

True Colors: Authenticity in Social and Economic Interactions

Francis Bloch and Rachel Kranton*

Preliminary Draft December 2025

Abstract: This paper considers settings where individuals benefit from choosing the same actions as others but also desire to be authentic to their own background or identity. Analyzing a *matching colors game*, the outcomes shed light on the conflicting interests between agents of the same background but have different intensities of preferences, and the gains and losses when agents can commit to their preferred action rather than conform. In the base case, agents randomly meet and the social category of counterparts and their intensities of preferences for associated actions, which we call “colors” are not observed. The unique equilibrium involves a “tyranny of the majority;” minority agents with the lowest cost play the majority-preferred color, which skews the play towards the majority and gives further incentive for minority agents to adopt the majority behavior. However, in a game when agents can commit to their preferred color or remain flexible, there are two equilibria, one in which the majority prevails and one in which the minority color prevails. All agents are better off except the “diehard” agents in the majority who no longer enjoy meeting as many people in the population who conform to the majority color. We also consider a game where one agent uses commitment status to select among a group of contenders and show that, as the number of contenders increases, the selector’s incentive to commit increases and the contenders’ incentives to commit decrease.

Keywords: social identity, conformity, social norms, tyranny of the majority, commitment, selection

JEL Classification Numbers: D80, D60, I30

* Bloch: University Paris 1 Pantheon Sorbonne and Paris School of Economics, francis.bloch@univ-paris1.fr; Kranton: Department of Economics, Duke University, Durham NC, rachel.kranton@duke.edu; We thank workshop and conference participants at Duke University, ERINN, Oxford University, Paris School of Economics, University of Auckland, University of Glasgow, University of Liverpool for comments and suggestions.

1 Introduction

Conforming. Covering. Passing. This paper considers settings where individuals benefit from choosing the same behaviors as others but, due to difference in identities, backgrounds, or tastes, they can also suffer or incur costs to do so. Examples range from adopting different measurement standards, to watching the same or different entertainment to get along better with colleagues, to expressing openly or not religious or political beliefs, to fertility. “Covering” at work – adapting individual behaviors in order to fit in professionally and socially – is a much discussed phenomenon in the United States (Yoshino (1997), Yoshino (2007)), and in an increasingly diverse workplaces, management studies elaborate on the prevalence of this phenomenon and policies that would instead promote authenticity.¹

This paper studies such settings and our main objective is to understand when and to what extent the minority adopts the preferred action of the majority, in different policy scenarios which allow or not agents to openly undertake and commit to their preferred action (e.g., “diversity, equity, and inclusion” versus “don’t ask, don’t tell”). We study the strategic choice of actions in meetings between people who desire to match their actions to each other but also desire to be authentic to their own preferences or identities. People’s preferences for different actions could also more or less intense, as it can be easier for some people to adjust their behavior or it can be less aversive from some to do so, and we consider the implications of “die-hard” agents who ultimately never adjust their behavior. We call the basic interaction an *matching colors game*, wherein people have benefits from matching their actions to others but also from being authentic, to expressing their true colors. We further study the possibility that people can commit to an action, such as taking an immutable action that is in essence a choice of action like getting a tattoo, wearing or not religious symbols, buying and wearing certain clothing, which ties a person to one set of actions. The decision to take such an action

¹See, for example, in the management literature Deloitte (2023) and Creary, Stephanie (2023), and blog posts such as Cunningham (<https://www.welcometothejungle.com/en/articles/how-to-be-authentic-at-work>). Related discussions concern “invisible mending” (Wells & MacAuley (2024)) and how the exertion of effort to adjust one’s behavior is a form of emotional labor (Hochschild (1983)).

involves a trade-off between maintaining flexibility to adjust to others and putting others in the position to accommodate one's preferences. We study this possibility in a random meeting setting and in the context of an employer hiring workers.

Our first set of results contrasts a benchmark setting where people have complete information about their counterparts' preferences with a setting where people's preferences are private information. The main finding concerns the presence of die-hard agents in the majority and the externalities imposed by those in the minority on other members of the minority. When preferences are known, the main issue is for people to match on the same action and multiple outcomes are possible. When the preferences are not known, however, there is a single outcome which involves a "tyranny of the majority." An equilibrium in which the majority agents play the minority action is not possible, since those with higher costs always benefit from the chance of matching with a die-hard majority member. In the unique equilibrium outcome, then, those in the minority with low costs adjust their action in the expectation their counterpart is from the majority. These strategies give higher cost minority agents the incentive to adjust to the majority as well since they are less likely to encounter someone who is playing the minority action. The larger the majority and the greater proportion of majority diehards, the greater is this effect on minority agents.

Our second set of results asks whether the possibility to commit to an action can mitigate or reverse this dynamic. We find that commitment completely changes the possible outcomes; the die-hard agents and those with high costs of inauthenticity commit to play their true color while other agents remain flexible. This commitment reduces the mismatching of colors. There are two stable equilibria: In one equilibrium, the high cost minority agents commit to play their preferred color and all other agents remain flexible. The flexible agents play the majority color with each other and accommodate the high cost minority's preferred color. In this way, the high cost minority agents block the "tyranny of the majority," and all minority agents are better off. Moderate majority agents, with low cost of inauthenticity are also better off. The only agents who might suffer from the agents' ability to commit are majority agents with

high cost of inauthenticity. In the second equilibrium, high cost majority agents commit to their preferred color and all other agents remain flexible. The flexible agents play the minority action with each other and accommodate the high cost majority's preferred action. In this equilibrium as well minority agents and majority moderate agents are also better off than in a game without commitment

Finally we consider the implications of a sort of "market power" on people's choices to commit or not to their preferred action. If people are hiring others for a job, for example, or seeking employment, how does this setting change the incentives to commit to an action or to remain flexible, for both the employer and potential employees? As in the basic commitment game, we consider two possible outcomes in the ultimate matching colors game - two flexible agents play either the majority or the minority action. We consider two case where the majority action prevails in the ultimate matching colors game. Comparing no competition to competing against another agent, we find that a majority agent who is making the choice whom to hire remain flexible just as in the case of no competition. On the other hand, a minority agent with market power commits more than otherwise, since she can possibly choose of an agent who will accommodate her preferred action. Minority contenders, though, are more flexible in this setting, in order to increase their probability of being hired, even though they incur a cost to accommodate their employer. This outcome reflects the struggles of minority employees who feel they must change their behavior in order to succeed at work (see citations above).

This paper both departs from and contributes to the now large economics literature on conformity, social norms, and conventions. Much of the work posits that people are concerned about their reputation, i.e., what other people think of them, perhaps starting with Akerlof (1980). Bernheim (1994) provides a model of conformity, where agents choosing an action is an equilibrium signal of an unobserved but socially desirable attribute or characteristic. Such modeling has been used to understand phenomena in particular settings, such as Muslim women wearing head scarfs (Carvalho (2013)).

In our analysis, people are not concerned about their reputation but derive a direct benefit

from doing the same action as other people. In this way, the present paper is related to the study of conventions (Young (1993)) and to situations where individuals are assumed to benefit from adopting actions closer to that in the population (e.g., Michaeli & Spiro (2015), Acemoglu & Jackson (2017)) or their social group or neighbors (e.g., Manski & Mayshar (2003), Jackson & Storms (2024)), possibly facing trade-offs given their idiosyncratic preferences. Bilancini & Boncinelli (2018) studies coordination games in which agents observe another player’s type after they interact once and condition their subsequent actions and connections to agents accordingly. In Genicot (2022) agents have heterogeneous ideal behaviors (“identities”) and different tolerance levels, and each chooses both a single action and interaction partners. A main result is that agents endogenously sort into groups, with sufficiently tolerant agents acting as bridges between otherwise disconnected groups. Michaeli & Spiro (2017) consider a population with diverse preferences over actions and agents meet pairwise, with the preferences to choose actions which are similar to those whom they meet. A main result is that in equilibrium agents choose the same action (i.e., there is a “norm”) when the marginal benefits of choosing the same action are decreasing in the distance between the norm and one’s preferred action. Relative to those models, the analysis in the present paper considers two exogenously distinct groups of agents who have conflicting preferences, and each group has a set of die-hard agents who never want to choose another action. Moreover, the analysis considers that agents can choose their action depending on whom they meet, which allows for study of a setting where agents can demonstrably commit to an action and other agents choose whether to adjust or not. This possibility reflects real-world social interactions and policies, and the analysis derives the welfare gains and losses in the equilibria that emerge.

Coordination games with a flexible option have been introduced by Galesloot & Goyal (1997) and Goyal & Janssen (1997). They extend a classical coordination game with different Pareto dominant and risk-dominant equilibria by adding a third option where agents, at a cost, can choose to remain flexible. These games are now called “bilingual games,” following the linguistic interpretation where agents choose either to speak a single language or learn

both languages. The objective of these papers is to study an evolutionary process where agents revise their strategies after observing actions chosen in their local neighborhood. These papers differ from ours on several grounds. Galesloot & Goyal (1997) and Goyal & Janssen (1997) assume that all agents are identical, so there is no heterogeneity and the game is a game of complete information. Furthermore, they introduce a cost of flexibility, whereas in our model flexibility is simply strategic with no additional cost. Finally, they focus attention on evolutionary stable equilibria whereas we consider stable Bayesian perfect equilibria. Naono (2022) considers heterogeneous agents who match to play a complete information bilingual game, with costs of flexibility. Our paper involves no cost of flexibility and the focus is on incomplete information and the implications for conformity to the majority, which are possibly mitigated by ability to commit to an action in a pre-play stage.

Coordination games with pre-play communication have been studied since Farrell (1987) who showed that a first stage of cheap talk can help players coordinate. In a recent contribution, Ganguly & Ray (2023) extend the analysis to a setting with incomplete information providing conditions under which full revelation arises. Our paper differs from this literature by considering a richer second period game where players may play either a coordination game or a battle-of-the-sexes game depending on their type, and by focusing attention on commitment rather than pre-play communication.

With applications to the interaction between people in different social groups, this paper relates to the growing work on social identity and norms for behavior. Some of this work follows in the vein of the theories of conformity; Austen-Smith & Fryer Jr (2005)'s theory of African American's "acting white" is based on a dual audience signaling problem; in equilibrium higher levels of education signal both higher work ability and lower concern for socializing with ones' peers. In recent work, Tirole (2023) considers the possibility that there is no general agreement in the population on which underlying attributes are desirable, i.e., people's attributes or opinions, such as their political or religious beliefs, might conflict. Agents concerned about their reputations then decide whether or not to act on their beliefs or retreat to "safe spaces"

populated by others with the same opinions. In the model we consider, agents care only about matching actions with others, not other’s types or their own reputation. We consider the setting when one group is the majority, and the majority and minority have different preferences for actions. Akerlof & Kranton (2000) considers only one side of this problem, when agents in the minority have an incentive to adopt the behavior of the majority in order to obtain higher returns to employment. Kuran & Sandholm (2008) posits that people have stronger incentives to match their actions when meeting someone in their social groups and consider that agents adjust their preferences over time and preferences converge as people interact with those in different groups.

The setting in the present paper is basic and straightforward and meant to capture day-to-interactions in a heterogeneous society. A key feature which departs from previous work is that people can adjust their behavior depending on whom they meet. When agents can demonstrably commit to an action, there are then fewer mismatches. The ability to commit leads to benefits for all but the highest cost majority members who are best off in a “don’t ask, don’t tell” setting of incomplete information when their group has a large majority. The change in their payoffs in the two settings captures the backlash of majority members to policies promoting authenticity.

The paper proceeds as follows. Section 2 provides the basic model of interaction – what we call the *matching colors game* – and solves for the equilibria of the game when agents choose their action knowing the categories and costs of their counterparts. In Section 3 we consider equilibria when agents meet but categories and individual costs are private information. Section 4 considers outcomes when agents can commit to their action, and Section 5 adds the possibility that one agent could have the ability to select a partner from among contenders. The Conclusion considers avenues for future research.

2 Matching colors game with complete information

Consider a continuum of agents of measure one. Each agent is either Green or Red, indicated with capital letters G and R, which we call an agent's (*social*) *category*, and we assume a proportion $\gamma \in (\frac{1}{2}, 1)$ of the agents are Green. Two agents are selected at random to interact. We suppose they each simultaneously choose a color, green (g) or red (r). For notation, we write an agent's category with capital letters and their actions with small letters; i.e., each agent is either G or R and chooses either g or r . An agent who chooses a color that does not correspond to their category incurs a cost, which we call an inauthenticity cost, denoted c , which is uniformly distributed on $[0, 1]$ for both Green and Red agents. If the two agents choose different colors, they each receive a benefit of 0. If two agents choose the same color, they each earn a benefit b , where $b \in (0, 1)$. We call those with costs strictly above b *die-hard* agents. The remaining agents with costs below b would be willing to adopt the action of the other category.

As a benchmark, consider first a complete information game, where each agent knows their own category and the category of the other agent as well as each agent's inauthenticity cost. We note first that for any die-hard agent, it is a strictly dominant strategy to choose the color corresponding to category.

If the two agents have the same category, as illustrated in Figure 1 where both agents are Green, this game is similar to a coordination game. There is a unique Nash equilibrium if one or both players are die-hard, which involves the color matching their category. Otherwise, there are two pure strategy Nash equilibria: both agents choose green or both agents play red. One of the equilibria is Pareto-dominated and involves a mismatching of colors. There is also a mixed strategy equilibrium where agents randomize between colors.

If two agents of opposite categories meet, as illustrated in Figure 2, the game is similar to Battle of the Sexes except that agents always incur a cost for choosing their not-preferred action. If both agents are die-hard, there is a unique equilibrium where both agents choose

		Player 2 GREEN	
		green	red
Player 1 GREEN	green	b, b	$0, -c_2$
	red	$-c_1, 0$	$b - c_1, b - c_2$

Figure 1: Complete Information Game between two Green agents

their own color. At the other extreme, when neither is die-hard, there are again two pure strategy Nash equilibria $\{\text{green}, \text{green}\}$ and $\{\text{red}, \text{red}\}$ as well as mixed strategy equilibrium. In the remaining case, when one agent is die-hard and the other is not, there is a unique equilibrium where both agents play the color corresponding to die-hard agent's category.

		Player 2 GREEN	
		green	red
Player 1 RED	green	$b - c_1, b$	$-c_1, -c_2$
	red	$0, 0$	$b, b - c_2$

Figure 2: Complete Information Game between a Red and a Green agent

3 Matching colors game with incomplete information

3.1 Game and Bayesian Nash Equilibrium

We now consider our main setting where two agents randomly meet and simultaneously choose a color but agents' categories and costs are private information. (Formally, an agent's type in this game is their category and their cost c .) We assume that the category and cost

distributions are common knowledge and consider Bayesian Nash equilibria, henceforth simply *equilibria*. We restrict attention to the equilibria which are stable to small perturbations in agents' play following Vives (1990) and Vives (2005) studies of equilibrium selection in games with strategic complementarities; details are provided in the Appendix.

In this setting, agents cannot condition their action on their counterpart's category or cost. We show that the unique (and stable) equilibrium privileges the preferences of the majority. First, there cannot be an equilibrium in which all Greens play red (or all Reds play green). With the presence of a die-hard players, the high cost (but not die-hard) players always have an incentive to play their own color to match with them. Second, the equilibrium involves all Greens playing green and some Reds playing green. The low-cost Red agents who play green give the incentive for other Red agents to play green, eventually giving a cut-off cost, where Red agents with costs below this critical value play green, and those with costs above this value play red.

We construct the strategies of each agent category and cost as follows. We show (in the Appendix) the equilibrium is characterized by cut-off costs such that agents with higher costs always play their preferred color. Let c_g and c_r denote these thresholds for the Green agent to play g and the Red agents to play r , respectively. The proportion of all agents playing green is then given by $\gamma(1 - c_g) + (1 - \gamma)c_r$, and the proportion of agents playing red by $\gamma c_g + (1 - \gamma)(1 - c_r)$. We can thus write, for any agent, the expected gross benefits of choosing green, denoted π_g , as

$$\pi_g = b[\gamma(1 - c_g) + (1 - \gamma)c_r], \tag{1}$$

and the expected gross benefits of choosing red, denoted π_r , as

$$\pi_r = b[\gamma c_g + (1 - \gamma)(1 - c_r)]. \tag{2}$$

In equilibrium, c_g is the cost such that a Green agent is indifferent between playing green

and red, $c_g = \pi_r - \pi_g$, and c_r the cost such that a Red agent is indifferent between playing red and green, $c_r = \pi_g - \pi_r$.

A preliminary analysis provides the necessary conditions on cutoff costs c_g and c_r . First, the equilibrium cut-off costs c_g and c_r must be less than or equal to b , since choosing own color is a dominant strategy for die-hard agents. Second, it cannot be that both threshold costs are positive, as this would imply that both $\pi_g - \pi_r$ and $\pi_r - \pi_g$ are positive. Hence, at least one of the two threshold costs must be equal to 0; either all Red agents play red or all Green agents play green. Finally, if $\gamma \neq \frac{1}{2}$, in equilibrium, we cannot have $c_g = c_r = 0$ as the zero-cost agent would then strictly prefer to play green and cannot be indifferent between the two colors. Together these imply that either $c_g = 0$ and $c_r > 0$ or $c_g > 0$ and $c_r = 0$; that is, in equilibrium, one category, and only one category, ever chooses the other color.

Lemma 1 *With incomplete information, any equilibrium of the matching colors game is characterized by cutoff costs c_g and c_r above which the agents choose their preferred color and below which they choose the other color such that (i) $c_g \leq b$ and $c_r \leq b$, (ii) either $c_r = 0$ or $c_g = 0$ or both, and (iii) if $\gamma \neq \frac{1}{2}$, either $c_r > 0$ or $c_g > 0$.*

Proof of Lemma 1: All proofs are provided in the Appendix.

We show next that there is a unique equilibrium in the matching colors game. In the equilibrium, all Green (majority) agents play green, $c_g = 0$, and some Reds also play green $c_r \geq 0$. The uniqueness of the equilibrium follows from the assumption that Reds are the minority and Greens are the majority. There is a fixed point proportion of agents choosing the opposite color only for the minority.

To see this, first consider the possible equilibrium where all Greens play green and some Reds play green. Consider possible values of c_r , the proportion of Reds who play green, and the difference $\pi_g - \pi_r$ which is the incentive to play green, show in 3 for $c_g = 0$.

$$\begin{aligned}
\pi_g - \pi_r &= b[\gamma + (1 - \gamma)c_r] - b[(1 - \gamma)(1 - c_r)] \\
&= 2b(1 - \gamma)c_r + b(2\gamma - 1)
\end{aligned} \tag{3}$$

At $c_r = 0$, since $\gamma > 1/2$, the gross payoffs of playing green exceed that of playing red, $\pi_g - \pi_r > 0$. As more Reds play green, the difference $\pi_g - \pi_r$ increases by the rate of $2b(1 - \gamma)$; more low-cost Reds play green, more higher cost Reds also have the incentive to play green. But, for $b < 1$ and $\gamma > 1/2$, this gain increases less than the cost c_r , and at $c_r = b$, $\pi_g - \pi_r < b$, hence there is an intermediate cost $c_r > 0$ such that $c_r = \pi_g - \pi_r$.

On the other hand, consider the possible equilibrium where all Reds play green, $c_r = 0$, and possible values of c_g , the proportion of Greens who play red, where by Lemma 1 $c_g \leq b$. The difference $\pi_r - \pi_g$ is the incentive to play red, show in ?? for $c_r = 0$.

$$\begin{aligned}
\pi_r - \pi_g &= b[\gamma c_g + (1 - \gamma)] - b[\gamma(1 - c_g)] \\
&= 2b\gamma c_g + b(1 - 2\gamma)
\end{aligned} \tag{4}$$

If $c_g = 0$, since $\gamma > 1/2$, the gross payoffs of playing red are smaller than that of playing green, $\pi_r - \pi_g < 0$. The incentive to play red is increasing linearly in c_g , at a rate $2b\gamma$, which is less than one since $\gamma > 1/2$. Even at $c_g = b$, we have $\pi_r - \pi_g < c_g$. Essentially, when even when all Greens with costs below b play red, there would be an incentive for the higher cost Green to deviate and play green. Hence, given $c_r = 0$, there is no cost $c_g \in [0, b]$ such that $\pi_r - \pi_g = c_g$.

We then consider the unique equilibrium, in which all Greens play green and some Reds play green, and construct the cutoff cost for Red agents, c_r , as a function of the size of the majority γ and the benefits of matching colors b . This value is given by the solution to the

equation $c_r = \pi_g - \pi_r$ given $c_g = 0$ as follows:

Proposition 1 *The unique equilibrium in the matching colors game is stable and involves the cut-off costs (c_g^*, c_r^*) where $c_g^* = 0$ and c_r^* where*

$$c_r^* = \frac{b(2\gamma - 1)}{1 - 2b(1 - \gamma)} \quad (5)$$

Figure 3 illustrates this equilibrium. The horizontal axis gives the division of the population into Green and Red categories. The vertical axes give the individual costs of playing the other color. The color of the shaded area gives the color choice of each set of agents. For costs above b , Green agents choose green and Red agents choose red, since playing own color is a dominant strategy. Green agents with cost below b also choose green, and Red agents with costs above c_r^* choose red and those below this cost choose green.

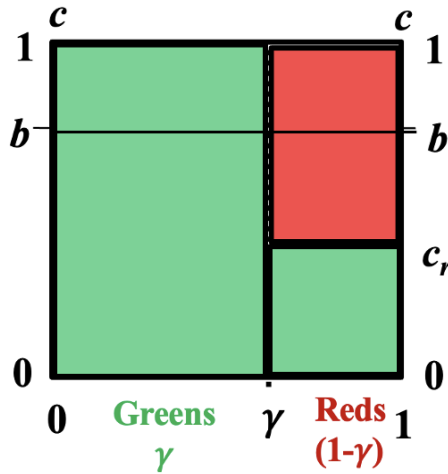


Figure 3: Equilibrium Strategies in Incomplete Information Game

We see that c_r^* is increasing and concave in proportion of the population that is Green, γ , and increasing and convex in the benefits of matching colors b so that there are fewer die-hard agents.

Proposition 2 *In the incomplete information game, ((i) the proportion of minority agents who play the majority color is increasing and concave in γ . As γ approaches $1/2$, c_r^* approaches*

0 and as γ approaches 1, c_r^* approaches b . (ii) the proportion of minority agents who play the majority color is increasing and convex in b .

The complementarities among the Red players are behind these comparative statics. When γ is close to $1/2$, the Reds are in the minority but are still a relatively large part of the population. The low-cost Reds who play green give an incentive for the other, relatively numerous, Reds to play green, leading to a large increase in c_r^* . When γ is closer to one, Reds are in the minority and are small part of the population. The complementarities are weaker in the sense that there are fewer Red players who provide the incentive to play green. Hence, the proportion of Reds who play green increases more in γ when γ is low than when γ is high.

As for b , there are two effects. The benefits of matching colors with the Green majority are directly increasing in b . In addition, $(1 - b)$ gives the proportion of die-hard Reds. When there are fewer such players, there is more incentive for Reds to play green, and the complementarities among Reds to play green are stronger.

The uniqueness of equilibrium in the game with incomplete information contrasts with the multiplicity of equilibria in the game with complete information. This result is reminiscent of equilibrium selection in global games introduced by Carlsson & Van Damme (1993) and surveyed in Morris & Shin (2003). As in global games, but with a different construction, uncertainty about players' payoffs coupled with the existence of extreme players with a dominant strategy enable us to prove uniqueness of equilibrium.

3.2 First-best: Comparison

Here we consider the thought experiment of the first-best - the assignment of actions in the population that would maximize aggregate expected payoffs and compare those actions and to the equilibrium outcome.

Given the cost distributions are the same for Green and Red agents, we first show that since Greens are the majority, the first best entails all Greens playing green and Reds with

cost below some x playing green and those with higher costs playing red. Essentially, any configuration in which some proportion of higher cost agents play the opposite color can be improved by switching their play with an equal proportion of lower cost agents of the same category. Letting W denote the aggregate expected payoffs, we have

$$W = \left[(\gamma + (1 - \gamma)x)^2 b - (1 - \gamma) \int_0^x c dc \right] + ((1 - \gamma)(1 - x))^2 b.$$

where the first two terms are the aggregate expected benefits of the Greens who play green and the Reds who play green minus the aggregate costs to the Reds, and the last term is the aggregate expected benefits to the remaining Reds of playing red. Maximizing this expression yields c_r^O , the first-best proportion of Reds who play green given in Lemma 2.

Lemma 2 *In the first-best configuration of play: all Greens play green, Reds with costs below some c_r^O play green, and the remaining Reds play red. For $b \geq \frac{1}{2}$, all Reds should play green; $c_r^O = 0$ since the benefits of matching colors exceeds the cost of the average Red agent. For $b < 1/2$, some Reds should play red, and*

$$c_r^O = \frac{2b(2\gamma - 1)}{1 - 4b(1 - \gamma)} > 0$$

.

Comparing the first-best c_r^O and the equilibrium outcome c_r^* , we easily see that the Reds play green at a lower rate in the equilibrium outcome than in the first-best for all b .

This divergence is due to the externalities which are considered in the first-best calculation but not considered by individual agents in the strategic setting. Red agents' incentives do not include the synergies that would arise from their color choices, and they are not compensated for the benefits a choice to play green would imply for others.

Proposition 3 *In the incomplete information game, fewer Red agents play green than in the first-best configuration which maximizes aggregate expected payoffs.*

4 Two-stage game: commitment then matching colors

In this section, we consider the possibility that agents can commit to a color or choose to remain flexible before playing the matching colors game. This commitment decision is observed by agents and involves trade-offs. Commitment could be beneficial if an agent meets someone who is flexible and who will accommodate their color choice. But commitment could be detrimental if an agent meets someone who has committed to a different color, and hence there is no possibility to match colors. We consider equilibria of this game and in particular whether the “tyranny of the majority” outcome is mitigated by the opportunity for agents to commit. We further consider which agents are better or worse off in this setting.

4.1 The Two-Stage Game

The two stage game is specified as follows: In the first stage agents simultaneously choose whether to commit or not to a particular color; the strategy space in the first stage is $(G)reen$, $(R)ed$, $(F)lexible$. In the second stage, two agents meet at random. They each observe the commitment decision of their counterpart (G , R , or F). They then play the matching colors game: Agents who chose G in the first stage are constrained to play green, and those who chose R are constrained to play red. Agents who chose F in the first stage choose between playing green or red, and their strategy is contingent on whether their counterpart has chosen G , R , or F .

We solve for the perfect Bayesian equilibria, henceforth simply “equilibria.” The two-stage game admits many Bayesian perfect equilibria, reflecting the fact that agents with high inauthenticity costs, $c \geq b$, always choose their own color in the second stage of the game and are hence indifferent between committing or not in the first stage of the game. To select among equilibria, we rule out weakly dominated strategies and in addition restrict attention to stable equilibria in the second-stage matching colors game (as discussed further below).

4.2 Equilibria of the Two-Stage Game

4.2.1 Preliminary results for agents' commitment strategies

Similar to the analysis of the incomplete information game, we first establish preliminary results concerning the strategies of agents with different costs, here for the first-stage decision whether or not to commit to a color and which one. We show the set of agents with these strategies are intervals in the space of inauthenticity costs. Hence, we can characterize the first-stage strategies of Red agents by two thresholds, which we denote c_G^R and c_R^R such that a Red agent chooses G if $0 \leq c < c_G^R$, chooses F if $c_G^R \leq c \leq c_R^R$ and chooses R if $c_R^R \leq c \leq 1$. Similarly, we can characterize the first-stage equilibrium strategies of Green agent by two thresholds c_R^G and c_G^G . In an equilibrium in weakly undominated strategies, all agents with inauthenticity costs above b commit to play their own color; i.e., the thresholds c_R^R and c_G^G must be weakly less than b . There is no gain from maintaining flexibility, since the benefit of playing the other color always exceeds the cost. We record these results below, with all formal proofs in the Appendix.

Lemma 3 : *Any equilibrium of the two-stage game where agents play weakly undominated strategies is characterized by cost intervals $[c_R^G, c_G^G]$ and $[c_G^R, c_R^R]$ in which, respectively, Green agents choose F and Red agents choose F , and $c_R^R \leq b$ and $c_G^G \leq b$.*

We now solve for the equilibria of the full game. We solve backwards: first we consider the matching colors game and then we consider agents' decisions to commit to a color or to remain flexible.

4.2.2 Matching colors game after commitment decisions

Consider two agents who meet and play the second-stage matching colors game. There are three possible kinds of encounters. (i) Both agents have committed to a color, in which case each agent simply plays their committed-to color. (ii) One agent has committed to a color and one agent has chosen to be flexible. The committed agent plays the committed-to color,

and, since, by Lemma 3 the flexible agent has a cost less than b , the flexible agent also plays this color. (iii) Both agents have chosen to be flexible. In this case, each agent has a strategy of whether to play green or red. We consider next the stable Bayesian Nash equilibria in such a “flexible-flexible” matching colors sub-game.

The sub-game in which both agents are flexible is similar to the incomplete information game but with a critical difference. Both Green and Red agents have costs which are in a truncated interval determined by the commitment decisions. Rather than the full distribution of costs $[0, 1]$, the costs of a Green agent are uniformly distributed over the interval $[c_R^G, c_G^G]$ and the costs of a Red agent are uniformly distributed over $[c_G^R, c_R^R]$. By Lemma 3, all agents have costs below b in these intervals, and hence no agent has a dominant strategy to play their own color. As we show next, the sub-game is a pure coordination game, and there are two stable equilibria - one in which all agents choose green and one in which all agents choose red.

To analyze this sub-game, we introduce the following notation: Let γ_G and γ_R denote the fraction of Green and Red agents in the pool of agents playing the sub-game, namely $\gamma_G \equiv \frac{\gamma(c_G^G - c_R^G)}{\gamma(c_G^G - c_R^G) + (1-\gamma)(c_R^R - c_G^R)}$ and $\gamma_R \equiv 1 - \gamma_G$. With a slight abuse of notation, let c_g and c_r denote the threshold values of costs above which Green agents and Red agents, respectively, choose their own color, where $c_g \in [c_R^G, c_G^G)$ and $c_r \in [c_G^R, c_R^R)$.

Our first result replicates Lemma 1 to show that in any equilibrium of the matching colors game either all Green agents play green or all Red agents play red, since the zero-cost Green and Red agents must choose the same color, and by Lemma 3, the set of agents selecting a color or remaining flexible forms an interval in inauthenticity costs.

Lemma 4 *In any equilibrium of the flexible-flexible matching colors sub-game, either $c_g = c_R^G$ or $c_r = c_G^R$.*

Given Lemma 4, there are only four possible equilibrium configurations:

- (1) All agents play green ($c_g = c_R^G$ and $c_r = c_R^R$).
- (2) All agents play red ($c_g = c_G^G$ and $c_r = c_G^R$).

(3) All Green agents play green, high cost Reds play red and low cost Red plays green ($c_g = c_R^G$ and $c_r \in [c_G^R, c_R^R)$).

(4) High cost Green plays green and low cost Green plays red, and all Red agents play red ($c_g \in [c_R^G, c_G^G)$ and $c_r = c_G^R$).

In the first two equilibria, the agents all choose the same color. These equilibria both exist for any value of γ_G and γ_R because of the strategic complementarities of agents' choices. If all other agents play a given color, any agent with an inauthenticity cost smaller than b has an incentive to play the same color. These corner equilibria are stable, again due to the strategic complementarities.

The last two equilibria are interior equilibria and involve coordination failures; they are not stable. In each, one category of agent plays their true color, and the other category of agents plays their true color only if their inauthenticity costs are high. These equilibria exist when the fraction of agents who play their true color is smaller than the fraction of agents who play the opposite color; that is, each equilibrium involves a “tyranny of the minority.” For each of them, a small perturbation in the measure of agents playing each color leads away from the equilibrium, since the complementarities favor moving towards the play of the majority.

Proposition 4 *In the two-stage commitment game, there exists two stable equilibria in the matching colors game between two flexible agents. In one equilibrium, which we call the “green outcome,” all agents choose green (i.e., $c_g = c_R^G, c_r = c_R^R$), and in the second equilibrium, which we call the “red outcome,” all agents choose red (i.e., $c_g = c_G^G, c_r = c_G^R$).*

4.2.3 Equilibria in the full game

We now characterize the equilibria of the two-stage game. We consider the agents' commitment strategies given one or the other of these equilibria arises in the subsequent matching colors games between two flexible agents – the green or the red outcome.

We first show that playing F is a weakly dominant strategy for all agents with costs below

b when their color is played in the F - F sub-game.

Lemma 5 : *In any equilibrium of the two-stage game: (i) with the green outcome, all Green agents with cost lower than b choose F (i.e. $c_R^G = 0, c_G^G = b$) (ii) with the red outcome, all Red agents with cost lower than b choose F ($c_G^R = 0, c_R^R = b$).*

To understand Lemma 5, consider the green outcome in the F - F sub-game. Consider the first-stage choice of a Green agent i with cost $c \leq b$. Suppose first agent i chooses F : i either meets (1) another agent who has chosen F , in which case i earns b , (2) an agent who has chosen G , in which case i plays green and earns b , or (3) an agent who has chosen R in which case i plays red and earns $b - c$. Suppose instead agent i chooses G : i earns the same payoffs as the agent who chooses F in situations (1) and (2) but earns 0 in situation (3). Hence, F weakly dominates G . Consider then i choosing R : agent i also earns lower payoffs than when choosing F , since i (1) earns $b - c$ when meeting an agent who chose F , (2) earns 0 when meeting an agent who chose G , and (3) earns $b - c$ when meeting an agent who chose R .

Combining Lemmas 3 and 5 simplifies greatly the analysis, showing that no agent commits to the opposite color in equilibrium. By Lemma 5, the agents with cost $c = 0$ (either Green or Red) choose F in any equilibrium. Since the commitment strategies involve intervals in the costs (Lemma 3), we can set $c_R^G = c_G^R = 0$. An equilibrium will thus only be characterized by the threshold values which we denote $c_R \equiv c_R^R$ and $c_G \equiv c_G^G$ above which a Red and, respectively Green, agent commits to their own color. Figure 4 illustrates these commitment strategies for the green outcome on the left and those for the red outcome on the right.

We consider next in turn equilibrium first stage strategies given either the green outcome or the red outcome in the last stage.

Last-stage green outcome. Working backwards, suppose that all flexible agents play green in the second-stage matching colors game. Consider then the commitment stage.

For Green agents, by Lemma 3 all agents with costs $c \geq b$ play G , and by Lemma 5 all agents with lower costs play F . Hence $c_G = b$.

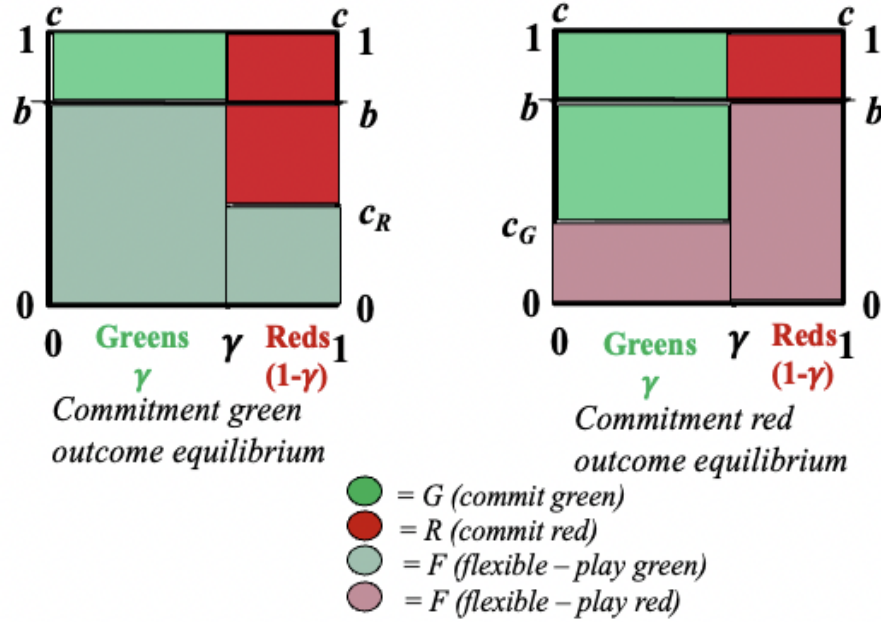


Figure 4: Equilibrium Strategies in the Two-Stage Commitment Game

For Red agents, by Lemma 3 all agents with costs $c \geq b$ play R , but Red agents with costs below b face a trade-off between F and R : By choosing F , a Red agent i can in the second stage choose to play green when it is profitable to do so and play red when it is profitable to do so. That is, i earns $b - c$ if she meets another agent who chose F , $b - c$ if she meets an agent who chose G , and earns b if she meets an agent who chose R . On the other hand, by choosing R a Red agent i can gain when other agents accommodate her commitment; she earns b if she meets an agent is flexible, earns 0 if she meets an agent who chose G , and earns b if she meets another agent who chose R . Thus, Red agents with sufficiently low costs choose F rather than R .

To determine the cut-off cost c_R following the discussion above: A Red agent who chooses R earns b except when she meets a Green agent who has chosen G , so the expected payoff is $b[1 - \gamma(1 - b)]$. Expected payoffs from F are $(b - c)[1 - (1 - \gamma)(1 - c_R)] + (1 - \gamma)(1 - c_R)b$. Setting $c = c_R$ to find the indifferent Red agent, we obtain the quadratic equation:

$$(1 - \gamma)c_R^2 + \gamma c_R - \gamma b(1 - b) = 0$$

This equation has a unique solution $c_R^* \in (0, b)$; at $c_R = 0$, the expression is strictly negative, at $c_R = b$, the expression is strictly positive, and the expression is increasing in c_R . The solution then yields the cut-off commitment cost for Red agents.

Last-stage red outcome. Given all flexible agents choose red, a parallel set of arguments establishes that (1) for all Red agents with costs below b , F is a dominant strategy. Hence, $c_R = b$. (2) Green agents face a trade-off between F and G , and c_G^* is a unique solution to a mirror quadratic equation.

We record these outcomes in the following Proposition:

Proposition 5 *(i) The two-stage game of commitment then matching colors admits exactly two stable perfect Bayesian equilibria:*

(a) an equilibrium in which Green agents with costs above b choose G and those with lower costs choose F , Red agents with costs above c_R^ choose R and those with lower costs choose F , and all agents choose green in a matching colors sub-game between two flexible agents, where c_R^* is the unique solution in $(0, b)$ to*

$$(1 - \gamma)c_R^2 + \gamma c_R - \gamma b(1 - b) = 0. \tag{6}$$

(b) an equilibrium in which Red agents with costs above b choose R and those with lower costs choose F , Green agents with costs above c_G^ choose G and those with lower costs choose F , and all agents choose red in a matching colors sub-game between two flexible agents, where c_G^* is the unique solution in $(0, b)$ to*

$$\gamma c_G^2 + (1 - \gamma)c_G - (1 - \gamma)b(1 - b) = 0. \tag{7}$$

We first remark that, in contrast to the game of Section 3 which has a single equilibrium,

when agents can choose to commit to a color or remain flexible, the game has two stable equilibria. To understand this, note that all the agents with inauthenticity costs $c \geq b$ commit to their own color (Lemma 3). Hence, two flexible agents in the matching colors games both have costs smaller than b , and the game admits two stable pure coordination equilibria; all flexible agents play either green or red. Notably, both equilibria exist even though Greens are the majority in the population. When agents cannot commit to a color, the only equilibrium involves a tyranny of the majority. However, when agents can commit to a color, the outcome might or might not privilege the Green majority.

The following Proposition establishes several properties of the equilibrium thresholds.

Proposition 6 *The equilibrium threshold level c_R^* is increasing in γ , while c_G^* is decreasing in γ . In addition, $c_R^* > c_G^*$, and $c_R^* > c_r^*$ if and only if*

$$(1 - \gamma)b(2\gamma - 1)^2 + \gamma(2\gamma - 1)(1 - 2b(1 - \gamma)) < \gamma(1 - b)(1 - 2b(1 - \gamma))^2.$$

We see that c_R^* is increasing in γ , equal to $\frac{\sqrt{1+4b(1-b)}-1}{2}$ when $\gamma = \frac{1}{2}$ and to $b(1-b) < b$ when γ converges to 1. Recall the trade-off faced by a Red agent in the green outcome equilibrium. By remaining flexible, she increases her chance to match with a committed Green agent; by committing red, she convinces a flexible Red agent to play red. As the fraction of Green agents grows, the payoff of flexibility goes up while the payoff of committing Red goes down. For the Red agent to be indifferent between F and R, the cost of inauthenticity must go up, explaining why c_R^* is increasing in γ . It is bounded at $b(1-b) < b$ because even at high levels of γ , Red agents with costs below b have some incentive to commit red since flexible agents will accommodate their color.

With a parallel argument to that above, c_G^* is increasing in $(1 - \gamma)$. The threshold value c_G^* is decreasing from $\frac{\sqrt{1+4b(1-b)}-1}{2}$ to 0 as γ increases from $\frac{1}{2}$ to 1.

The fraction of Red agents who choose R in the green outcome equilibrium is smaller than the fraction of Green agents who choose G in the red outcome equilibrium. This can

be seen from Figure 4. Since $\gamma > 1/2$, the payoff of a Red agent committing red in the green equilibrium (the size of the red rectangle in the left panel of the Figure) is lower than the payoff of a Green agent committing green in the red equilibrium (the size of the green rectangle in the right panel of the Figure). Furthermore, the payoff of flexibility of a Red agent in the green outcome equilibrium (the size of the gray area in the left panel) is higher than the payoff of flexibility of a Green agent in the red outcome equilibrium (the size of the gray area in the right panel). Thus, the cost of the Red agent who is indifferent between flexibility and committing red, must be higher than the cost of the Green agent who is indifferent between flexibility and committing green.

Finally, while there is no unambiguous way to compare c_R^* and c_r^* we can determine the relationship between them at the extreme values of γ . When γ is close to $\frac{1}{2}$, $c_R^* > c_r^*$, so that there are fewer Red agents committing red in the game with commitment than Red agents choosing red in the game without commitment, but when γ is close to 1, $c_R^* < c_r^*$, so that there are more Red agents committing red in the game with commitment than Red agents choosing red in the game without commitment.

4.3 Welfare comparisons

In this section, we consider the payoffs of different types of agents in the equilibria of the commitment and no-commitment games. We first consider the differences between the two stable equilibria of the commitment game. We then consider how different types of agents fare in the game with commitment relative to the game without commitment.

Welfare comparison: green outcome and red outcome equilibria

We first consider who would gain and who would lose in the green vs. red outcome equilibria of the commitment game. The proportion of agents who commit red in the green outcome equilibrium is higher than the proportion of agents who commit red in the red outcome equilibrium. This implies that die-hard Green agents are actually strictly better off in the red

outcome equilibrium (since more players will accommodate their color). Similarly die-hard Red agents are strictly better off in the green outcome equilibrium.

On the other hand, low-cost Green agents who remain flexible both in the green outcome and red outcome equilibria are more likely to play green in the green outcome equilibrium than in the red outcome equilibrium and hence are strictly better off in the green outcome equilibrium. Similarly, low-cost Red agents are better off in the red outcome equilibrium.

We can compute threshold values of the costs such that Green agents prefer the red outcome equilibrium to the green outcome equilibrium if and only if $c \geq \tilde{c}_G$ and Red agents prefer the green outcome equilibrium to the red outcome equilibrium if and only if $c \geq \tilde{c}_R$.

Proposition 7 *There are two cost thresholds $0 < \tilde{c}_G = \frac{b(1-b)}{1-c_R^*} < c_G^*$ and $0 < \tilde{c}_R = \frac{b(1-b)}{1-c_G^*} < c_R^*$ such that Green agents have higher payoffs in the red outcome equilibrium than in the green outcome equilibrium if and only if $c \geq \tilde{c}_G$ and Red agents have higher payoffs in the green outcome equilibrium than in the red outcome equilibrium if and only if $c \geq \tilde{c}_R$.*

Welfare Comparison: incomplete information game vs. commitment game

We next compare the payoffs of different types of agents in the equilibrium in the incomplete information game, pictured again for convenience on the left of Figure 5 with those in the commitment game green-outcome equilibrium, depicted again for convenience in the middle of Figure 5.

All Red agents are strictly better off in the commitment game in this case. Consider first a die-hard Red agent. In the commitment setting, since the flexible agents accommodate i 's commitment to red, i earns b when meeting any agent except those who chose G (die-hard Greens). In contrast, as can be seen in Figure 3, i would match colors and earn b only when meeting Red agents with costs above c_r and earn 0 in all other meetings. Consider next a Red agent j with cost below b and who commits red: j also earns b in all the same meetings. Relative to the incomplete information game, j matches colors more often with counterparts and does not incur any cost. Finally, consider a Red agent l who is flexible: l must be earning

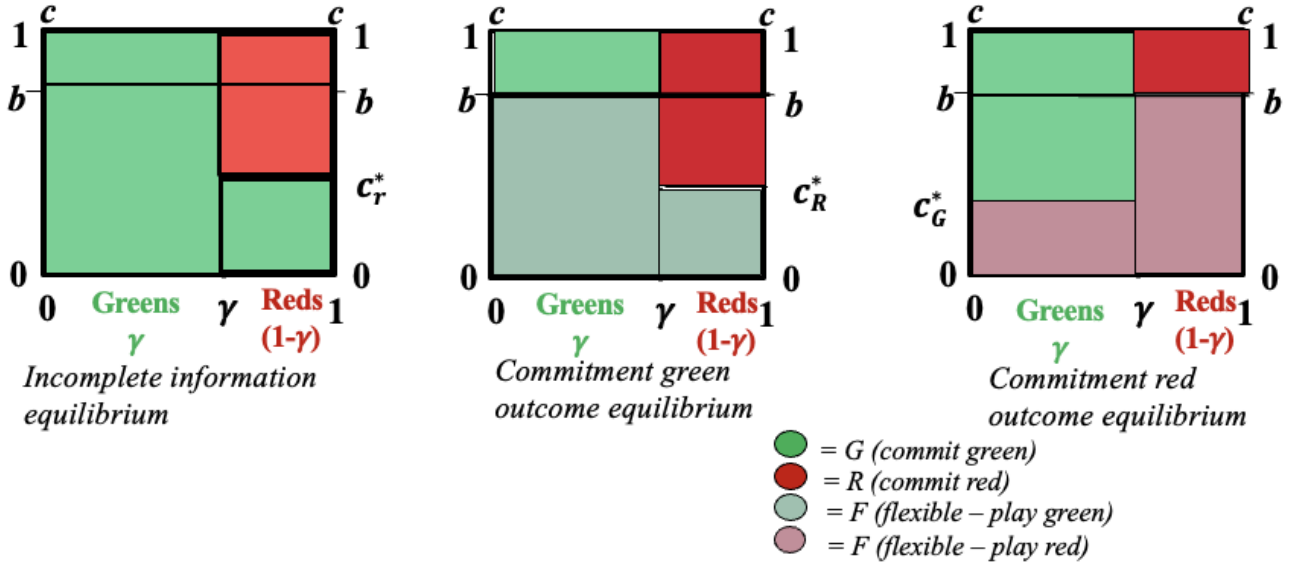


Figure 5: Equilibrium in Incomplete Information vs. Commitment Games

more by playing flexible than by committing, since l values the opportunity to accommodate committed Green agents. Hence since j earns more than in the incomplete information game, l must be earning more as well.

Comparing the payoffs for Green agents is more nuanced and requires comparing the proportion of Reds who ultimately play red in the matching colors game in each setting. Consider a die-hard Green agent i . As seen in the left panel of Figure 4, in the commitment game, i earns b when meeting any agent but Red agents with costs above c_R . As seen in Figure 3, in the incomplete information game, i earns b when meeting any agent but Red agents with costs above c_r . Agent i then earns less in the commitment game when $c_R^* < c_r^*$. The comparison of the payoffs of high cost Green agents who play flexible depends similarly on the relative magnitudes of c_R^* and c_r^* .

We can see that c_r^* is greater than c_R^* for γ sufficiently high. As shown above, both c_r^* and c_R^* are increasing in γ , but c_r^* ranges from 0 to b and c_R^* is bounded above 0 and bounded below b . Hence there exists a γ sufficiently high so that $c_R^* < c_r^*$ and high cost Green agents

earn less in the commitment game than in the incomplete information game.

On the other hand, lowest cost Green agents are always better off in this commitment game. In Figure 4 left panel consider a Green agent j with low cost who plays flexible: j plays green and earns b when meeting another flexible agent or a committed green agent and plays red and earns $b - c$ when meeting a committed red agent. That is, there is no mismatch between agent j and their counterpart in any meeting. In contrast, as seen in Figure 3, in the incomplete information game, j would not match colors with Red agents whose cost is above c_r . For cost c sufficiently small, j 's payoffs are higher in the commitment game.

Hence, we conclude that for low γ such that $c_R^* > c_r^*$ all Green agents are better off in the commitment game. But for high γ such that $c_R^* < c_r^*$, there is threshold value of the cost, $\hat{c} < b$, such that Green agents with cost $c > \hat{c}$ prefer the equilibrium without commitment and Green agents with cost $c < \hat{c}$ prefer the equilibrium with commitment.

Overall, the intuition is that at high levels of γ , Red agents are unlikely to meet each other, and in the incomplete information game have little incentive to play red. This benefits all Green agents. In the commitment game, however, even at high levels of γ Red agents have some incentive to commit red since flexible agents will accommodate their color. This pocket of Red committed agents then makes the high cost Green agents worse off.

We next consider the incomplete information game with the commitment game red-outcome equilibrium, depicted again for convenience on the right of Figure 5.

All Red agents earn higher payoffs in this commitment game relative to incomplete information. Consider first a die-hard Red agent i . In the commitment setting in the color matching game, since the flexible agents accommodate i 's commitment to red, as can be seen in Figure 4 i earns b when meeting any agent except those who committed green (Green agents with costs above c_G^*). In contrast, as can be seen in Figure 3, i would match colors and earn b only when meeting Red agents with costs above c_r and earn 0 in all other meetings. Consider next a Red agent j with cost below b and who has chosen F : j also earns b in all the same meetings except for those who committed Green in which j earns $b - c$. Relative

to the incomplete information game, j matches colors more often with their counterpart and pays the cost in fewer instances.

All Green agents also earn higher payoffs in this commitment game relative to incomplete information. Consider a Green agent i who has chosen G . Since flexible agents accommodate the commitment, this agent earns b in all encounters except for a Red who has chosen R (Reds with costs above b as seen in Figure 4). As shown in Figure 3, in the incomplete information case, agent i does not match colors with a larger group of Red agents since $c_r^* < b$. The same is true for any Green agent j with costs below b but who chose G . Finally consider a Green agent l who chooses F ; l must be earning at least as much as agent j who chooses G . Hence l is also better off in the commitment case; l takes advantage of their low cost and plays red against flexible agents and accommodates any agent who has committed to a color.

Summary of Welfare Comparison between Incomplete Information and Commitment Games

In sum, all Red agents prefer a setting where they can commit to a color, but there is a possible conflict between Green agents. Green agents are all better off in a setting with commitment when their majority is sufficiently small. Even if flexible agents were to play red, a Green agent can commit to green and earn at least much as when commitment is not possible. When the majority is sufficiently large, however, the incomplete information setting is best for the die-hard Green agents since only a small proportion of Red agents choose red. But when commitment is possible, a larger proportion of Red agents play red by committing to do so, which harms the die-hard and other high cost Greens.

We record this outcome in the following proposition:

Proposition 8 *All Red agents earn higher payoffs in the commitment game than in the incomplete information game. All Green agents earn higher payoffs in the commitment game except when γ is sufficiently high. In this case, there exists a cost $\hat{c} < b$, such that Green agents with cost $c > \hat{c}$ earn higher payoffs in the incomplete information game than in the*

green outcome equilibrium of the commitment game.

5 Three-stage game: commitment, selection then matching colors

In this section, we extend the game to consider commitment in asymmetric settings: after agents commit or not to a color, one agent chooses a counterpart from among the others. The modeling represents, for example, an employer who sets the tone for the workplace and then selects an employee. The game then has three stages. First, agents commit or not to a color. Second, the selector chooses a counterpart. Finally, the selector and the counterpart play the matching colors game. We consider perfect Bayesian equilibria of the game, which is formally elaborated below.

5.1 The Three-Stage Game

There are $d+1$ agents, one of whom is designated as a *selector* and the d others are designated as *contenders*. Their interaction proceeds as follows.

- (i) The selector and each contender simultaneously choose whether to commit green, commit red, or remain flexible. As above, the strategy space is $(G)reen$, $(R)ed$, $(F)lexible$
- (ii) agents' choices in the first stage are observed, and the selector chooses one of the contenders.
- (iii) The selector and the chosen contender simultaneously choose actions in a matching colors game; an agent who chose G or R in the first stage plays the corresponding color, and an agent who chose F picks either green or red.

We study the symmetric Perfect Bayesian Equilibria of the three-stage game.

5.2 Equilibria of the three-stage game

We solve the game by backward induction. We start with the matching colors game in the third stage. The outcome is deterministic when either one of the agents has committed to a color. We thus examine the game between two flexible agents. We extend Lemma 7 to show that (i) agents who choose F form an interval in the magnitude of costs, and (ii) agents with costs higher than b weakly prefer to commit to their own color. We then use Proposition 4 to show that there are only two stable equilibria in the game played by flexible agents, which we again call the *green outcome* and the *red outcome*.²

We next consider the selector's choice of a counterpart in the second stage of the game.

Without loss of generality, consider a Green selector.³ *Green outcome*: If the Green selector anticipates the green outcome in the final stage, she is indifferent between picking any of the contenders who play G or F , and she prefers them to contenders who play R . *Red outcome*: If the Green selector anticipates that the red outcome in the final stage, her selection rule depends on whether she has chosen G or F . A Green selector playing G is indifferent between selecting any contender who plays G or F , and she prefers them to contenders who choose R . A Green selector playing F prefers any contender who plays G and is indifferent among those who play F or R .

Whenever the Green selector is indifferent between two groups of contenders, we assume that she employs a *uniform* selection rule, which assigns the same probability to all members of the two groups. The uniform selection rule is non-discriminatory, and treats all the contenders who confer the same payoffs equally.⁴

²All formal results and proofs for Section 5 are given in the Appendix.

³The selection rule of a Red selector is obtained by reversing the two colors.

⁴Clearly, when the selector is indifferent, she may also choose to discriminate and pick different contenders with different probabilities. In the Appendix, we study an alternative selection rule, the *hierarchical selection rule* where the Green (Red) selectors prioritize contenders who play G (respectively R) over contenders who play F and contenders who play F over contenders who play R (respectively G).

5.2.1 Equilibria of the three-stage game

We now turn to the first stage of the game and characterize the stable Perfect Bayesian equilibria assuming first that the green outcome is played in the final stage of the game.

Green Outcome Equilibrium

Green Outcome: Selector's Incentives

We first show that a Green selector with $c < b$ always prefers to play F : (i) The selector earns the same payoffs from choosing G or F when she faces a committed green or flexible contenders. If all contenders choose R , however, the Green selector prefers F , as this gives her a payoff of $(b - c)$ whereas G results in a payoff of 0. (ii) Playing R for a Green selector never higher payoffs than G or F . If all contenders are committed red, she receives the same payoff as choosing F . If some contenders are committed green or flexible, she would obtain the highest payoff of b from choosing F . If instead she chooses R , she either obtains a payoff of $-c$ (when all contenders play G) or a payoff of $b - c$ when there exists a contender playing F .

We next consider the incentives of a Red selector. We first note that a Red selector will not commit to green. As the zero-cost selector prefers to remain flexible over committing green, all Red selectors prefer to remain flexible over committing green. To compute the optimal strategy of the Red selector choosing F or R , let α_G and α_R be the measures of contenders who play G and R in the symmetric Perfect Bayesian equilibrium. By committing red, the Red selector will end up playing the same color as a contender except when all contenders play G , so the expected payoff of the selector is given by

$$\Pi_R^s = b(1 - \alpha_G^d)$$

Remaining flexible, the Red selector will match colors with probability one, but pay the cost c whenever there is no contender who plays R . Hence her expected payoff is given by

$$\Pi_F^s = b - c(1 - \alpha_R)^d.$$

We conclude that a Red selector plays R if and only if

$$c \geq c^s \equiv b \frac{\alpha_G^d}{(1 - \alpha_R)^d}$$

Interestingly, since $\alpha_G + \alpha_R < 1$, for a fixed strategy of the contenders, the threshold value of the cost over which a Red selector commits given the green outcome in the final stage is strictly decreasing in d . When competition increases, the Red selector is more likely to face a contender who does not play G , and hence has a stronger incentive to commit in order to induce flexible contenders to play red. In the limit, as d goes to infinity, the threshold value of the cost goes to zero. The Red selector always commits when the number of contenders becomes arbitrarily large.

Green Outcome: Contenders' incentives

Consider now the incentives of a Green contender with cost $c < b$. Under the uniform selection rule, the probability of being selected by a Green selector or a flexible Red selector and obtaining the payoff b is the same if she plays G or F . However, if the selector commits red, the contender is selected with higher probability and obtains the positive payoff $b - c$ when she plays F than when she plays G . We conclude that a Green contender with cost $c < b$ has no incentive to play G , and hence no Red contender has an incentive to play G either.

Next, we need to compare the payoff of a Green (respectively Red) contender when she plays R versus F . Using the uniform selection rule, a contender who plays R is only chosen by a Green contender when all contenders play R , is chosen by a flexible Red selector with uniform probability among the contenders who play R , and is chosen by a committed Red selector with uniform probability among the contenders who either play R or F . Hence, the expected payoff of a contender committing red (gross of the inauthenticity cost) is given by

$$\begin{aligned}\Pi_R = & \frac{b}{d}[\gamma b \alpha_R^{d-1} + d(1-\gamma)c^s \sum_{k=0}^{d-1} \frac{1}{k+1} \binom{d-1}{k} \alpha_R^k (1-\alpha_R)^{d-1-k} \\ & + d(1-\gamma)(1-c^s) \sum_{k=0}^{d-1} \frac{1}{k+1} \binom{d-1}{k} (1-\alpha_G)^k \alpha_G^{d-1-k}]\end{aligned}$$

This expression can be simplified by using the following binomial formula, proven in the Appendix

Lemma 6 For any p, k ,

$$\sum_{k=0}^{d-1} \frac{1}{k+1} \binom{d-1}{k} (1-p)^k p^{d-1-k} = \frac{1-p^d}{d(1-p)}.$$

to give the expected payoff of playing R as:

$$\Pi_R = \frac{b}{d}[\gamma b \alpha_R^{d-1} + (1-\gamma)c^s \frac{1-(1-\alpha_R)^d}{\alpha_R} + (1-\gamma)(1-c^s) \frac{1-\alpha_G^d}{1-\alpha_G}].$$

A contender who plays F is chosen by a Green selector uniformly among the contenders who play G or F , by a flexible Red selector uniformly among all contenders if there is no contender playing R and by a committed Red selector uniformly among the contenders who either play R or F . Hence the expected payoff of remaining flexible, gross of the inauthenticity cost, can be computed as

$$\Pi_F = \frac{b}{d}[\gamma \frac{1-\alpha_R^d}{1-\alpha_R} + (1-\gamma)c^s (1-\alpha_R)^{d-1} + (1-\gamma)(1-c^s) \frac{1-\alpha_G^d}{1-\alpha_G}].$$

Green Outcome: Equilibrium for $d \rightarrow \infty$

As d goes to infinity, c^s goes to zero, and $d\Pi_R$ and $d\Pi_F$ converge to

$$d\Pi_R = b(1-\gamma) \frac{1}{1-\alpha_G} < d\Pi_F = b[\gamma \frac{1}{1-\alpha_R} + (1-\gamma) \frac{1}{1-\alpha_G}].$$

As $\pi_R < \pi_F$, the zero-cost agent prefers F , and the threshold value for a Red selector to choose R is given by

$$(b - c^e) \frac{\gamma}{1 - (1 - \gamma)(1 - c^e)} = 0,$$

so that c^e converges to b : all contenders with cost $c < b$ choose F . We record this finding

Proposition 9 *Any Green Selector with cost $c < b$ remains flexible. As the number of contenders grows large, any Red selector chooses to commit whereas all Contenders choose to remain flexible.*

As the Red selector is more and more likely to choose R , Red contenders have less incentive to choose R . They prefer F , as this will not affect the selection decision of a committed Red selector, but increases the probability of being selected by a Green selector under the uniform selection rule.

Green Outcome: Equilibrium for d finite

For finite, arbitrary values of d , the computation of equilibrium thresholds is much more complex. Even when $d = 2$, the threshold values of the costs are solutions to a system of cubic polynomial equations.

For illustrative purposes, we compare outcomes for small values of d . When $d = 2$, we prove in the Supplementary Appendix existence of a unique equilibrium, and show that contenders are more likely to remain flexible when $d = 2$ than when $d = 1$. This outcome is also illustrated in a numerical analysis which provides the threshold values for the selector c^s and for the contenders c^e when $b = 0.6$ for different γ and values of $d \in \{1, 2, 3\}$.

Figure 6 displays the equilibrium threshold for the selector and for a contender as a function of γ for $d = 1$ (in blue), $d = 2$ (in orange) and $d = 3$ (in green). As the number of contenders increases, the threshold of the selector goes down and the threshold of a contender goes up. That is, when competition increases, the selector is more likely to commit while contenders are more likely to be flexible. We also observe that, for $d \geq 2$, the thresholds of the selector

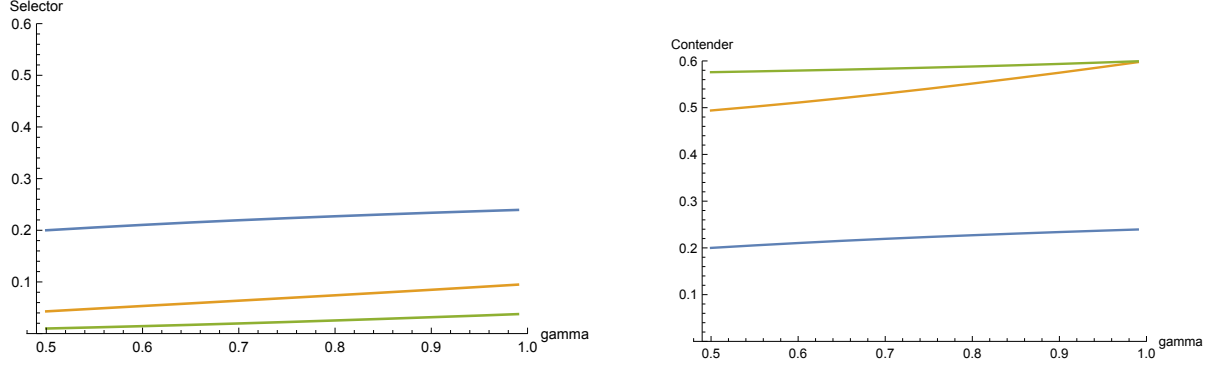


Figure 6: Threshold values for the selector and the contender as a function of γ for $d = 1, 2, 3$

and contenders are increasing in γ : as the fraction of Green agents goes up, both the Red selector and Red contenders are less likely to choose R .

Red Outcome in Final Stage If the flexible agents play the red outcome in the final stage of the game, the analysis of the Perfect Bayesian equilibria requires reversing the incentives of the Green and Red contenders.

The Red selector only has an incentive to commit when $c > b$ and remains flexible for all $c \leq b$. The Green selector chooses G if and only if

$$c \geq c^s \equiv b \frac{\alpha_R^d}{(1 - \alpha_G)^d}$$

Committing red is always dominated by remaining flexible, so that contenders must choose between choosing G and F , which results in gross payoffs:

$$\begin{aligned} \Pi_G &= \frac{b}{d} \left[(1 - \gamma b) \alpha_G^{d-1} + \gamma c^s \frac{1 - (1 - \alpha_G)^d}{\alpha_G} + \gamma (1 - c^s) \frac{1 - \alpha_R^d}{1 - \alpha_R} \right], \\ \Pi_F &= \frac{b}{d} \left[(1 - \gamma) \frac{1 - \alpha_G^d}{1 - \alpha_G} + \gamma c^s (1 - \alpha_G)^{d-1} + \gamma (1 - c^s) \frac{1 - \alpha_R^d}{1 - \alpha_R} \right]. \end{aligned}$$

As d goes to infinity, c^s goes to zero, $\pi_G < \pi_F$ so that the zero-cost agent prefers to remain flexible, and the threshold cost of the green agent above which G is chosen is given by the

solution to

$$(b - c^e) \frac{(1 - \gamma)}{1 - \gamma(1 - c^e)} = 0,$$

so that $c^e = 0$: all Green contenders choose to remain flexible. Hence, in both the red outcome and the green outcome equilibrium, as the number of contenders grows large, the selector and contenders whose color is privileged in the final stage remain flexible, the selector whose color is not played in the third stage equilibrium chooses to commit to her color, while the contenders whose color is not played in the third stage equilibrium prefer to remain flexible. The analysis of the three-stage game with red outcome for finite d is given in the Supplementary Appendix.

6 Conclusion

This paper builds and analyzes a simple model of social interactions in a heterogeneous population. Agents benefit from matching their actions but agents in different groups find it more or less costly to adopt the preferred actions of others. A key, new feature of the model is that agents can choose their action depending on whom they meet. We can then compare outcomes under different scenarios of the visibility of agent's categories and the possibility of committing to certain actions in the hopes that others will accommodate.

The analysis yields three main insights. First, when agents do not know each other's categories and costs and no-one can commit to an action, the unique equilibrium involves the majority-preferred action. The analysis thus identifies the worst-case conditions for a minority; the smallest proportion chooses authentic actions. Second, when agents can commit to an action, there are two equilibrium outcomes, one of which privileges the majority action but one of which privileges the minority action. Almost all agents are better, as there is less mismatch of actions overall. However, the die-hard majority agents and others with very high costs of playing the minority action can be worse off. Finally, minority members are best off when they have capability to both commit and choose among contenders for interactions. A

minority employer, for example, has a greater incentive to commit to her preferred action as the number of contenders grows very large, since she is almost surely will be able to choose a contender who will accommodate her preference or has committed to the same action.

Future research could consider the possibility of policies that change agents' incentives as well as more complex social and economic settings. As mentioned in the Introduction, companies, for example, can institute programs which encourage workers to be authentic. This emphasis on authenticity could lower the costs of the minority of taking certain actions or shift the distribution. Rather than randomly match, agents of different categories could take actions to make themselves more visible and easy to find, inducing homophily in the network of social relations. The features of a network of interaction would change agents' incentives to commit to certain actions. A fuller model would consider the trade-offs of diverse interactions and segregation.

References

- Acemoglu, Daron, and Matthew O Jackson.** 2017. "Social norms and the enforcement of laws." *Journal of the European Economic Association*, 15(2): 245–295.
- Akerlof, George A.** 1980. "A theory of social custom, of which unemployment may be one consequence." *The quarterly journal of economics*, 94(4): 749–775.
- Akerlof, George A, and Rachel E Kranton.** 2000. "Economics and identity." *The quarterly journal of economics*, 115(3): 715–753.
- Austen-Smith, David, and Roland G Fryer Jr.** 2005. "An economic analysis of "acting white"." *The Quarterly Journal of Economics*, 120(2): 551–583.
- Bernheim, B Douglas.** 1994. "A theory of conformity." *Journal of political Economy*, 102(5): 841–877.

- Bilancini, Ennio, and Leonardo Boncinelli.** 2018. “Social coordination with locally observable types.” *Economic Theory*, 65(4): 975–1009.
- Carlsson, Hans, and Eric Van Damme.** 1993. “Global games and equilibrium selection.” *Econometrica: Journal of the Econometric Society*, 989–1018.
- Carvalho, Jean-Paul.** 2013. “Veiling.” *The Quarterly Journal of Economics*, 128(1): 337–370.
- Creary, Stephanie.** 2023. “Why Covering can Harm Diversity in the Workplace.” <https://knowledge.wharton.upenn.edu/podcast/knowledge-at-wharton-podcast/why-covering-can-harm-diversity-in-the-workplace>.
- Cunningham, Kim.** <https://www.welcometothejungle.com/en/articles/how-to-be-authentic-at-work>. “Be real, get ahead: The power of authenticity in your career.” *Welcome to the Jungle*, April 24, 2024.
- Deloitte.** 2023. “Uncovering culture: A call to action for leaders.” <https://www2.deloitte.com/us/en/pages/about-deloitte/articles/uncovering-culture.html>.
- Farrell, Joseph.** 1987. “Cheap talk, coordination, and entry.” *The RAND Journal of Economics*, 34–39.
- Galesloot, Bob M, and Sanjeev Goyal.** 1997. “Costs of flexibility and equilibrium selection.” *Journal of Mathematical Economics*, 28(3): 249–264.
- Ganguly, Chirantan, and Indrajit Ray.** 2023. “Information revelation and coordination using cheap talk in a game with two-sided private information.” *International Journal of Game Theory*, 52(4): 957–992.
- Genicot, Garance.** 2022. “Tolerance and Compromise in Social Networks.” *Journal of Political Economy*, 130(1): 94–120.

- Goyal, Sanjeev, and Maarten CW Janssen.** 1997. “Non-exclusive conventions and social coordination.” *journal of economic theory*, 77(1): 34–57.
- Hochschild, Arlie Russell.** 1983. *The managed heart: commercialization of human feeling*. Berkeley: University of California Press.
- Jackson, Matthew, and Evan Storms.** 2024. “Safe spaces: shelters or tribes?” mimeo.
- Kuran, Timur, and William H Sandholm.** 2008. “Cultural integration and its discontents.” *The Review of Economic Studies*, 75(1): 201–228.
- Manski, Charles F, and Joram Mayshar.** 2003. “Private incentives and social interactions: Fertility puzzles in Israel.” *Journal of the European economic association*, 1(1): 181–211.
- Michaeli, Moti, and Daniel Spiro.** 2015. “Norm conformity across societies.” *Journal of public economics*, 132: 51–65.
- Michaeli, Moti, and Daniel Spiro.** 2017. “From peer pressure to biased norms.” *American Economic Journal: Microeconomics*, 9(1): 152–216.
- Morris, Stephen, and Hyun Song Shin.** 2003. “Global games: Theory and applications.” *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, 56–114, Cambridge University Press.
- Naono, Miharuru.** 2022. “Cost heterogeneity and the persistence of bilingualism.” *Games and Economic Behavior*, 136: 325–339.
- Tirole, Jean.** 2023. “Safe spaces: shelters or tribes?” mimeo.
- Villas-Boas, J Miguel.** 2002. “Comparative statics of fixed points.” *Journal of Economic Theory*, 73(1): 769–930.

- Vives, Xavier.** 1990. “Nash equilibrium with strategic complementarities.” *Journal of Mathematical Economics*, 19(3): 305–321.
- Vives, Xavier.** 2005. “Complementarities and games: New developments.” *Journal of Economic Literature*, 43(2): 437–479.
- Wells, Janelle E, and Doreen MacAuley.** 2024. “Invisible Mending: The Silent Struggle of Conforming at Work.” *Psychology Today*.
- Yoshino, Kenji.** 1997. “Covering.” *Yale Law Review*, 4(11): 183–198.
- Yoshino, Kenji.** 2007. *Covering: The hidden assault on our civil rights*. Random House Trade Paperbacks.
- Young, H Peyton.** 1993. “The evolution of conventions.” *Econometrica: Journal of the Econometric Society*, 57–84.

7 Appendix

7.1 Definition of Stable Equilibrium

In the simultaneous move incomplete information game, let μ_g and μ_r denote the measures of agents playing green and red, respectively, in an equilibrium of a matching colors game. Consider a small exogenous change such that μ_g increase by ϵ and μ_r decreases by ϵ . If, following this exogenous change, the measure of agents whose best response is to play green is higher by *more than* ϵ , we say the equilibrium is *unstable*. If instead the measure of agents whose best response is to play green increases *less than* ϵ , we say the equilibrium is *stable*. This definition, written formally below, mirrors that of Vives (1990) and Vives (2005).

Definition 1 *Consider an equilibrium of the matching colors game with the thresholds (c_g, c_r) such that all Green agents with costs above c_g play green, all Red agents with costs above c_r play red, and those with costs below these cut-off costs play the opposite color,. The equilibrium is stable if there exists $\bar{\epsilon}$ such that, for all $\bar{\epsilon} > \epsilon > 0$,*

$$\begin{aligned} \gamma[c_g(\mu_g, \mu_r) - c_g(\mu_g + \epsilon, \mu_r - \epsilon)] + (1 - \gamma)[c_r(\mu_g + \epsilon, \mu_r - \epsilon) - c_r(\mu_g, \mu_r)] &< \epsilon, \\ \gamma[c_g(\mu_g - \epsilon, \mu_r + \epsilon) - c_g(\mu_g - \mu_r)] + (1 - \gamma)[c_r(\mu_g, \mu_r) - c_r(\mu_g - \epsilon, \mu_r + \epsilon)] &< \epsilon \end{aligned}$$

7.2 Proofs of Section 3

Proof of Lemma 1. We first show that the equilibrium is characterized by cutoff costs. Consider a Red agent with inauthenticity cost c . If the agent plays green, she earns $\pi_g - c$ and she earns π_r if she plays red. If it is optimal for the agent to play red, then $c \geq \pi_g - \pi_r$ and all Red agents with cost $c' > c$ also prefer to play red. If it is optimal for the agent to play green, $c \leq \pi_g - \pi_r$ and all Red agents with cost $c' < c$ also prefer to play green. Hence there is a cutoff cost c_r above which a Red agent always chooses to play red and below which a red agent always chooses to play green.

We next consider the three necessary conditions on the cutoffs.

- (i) Immediate, as it is a dominant strategy for an agent with cost $c \geq b$ to play her own color.
- (ii) If $c_r > 0$ and $c_g > 0$, then $\pi_g - \pi_r > 0$ and $\pi_r - \pi - g > 0$, a contradiction.
- (iii) Suppose that $c_r = c_g = 0$. Then the zero cost agent obtains a payoff $\pi_g = \gamma > 1 - \gamma = \pi_r$, so that she cannot be indifferent between playing green and red, contradicting the fact that $c_r = c_g = 0$. ■

Proof of Proposition 1. First, we show that in any equilibrium $c_g = 0$ and $c_r \geq 0$. Suppose not and, by Lemma 1, let $c_r = 0$ and $c_g \geq 0$.

The equilibrium cut-off $c - g$ is given by the solution to $\pi_g - (\pi_r - c) = 0$. Rearranging and letting $c_r = 0$, we obtain

$$\pi_g - (\pi_r - c_g) = b[(2\gamma - 1) + c_g(1 - 2b\gamma)].$$

Now, if $1 - 2b\gamma > 0$, as $2\gamma - 1 \geq 0$, $\pi_g - (\pi_r - c_g) > 0$, resulting in a contradiction. If instead $1 - 2b\gamma \leq 0$, then

$$\pi_g - (\pi_r - c_g) = b[(2\gamma - 1) - c_g(2b\gamma - 1)].$$

But as $b < 1$ and $c_g \leq b$, we obtain again $\pi_g - (\pi_r - c_g) > 0$, resulting in a contradiction.

Second, for $c_g = 0$, we solve for the cut-off c_r . Consider a Red agent with cost c and compute $\pi_r - (\pi_g - c) = 0$ when $c_g = 0$. We obtain:

$$c_r^* = \frac{b(2\gamma - 1)}{1 - 2b(1 - \gamma)}. \tag{8}$$

For $\gamma > 1/2$, the numerator and denominator are both strictly positive, hence $c_r^* > 0$. At $\gamma = 1/2$, $c_r^* = 0$.

We next check the stability of this equilibrium. In the equilibrium we have

$$\begin{aligned}c_g &= 0, \\c_r &= (\mu_g - \mu_r)b\end{aligned}$$

where, by the definition of stability, μ_g and μ_r are the measures of agent playing red and green. For ϵ small enough, the threshold values of c_g and c_r also involve $c_g = 0$ and $c_r \in (0, 1)$. For an increase in μ_g we have

$$\gamma[c_g(\mu_g, \mu_r) - c_g(\mu_g + \epsilon, \mu_r - \epsilon)] + (1 - \gamma)[c_r(\mu_g + \epsilon, \mu_r - \epsilon) - c_r(\mu_g, \mu_r)] = 2(1 - \gamma)b\epsilon < \epsilon$$

where the last inequality stems from the fact that $\gamma > \frac{1}{2}$ and $b < 1$. Similarly, for an increase in μ_r ,

$$\gamma[c_g(\mu_g - \epsilon, \mu_r + \epsilon) - c_g(\mu_g - \mu_r)] + (1 - \gamma)[c_r(\mu_g, \mu_r) - c_r(\mu_g - \epsilon, \mu_r + \epsilon)] = 2(1 - \gamma)b\epsilon < \epsilon,$$

showing that the equilibrium is indeed stable. ■

Proof of Proposition 2: Differentiating c_r^* with respect to γ :

$$\frac{\partial c_r^*}{\partial \gamma} = \frac{4b(1 - b)}{(1 - 2b(1 - \gamma))^2}, \tag{9}$$

which is strictly positive since $b < 1$. Taking the second derivative, we have

$$\frac{\partial^2 c_r^*}{\partial \gamma^2} = -\frac{16b^2(1 - b)}{(1 - 2b(1 - \gamma))^3} < 0. \tag{10}$$

Differentiating c_r with respect to b we obtain:

$$\frac{\partial c_r^*}{\partial b} = \frac{(2\gamma - 1)}{(1 - 2b(1 - \gamma))^2}, \quad (11)$$

which is strictly positive since $\gamma \in (\frac{1}{2}, 1)$ and $b < 1$. Taking the second derivative, we obtain

$$\frac{\partial^2 c_r^*}{\partial b^2} = \frac{4(1 - \gamma)(2\gamma - 1)}{(1 - 2b(1 - \gamma))^3} > 0, \quad (12)$$

completing the proof of the Proposition.

■

Proof of Lemma 2: Let π_g be the benefits of playing green and let π_r be the benefits of playing red. A Green agent who plays green earns π_g and a Red agent with cost c who plays green earns $\pi_g - c$. (A) We first note that the proportion of Greens who play green and Reds who play green must be connected intervals in the cost space. Suppose not. Suppose that some measure of higher cost Reds play green and some measure of lower cost Reds play red. It is then possible to switch assigned play so that some higher cost Reds play red and some lower costs Reds play green keeping the aggregate measures of agents choosing green and red the same. The payoffs π_g and π_r are not changed but aggregate costs are lower. (B) We then note that in the first best, either all Reds are assigned to play red or all Greens are assigned to play green. Suppose not. Building on (A) so that agents are assigned their colors in connected intervals, consider a configuration where Greens with costs below $\bar{c}_G > 0$ play red and Reds with costs below $\bar{c}_R > 0$ play green. The play can then be reassigned so that some of these Greens play green and some of the Reds play red, keeping the aggregate measures of agents playing green and red the same. The payoffs π_g and π_r are unchanged but aggregate costs are lower. (C) We finally prove that the first best involves all Greens playing green, Reds with costs below some x playing green, and Reds with higher costs playing red. To see this, note that from (A) and (B) we can conclude that the first best involves either (1) all Greens play green, and Reds with costs below some level x play green and those with higher costs play red

or (2) all Reds play red, and Greens with costs below some level y play red and those with higher costs play green. Consider the latter configuration. The expected benefits of playing red are $((1 - \gamma) + \gamma y)$ and the expected benefits of playing green are $\gamma(1 - y)$. Aggregate payoffs are then

$$((1 - \gamma) + \gamma y)^2 b + (\gamma(1 - y))^2 b - \gamma \frac{y^2}{2}$$

Now consider switching the configuration so that all Greens play green, all Reds with costs below y play green and those with higher costs play red. Overall welfare is then

$$(\gamma + (1 - \gamma)y)^2 b + ((1 - \gamma)(1 - y))^2 b - (1 - \gamma) \frac{y^2}{2}.$$

The total costs are lower since $\gamma > 1/2$. The comparison of the benefits is $(1 - 2\gamma)2y(y - 1)$ which is positive since $\gamma > 1/2$ and $0 \leq y \leq 1$.

The question then becomes the optimal level of x . In this configuration, a proportion of the population $(\gamma + (1 - \gamma)x)$ plays green and the remaining $((1 - \gamma)(1 - x))$ play red. With random meetings among the agents, an arbitrary agent who plays green earns expected benefits $\pi_g = (\gamma + (1 - \gamma)x)b$. An arbitrary agent who plays red earns expected benefits $\pi_r = ((1 - \gamma)(1 - x))b$. Letting W denote the aggregate expected payoffs, since Reds with costs below x play green, we have

$$W = (\gamma + (1 - \gamma)x)\pi_g - (1 - \gamma) \int_0^x cdc + ((1 - \gamma)(1 - x))\pi_r$$

Substituting for π_g and π_r , and with the uniform cost distribution we have

$$W = (\gamma + (1 - \gamma)x)^2 b - (1 - \gamma) \frac{x^2}{2} + ((1 - \gamma)(1 - x))^2 b$$

This function is quadratic in x , and differentiating with respect to x under the constraint $0 \leq x \leq 1$ yields the optimal proportion of Reds who play green, c_r^O . Whenever $b \geq \frac{1}{4(1-\gamma)}$, the quadratic function is increasing in x for $x \geq 0$. Whenever $\frac{1}{4(1-\gamma)} > b > 1/2$, the quadratic function is increasing in x for $x \in [0, 1]$. Hence, $c_r^O = 1$ whenever $b \geq \frac{1}{2}$. Whenever $b < 1/2$,

the quadratic function attains a maximum in $[0, 1]$ given by

$$c_r^O = \frac{2b(2\gamma - 1)}{1 - 4b(1 - \gamma)},$$

. which is zero at $b = 0$ and increasing in b . ■

Proof of Proposition 3: Immediate by comparing c_r^* and c_r^O . ■

7.3 Proofs of Section 4

Proof of Lemma 3: We first show that the commitment strategies are intervals in the space of costs. For this proof, we consider a Red agent i . The proof does not rely on the fact that $\gamma > \frac{1}{2}$ and hence can be replicated to prove the statement for a Green agent.

Let π_R, π_G and π_F denote the payoff of committing red, committing green and remaining flexible. Notice that π_R is independent of c . On the other hand, when agent i commits to play green, she pays the cost c with probability 1. When agent i remains flexible, she pays the cost c with a probability $p = \Pr[j \text{ commits green or } j \text{ remains flexible and } i \text{ plays green against a flexible agent}]$. Hence $\pi_G = a_1 - c$ whereas $\pi_F = a_2 - pc$ for some positive a_1, a_2 . We thus observe that π_G is strictly decreasing in c , π_F is weakly decreasing in c and $\pi_G - \pi_F = (a_1 - a_2) - (1 - p)c$ is weakly decreasing in c .

Now suppose that agent i commits to play red at cost c and suppose that $c' > c$. We then have

$$\pi_R \geq \Pi_F(c) \geq \pi_F(c')$$

$$\pi_R \geq \pi_G(c) > \pi_G(c'),$$

where the first inequalities stem from the fact that agent i weakly prefers to commit red, and the second inequalities from the fact that both π_F and π_G are decreasing in c .

Next suppose that agent i commits to play green at cost c and suppose that $c' < c$. We then have

$$\begin{aligned}\pi_G(c') &> \pi_G(c) \geq \pi_R, \\ \pi_G(c') - \pi_F(c') &\geq \pi_G(c) - \pi_F(c) \geq 0,\end{aligned}$$

where the first inequalities stem from the fact that π_G and $\pi_G - \pi_F$ are decreasing in c and the second inequalities from the fact that agent i weakly prefers to commit green.

We next show that the upper bound of the cost thresholds cannot be higher than b . For this proof, we again consider a Red agent. The proof does not rely on the fact that $\gamma > \frac{1}{2}$ and hence can be replicated to prove the statement for a Green agent. Suppose by contradiction that $c_R^R > b$ and consider an agent with cost $b < c < c_R^R$. In an equilibrium of the two-stage game, the agent cannot commit to play green (which gives a strictly negative payoff) and hence must choose to remain flexible. By the tie-breaking rule, this implies that $\pi_F(c) > \pi_R$.

Because $c > b$, the agent always plays red in the matching colors game and hence $\pi_F(c) = \pi_F$ is independent of c and only depends on the probability that the agent is matched to another agent playing red in equilibrium. Hence $\pi_F > \pi_R$ implies that the probability that the Red agent is matched to another agent playing red is *strictly higher* when the agent is flexible than when the agent commits to play red.

If the agent commits red, she matches colors with an agent playing red whenever the other agent is either committed red or flexible. When the agent is flexible, she matches colors with an agent playing red when the other agent is committed red or flexible and plays red in the matching colors game. Hence, the probability that a Red agent meets an agent playing red is at least as high when the Red agent commits red than when she remains flexible, a contradiction. ■

Proof of Lemma 4: Similar to the proof of (iii) of Lemma 1; ■

Proof of Proposition 4: We first construct the four possible equilibria of the sub-game. Consider first the equilibria where all agents play the same strategy. Without loss of generality, suppose that all agents play green, $c_r = c_R^R$ and $c_g = c_R^G$.

The agent with highest incentive to deviate is the Red agent with the cost c_R^R . This agent earns $b - c_R \geq 0$ in equilibrium. If this agent were to deviate and play red, she would earn 0 payoffs, since all agents in the flexible-flexible sub-game play green. Hence the agent has no incentive to deviate. A parallel argument holds for $c_r = c_G^R$ and $c_g = c_G^G$.

Now consider an equilibrium where all Red agents play red and some Green agents play green, i.e. with $c_r = c_G^R$ and $c_R^G \leq c_g < c_G^G$. Consider the gross payoff of playing red and green in the matching colors game,

$$\begin{aligned}\pi_r &= \frac{((1 - \gamma)(c_R^R - c_G^R) + \gamma(c_g - c_R^G))b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}, \\ \pi_g &= \frac{\gamma(c_G^G - c_g)b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}\end{aligned}$$

For any Green agent, the difference between playing green and red is given by

$$\pi_g - \pi_r + c = \frac{(\gamma c_G^G + \gamma c_R^G - (1 - \gamma)(c_R^R - c_G^R) - 2\gamma c_g)b + c(\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R))}{\gamma c_G + (1 - \gamma)c_R}.$$

An equilibrium will thus exist if and only if there exists a unique solution $c \in [c_R^G, c_R^R]$ to

$$H(c) \equiv (\gamma c_G^G + \gamma c_R^G - (1 - \gamma)(c_R^R - c_G^R))b - c(2\gamma b - \gamma(c_G^G - c_R^G) - (1 - \gamma)(c_R^R - c_G^R)) = 0.$$

Now at $c = c_R^G$,

$$H(c_R^G) = (\mu_g - \mu_r) - c_R^G(\mu_g + \mu_r).$$

So $H(c_R^G) > 0$ if and only if $\gamma_G - \gamma_R > c_R^G$. On the other hand, $H(c_G^G) = -b + c_G^G < 0$. As $H(\cdot)$ is a linear function of c , it will have a unique solution if and only if $H(c_R^G) > 0$.

Next consider an equilibrium with $c_g = c_R^G$ and $c_G^R < c_r < c_R^R$. In this equilibrium, the gross payoff of playing red and green in the matching colors game are

$$\begin{aligned}\pi_r &= \frac{(1 - \gamma)(c_R^R - c_r)b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}, \\ \pi_g &= \frac{(\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_r - c_G^R))b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}\end{aligned}$$

An equilibrium exists if there is a unique solution c in $[c_G^R, c_R^R]$ to

$$D(c) = ((1 - \gamma)(c_R^R - c_G^R) - \gamma(c_G^G - c_R^G))b - c[2(1 - \gamma)b - \gamma(c_G^G - c_R^G) - (1 - \gamma)(c_R^R - c_G^R)] = 0$$

Now at $c = c_G^R$,

$$D(c_G^R) = (\mu_r - \mu_g) - c_G^R(\mu_g + \mu_r).$$

So $D(c_G^R) > 0$ if and only if $\gamma_R - \gamma_G > c_G^R$. On the other hand, $D(c_R^R) = -b + c_R^R < 0$. As $D(\cdot)$ is a linear function of c , it will have a unique solution if and only if $D(c_G^R) > 0$. ■

7.4 Proof of Unstable Equilibria in Matching Colors Game

Following the definition of stability from above, let μ_g and μ_r denote the measures of agents playing green and red in the matching colors game. Fix the total size. For any fixed measures (μ_g, μ_r) , we compute the corresponding thresholds as follows:

If $\mu_g > \mu_r$,

$$\begin{aligned}c_g(\mu_g, \mu_r) &= c_R^G, \\ c_r(\mu_g, \mu_r) &= \min\left\{\frac{(\mu_g - \mu_r)b}{\mu_g + \mu_r}, c_R^R\right\}\end{aligned}$$

where the latter is from an indifference condition between a Red person playing green and

incurring cost c vs. playing red and not incurring the cost. If $\mu_r > \mu_g$,

$$\begin{aligned} c_r(\mu_g, \mu_r) &= c_G^R, \\ c_g(\mu_g, \mu_r) &= \min\left\{\frac{(\mu_r - \mu_g)b}{\mu_g + \mu_r}, c_G^G\right\} \end{aligned}$$

A pair of threshold strategies (c_g, c_r) forms a Nash equilibrium of the matching colors game if and only if the measures and thresholds are consistent, i.e.

$$\begin{aligned} \mu_g &= \gamma(c_G^G - c_g(\mu_g, \mu_r)) + (1 - \gamma)(c_r(\mu_g, \mu_r) - c_G^R), \\ \mu_r &= \gamma(c_g(\mu_g, \mu_r) - c_R^G) + (1 - \gamma)(c_R^R - c_r(\mu_g, \mu_r)) \end{aligned}$$

Now, as per the definition of stability, fix an equilibrium and consider a small exogenous change in the measures of agents playing green and red, which leaves the total measure of agents constant. Suppose first that the measure of agents playing green increases by ϵ so that necessarily the measure of agents playing red decreases by ϵ . If the measure of agents finding it optimal to play green, given $\mu_g + \epsilon$, is less than $\mu_g + \epsilon$, the equilibrium is stable, as the measure of agents playing green and red will eventually converge back to the initial equilibrium. The same must true for μ_r and a perturbation of ϵ .

We show that (i) The equilibrium where all agents play green ($c_g = c_R^G, c_r = c_R^R$) is stable except when $c_R^R = b$ and $2(1 - \gamma)b > \gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)$. (ii) The equilibrium where all agents play red ($c_g = c_G^G, c_r = c_G^R$) is stable except when $c_G^G = b$.

Consider the equilibrium where all agents play green. The only relevant perturbation is to increase the measure of agents playing red from 0 to ϵ , and decrease the measure of agents playing green from $\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)$ to $\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R) - \epsilon$. For ϵ

sufficiently small, $\mu_g - \epsilon > \mu_r + \epsilon$, so that

$$\begin{aligned} c_g(\mu_g - \epsilon, \mu_r + \epsilon) &= c_R^G, \\ c_r(\mu_g - \epsilon, \mu_r + \epsilon) &= \min\left\{\frac{(\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R - 2\epsilon)b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}, c_R^R\right\} \end{aligned}$$

Now, as long as $c_R^R < b$, we may fix

$$\bar{\epsilon} = \frac{(\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R))(b - c_R^R)}{2},$$

so that for all $\bar{\epsilon} > \epsilon > 0$,

$$c_r(\mu_g - \epsilon, \mu_r + \epsilon) = c_R^R$$

We then have

$$\begin{aligned} c_g(\mu_g - \epsilon, \mu_r + \epsilon) &= c_g(\mu_g, \mu_r) = 0, \\ c_r(\mu_g - \epsilon, \mu_r + \epsilon) &= c_r(\mu_g, \mu_r) = c_R^R, \end{aligned}$$

so that

$$\gamma[c_g(\mu_g - \epsilon, \mu_r + \epsilon) - c_g(\mu_g, \mu_r)] + (1 - \gamma)[c_r(\mu_g, \mu_r) - c_r(\mu_g - \epsilon, \mu_r + \epsilon)] = 0 < \epsilon$$

showing that the equilibrium is stable.

If now $c_R^R = b$, then for all ϵ ,

$$c_r(\mu_g - \epsilon, \mu_r + \epsilon) = \left[1 - \frac{2\epsilon}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}\right]b$$

and

$$\gamma[c_g(\mu_g - \epsilon, \mu_r + \epsilon) - c_g(\mu_g, \mu_r)] + (1 - \gamma)[c_r(\mu_g, \mu_r) - c_r(\mu_g - \epsilon, \mu_r + \epsilon)] = \frac{2(1 - \gamma)\epsilon b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)}$$

Hence, the equilibrium is stable if and only if

$$\frac{2(1 - \gamma)\epsilon b}{\gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R)} < \epsilon$$

or

$$2(1 - \gamma)b < \gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R).$$

A parallel argument can be made for the equilibrium where all agents play red. In that case, as $\gamma \geq 1 - \gamma$ and $b \geq c_G^G, b \geq c_R^R$, we always have

$$2\gamma b \geq \gamma(c_G^G - c_R^G) + (1 - \gamma)(c_R^R - c_G^R).$$

so that whenever $c_G^G = b$, the equilibrium where all agent play red is unstable.

We next show that (i) The equilibrium $c_r = c_R^R$ and $c_g \in [c_R^G, c_G^G]$ (which exists if and only if $\gamma_G - \gamma_R \geq c_G^R$) is unstable (ii) The equilibrium $c_g = c_G^G$ and $c_r \in [c_G^R, c_R^R]$ (which exists if and only if $\gamma_R - \gamma_G \geq c_G^R$) is also unstable.

Suppose that $\gamma_G > \gamma_R$ and consider the equilibrium $c_r = c_G^R, c_g \in [c_R^G, c_G^G]$. In this equilibrium, as the agent with cost $c = c_G^R$ plays red, we must have $\mu_r > \mu_g$. Pick an ϵ small enough so that $\mu_r - \epsilon > \mu_g + \epsilon$. Then

$$\begin{aligned}
c_r(\mu_g, \mu_r) &= c_G^R, \\
c_g(\mu_g, \mu_r) &= \frac{(1-\gamma)(c_R^R - c_G^R) - \gamma c_R^G + \gamma c_g(\mu_g, \mu_r))b}{\gamma(c_G^G - c_R^G) + (1-\gamma)(c_R^R - c_G^R)} \\
c_r(\mu_g + \epsilon, \mu_r - \epsilon) &= c_G^R, \\
c_g(\mu_g + \epsilon, \mu_r - \epsilon) &= \frac{(1-\gamma)(c_R^R - c_G^R) - \gamma c_R^G + \gamma c_g(\mu_g, \mu_r) - 2\epsilon)b}{\gamma(c_G^G - c_R^G) + (1-\gamma)(c_R^R - c_G^R)}
\end{aligned}$$

Hence,

$$\gamma[c_g(\mu_g, \mu_r) - c_g(\mu_g + \epsilon, \mu_r - \epsilon)] + (1-\gamma)[c_r(\mu_g + \epsilon, \mu_r - \epsilon) - c_r(\mu_g, \mu_r)] = \frac{2\gamma\epsilon b}{\gamma(c_G^G - c_R^G) + (1-\gamma)(c_R^R - c_G^R)}$$

Now, as $\gamma_G - \gamma_R \geq c_R^G \geq 0$, $\gamma(c_G^G - c_R^G) > (1-\gamma)(c_R^R - c_G^R)$ and

$$\frac{2\gamma\epsilon b}{\gamma(c_G^G - c_R^G) + (1-\gamma)(c_R^R - c_G^R)} > \frac{2\gamma\epsilon b}{2\gamma(c_G^G - c_R^G)} = \frac{\epsilon b}{(c_G^G - c_R^G)},$$

and as $b \geq c_G^G \geq c_G^G - c_R^G$,

$$\frac{\epsilon b}{(c_G^G - c_R^G)} \geq \epsilon,$$

showing that the equilibrium is unstable. A parallel argument applies to the equilibrium

$c_g = c_R^G$, $c_r \in [c_G^R, c_R^R)$ which exists when $\gamma_R - \gamma_G \geq c_G^R \geq 0$. ■

Proof of Lemma 5: The proof follows from the text. ■

Proof of Proposition 5: The proof follows from the text. ■

Proof of Proposition 6: By implicit differentiation, we obtain

$$\frac{\partial c_R^*}{\partial \gamma} = \frac{b(1-b) - c_R^*(1-c_R^*)}{2-\gamma)c_R^* + \gamma}.$$

Using the first order condition

$$b(1 - b) = c_R^* + \frac{1 - \gamma}{\gamma} c_R^{*2}$$

so that

$$b(1 - b) - c_R^*(1 - c_R^*) = \frac{c_R^{*2}}{\gamma} > 0.$$

A parallel arguments show that c_G^* is decreasing in γ .

Next, using equations (6) and (7) we obtain

$$\begin{aligned} \gamma^2 c_G^{*2} + \gamma(1 - \gamma)c_G^* &= (1 - \gamma)^2 c_R^{*2} + \gamma(1 - \gamma)c_R^*, \\ (\gamma c_G^* + (1 - \gamma)c_R^*)(\gamma c_G^* - (1 - \gamma)c_R^*) &= \gamma(1 - \gamma)(c_R^* - c_G^*) \end{aligned}$$

Suppose by contradiction that $c_R^* < c_G^*$. Then we must have

$$\gamma c_G^* < (1 - \gamma)c_R^*$$

which implies, as $\gamma > 1 - \gamma$ that $c_G^* < c_R^*$, a contradiction.

Finally, replacing c_r^* by $\frac{b(2\gamma-1)}{1-2b(1-\gamma)}$ and plugging back in equation (6), we obtain the condition for $c_R^* > c_r^*$ given in the Proposition. ■

Proof of Propositions 7, and 8 : The proofs follow from the text. ■

7.5 Proofs of Section 5

Lemma 7 : *Any equilibrium of the three-stage game where agents play weakly dominant strategies under the uniform selection rule is characterized by cost intervals $[c_R^G, c_G^G)$ and $[c_R^R, c_G^R)$ in which, respectively, Green agents choose F and Red agents choose F, and $c_R^R \leq b$ and $c_G^G \leq b$*

Proof of Lemma 7: Consider a Red selector. By committing Red, she never incurs the cost c , by committing green she incurs the cost c with probability 1 and by remaining flexible, she incurs the cost c with probability

$$p = \Pr[\text{all contenders commit to play green or there is no contender who plays red,} \\ \text{at least one contender remains flexible and } i \text{ plays green against a flexible contender}]$$

We repeat the arguments of the proof of Lemma 3 to show that if the selector with cost c commits red, then any selector with cost $c' > c$ also commits red, and if the Selector with cost c commits green, the any selector with cost $c' < c$ also commits green.

Consider next a Red contender. Again, by committing Red, she never incurs the cost c , by committing green she incurs the cost c with probability 1 and by remaining flexible, she incurs the cost c with probability

$$p = \Pr[\text{The selector chooses } i \text{ and the selector commits to green, or remains flexible} \\ \text{and agent } i \text{ plays green against a flexible selector}]$$

Repeating the arguments of the proof of Lemma 3 if a contender with cost c commits red, then any contender with cost $c' > c$ also commits red, and if a contender with cost c commits green, the any contender with cost $c' < c$ also commits green.

We next show that, under the uniform selection rule, all agents with cost $c > b$ weakly prefer to commit.

Consider a Red selector. As in the proof of Lemma 3 suppose by contradiction that $c_R^R > b$ and consider a selector with cost $c > b$ who prefers to remain flexible. This implies that the probability of being matched to a contender playing red must be higher when the selector

remains flexible than when she commits to red.

However, when a Red selector commits to red, she is matched with a contender playing red whenever there exists a contender committing red or remaining flexible. When a Red selector remains flexible, she is matched with a contender playing red either when there exists a contender committing red or when a flexible contender playing red. The probability that there exists a red agent committing red or remaining flexible is larger or equal to the probability that there exists a red contender committing red or a flexible contender playing red, a contradiction.

Consider next a Red contender. By the same argument, the probability of being selected and matched with a selector playing red must be higher when the contender remains flexible than when she commits red. If the selector commits red, as flexible contenders may play green, she weakly prefers to select a committed red contender to a flexible red contender. Similarly, if the selector is flexible and plays red, she weakly prefers to select a committed red contender to a flexible red agent. Hence, the probability of being selected and matched to a selector playing red is weakly higher when the contender commits to play red than remains flexible, a contradiction. ■

Proof of Lemma 6: Notice first that

$$\frac{1}{k+1} \binom{d-1}{k} = \frac{1}{d} \binom{d}{k+1}.$$

Hence,

$$\sum_{k=0}^{d-1} \frac{1}{k+1} \binom{d-1}{k} (1-p)^k p^{d-1-k} = \sum_{k=0}^{d-1} \frac{1}{d} \binom{d}{k+1} (1-p)^k p^{d-1-k}.$$

Now,

$$\begin{aligned}\frac{1}{d} \sum_{k=0}^{d-1} \binom{d}{k+1} (1-p)^k p^{d-1-k} &= \frac{1}{d(1-p)} \sum_{k=0}^{d-1} \binom{d}{k+1} (1-p)^{k+1} p^{d-1-k} \\ &= \frac{1}{d(1-p)} \sum_{k=1}^d \binom{d}{k} (1-p)^k p^{d-k} \\ &= \frac{1}{d(1-p)} \left[\sum_{k=0}^d \binom{d}{k} (1-p)^k p^{d-k} - p^d \right] \\ &= \frac{1-p^d}{d(1-p)},\end{aligned}$$

completing the proof of the Lemma. ■

Proof of Proposition 9 The proof follows from the text. ■

8 Supplementary Appendix

8.1 Three stage game with green outcome: finite d

Lemma 8 *In the three-stage game under the uniform selection rule, if agents play the green outcome equilibrium in the third stage, in a symmetric equilibrium, the zero-cost contender remains flexible.*

Proof of Lemma 8 : Compute the difference in payoff between committing red and remaining flexible for the zero-cost contender as

$$G = -b[\gamma(\sum_{k=0}^{d-1} \alpha_R^k - b\alpha_R^{d-1}) - (1-\gamma)c^s(\sum_{k=0}^{d-1} (1-\alpha_R)^k - (1-\alpha_R)^{d-1})]$$

Suppose by contradiction that $G > 0$, then $\alpha_R = (1-\gamma) + \gamma c_G^R \geq 1-\gamma$ and $1-\alpha_R = \gamma(1-c_G^R) \leq \gamma$. Now,

$$\begin{aligned} G &\leq -b[\gamma \sum_{k=0}^{d-2} \alpha_R^k - (1-\gamma)c^s \sum_{k=0}^{d-2} (1-\alpha_R)^k] \\ &\leq -b[\gamma \sum_{k=0}^{d-2} \alpha_R^k - (1-\gamma) \sum_{k=0}^{d-2} (1-\alpha_R)^k] \end{aligned}$$

As $\alpha_R \geq 1-\gamma$ and $1-\alpha_R \leq \gamma$,

$$\begin{aligned} \gamma \sum_{k=0}^{d-2} \alpha_R^k - (1-\gamma) \sum_{k=0}^{d-2} (1-\alpha_R)^k &\geq \gamma \sum_{k=0}^{d-2} (1-\gamma)^k - (1-\gamma) \sum_{k=0}^{d-2} \gamma^k \\ &= 1 - (1-\gamma)^{d-1} - (1-\gamma)^{d-1}, \\ &= \gamma^{d-1} - (1-\gamma)^{d-1}. \end{aligned}$$

As $\gamma \geq 1 - \gamma$, we conclude that $G \leq 0$, a contradiction which shows that the zero-cost contender prefers to remain flexible than to commit red. ■

With two contenders, the thresholds of the Red selector and contenders are given by the solutions to the equations:

$$\Delta_2^s(c^s, c^e) \equiv c^s(1 - (1 - \gamma)(1 - c^e))^2 - b(\gamma(1 - b))^2 = 0 \quad (13)$$

$$\begin{aligned} \Delta_2^e(c^e, c^s) &\equiv -(b - c^e)[\gamma(1 + (1 - \gamma)(1 - c^e)) + (1 - \gamma)c^s(1 - (1 - \gamma)(1 - c^e))] \quad (14) \\ &+ b[\gamma b(1 - \gamma)(1 - c^e) + (1 - \gamma)c^s(1 + 1 - (1 - \gamma)(1 - c^e))] \\ &= 0 \end{aligned}$$

Proposition 10 *In a perfect Bayesian equilibrium of the three-stage game where agents play the green F-F equilibrium in the third stage, a Green selector and Green contenders all choose to be flexible ($c_G^s = c_G^e = b$) and a Red selector and Red contenders commit, respectively, if and only if $c \geq c_2^s$ and $c \geq c_2^e$, where c_2^s and c_2^e are solutions to equations (13) and (14) in $[0, b]$. Compared to the case with a single contender, in any perfect Bayesian equilibrium, a Red selector commits more $c_2^s < c_1^s$ while Red contenders commit less, $c_2^e > c_1^e$.*

Proof of Proposition 10: We show that, in a Perfect Bayesian equilibrium of the game, $c_2^s < c_1^s$ and $c_2^e > c_1^e$. Consider first the selector and let ϕ_2^s be the implicit function defined by $\Delta_2^s = 0$, i.e. $c^s = \phi_2^s(c^e)$. (We similarly define ϕ_1^s to be the implicit function defined by $\Delta_1^s = 0$ when $d = 1$.) We compute

$$\frac{\partial \Delta_2^s}{\partial c^e} = 2(1 - \gamma)c^s(1 - (1 - \gamma)(1 - c^e)) > 0$$

and

$$\frac{\partial \Delta_2^s}{\partial c^s} = (1 - (1 - \gamma)(1 - c^e))^2 > 0$$

We thus have

$$\frac{\partial \phi_2^s}{\partial c^e} = -\frac{\partial \Delta_2^s / \partial c^e}{\partial \Delta_2^s / \partial c^s} < 0,$$

showing that the function $\phi_2^s(c^e)$ is decreasing. Next note that, when there is only one contender, the best response of the selector is given by the solution to the equation

$$\Delta_1^s(c^s, c^e) \equiv c^s[1 - (1 - \gamma)(1 - c^e)] - b\gamma(1 - b) = 0.$$

We observe that

$$\Delta_2^s(c^s, c^e) = \Delta_1^s(c^s, c^e)(1 - (1 - \gamma)(1 - c^e)) + b\gamma(1 - b)(\gamma b + (1 - \gamma)c^e).$$

We thus have, at the point $\phi_1^s(c^e)$ where $\Delta_1^s(c^s, c^e) = 0$,

$$\Delta_2^s(\phi_1^s(c^e), c^e) > 0,$$

and as Δ_2^s is strictly increasing in c^s , the unique point at which $\Delta_2^s(c^s, c^e) = 0$ must be strictly smaller than $\phi_1^s(c^e)$. Hence, for all c^e ,

$$\phi_2^s(c^e) < \phi_1^s(c^e)$$

.

Next, we consider a contender. We compute

$$\begin{aligned}
\frac{\partial \Delta_2^e}{\partial c^e} &= \gamma(1 + (1 - \gamma)(1 - c^e)) + (1 - \gamma)c^s(1 - (1 - \gamma)(1 - c^e)) \\
&+ (1 - \gamma)(b - c^e)(1 + (1 - \gamma)c^s) - \gamma(1 - \gamma)b^2 + (1 - \gamma)^2 c^s \\
&> 0.
\end{aligned}$$

This establishes that there is a unique solution to the equation

$$\Delta_2^e(c^e, c^s) = 0.$$

This allows to define the implicit function given by $\Delta_2^e(c^e, c^s) = 0$, as $c^e = \phi_2^e(c^s)$.

We next show that $\frac{\partial \Delta_2^e}{\partial c^s} > 0$.

$$\begin{aligned}
\frac{\partial \Delta_2^e}{\partial c^s} &= -(1 - \gamma)(b - c^e)(1 - (1 - \gamma)(1 - c^e)) \\
&+ b(1 - \gamma)(1 + (1 - (1 - \gamma)(1 - c^e))) \\
&= (1 - \gamma)(b - c^s(1 - (1 - \gamma)(1 - c^e))) \\
&> 0.
\end{aligned}$$

Hence, as $\frac{\partial \phi_2^e}{\partial c^s} = -\frac{\partial \Delta_2^e / \partial c^s}{\partial \Delta_2^e / \partial c^e} < 0$, the function $\phi_2^e(c^s)$ is decreasing in c^s .

We finally need to show that $\phi_2^e(c^s)$ belongs to the interval $[0, b]$. Since Δ_2^e is increasing in c^s , if $\Delta_2^e(0, b) < 0$ then $\Delta_2^e(0, c^s) < 0$ for all c^s . Similarly, if $\Delta_2^e(b, 0) > 0$ then $\Delta_2^e(b, c^s) > 0$ for all c^s . Now

$$\begin{aligned}
\Delta_2^e(0, b) &= -b(\gamma(2 - \gamma) + b\gamma(1 - \gamma) + b[\gamma b(1 - \gamma) + b(1 - \gamma)(2 - \gamma)]) \\
&= b(2 - \gamma)(-\gamma + b(1 - \gamma)) < 0,
\end{aligned}$$

where the last inequality is due to the fact that $\gamma \geq 1 - \gamma$. Hence, for any c^s , $\Delta_2^e(0, c^s) < 0$.
Now,

$$\Delta_2^e(b, 0) = b[\gamma b(1 - \gamma)(1 - b) + (1 - \gamma)b(2 - n(1 - \gamma)(1 - b))] > 0,$$

establishing that , for any c^s , $\Delta_2^e(b, c^s) > 0$, so that the unique value $\phi_2^e(c^s)$ belongs to the interval $[0, b]$

Next recall that

$$\Delta_1^e(c^s, c^e) \equiv c^e[1 - (1 - \gamma)(1 - c^s)] - b\gamma(1 - b) = 0$$

so that

$$\begin{aligned} \Delta_2^e(c^s, c^e) &= \Delta_1^e(c^s, c^e) - [(b - c^e)(1 - \gamma)(1 - c^e)(\gamma - (1 - \gamma)c^s)] \\ &\quad - b[(1 - (1 - \gamma)(1 - c^e))(b\gamma - (1 - \gamma)c^s)] \end{aligned}$$

We thus have, at the point $\phi_1^e(c^s)$ where $\Delta_1^e(c^s, c^e) = 0$, as $\gamma \geq 1/2$ and $1 > b \geq c^s$, $\gamma - (1 - \gamma)c^s > b\gamma - (1 - \gamma)c^s \geq 0$ and

$$\Delta_2^e(c^s, \phi_1^e(c^s)) < 0.$$

As Δ_2^e is strictly increasing in c^e , the unique point at which $\Delta_2^e(c^s, c^e) = 0$ must be strictly greater than $\phi_1^e(c^s)$. Hence, for all c^s ,

$$\phi_2^e(c^s) > \phi_1^e(c^s).$$

The equilibrium cut-off levels (c_2^s, c_2^e) are given by the solutions to the system of two

equations

$$\begin{aligned}c_2^s &= \phi_2^s(c_2^e), \\c_2^e &= \phi_2^e(c_2^s)\end{aligned}$$

Next, notice that the equilibrium level c_2^e is given by the fixed point of $\phi_2^e \circ \phi_2^s$ and the equilibrium level c_2^s by the fixed point of $\phi_2^s \circ \phi_2^e$.

Now for any c , as $\phi_1^s(c) > \phi_2^s(c)$ and $\phi_1^e(\cdot)$ is decreasing,

$$\phi_1^e[\phi_1^s(c)] < \phi_1^e[\phi_2^s(c)].$$

Furthermore as, for any c , $\phi_1^e(c) < \phi_2^e(c)$,

$$\phi_1^e[\phi_2^s(c)] < \phi_2^e[\phi_2^s(c)]$$

so that, for any c ,

$$(\phi_1^e \circ \phi_1^s)(c) < (\phi_2^e \circ \phi_2^s)(c)$$

Let $T_1(c) \equiv (\phi_1^e \circ \phi_1^s)(c)$ and $T_2(c) \equiv (\phi_2^e \circ \phi_2^s)(c)$ be two continuous functions defined on $[0, b]$. By Theorem 1, p. 184 in Villas-Boas (2002), the highest fixed point of T_1 is lower than the highest fixed point in T_2 and the lowest fixed point in T_2 is higher than the lowest fixed point in T_1 . Call \underline{c}_2^e the lowest fixed point of T_2 and c_1^e the unique fixed point of T_1 . Then $c_1^e < \underline{c}_2^e$ showing that in equilibrium $c_1^e < c_2^e$.

Similarly, we can show that

$$(\phi_2^s \circ \phi_2^e)(c) < (\phi_1^s \circ \phi_1^e)(c).$$

and we conclude, using Theorem 1 in Villas-Boas (2002) again, and letting \overline{c}_2^s denote the highest fixed point of $(\phi_2^s \circ \phi_2^e)$ and c_1^s the unique fixed point of $(\phi_1^s \circ \phi_1^e)$ that $\overline{c}_2^s < c_1^s$, showing

that $c_2^s < c_1^s$. ■

8.2 Three stage game with red outcome: finite d

If the flexible players play the red outcome equilibrium in the third stage of the game, the analysis of the Perfect Bayesian equilibria requires reversing the incentives of the Green and Red contenders.

The Red selector only has an incentive to commit when $c > b$ and remains flexible for all $c \leq b$. The Green selector commits green if and only if

$$c \geq b \frac{\alpha_R^d}{(1 - \alpha_G)^d}$$

Committing red is always dominated by remaining flexible, so that contenders must choose between committing green and remaining flexible, which results in gross payoffs:

$$\begin{aligned} \Pi_G &= \frac{b}{d} \left[(1 - \gamma b) \alpha_G^{d-1} + \gamma c^s \frac{1 - (1 - \alpha_G)^d}{\alpha_G} + \gamma (1 - c^s) \frac{1 - \alpha_R^d}{1 - \alpha_R} \right], \\ \Pi_F &= \frac{b}{d} \left[(1 - \gamma) \frac{1 - \alpha_G^d}{1 - \alpha_G} + \gamma c^s (1 - \alpha_G)^{d-1} + \gamma (1 - c^s) \frac{1 - \alpha_R^d}{1 - \alpha_R} \right]. \end{aligned}$$

As opposed to the situation where the green outcome equilibrium is played in the third stage, it is no longer true that the zero-cost player always prefers to remain flexible than commit green. In fact, numerical simulations show that there may exist multiple equilibria, including an equilibrium where some Red contenders commit green when the fraction of Green players, γ , is sufficiently large. However, a numerical analysis shows the existence of an equilibrium where the zero-cost player remains flexible for all values of γ and d . In this equilibrium, the value of the thresholds are given by

$$b[(1 - \gamma)b\alpha_G^{d-1} + \gamma c^s \frac{1 - (1 - \alpha_G)^d}{\alpha_G}] = (b - c^e)[(1 - \gamma)\frac{1 - \alpha_G^d}{1 - \alpha_G} + \gamma c^s(1 - \alpha_G)^{d-1}]. \quad (15)$$

with

$$c^s = b \frac{\alpha_R^d}{(1 - \alpha_G)^d}, \alpha_G = \gamma(1 - c^e), \alpha_R = (1 - \gamma)(1 - b).$$

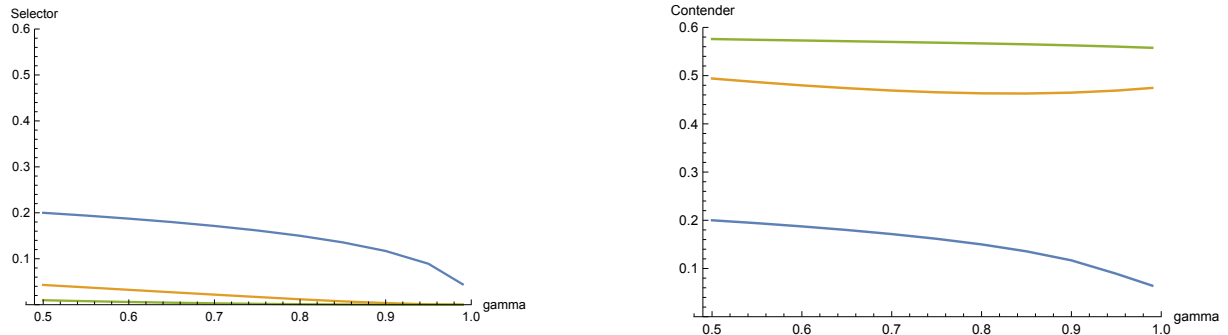


Figure 7: Threshold values for the selector and the contender as a function of γ for $d = 1, 2, 3$

Figure 7 illustrates the threshold values of the selector and contenders when the red outcome equilibrium is played in the third stage. The effect of an increase in the number of contenders is the same as when the green outcome equilibrium is played: the Green selector is more likely to commit and Green contender more likely to remain flexible as the number of contenders grows from 1 to 2 and 3. The effect of an increase in γ on the Green players' incentives to commit is less clear than in the green outcome equilibrium case - while a stronger Green majority increases the selector's incentive to commit, it has a non-monotonic effect on the contender's incentive to commit.

8.3 Three stage game with hierarchichal selection rule

We no longer assume that the selector places equal probabilities on flexible and committed contenders. We assume instead that she employs a *hierarchichal* selection rule, where the Green

(respectively Red) selector prioritizes contenders who committed green (red) over flexible contenders and prioritizes flexible contenders over those who committed red (green). While both selection rules are optimal strategies of the Green selector in the perfect Bayesian equilibria of the three-stage game, the hierarchical selection rule weakly dominates the uniform selection rule.

Notice that the selector's incentives to commit are independent of the selection rule. Hence, if the F-F green equilibrium is played in the third stage, Green selectors with cost $c < b$ remain flexible and Red selectors commit red whenever

$$c \geq c^s = b \frac{\alpha_G^d}{(1 - \alpha_R)^d}.$$

If the red equilibrium is played in the third stage, Red selectors with cost $c < b$ remain flexible and Green selectors commit green whenever

$$c \geq b \frac{\alpha_R^d}{(1 - \alpha_G)^d}.$$

We next turn to the contender's incentives to commit, and first assume that the green equilibrium is played in the third stage.

8.3.1 Equilibria in the green outcome

The contenders' incentives to commit are more complex than when the selector employs a uniform selection rule, as commitment decisions directly affect the probability of being selected.

The expected payoff of committing green (gross of the inauthenticity cost) is now given by

$$\Pi_G = \frac{b}{d} \left[\gamma \frac{1 - (1 - \alpha_G)^d}{\alpha_G} + (1 - \gamma) c^s \alpha_G^{d-1} \right]$$

The gross expected payoff of remaining flexible is now

$$\Pi_F = \frac{b}{d} \left[\gamma \frac{(1 - \alpha_G)^d - \alpha_R^d}{1 - \alpha_G - \alpha_R} + (1 - \gamma) \frac{(1 - \alpha_R)^d - \alpha_G^d}{1 - \alpha_G - \alpha_R} \right].$$

It is no longer the case that remaining flexible dominates committing green for all Green agents with cost $c < b$. Committing green now increases the probability of being chosen by a Green selector, and this effect implies that there is a threshold value of the cost $0 \leq c_G < b$ above which green agents prefer to commit green.

The gross expected payoff of committing red is given by

$$\Pi_R = \frac{b}{d} \left[\gamma b \alpha_R^{d-1} + (1 - \gamma) \frac{1 - (1 - \alpha_R)^d}{\alpha_R} \right].$$

There is a threshold value $0 \leq c_R \leq b$ above which red agents prefer to commit red.

As opposed to the case of a uniform selection rule, an equilibrium is now characterized by two threshold values c_G and c_R . Another difficulty stems from the fact that it is no longer the case that the zero agent prefers to remain flexible. This implies that the indifferent contenders at c_G and c_R may either be Green or Red, inducing several possible regimes.

When $d = 2$, we observe that committing red is always dominated by remaining flexible as the difference in payoffs is given by

$$\gamma(1 - \alpha_G + \alpha_R(1 - b)) - (1 - \gamma)(1 - \alpha_G) = (1 - \alpha_G)(2\gamma - 1) + \alpha_R(1 - b) > 0.$$

We conclude that an agent with zero cost *either chooses to remain flexible or to commit green*. The difference in payoffs between remaining flexible and committing green is given by:

$$\Delta_{FG} = -(2\gamma - 1)(1 - \alpha_R) + (1 - \gamma)\alpha_G(1 - c^s).$$

The difference Δ_{FG} is decreasing in γ . When $\gamma = \frac{1}{2}$, $\Delta_{FG} = \frac{1}{2}\alpha_G(1 - c^s) > 0$ and when $\gamma = 1$, $\Delta_{FG} = -(1 - \alpha_R) < 0$. Hence, when the size of the majority γ is sufficiently small,

the zero-cost agent prefers to remain flexible, but when the majority becomes very large, she prefers to commit green. We conclude that there are three possible regimes:

1. **Regime 1** The zero-cost agent remains flexible. There are two thresholds $c_G \leq b$ and $c_R \leq b$ above which Green (respectively Red) contenders choose to commit green (respectively red)
2. **Regime 2** The zero-cost agent commits green. All Green contenders commit green, and there are two thresholds $0 < c_G < c_R \leq b$ such that Red contenders with costs $c \in [0, c_G]$ commit green, Red contenders with costs $c \in [c_G, c_R]$ remain flexible, and Red contenders with costs $c \in [c_R, 1]$ commit red
3. **Regime 3** The zero-cost agent commits green. All Green contenders commit green, and there is a single threshold c^e such that all Red contenders with cost $c \in [0, c^e]$ commit green and all Red contenders with cost $c \in [c^e, 1)$ commit red. (No contender remains flexible)

The first panel of Figure 9 shows that the three different regimes emerge for different values of the parameters b and γ . Regime 1 appears when the size of the majority is small. Regime 2 emerges for low values of the coordination gain b and intermediate values of γ . Finally, regime 3 (with no flexible contender) appears when the size of the majority is sufficiently large.

The last three panels of figure ?? compute the threshold values above which a Red selector commits red, and the threshold values of Green and Red contenders as a function of γ for $b = 0.6$. The top-left panel shows that the Red selector chooses to remain flexible whenever regime 3 is reached and all contenders choose to commit, as there is no longer any incentive to induce a flexible Red contender to play red. The bottom-right panel shows that, When γ is small (regime 1), Green contenders have an incentive to commit, up to the point where the threshold reaches zero and all Green contenders commit (switch from regime 1 to regime 2). The bottom-left panel illustrates the behavior of Red contenders. For low values of γ (regime

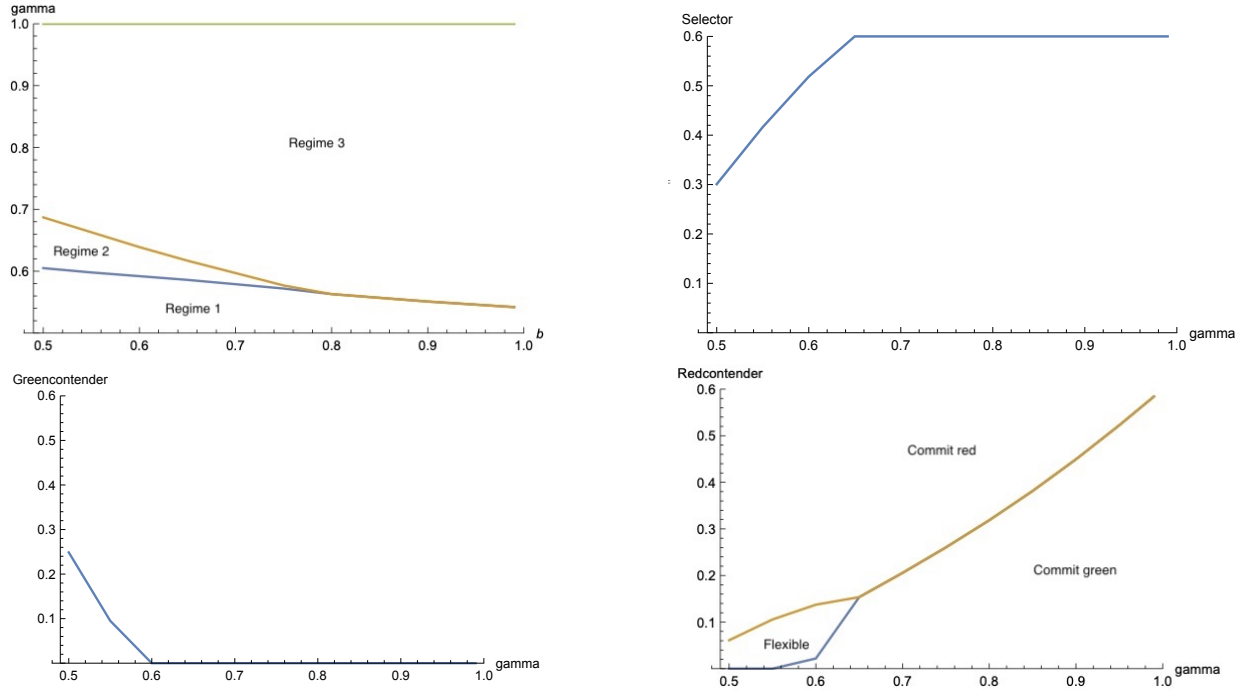


Figure 8: Regimes and threshold values for the selector and the contenders as a function of γ

1), Red contenders either remain flexible or commit red. As γ increases, the incentive to commit red goes down, and the threshold c_R increases. However, as soon as regime 2 appears, Red contenders with low cost have an incentive to commit green, and the threshold c_G goes up with γ , so that more and more Red contenders commit green. As regime 3 is reached, there are no longer any flexible Red contenders and, as γ increases, the fraction of Red contenders committing green goes up and the fraction of Red contenders committing red goes down.

Finally, we note that, as d goes to infinity, the contenders' incentives to commit are no longer driven by the expected change in the behavior of the selector (all Red selectors commit and all Green selectors with cost $c < b$ remain flexible). Instead, the incentives to commit are driven by the probability of being selected. Under the hierarchical selection rule, as d becomes large, flexible agents will never be selected as there will always be committed green and red contenders available. Hence, the incentives to remain flexible disappear, and in the limit, all Green contenders commit green and all Red contenders commit red.

8.3.2 Equilibria in the red outcome case

If agents play the red outcome equilibrium, the roles of Red and Green agents are reversed and the expected gross payoffs of committing green, remaining flexible and committing red are given by

$$\begin{aligned}\Pi_G &= \frac{b}{d} \left[\gamma \frac{1 - (1 - \alpha_G)^d}{\alpha_G} + (1 - \gamma) b \alpha_G^{d-1} \right], \\ \Pi_F &= \frac{b}{d} \left[\gamma \frac{(1 - \alpha_G)^d - \alpha_R^d}{1 - \alpha_G - \alpha_R} + (1 - \gamma) \frac{(1 - \alpha_R)^d - \alpha_G^d}{1 - \alpha_G - \alpha_R} \right], \\ \Pi_R &= \frac{b}{d} \left[\gamma c^s \alpha_R^{d-1} + (1 - \gamma) \frac{1 - (1 - \alpha_R)^d}{\alpha_R} \right].\end{aligned}$$

As in the green outcome equilibrium case, it is no longer the case that committing red is dominated by being flexible, so that an equilibrium is characterized by two thresholds, corresponding to the contenders indifferent between committing green and remaining flexible, and indifferent between committing red and remaining flexible.

A new difficulty arises in the study of the three stage game with hierarchical selection rule when the red outcome equilibrium arises in the third stage. It is no longer necessarily true that Red contenders with cost $c > b$ prefer committing red over remaining flexible. The intuition is as follows. Red contenders with cost $c > b$ play the same strategy in the matching colors games, so their payoffs are identical. But the probability of being selected differs according to their commitment decision. If the selector is Red, the probability of being selected is higher under commitment than under flexibility, whereas the probability is higher under flexibility than under commitment if they meet a flexible Green selector. If the fraction of Green selectors is sufficiently large, and the number of contenders sufficiently low, the probability of meeting a flexible Green selector may exceed the probability of meeting a Red selector, and high cost Red contenders prefer remaining flexible over committing red.⁵

⁵Notice that in that case, even though $c_R^R > b$, we always have $c_G^G \leq b$ – all flexible Green contenders have a cost smaller than b . Hence, the F-F red equilibrium exists and is stable in the third stage of the game.

We analyze the existence of different equilibrium configurations when $d = 2$. As in the green outcome equilibrium case, the zero-cost agent always prefers to remain flexible than commit red. This leaves us with three different regimes for equilibria in the three-stage game:

1. **Regime 1** The zero-cost agent remains flexible. There are two thresholds $c_G \leq b$ and $c_R \leq b$ above which Green (respectively Red) contenders choose to commit green (respectively red)
2. **Regime 2** The zero-cost agent commits green. All Green contenders commit green, and there are two thresholds $0 < c_G < c_R \leq b$ such that Red contenders with costs $c \in [0, c_G]$ commit green, Red contenders with costs $c \in [c_G, c_R]$ remain flexible, and Red contenders with costs $c \in [c_R, 1]$ commit red
3. **Regime 3** The zero-cost agent commits green. All Green contenders commit green, and there is a single threshold c^e such that all Red contenders with cost $c \in [0, c^e]$ commit green, all Red contenders with cost $c \in [c^e, b)$ remain flexible, and Red contenders with cost $c \in [b, 1]$ randomize over committing red and remaining flexible.

The top-left panel shows that, as the fraction of Green agents increases, the equilibrium switches from regime 1 to regime 2 and regime 3. For large majorities of Green agents, Red contenders either commit green or remain flexible (with some but not all high cost Red contenders committing Red). The top-right panel shows that the Green selector is more likely to commit as γ increases, and in fact always commits when the fraction of Green agent goes to 1. As the bottom-left panel shows, Green contenders are also more likely to commit as γ increases, up to the point where the regime switches from regime 1 to regime 2 and all Green contenders, including the zero-cost agents, prefer to commit green. Finally, the bottom-right panel shows that the fraction of Red agents committing green increases with γ , and the expected cost of Red contenders committing red decreases with γ . When γ goes to 1, Red contenders either commit green or remain flexible, and there is no longer any Red contender committing red.

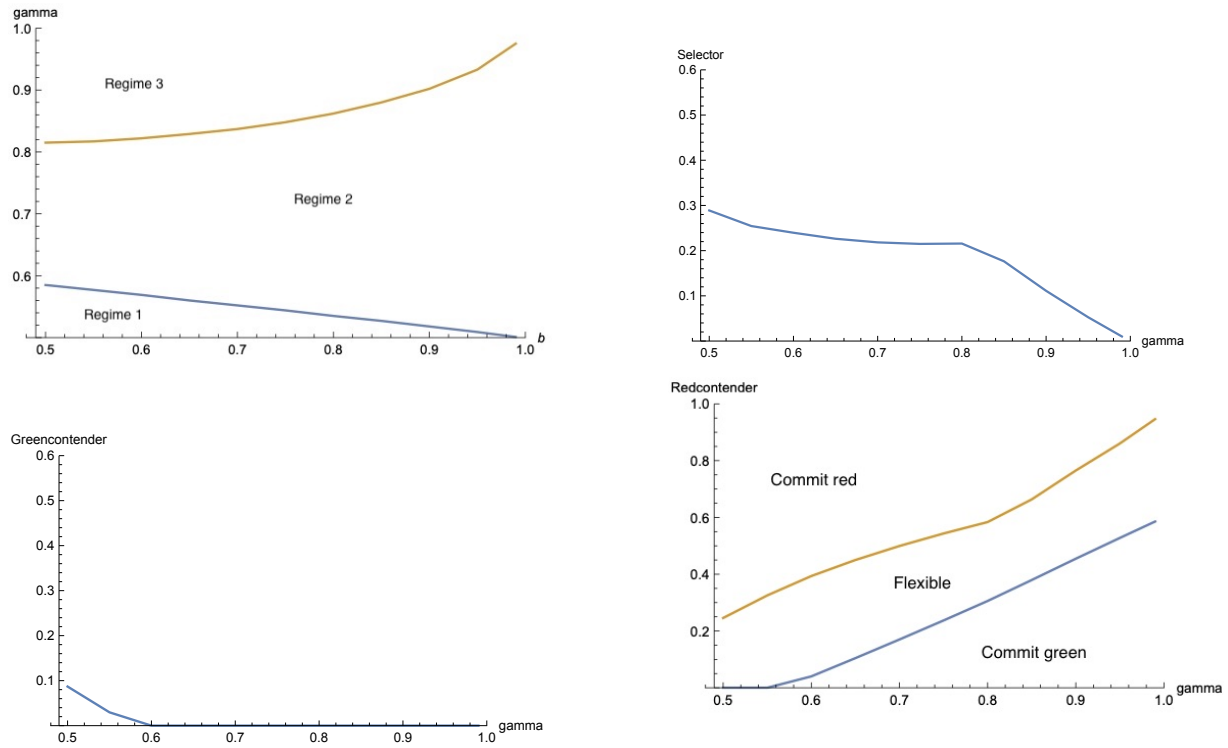


Figure 9: Regimes and threshold values for the selector and the contenders as a function of γ

When d goes to infinity, all Green selectors commit, all Red selectors with cost $c < b$ remain flexible. The incentives of contenders are the same as in the green outcome case. No contender has an incentive to remain flexible, and Green contenders commit green while Red contenders commit red.