than promoting cooperation. In contrast, committed unconditional cooperators significantly enhance cooperation under weak imitation and also trigger a critical mass effect.

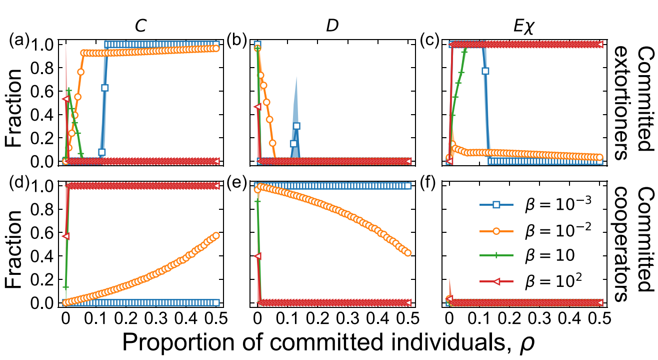

FIG. 12. The impact of committed individuals on cooperation using the Fermi rule in a well-mixed population. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. In the Fermi rule, strong imitation $\beta \to +\infty$ means that the player's strategy imitation depends on the difference in payoffs, while the individual performs random imitation under weak imitation $\beta \to 0$. The parameters are set to $b = 1.1$ and $\chi = 1.5$. The shaded areas represent the standard deviation of the outcome.

2. The influence of imitation intensity and extortion factor on committed individuals in structured populations

Based on the findings illustrated in Figs. 7–9, it was observed that the committed individuals had a similar impact on lattice, SW, and ER networks, in terms of imitation intensity. The presence of committed extortioners can facilitate the evolution of cooperation and exhibit a critical mass effect, which aligns with the results observed in mixed well-behaved populations. This suggests that under strong imitation, committed extortioners can promote cooperation, while committed unconditional cooperators work under weak imitation intensity. Furthermore, it was observed that there exists an optimal proportion of committed individuals that effectively facilitates cooperation in these networks.

For the SF network scenario, Fig. 10 demonstrates that committed extortioners exhibit critical mass effects in promoting cooperation under strong imitation intensity. Moreover, we observe that committed unconditional cooperators can also trigger a critical mass effect to facilitate the evolution of

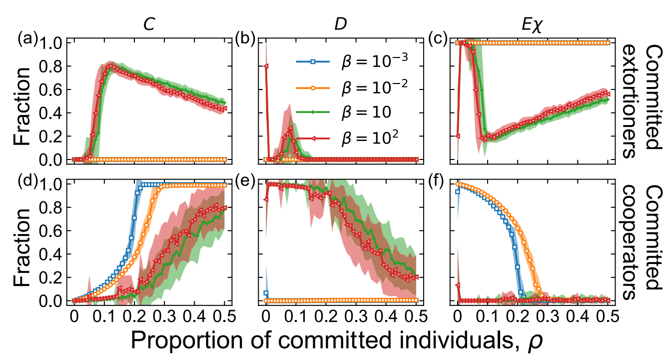

FIG. 14. The impact of committed individuals on cooperation using the Fermi rule in a Barabási-Albert scale-free network. Depicted are the average fractions of (a) and (d) C, (b) and (e) D, and (c) and (f) $E\chi$ among normal players as a function of the proportion of committed individuals ρ for committed extortioners and committed cooperators. The parameters are set to $b = 1.8$, $\chi = 1.5$, $N = 3000$, and $\langle K \rangle = 4$. The shaded areas represent the standard deviation of the outcome.

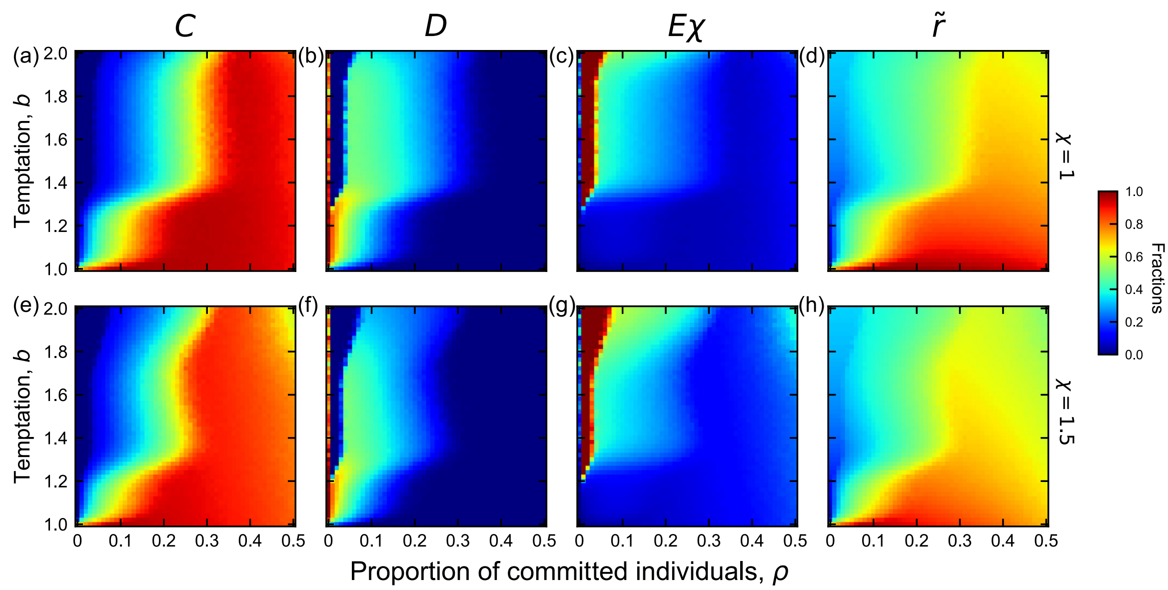

FIG. 15. The impact of committed individuals on cooperation using the Fermi rule in a lattice network. Depicted are the average fractions (color code) of C, D, $E\chi$, and the normalized average payoff $\tilde{r}$ for normal players as a function of the proportion of committed individuals ρ and temptation b at (a)–(d) $\chi = 1$ and (e)–(h) $\chi = 1.5$. The parameter β is set to 0.1.

cooperation among normal players, irrespective of whether the selection is strong or weak. These results reveal that regardless of the imitation intensity or the type of network, committed individuals have the potential to extensively trigger critical mass effects.

We also investigated the impact of the extortion factor on committed extortioners within a structured population. Figure 11 illustrates that as the extortion factor increases, the ability of committed extortioners to promote cooperation is reduced. However, in a scale-free network, committed extortioners continue to exhibit critical effects regardless of the value of the extortion factor.

3. The impact of the pairwise Fermi rule

Unlike the scenarios described in the main text, the individual's strategy update adheres to the replicator-like dynamics rule, where only the successful player can be imitated. Here, we explore the scenario with the Fermi rule [42], which captures the decision-making process of individuals with bounded rationality. In this scenario, a normal focal player i imitates the strategy of a randomly selected neighbor j with a probability of $p_{i\leftarrow j}$:

$$p_{i\leftarrow j} = \frac{1}{1 + \exp\{(r_i - r_j)/\beta\}}, \tag{A1}$$

where r_x represents the payoff of player x and β denotes the imitation intensity. $\beta \to 0$ indicates that player strategy imitation is completely dependent on differences in payoffs, while for $\beta \to +\infty$, player strategy imitation tends to become random and unrelated to payoffs. Typically, β is set to 0.1.

Figures 12–15 suggest that when player strategy updating is governed by the Fermi rule, the role of the committed individuals is robust.

[1] J.-P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, Rev. Mod. Phys. **57**, 617 (1985).

[2] M. Fruchart, R. Hanai, P. B. Littlewood, and V. Vitelli, Non-reciprocal phase transitions, Nature (London) **592**, 363 (2021).

[3] D. Achlioptas, R. M. D'Souza, and J. Spencer, Explosive percolation in random networks, Science **323**, 1453 (2009).

[4] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Critical phenomena in complex networks, Rev. Mod. Phys. **80**, 1275 (2008).

[5] G. J. Baxter, S. N. Dorogovtsev, K.-E. Lee, J. F. F. Mendes, and A. V. Goltsev, Critical dynamics of the k-core pruning process, Phys. Rev. X **5**, 031017 (2015).

[6] J. Gómez-Gardenes, S. Gómez, A. Arenas, and Y. Moreno, Explosive synchronization transitions in scale-free networks, Phys. Rev. Lett. **106**, 128701 (2011).

[7] C. Lagorio, M. Dickison, F. Vazquez, L. A. Braunstein, P. A. Macri, M. V. Migueles, S. Havlin, and H. E. Stanley, Quarantine-generated phase transition in epidemic spreading, Phys. Rev. E **83**, 026102 (2011).

[8] D. Yllanes, M. Leoni, and M. C. Marchetti, How many dissenters does it take to disorder a flock? New J. Phys. **19**, 103026 (2017).

[9] M. Mobilia, Does a single zealot affect an infinite group of voters? Phys. Rev. Lett. **91**, 028701 (2003).

[10] N. Crokidakis, V. H. Blanco, and C. Anteneodo, Impact of contrarians and intransigents in a kinetic model of opinion dynamics, Phys. Rev. E **89**, 013310 (2014).

[11] S. Galam and F. Jacobs, The role of inflexible minorities in the breaking of democratic opinion dynamics, Phys. A (Amsterdam) **381**, 366 (2007).

[12] P. Jensen, T. Matreux, J. Cambe, H. Larralde, and E. Bertin, Giant catalytic effect of altruists in Schelling's segregation model, Phys. Rev. Lett. **120**, 208301 (2018).
[13] S. Galam and S. Moscovici, Towards a theory of collective phenomena: Consensus and attitude changes in groups, Eur. J. Soc. Psychol. **21**, 49 (1991).
[14] S. Galam, Stubbornness as an unfortunate key to win a public debate: An illustration from sociophysics, Mind Soc. **15**, 117 (2016).
[15] J. Hofbauer and K. Sigmund, Evolutionary game dynamics, Bull. Am. Math. Soc. **40**, 479 (2003).
[16] G. Szabó and G. Fath, Evolutionary games on graphs, Phys. Rep. **446**, 97 (2007).
[17] J. Wu and R. Axelrod, How to cope with noise in the iterated prisoner's dilemma, J. Conflict Resol. **39**, 183 (1995).
[18] R. Axelrod and W. D. Hamilton, The evolution of cooperation, Science **211**, 1390 (1981).
[19] W. H. Press and F. J. Dyson, Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent, Proc. Natl. Acad. Sci. USA **109**, 10409 (2012).
[20] J. M. Pacheco, A. Traulsen, and M. A. Nowak, Coevolution of strategy and structure in complex networks with dynamical linking, Phys. Rev. Lett. **97**, 258103 (2006).
[21] F. C. Santos, J. M. Pacheco, and T. Lenaerts, Evolutionary dynamics of social dilemmas in structured heterogeneous populations, Proc. Natl. Acad. Sci. USA **103**, 3490 (2006).
[22] H. Gintis, *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior* (Princeton University Press, Princeton, NJ, 2000).
[23] M. Mobilia, A. Petersen, and S. Redner, On the role of zealotry in the voter model, J. Stat. Mech.: Theory Exp. **2007**, P08029 (2007).
[24] D. L. Arendt and L. M. Blaha, Opinions, influence, and zealotry: A computational study on stubbornness, Comput. Math. Organ. Theory **21**, 184 (2015).
[25] N. Masuda, Evolution of cooperation driven by zealots, Sci. Rep. **2**, 646 (2012).
[26] R. Matsuzawa, J. Tanimoto, and E. Fukuda, Spatial prisoner's dilemma games with zealous cooperators, Phys. Rev. E **94**, 022114 (2016).
[27] A. Cardillo and N. Masuda, Critical mass effect in evolutionary games triggered by zealots, Phys. Rev. Res. **2**, 023305 (2020).
[28] Y. Nakajima and N. Masuda, Evolutionary dynamics in finite populations with zealots, J. Math. Biol. **70**, 465 (2015).
[29] C. Shen, Z. He, L. Shi, Z. Wang, and J. Tanimoto, Simple bots breed social punishment in humans, arXiv:2211.13943.
[30] G. Sharma, H. Guo, C. Shen, and J. Tanimoto, Small bots, big impact: Solving the conundrum of cooperation in optional prisoner's dilemma game through simple strategies, J. R. Soc. Interface **20**, 20230301 (2023).
[31] H. Guo, C. Shen, S. Hu, J. Xing, P. Tao, Y. Shi, and Z. Wang, Facilitating cooperation in human-agent hybrid populations through autonomous agents, iScience **26**, 108179 (2023).
[32] C. Hilbe, M. A. Nowak, and K. Sigmund, Evolution of extortion in iterated prisoner's dilemma games, Proc. Natl. Acad. Sci. USA **110**, 6913 (2013).
[33] C. Adami and A. Hintze, Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything, Nat. Commun. **4**, 2193 (2013).
[34] C. Hilbe, T. Röhl, and M. Milinski, Extortion subdues human players but is finally punished in the prisoner's dilemma, Nat. Commun. **5**, 3976 (2014).
[35] X. Xu, Z. Rong, Z.-X. Wu, T. Zhou, and C. K. Tse, Extortion provides alternative routes to the evolution of cooperation in structured populations, Phys. Rev. E **95**, 052302 (2017).
[36] Z.-X. Wu and Z. Rong, Boosting cooperation by involving extortion in spatial prisoner's dilemma games, Phys. Rev. E **90**, 062102 (2014).
[37] A. Szolnoki and M. Perc, Evolution of extortion in structured populations, Phys. Rev. E **89**, 022804 (2014).
[38] K. Sigmund, H. De Silva, A. Traulsen, and C. Hauert, Social learning promotes institutions for governing the commons, Nature (London) **466**, 861 (2010).
[39] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature (London) **393**, 440 (1998).
[40] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
[41] P. Erdős and A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).
[42] A. Traulsen, J. M. Pacheco, and M. A. Nowak, Pairwise comparison and selection temperature in evolutionary game dynamics, J. Theor. Biol. **246**, 522 (2007).
[43] G. Szabó and C. Tőke, Evolutionary prisoner's dilemma game on a square lattice, Phys. Rev. E **58**, 69 (1998).
[44] D. Fudenberg, D. G. Rand, and A. Dreber, Slow to anger and fast to forgive: Cooperation in an uncertain world, Am. Econ. Rev. **102**, 720 (2012).
[45] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson, Rational cooperation in the finitely repeated prisoners' dilemma, J. Econ. Theory **27**, 245 (1982).
[46] A. Mao, L. Dworkin, S. Suri, and D. J. Watts, Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoner's dilemma, Nat. Commun. **8**, 13800 (2017).
[47] H. Fort and E. Sicardi, Evolutionary Markovian strategies in 2×2 spatial games, Phys. A (Amsterdam) **375**, 323 (2007).