Social Conflict and the Evolution of Unequal Conventions

Sung-Ha Hwang*
Suresh Naidu‡ Samuel Bowles†

January 16, 2018

Abstract

We propose a theory of unequal social norms, or conventions, where unequal practices persist over long periods of time despite being inefficient and not supported by formal institutions. We extend the standard asymmetric stochastic evolutionary game model to allow sub population sizes to differ and idiosyncratic rejection of a status quo convention to be intentional to some degree (rather than purely random as in the standard evolutionary models), consistent with historical cases. In this setting, if idiosyncratic play is sufficiently intentional and the subordinate class sufficiently large relative to the elite, then risk-dominated conventions that are both unequal and inefficient relative to alternative conventions can be stochastically stable and will persist for long periods. We show that the same is true in a general bipartite network of the population if most of the subordinate groups interactions are local, while the elite is more "cosmopolitan". We illustrate the model with a number of historical transitions.
JEL codes: D02 (Institutions), D3 (Distribution), C73 (Stochastic and Dynamic Games; Evolutionary Games)
Keywords: Conventions, evolution, stochastic stability, collective action, inequality

*Korean Advanced Institute of Science and Technology(KAIST), sungha@kaist.ac.kr. ‡Corresponding Author, Columbia University and NBER sn2430@columbia.edu. †Sante Fe Institute and the University of Siena, samuel.bowles@gmail.com. We also thank Robert Axtell, Ellora Derenoncourt, Jeff Jacobs, Oded Galor, Bentley Macleod, Tone Ognedal, Elisabeth Wood, Peyton Young, and seminar participants at Brown, MIT, CIFAR, and UC Berkeley for comments and the Behavioral Sciences Program of the Santa Fe Institute, National Research Foundation of Korea (NRF-2016S1A3A2924944), and the Social Sciences and Humanities Research Council of Canada, for support of this project.

1 Introduction

Unequal social norms have been ubiquitous over long periods of history. Systems of caste and racial exclusion, conventions governing labor markets, norms regulating relations between men and women, and linguistic practices such as honorific pronouns have emerged and persisted in highly varied environments, and appear to be favored in the historical dynamics by which social structures evolve. Unequal social conventions appear to be what the sociologist Talcott Parsons termed an "evolutionary universal" (Parsons, 1964). Two classes of reasons are often offered to explain why this might be so.

First, unequal social norms may contribute to the evolutionary success of individuals or groups that adopt them under the relevant technological, biological, or environmental constraints. For example, Becker (1981) argued that gender norms around the household division of labor may have been historically adapted to reflect comparative advantage. However, given the abundant evidence of adverse effects of historically durable cultural norms on economic development,[1] it is difficult to believe that efficiency considerations provide an adequate explanation.

Second, even in the absence of any resulting efficiency advantage, the persistence of unequal social norms may simply be the result of unequal laws and institutions, reflecting the political power and coordinated collective action of elites who benefit from the norms (Acemoglu and Robinson, 2008). But the very long term persistence of many unequal norms and institutions – crop sharing conventions (Young and Burke, 2001)and linguistic markers of superior and subordinate status (Clyne et al., 2009), for example – cannot adequately be explained as the result of deliberate *de jure* interventions from above by elites. The substantailly uncoordinated emergence of novel social norms are often harbingers of major changes in social inequality appearing well before being institutionalized formally in law or policy.

Here we suggest a third (possibly complementary) mechanism, showing that unequal social arrangements may emerge and persist over long periods *without* the intervention of elites, and in cases in which alternative institutions would be more efficient. Like the top

[1] See (Edgerton, 1992) for numerous examples of maladaptive practices in small-scale societies and (Nunn, 2014) for a review of the economics literature.

down approach, our model identifies conditions under which inefficient economic institutions that implement high levels of inequality persist in the long run. But we do not posit aggregate representative agents bargaining over institutions, and neither commitment problems nor concentration of political power in elite hands play any role in our approach.

Instead we study the way that customary and expected patterns of behavior, including asymmetric norms governing relations between racial groups, genders, and economic classes, can constitute a decentralized mechanism by which inequality can be implemented and perpetuated. These unwritten rules structure regular, asymmetric interactions, endure for long periods of time despite not being codified in formal laws and regulations and not conveying any efficiency advantages to individuals or groups. Further examples include community standards for female labor force participation (Alesina et al., 2011), customs governing inheritance (Goody et al., 1976), conventions of wage-setting (Bewley, 1999), informal economic norms of property (Silbey, 2010) and legal contracting (Gulati and Scott, 2012). In an empirical application of the bottom up approach developed here (Naidu et al., 2017) we study the evolution of linguistic conventions (illustrated by the asymmetrical use of "vous", "tu" and other pronouns), building on classic work in the sociolinguistics of inequality (Brown and Gilman, 1960).

In our model unequal informal conventions persist because they are stable Nash equilibria in an evolutionary asymmetric coordination game, adherence to which is a best response for members of both the privileged and the subordinate members of a society as long as most others do the same. Transitions to less unequal conventions may occur as a result of the chance bunching of challenges to the status quo by a sufficiently large fraction of the subordinate class. These new conventions then endure, as for example, in the case of fair wage norms: Hanes (1993) shows that nominal wage rigidity during the 1893 downturn was higher in U.S. cities that experienced larger strike waves in the mid-1880s. But because strikes and other acts of defiance are rare, we can show that transitions to more equal conventions are less likely to occur when the less well off class is more numerous than the elite, which empirically is often the case.

Evolutionary game theory has been used extensively for modeling and selecting among equilibrium conventions that emerge from decentralized interactions in a large network of

low-rationality agents (Foster and Young, 1990; Kandori et al., 1993). It is particularly well suited to model the evolution of norms, culture, and traditional informal contracts that are not maintained by the state or explicitly negotiated among a few collective actors, but instead emerge at the population level. In these models the long-run stability of a convention is determined by which equilibrium is more robust in the face of possible disruption due to idiosyncratic non-best-response play. This can depend on state-dependent idiosyncratic play as in Bergin and Lipman (1996), on network topology as in Young (2011) and Kreindler and Young (2011), and other local interactions, as in Ellison (1993), all of which not only provide a empirically grounded dynamic, but also accelerate the otherwise implausibly long waiting time for transitions between conventions.

But asymmetric social conventions may be better modelled by an enriched evolutionary dynamic more consistent with the history of stasis and transformation of unequal conventions. Our extension to the standard model concerns the representation of the idiosyncratic play that generates transitions among conventions. In the standard model deviance from the best response is analogous to mutation in a population genetics model that is, an undirected chance event.

For many, perhaps most, applications to historical transitions we find this formulation problematic, for two reasons. First, as we will see below, the implied process of change—which actors' deviant actions induce a transition—is empirically implausible. Second, historical transitions between conventions have been driven not so much by mistakes but instead by what might be termed intentional idiosyncratic play, namely actions that are not a best response to the status quo but that would benefit the actor if sufficiently many of the members of his class did the same.

The kinds of deviance that account for bottom up transformations of an unequal convention, and that we would like to model include Rosa Parks' refusal to give up her seat at the front of the bus, the Soweto school children who boycotted classes in protest of apartheid, and the 14th century British farmers who simply refused to perform the labor duties owed to their lord. These actions are idiosyncratic not because they are literally random but because they occur for reasons outside the model. We are not attempting to model the complex process of collective action in social movements, instead we examine what conventions are

stochastically stable when the process generating transitions between conventions resemble social movements.

We do not explore these reasons—intentional idiosyncratic play is a primitive of our modeling strategy—but they include outrage, a quest for personal dignity through opposition to injustice and other motives not necessarily tied to the objective of inducing a transition (Wood, 2003). The intentionality we introduce is simply directional: the 14th century farmers do not deviate by insisting on providing more than the conventional labor services to the lord.

To introduce this minimalist conception of intentionality (extending our in Bowles (2004) and Naidu, Hwang, and Bowles (2010)), we generalize the noise process accounting for idiosyncratic play to allow for a variable degree of intentionality ranging from entirely unintentional (as in the standard model) to entirely intentional, in which case an agent never plays idiosyncratically when their preferred convention is the status quo. This relatively simple extension of the standard framework increases the applicability of evolutionary models to many more historical examples of endogenous cultural change.

Taking account of the intentional nature of idiosyncratic play allows us to address what appears to be a restrictive and possibly anomalous aspect of the standard unintentional noise-like treatment of idiosyncratic play. In the standard case, transitions from unequal to equal conventions are always driven by the cumulative non best response play of the better-off group, who would collectively lose from the transition, not by the deviance from the unequal norm by the poor, who would benefit (Bowles, 2004). While one can imagine a transition resulting, for example, from landlords idiosyncratically asking their tenants to pay less than the conventional crop share, we think that the empirically more relevant case would be when the tenants idiosyncratically insist on paying less than the norm. In our model, transitions arise from cumulative deviations of the disadvantaged group, more in line with historical experience, as our examples will illustrate.

Two further apparent anomalies follow as a consequence: larger populations, and those with lower rates of non-best-response play are favored in this dynamic. The reason in both cases is that large population size and less frequent deviance means that a population's own idiosyncratic play is less likely to induce a transition which, if it occurred, would be

away from their favored convention. The conventional approach for modelling transitions seem inappropriate in contexts where agents understand their group interest, so that their idiosyncratic play is likely to be limited to those strategies that would yield a higher payoff were a sufficiently large number of others to do the same.

Introducing what we term "directed idiosyncratic play" alters the resulting dynamic and the equilibrium selection processes in bargaining games. In an asymmetric coordination game between members of different groups bargaining over conventional contracts (e.g. segregation norms or crop shares), Young (1998b) shows that conventional contracts that are selected in an evolutionary dynamic subject to small random shocks implements the Kalai-Smorodinsky solution. Exploiting the properties of risk-dominant payoffs in coordination games, his contract theorem has the striking implication that evolution favors conventions that are not only efficient, but also have the Rawlsian property that the payoffs of the least well off are maximized, relative to the maximum they could get in any of the feasible set of contracts.

In contrast, our main results give conditions for the stochastic stability of risk-dominated conventions, which can be both unequal and inefficient (in the sense of a lower joint surplus) relative to the alternate convention. Beginning with the random matching environment that is standard in this literature, Proposition 1 shows that when idiosyncratic play is sufficiently intentional and the relative size of the less well off group is sufficiently great, a risk-dominated convention will be stochastically stable: Unequal social norms can persist (in competition with less unequal conventions) without political inequality and without efficiency advantages. We then show that this result remains true under the uniform local matching protocol studied in Ellison (1993).

Building on the link between risk-dominance and the equity-efficiency result in Young (1998b), this shows that if the poor class has either a low idiosyncratic play rate or is relatively large, then conventions that are unequal and inefficient can persist for extremely long periods. More precisely, Proposition 1 shows that when idiosyncratic play is sufficiently intentional and group sizes and idiosyncratic play rates are sufficiently asymmetric, the risk-dominated convention is stochastically stable in a standard environment with either uniform or local random matching.

Finally, we provide a framework that can account for the strongly networked aspects of social movements which are prominent in many historical unequal-to-equal transitions(McAdam, 1986). We provide conditions on an interaction network under which risk-dominated conventions are stochastically stable when play is intentional and relative population size and rates of idiosyncratic play are highly asymmetric. Jackson and Watts (2002) show that network structure can drastically alter which states are stochastically stable, while Young (2011) shows that network topology determines the speed of adoption of the efficient strategy in a symmetric coordination game on networks. We generalize Young (2011) and the concept of network "cohesiveness"(Morris, 2000) to asymmetric coordination games.

2 Equilibrium Selection in Contract Games

2.1 Contracts

Our model is deliberately both abstract and simple. We consider distributional conventions such as feudal obligations between lords and serfs, segregation in the U.S. South, gender division of labor, and linguistic modes of address as strategies in a 2×2 game. What distinguishes conventions from simple divisions of output is that the payoff to non-cooperation is 0, as we think of the strategies as themselves "rules of the game", about which there must be agreement in order for any productive coordination to occur. Mis-coordination results in no productive joint activity. While one can imagine different groups facing different off-diagonal payoffs (benefits and costs of mismatch, for example due to violent enforcement of norms or government sanctions), we follow the literature on evolution in coordination games and leave a study of such asymmetry to future work. While extremely simple, this setup allows us to focus on the structure of idiosyncratic play and population interactions rather than traditional economic trade-offs. We will call these distributional conventions "contracts", which must be agreed upon by both "rich" and "poor" for any production to occur. This could be the norm around bus seating (as in the U.S. South), the number of days pledged in feudal obligations, or the distribution of household labor between men and women.

We consider a large population consisting of two groups whose members myopically play

	U	E
U	a_P, a_R	0,0
E	0,0	b, b

Table 1: **Payoffs in the Contract Game.** P gets the row payoffs, R gets the column payoffs. Note that because $a_P < b < a_R$ so Rs strictly prefer the unequal contract (U, U) while Ps strictly prefer the equal contract (E, E).

an asymmetric 2×2 contract game with 2 pure-strategy Nash equilibria. Both contract equilibria are Pareto-optimal but one has higher total payoff than the other. They differ as well in the distribution of the surplus that they implement. To illustrate the kinds of contracts among which decentralized selection may take place, suppose the Rs (column players in our game matrix) are landowners, men, whites, or employers while Ps (row players) are tenants, women, African-Americans, or workers. Contract E is a relatively equal sharecropping or profit-sharing contract, yielding payoff b to both poor and rich agents, while Contract U is an unequal fixed rental or wage contract, yielding payoffs a_P and a_R to the poor and the rich agents, respectively. We furthermore assume that $a_P < b < a_R$, which means that Rs prefer the unequal contract U while Ps prefer the equal contract E. The risk-potential of each contract is given by the product of the payoffs, so U is risk-dominant if $a_R a_P > b^2$. We can represent the payoffs from a contract as a 2×2 game matrix, as in Table 1, where $\pi_P(i, j)$ (or $\pi_R(i, j)$) is the payoff to a poor agent (or a rich agent) when the poor agent plays strategy i against the rich's strategy j.

The US Southern Jim Crow convention, which we return to below, would have U being the convention of unequal bus seating or modes of address, with whites playing U by sitting at the front or addressing black men as "boy", and blacks playing U by sitting at the back or addressing white men as "Sir". The E convention could be that seating would be allocated on first arrival. So if a white respected the norm of first arrival while the society was in the U equilibrium, it would prevent the bus from moving as readily as a black man sitting on the seat he encountered first, as described by the 0 off-diagonal payoffs.

2.2 Best-Response Dynamics with Intentional Deviations

To focus on main results and intuitions, we first introduce a simple convention selection dynamic for the contract game in Table 2.1, which we call a uniform interaction model (see also Section 4 for the local interaction model on bipartite networks). Suppose the rich population has size N^R and the poor population has size N^P.

We parameterize the *effective relative population size* with η, where $N^P = \eta \times N^R$ and ηN^R is assumed to be a positive integer for generic η. η captures both the relative population size of the poor as well as the possibly lower rate of non-best-response play. As we show in Naidu, Hwang, and Bowles (2010), population size and rate of idiosyncratic play are functionally equivalent from the perspective of stochastic stability. A small relative population of poor agents generates a high probability of a sufficiently large fraction deviating at the same time, inducing a transition, as does a fast rate of idiosyncratic play.

A state is given by a number for each population (X, Y), where X (or Y) is the number of poor (or rich) agents using strategy U. We assume that agents from each population are randomly matched to play the game, so that every agent in each population has an equal probability of interacting with every agent in the other population each round. For example, the bus segregation convention would be the convention where the P class gives up their place to a R agent, and so the payoffs to the P agent are lower than in the equal (unsegregated) convention E, which are in turn lower than the payoffs from the R agents in the segregated convention. Thus, the expected payoffs to a poor agent (U_P) and to a rich agent (U_R) using strategy i are given by

$$U_P(i, Y) = \pi_P(i, U)Y + \pi_P(i, E)(N^R - Y) \text{ and } U_R(j, X) = \pi_R(U, j)X + \pi_R(E, j)(\eta N^R - X)$$

respectively. The (myopic) best responses for agents, which we assume are uniquely determined, are denoted by $B_P(Y)$ and $B_R(X)$ and given by:

$$B_P(Y) := \arg \max_{i \in \{U,E\}} U_P(i, Y), \text{ and } B_R(X) := \arg \max_{j \in \{U,E\}} U_R(j, X)$$

Our primary theoretical contribution in this paper is to explore the implications of idiosyn-

cratic play that is "intentional" in a sense we make precise in this section. Social conventions may be upset by activists who deviate from the population-level best-response in a directed fashion namely by taking an action that would benefit all members of their group were sufficiently many to deviate in the same way. This restriction that we impose on noisy behaviors leads us to call them collective action shocks, despite consisting of independent draws across individuals.

We think of these as forms of decentralized social conflict (Scott, 1985) where one actor incurs a cost by playing a strategy which is not a best-response, but would yield a higher payoff were it to become an equilibrium. Thus we consider our collective action shocks a reduced form way of incorporating activities such as strikes and lockouts, legal prosecutions, land invasions and evictions. The non-best-response play can be considered collective action because it uses the strategy that would yield a higher payoff for the group, were all agents to play it, but the shocks are independent across individuals[2], so it is only when a large enough set of agents is simultaneously perturbed that the equilibrium changes.

We interpret the intentional ϵ as the probability of engaging in collective action by playing the strategy that would be best for that class were it to be played by both classes in equilibrium. We describe the stochastic process more fully and apply it to a more general class of bargaining games in Naidu, Hwang, and Bowles (2010). While unintentional idiosyncratic play is plausible for many changes in conventions, as we will show, it has some restrictive properties that make it ill-suited as a general model of transitions between equilibria. We thus generalize this to allow errors to be parameterized by a degree of intentionality ι. $\iota = 1$ will correspond to the standard unintentional idiosyncratic play structure. When ι is large, however, agents are much more likely to systematically play strategies that would yield larger payoff were they to become Nash equilibria. As we are interested in the long-run equilibrium selection implications of this historically plausible modification to the evolutionary model, we do not present detailed microfoundations for this intentionality here.[3]

[2]In reality, collective action shocks are likely to be at least weakly correlated across individuals, but we abstract from that in this part of the paper. Our model illustrates the dynamics of social change *without* a coordination mechanism, and thus serves as a benchmark model for extensions which do try to model more coordinated social action.

[3]Parallelling the analysis in van Damme and Weibull (2002), one could posit the existence of leadership or organizations for both populations, who would like to have all their members play their most preferred contract, but have control costs when trying to induce agents to deviate from their best-responses. Then the

Our specification of idiosyncratic play is also consistent with recent experimental work. Lim and Neary (2016) find that individual mistakes depend on the myopic best-response payoff and are directed in the sense of being group-dependent. The directed mistakes in their paper are intentional idiosyncratic behaviors of deviant agents and, for example, they find that 2.25% of subjects play mistakes when the best response is the preferred strategy, while 20.85% of subjects play mistakes when the best response is the less preferred strategy (Figure 5 (a) on pp. 19). In a lab experiment with evolutionary games, Mäs and Nax (2016, pp. 204) similarly find that the vast majority of decisions (96%) are myopic best responses, but deviations are sensitive to their costs. Specifically, they report that "deviation rates were significantly lower when subjects faced a decision where the MBR (myopic best response) was the subjects preferred option, which lends support to the assumption that deviations can be directed" (see also Hwang et al. (2018)).

To capture these intentional collective action shocks in a reduced-form way, we suppose that at each period, one agent is drawn from either the rich population or the poor population and this drawn agent has a chance to demand a new contract with the following probability:

$$P\text{ agent chooses }\begin{cases} B_P(Y) & \text{with probability } 1-\epsilon-\epsilon^\iota \\ U & \text{with probability } \epsilon^\iota \\ E & \text{with probability } \epsilon \end{cases} \qquad R\text{ agent chooses }\begin{cases} B_R(X) & \text{with probability } 1-\epsilon-\epsilon^\iota \\ U & \text{with probability } \epsilon \\ E & \text{with probability } \epsilon^\iota \end{cases} \tag{1}$$

where $0 := \epsilon^\infty$ and ι is the parameter characterizing the degree of intentionality as follows. First observe that when $\iota = 1$, the mistake model in 1 gives the "uniform mistake model" in which every mistake is equally likely—one of most popular mistake models in the standard stochastic evolutionary game theory models. Thus agents make mistakes independently of their preferences over contracts and we call this the unintentional model. Our formulation of the perturbations in (1), when $\iota > 1$, is the key difference between how noise is generated in our model and the standard stochastic evolutionary game theory models.[4]

idiosyncratic play probabilities are the optimal choices of the organizations/leaders subject to these control costs.

[4]Even though we use a specific parametrization of ϵ, a more flexible formulation is straightforward. For example, we can define the probability that P agents choose U (or R agents choose E) to be $\phi(\epsilon)$ such that $\epsilon\phi'(\epsilon)/\phi(\epsilon) \to \iota$ as $\epsilon \to 0$, with ι being interpreted the "elasticity" of intentionality with respect to the rate of idiosyncratic play ϵ as ϵ approaches 0.

2.3 Equilibrium Selection under Random Matching

To build intuition and as a benchmark case, we begin by determining which Nash equilibrium is stochastically stable when agents interact randomly with all members of the opposing population. To do this we call the state in which every agent plays U (or E) the U convention (or the E convention). These states are precisely the absorbing states in the unperturbed process. We show that the Markov chain defined by (1) admits a unique invariant measure in a more general setting in Lemma 1 below. The stochastically stable state is the state which has a positive mass of the invariant measure at $\epsilon \to 0$. Stochastic stability thus can be studied by determining the number of idiosyncratic players it takes to upset each convention (Young, 1993a; Kandori et al., 1993; Young, 1998b). In the literature, these costs of transition, also called *resistances*, govern the speed of transitions from one convention to the other.

While we provide formal definitions below, here we describe the general intuition. First suppose that the status quo convention is U, the unequal convention that favors the Rs. If sufficiently many idiosyncratically playing Ps demand contract E rather than the status quo contract U, best responding Rs will switch to offering contract E in the subsequent period. By letting p be the fraction of idiosyncratic players in the P class, and equating Rs' expected payoffs from sticking with the U contract or shifting to the E contract, we see that the R class agents will play E as their best responses if $(1-p)a_R < pb$.

Thus, the lowest fraction of deviant Ps sufficient to induce a transition is $p^* = \frac{a_R}{a_R+b}$. As a result, the minimum number of Ps deviating from the status quo to induce a switch from contract U to contract E is given by the first equation in the right hand side of (2). Unlike this resistance, the corresponding resistance for a R-induced transition from the U contract to the E contract (the second equation in (2)) is determined by two factors: the minimum number of rich agents that can induce a transition and the degree of intentionality of the rich's idiosyncratic behaviors. The critical number of the rich agents can be similarly computed as before and the more intentional the deviant behaviors are, the higher the resistance (or cost), and the more reluctant rich agents are to induce a transition from U. The terms in equation (3) can be interpreted similarly (Section 4 derives these costs more rigorously).

$$c(U,E) = \min(\underbrace{\left\lceil \eta N^R \frac{a_R}{a_R+b} \right\rceil}_{\text{poor}}, \iota \underbrace{\left\lceil N^R \frac{a_P}{a_P+b} \right\rceil}_{\text{rich}}) \tag{2}$$

$$c(E,U) = \min(\underbrace{\left\lceil N^R \frac{b}{b+a_P} \right\rceil}_{\text{rich}}, \iota \underbrace{\left\lceil \eta N^R \frac{b}{b+a_R} \right\rceil}_{\text{poor}}) \tag{3}$$

where $\lceil t \rceil$ is the lowest integer that is greater than or equal to t. Intuitively, the resistances (or the costs) measure the size of the basin of attraction of each convention and the convention with a larger resistance is harder to escape. Thus, the stochastically stable state is the convention i in which $c(i,j) > c(j,i)$, a convention which requires more non-best-response play to escape (see Young (1998a) and more rigorous discussion in Section 4).

Equations (2) and (3) can be used to define a locus of ι and η such that both populations are equally likely to induce a transition from U to E and E to U, respectively. That is, Locus 1 (or Locus 2) in Figure 1 is the set of all ι and η equating the two minimands in equation (2) (or equation (3)). We plot these loci in Figure 1, and designate the areas under which U and E are stochastically stable.

The intuition behind Locus 1's upward slope is that an increase in η makes the probability of the poor inducing a transition from U to E less likely, because there will be a smaller probability of enough idiosyncratic play to induce a transition, but an increase in ι makes the rich also less likely to induce a transition from U to E, because the rich will be less likely to deviate to the egalitarian convention that would hurt them were it to become an equilibrium. The two parameters affect the different populations' probability of inducing a transition independently, and thus do not interact, hence Locus 1 is a straight line.

The intuition behind Locus 2's downward slope is that an increase in η makes the likelihood of the poor inducing a transition from E to U less likely, again because of the smaller probabilty of sufficient idiosyncratic play to induce a transition. However in this case an increase in ι makes the probability of the rich inducing a transition *more* likely, as higher ι implies that the poor are innovating less and the rich are innovating more. In this case, ι and η complement each other, as the larger poor population size magnifies the impact of

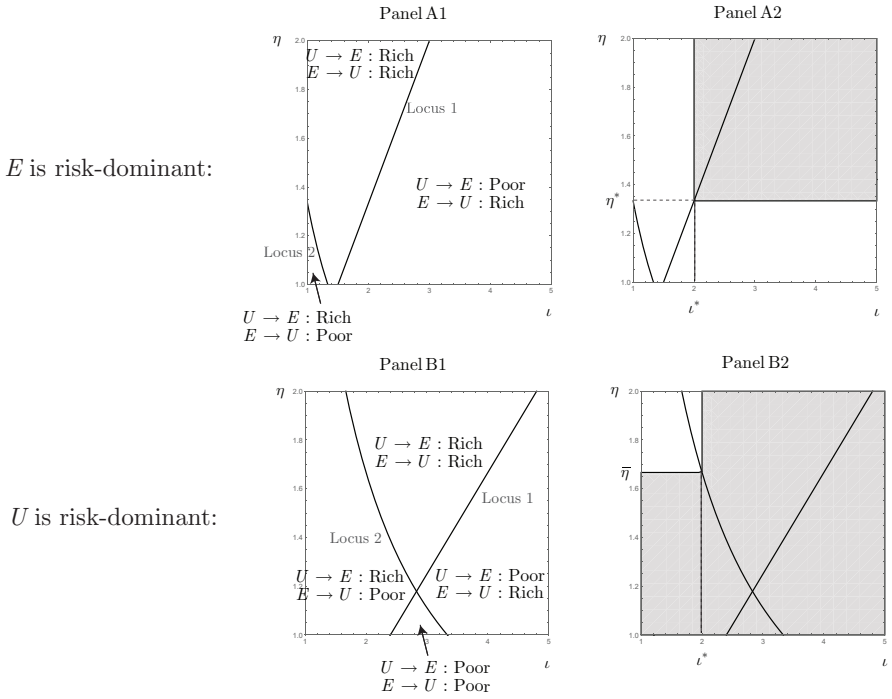

Figure 1: Panel A1 and A2 show the case where E is risk dominant. Panel A1 shows the combinations of ι and η under which each of the two populations is driving the transition. The shaded area in panel A2 shows the combination of ι and η under which U is stochastically stable. Panel B1 and B2 show the case where U is risk dominant, with Panel B1 showing again the combinations of ι and η under which the idiosyncratic play of the rich or the poor drive transitions. The shaded area in Panel B2 shows the combinations of ι and η where U is stochastically stable.

intentionality in reducing the likelihood of a transition by the poor. Hence, if transitions become more intentional, transitions driven by the rich from E to U become more likely, and so a smaller η is required to increase the odds of a transition driven by the rich in order to keep the probability of a transition driven by the two groups equal.

These loci and equations (2) and (3) show how transitions are induced by the idiosyncratic play of the group for which the least number are required to induce the best responders in the other group to switch strategies. If $\iota = 1$ and $\eta = 1$, as in the standard evolutionary models, resistances that drive transitions are identified by letting the degree of *unintentional* idiosyncratic behavior become arbitrarily small, and in contract games this results in idiosyncratic play of the losing population driving the transitions (Binmore et al., 2003).

Inspection of (2) shows that for large N^R we have $c(U,E) \approx \frac{a_P}{a_P+b}$, so that it is the resistance of the poor that must be overcome by the idiosyncratic play of rich in order to transition from the unequal contract to the equal contract (where the rich do worse). Similarly (3) shows that $c(E,U) \approx \frac{b}{a_R+b}$ so that it is the resistance of the rich that must be overcome by idiosyncratic play of the poor in order to transition from the pro-poor E contract to the pro-rich U contract.

This counterintuitive feature of transitions in the standard model motivates allowing for the varying degree of intentionality in idiosyncratic behaviors to account for historical and empirical plausibility. First, when ι is sufficiently large with the population size still being equal ($\eta = 1$), resistances are the least number of *intentional* idiosyncratic deviations required to induce a transition by those who would benefit from the transition occur (see the left panels in Figure 1). Thus as idiosyncratic behaviors become more intentional, transitions are more likely induced by those who stand to benefit. In the extreme case where $\iota = \infty$, the transition from the unequal convention is always induced by the poor with resistance $\left\lceil \eta N^R \frac{a_R}{a_R+b} \right\rceil$, while the transition from the equal convention is always induced by the rich with resistance $\left\lceil N^R \frac{b}{b+a_P} \right\rceil$. In this case, when the poor are more numerous, more idiosyncratic poor players are required to escape convention U, thus convention U is harder to escape and is stochastically stable. The numerous poor are at a disadvantage in this case.

Second, when the size of the poor is large or the poor have low rates of idiosyncratic play, the poor are less effective in transitions and all transitions are induced by the rich (observe that the regions of $(U \to E : \text{Rich})$ and $(E \to U : \text{Rich})$ enlarge as η increases in the left panels of Figure 1). That is, the larger a group, the less effective it is in triggering a transition and the smaller the group, the more likely this group is able to generate transitions. Thus, in the extreme case of $\eta = \infty$, all transitions are induced by the rich with the resistances from conventions U and E being $\iota \left\lceil N^R \frac{a_P}{a_P+b} \right\rceil$ and $\left\lceil N^R \frac{b}{b+a_P} \right\rceil$, respectively. If the rich's idiosyncratic behaviors are sufficiently intentional, the rich agents deviate from convention U less frequently (with the probability ϵ^ι) and thus it becomes more difficult to escape from convention U, and U again becomes stochastically stable.

The proposition below formalizes this discussion and presents the sufficient and necessary conditions for the stochastic stability of each convention in both the case where E is risk-

dominant and where U is risk-dominant. We assume that N^R is sufficiently large, ignoring integer problems.

Proposition 1. *There exist $\iota^* > 1, \bar{\eta} > 1$ and $\eta^* > 1$.*
(i) Suppose that E is risk-dominant. Then U is stochastically stable if and only if $\iota > \iota^*$ and $\eta > \eta^*$.
(ii) Suppose that U is risk-dominant. Then E is stochastically stable if and only if $\iota < \iota^*$ and $\eta > \bar{\eta}$.

Proof. The proof consists in checking whether equation (2) is greater than or less than (3) under the two cases. We present the details in the Appendix. □

In sum, the intentionality determines the direction of a group's idiosyncratic play rate while the relative group size determines its relative speed. Intentional idiosyncratic play of R pushes towards U, and a large η means that those transitions are relatively quite rapid, compared to transitions driven by the idiosyncratic play of P pushing toward E. When both the degree of intentionality and the relative group size are sufficiently high, these two effects jointly contribute to the long-run persistence of the unequal convention. Indeed Figure 1 shows that when ι and η are large, U is stochastically stable if

$$c(U,E) = \min\left\{\left\lceil \eta N^R \frac{a_R}{a_R+b}\right\rceil, \iota\left\lceil N^R\frac{a_P}{a_P+b}\right\rceil\right\} > \left\lceil N^R\frac{b}{b+a_P}\right\rceil = c(E,U) \tag{4}$$

irrespective of risk dominance, so long as $\iota > \iota^*$ and $\eta > \eta^*$. This result is unlike the literature, for example Young (1993b), who shows that only risk-dominant conventions are stochastically stable in 2×2 games.

Inequality (4) is at the heart of our model. Our extensions below will modify the inequality so that properties of the interaction structure, rather than N_R, define the conditions under which E or U is stochastically stable.

We can also investigate how the level of equality and efficiency of a contract, together with the population structure, affects the persistence of the associated convention. To do this, we parameterize the payoffs in the U contract with ρ and θ, which measure the joint surplus and inequality under the unequal contract, respectively. More precisely, the unequal

contract U gives joint surplus $\rho := a_P + a_R$, of which a share θ goes to the Ps and the remainder $(1-\theta)$ goes to the Rs, where $0 \leq \theta \leq 1/2$. We set $b = 1$ for simplicity: i.e.,

$$\frac{b}{b+a_P} = \frac{1}{1+\theta\rho}, \quad \frac{b}{b+a_R} = \frac{1}{1+(1-\theta)\rho}$$

In this case risk-dominance of U is equivalent to $\rho^2(1-\theta)\theta > 1$. Clearly contracts with sufficiently high surplus ρ and low inequality (higher θ) will have higher risk-potential than the E contract, and thus be stochastically stable in the traditional model. This relationship between stochastic stability, risk-dominance, and efficiency-equity is behind Young's contract theorem discussed in the introduction.

It is simple to check that if $\eta = 1$ and $\iota = 1$—that is, the classes are equally numerous and idiosyncratic play is unintentional—the stochastically stable state (for all ι) is risk-dominant. In the 2×2 contract game, this will be the contract that maximizes the product of the payoffs of the two classes, namely $\rho^2(1-\theta)\theta$ for convention U and 1 by assumption for convention E. Thus, if $\rho^2(1-\theta)\theta > 1$ then $c(U,E) > c(E,U)$, and U will be selected, otherwise E is selected (and both are stochastically stable in the case of a tie). We can also see that increased inequality in the division of the surplus will destabilize the unequal contract. Greater inequality (lower θ since $\theta < \frac{1}{2}$) in the unequal contract decreases both resistances in (2) but it decreases $c(E,U)$ less than it decreases $c(U,E)$, lowering the relative probability of a transition from E to U. But while inequality lowers the risk-dominance of a contract, efficiency, as measured by a higher ρ, raises it. This is the sense in which the stochastically stable equilibrium implements equal and efficient outcomes. But when ι and η are sufficiently large, this equivalence between stochastic stability, risk-dominance, and equal and efficient outcomes is broken: stochastically stable contracts can be risk-dominated, and thus inefficient as well as inegalitarian.

Ignoring integer considerations and setting $c(U,E) = c(E,U)$ from (2) and (3) allows us to determine the levels of efficiency and inequality of alternative contracts (or, equivalently, the risk-potential) such that the population would spend approximately half of the time at the unequal and half the time at the equal contract. In order to look at this simple case, define $\eta^{*,\infty}$ to be the critical fraction η equating the two resistances, $c(U,E) = c(E,U)$, when

$\iota = \infty$ and therefore satisfying

$$\eta^{*,\infty}(\rho, \theta) = \frac{1 + (1-\theta)\rho}{(1-\theta)\theta\rho^2 + (1-\theta)\rho}. \tag{5}$$

This expression yields informative comparative statics: If $\rho < 2$, so that the unequal convention is less efficient than the equal convention, then $\eta^{*,\infty}(\rho, \theta) > 1$ and the unequal convention is risk-dominated (since $(1-\theta)\theta\rho^2 < 1$). We also have $\frac{d\eta^{*,\infty}}{d\rho} < 0$. Thus, inefficient conventions that are also unequal and thus risk-dominated (i.e. those with $\theta < \frac{1}{2}$ and sufficiently low ρ) require strictly larger populations of poor agents to be stochastically stable with intentional dynamics.

The reason is not the free-rider logic inspired by Olson (1965); nor is it related to the fact that increased supply of a factor of production may disadvantage its owners in markets. Rather, the advantage of small size arises simply because smaller groups are more likely to experience realizations of collective action (that is, simultaneous deviations from the status quo contract) large enough to induce a transition.

3 Interactions in Bipartite Networks

In this section, we show that our main result, that intentional idiosyncratic play together with a relatively larger poor population stabilizes inefficient and unequal (risk-dominated) conventions, holds under more general assumptions on interaction structure.

The random matching model is a useful benchmark, which clearly illuminates the main mechanism for the persistence of the unequal convention, but it assumes that every agent in one population is matched with all agents in the other population and vice versa. But interactions between agents in the situations we are modelling are inherently local because agents are separated by spatial, institutional, and cultural distances. In addition, the assumption of interactions with all agents in the population implies that the expected waiting time for a transition is enormous when the size of the total population is large (Ellison, 1993). This is because when an agent responds to the whole population distribution, the threshold number of deviant agents necessary to induce transitions is of the same order as population size and the expected waiting time increases exponentially in population size.

For these reasons, in our model of two classes, it is natural to consider a bipartite network in which each population occupies one of two sets of vertices of a graph and agents in one group have a limited number of interactions with agents in the other group.

3.1 Simple bipartite networks

We begin with the simplest possible bipartite network, which is analogous to a circle model as in Ellison (1993) (see Figure 2). We call this bipartite network a 1-d bipartite network. Recall that a graph Λ consists of a set of vertices and a set of edges. A graph Λ is bipartite if its vertex set can be partitioned into two sets, Λ_P and Λ_R, such that every edge of Λ is incident to one vertex in Λ_P and one vertex in Λ_R. Suppose that the poor agents and the rich agents are located in Λ_P and Λ_R, respectively. To accommodate the asymmetry in the sizes of two populations, we suppose that a rich agent interacts with a subset of poor agents—named the "village" of poor agents (see the dotted circles in Figure 2). The rich agents and the poor villages, each of which consists of η poor agents, together form the vertices of Λ_R and Λ_P, respectively. Here, we assume that every poor village has the same number of poor agents, and hold the rate of non-best response play constant across rich and poor, but our results still hold for the case where poor villages consist of different numbers of poor agents with slight modification. We write $\sigma_1(x)$, $\sigma_2(x) \cdots$, $\sigma_\eta(x) \in \{U, E\} =: S$ as the strategies of the individual poor agents within the village $x \in \Lambda_P$ and $\sigma(y)$ be the strategy of a rich agent $y \in \Lambda_R$. We also write

$$\sigma(x) = (\sigma_1(x), \sigma_2(x), \cdots, \sigma_\eta(x)) \text{ for all } x \in \Lambda_P \text{ and } \sigma = (\{\sigma(x)\}_{x\in\Lambda_P}, \{\sigma(y)\}_{y\in\Lambda_R})$$

and call σ a population state. Let Ξ be the set of all possible population states.

We denote by $N_x \subset \Lambda_R$ the neighbors of the poor village at $x \in \Lambda_P$ and by $N_y \subset \Lambda_P$ the neighbors of the rich agent at $y \in \Lambda_R$. Note that in bipartite networks, N_x (for $x \in \Lambda_P$) consists of rich agents, while N_y (for $y \in \Lambda_R$) consists of poor villages. Also if x is a neighbor of y, then y is also a neighbor of x; i.e., $x \in N_y \iff y \in N_x$. The payoffs to a poor agent at village x demanding contract i at state σ, $u_P(x, i, \sigma)$, and to a rich agent at site y demanding contract j at state σ, $u_R(y, j, \sigma)$, are given by:

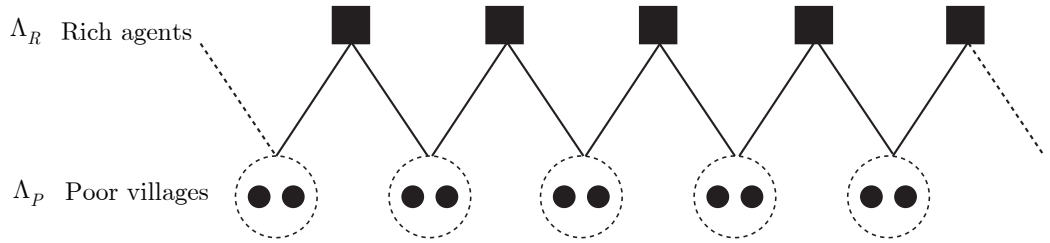

Figure 2: **A Simple Bipartite Network**. In this bipartite network, the black squares represent rich agents and the dotted circles represent poor villages, in each of which two ($= \eta$) poor agents are located. We connect the first poor village to the last rich agent, which is equivalent to imposing a periodic boundary condition. Each rich agent has interactions with two poor villages, each of which is itself interacting with one other rich agent. Similarly each poor village has interactions with two rich agents, each of which is interacting with one other poor village.

$$u_P(x,i,\sigma) := \sum_{y\in N_x} \pi_P(i,\sigma(y)), \text{ and } u_R(y,j,\sigma) := \sum_{x\in N_y}\sum_{\tilde{\eta}=1}^{\eta} \pi_R(\sigma_{\tilde{\eta}}(x),j). \tag{6}$$

Similarly to the random matching model, we write $\beta_x(\sigma)$ and $\beta_y(\sigma)$ for the best responses for the poor agents at village x and the rich agent at site y:

$$\beta_x(\sigma) := \arg\max_{i\in\{U,E\}} u_P(x,i,\sigma) \text{ and } \beta_y(\sigma) := \arg\max_{j\in\{U,E\}} u_R(y,j,\sigma)$$

where again we assume that the best responses are uniquely determined. Note that all poor agents in the same village have the same neighboring rich agent and hence the same payoffs, implying that their best responses are the same.

Evolutionary dynamics under local interactions, similarly to those under random matching, are defined as follows. At each period, either a poor agent or rich agent can randomly change their contract strategy according to the following probabilities:

$$P \text{ agent chooses } \begin{cases} \beta_x(\sigma) & \text{with probability } 1-\epsilon-\epsilon^\iota \\ U & \text{with probability } \epsilon^\iota \\ E & \text{with probability } \epsilon \end{cases} \quad R \text{ agent chooses } \begin{cases} \beta_y(\sigma) & \text{with probability } 1-\epsilon-\epsilon^\iota \\ U & \text{with probability } \epsilon \\ E & \text{with probability } \epsilon^\iota \end{cases} \tag{7}$$

To describe transitions between states, we denote by $\sigma^{x,\tilde{\eta},i}$ the state induced from a state σ by the switching of the $\tilde{\eta}$-th poor agent at village $x \in \Lambda_P$ to strategy i. Similarly, we denote by $\sigma^{y,i}$ the state induced from a state σ by the switching of the rich agent at site $y \in \Lambda_R$ to strategy i. Then equation (7) defines transition probabilities $\{P^\epsilon(\sigma,\sigma')\}_{\sigma'=\sigma^{x,\tilde{\eta},i},\sigma^{y,i}}$, which fully specifies the stochastic convention evolution. Again, the chain defined by (7) admits a unique invariant measure, which enables us to study stochastic stability.

Lemma 1. *The Markov chain defined by* (7) *has a unique invariant measure for all* ι*, including* $\iota = \infty$*.*

Proof. See Appendix A.2. □

As in the uniform interaction case, the rarity of a transition between two states is measured by the resistance (which we also call cost) of a transition, defined by $c(\sigma,\sigma') := \lim_{\epsilon\to 0} \frac{\ln P^\epsilon(\sigma,\sigma')}{\ln \epsilon}$ and this is given in our model as follows:

$$c(\sigma,\sigma') = \begin{cases} 0 & \text{if } \sigma' = \sigma^{x,\tilde{\eta},\beta_x(\sigma)} \\ \iota & \text{if } \sigma' = \sigma^{x,\tilde{\eta},U}, \beta_x(\sigma) = E \\ 1 & \text{if } \sigma' = \sigma^{x,\tilde{\eta},E}, \beta_x(\sigma) = U \end{cases} \qquad c(\sigma,\sigma') = \begin{cases} 0 & \text{if } \sigma' = \sigma^{y,\beta_y(\sigma)} \\ 1 & \text{if } \sigma' = \sigma^{y,U}, \beta_y(\sigma) = E \\ \iota & \text{if } \sigma' = \sigma^{y,E}, \beta_y(\sigma) = U \end{cases} \tag{8}$$

Observe that the model under random matching is simply a model on a complete bipartite network—a bipartite network in which all agents in one group are connected to all the agents in the other group and vice versa.

To study the problem of stochastic stability, as is explained intuitively in Section 3, we first need to identify the absorbing states for the unperturbed process—i.e., the process with $\epsilon \to 0$. These absorbing states are called stable states because every agent plays their best response at these states. To proceed, we denote by $\mathcal{E}$ a generic stable state, with $\mathcal{E}_U$ the U-convention (where all agents play U), and by $\mathcal{E}_E$ the E-convention (where all agents play E). We call a finite sequence of states $\gamma = (\sigma_1, \sigma_2, \cdots, \sigma_T)$ a path if σ_{t+1} is obtained from σ_t by a single agent's switching from one strategy to another. The cost for a path, $c(\gamma)$, is defined to be the sum of all costs of transitions associated with the path: i.e., $c(\gamma) := \sum_t c(\sigma_t, \sigma_{t+1})$ and the cost between absorbing states, $C(\mathcal{E}_1, \mathcal{E}_2)$, is defined to be

$$C(\mathcal{E}_1,\mathcal{E}_2) := \min\{c(\gamma) : \gamma \text{ is a path from } \mathcal{E}_1 \text{ to } \mathcal{E}_2\} \tag{9}$$

How does the unequal convention persist in the network interaction model? There are two distinctive features of the network interactions from the random matching model. The first concerns the stable states. In the global interaction model, the only two stable generic states correspond to the two pure Nash equilibria. In the network variant of the model, since interactions are local and limited to neighbors, there are many possible stable states. This is true even in the case of the simple bipartite network in Figure 2, restricted to only 10 nodes, with 5 nodes of each class, with each poor node being a village containing 2 poor agents. For example, the state given by:

$$\mathcal{E}_1 = (\underbrace{E}_{\text{rich}}, \underbrace{E}_{\text{poor}}, \underbrace{U}_{\text{rich}}, \underbrace{U}_{\text{poor}}, \underbrace{U}_{\text{rich}}, E, E, E, E, E) \tag{10}$$

is stable. The first entry in $\mathcal{E}_1$ shows the strategy of first rich agent, the second entry the strategy played by all agents in the first poor village, etc. This state thus exhibits two contiguous blocks of different strategies, one block playing U and one block playing E, which is preserved under best response. The stability of the state, $\mathcal{E}_1$, can easily be checked. The poor agents in the the second (or sixth) site best respond with E because $b > a_P$ and half of their neighbors are playing E. Also, the poor agents in the village at the fourth site best respond with U because $a_P > 0$ and both of their neighbors are playing U. The rich agents at the third or fifth site also best respond with U because $a_R > b$. Thus there is the possibility of global heterogeneity despite local uniformity, as in Young and Burke (2001) and our serfdom and desegregation examples below.

Recall that in comparing the basin of attraction of conventions U and E under random matching, we rely on the resistances (or the cost of transitions) from convention U *directly* to E and vice versa (in equation (4)), since there are only two stable states. By contrast, in the network model when we consider the minimum number of idiosyncratic players which can upset each convention, these numbers are determined by transitions to *intermediate stable states*, $\mathcal{E}'$, as the above example shows. That is, for the simple bipartite network model we will show that

$$\min_{\mathcal{E}'} C(\mathcal{E}_U, \mathcal{E}') \geq \min(\iota \left\lceil 2\frac{a_P}{b+a_P} \right\rceil, \left\lceil 2\eta\frac{a_R}{b+a_R} \right\rceil) > \left\lceil 2\frac{b}{b+a_P} \right\rceil \geq \min_{\mathcal{E}'} C(\mathcal{E}_E, \mathcal{E}') \tag{11}$$

for large ι and η. The left-hand side is the lowest cost transition out of $\mathcal{E}_U$ to any intermediate stable state $\mathcal{E}'$, while the far right-hand side is the lowest cost transition out of $\mathcal{E}_E$ to any intermediate stable state. Equation (11) can be regarded as a generalization of equation (4) to the local interaction model with N^R in equation (4) being replaced by 2—the number of neighboring vertices of poor agents in the simple bipartite local interaction network.

The second factor is the propagation of conventions via overlapping neighborhoods. Each neighborhood overlaps with others and the transition from one convention to another occurs due to the propagation of a strategy through clusters of neighbors —hence through the intermediate stable states, states in which some agents demand U while others demand E (e.g., see Ellison (1993, 2000)). To give an example of a fast propagation mechanism in our simple bipartite model, consider the following stable states, $\mathcal{E}_1$ as in (10) (presented below again) and $\mathcal{E}_2$ such that:

$$\mathcal{E}_1 = (\underbrace{E}_{\text{rich}}, \underbrace{E}_{\text{poor}}, \underbrace{U}_{\text{rich}}, \underbrace{U}_{\text{poor}}, \underbrace{U}_{\text{rich}}, E, E, E, E, E) \qquad \mathcal{E}_2 = (E, E, U, U, U, U, U, E, E, E) \quad (12)$$

where again there are ten sites in total and each site alternates between the rich agent and poor village sites. It is also easy to see that from $\mathcal{E}_1$, one idiosyncratic play by the rich agent at the seventh site induces the poor village at the sixth to best respond with U, the population state becomes stabilized at $\mathcal{E}_2$. Thus if deviant play causes a transition from convention E to state $\mathcal{E}_1$, we will have the following propagation mechanism of contract U:

$$\mathcal{E}_E \to \mathcal{E}_1 \to \mathcal{E}_2 \to ... \quad (13)$$

where $C(\mathcal{E}_1, \mathcal{E}_2) = 1$, $C(\mathcal{E}_2, \mathcal{E}_3) = 1$, and so on. In the next section, we show that general versions of (11) and a condition guaranteeing that $C(\mathcal{E}', \mathcal{E}'') = 1$ for all transitions between $\mathcal{E}'$ and $\mathcal{E}''$ on the path between the E and U conventions are together sufficient for the stochastic stability of the unequal convention in the simple bipartite network. This yields the following proposition as a result:

Proposition 2. *Consider the local interaction model in the bipartite network given in Figure 2. Then there exist ι^* and η^* such that for all $\iota > \iota^*$ and $\eta > \eta^*$, $\mathcal{E}_U$ is stochastically stable.*

Proof. According to Definition 2 below, the 1-d bipartite network in Figure 2 is *P-fragile* (1, 1/2), and the result follows from Theorem 2 below. □

This local interaction model, as in Young and Burke (2001), can generate transition paths marked by heterogeneity in conventions. Applied to the two population case, it can represent the phenomenon noted by Oberdorfer (1963) during the civil rights movement (discussed further below), where, within the same city, lunch counters would be desegregated while hotels remained segregated, as well as the piecemeal transition to desegregation across cities prior to 1964. Similarly, the diversity of tenancy relationships across England after the Black Death shows heterogeneity consistent with the networked version of the model, where manorial obligations disappeared in some parts of England very quickly while others took much longer. However, the network structure in this version of the model is highly stylized, and looks little like real world social networks. Generalizing the results from the local interaction model to arbitrary bipartite graphs is what we turn to in the next section.

3.2 General Bipartite Networks

We now prove our most general result about the stochastic stability of the U convention. We show that a similar characterization and intuition hold for a broad range of bipartite networks satisfying a condition we call *P-fragility*. Loosely, *P-fragility* guarantees that enough of the P population will respond to the idiosyncratic play of a small cluster of R players, so that the rest of the R population has to best-respond by switching strategies as well. *P-fragility* rules out networks that are robust to idiosyncratic play of the rich, no matter how intentional and how frequent.

Recall that a bipartite graph $(\Lambda_R, \Lambda_P, E)$ is a graph in which a set of vertices $\Lambda := \Lambda_R \cup \Lambda_P$ is partitioned into two sets, Λ_R, Λ_P such that every edge in edge set E is incident to one vertex in Λ_P and another vertex in Λ_R. Recall also that a path in a graph is a sequence of distinct vertices in the edge set. We say that a graph Λ is connected if every pair of vertices can be joined by a path and we consider only connected bipartite networks. If we have disconnected networks, the analysis can be done separately for each connected component. Recall that N_z is the neighbors of site z and similarly we let $N(S)$ be the neighbors of set

S: i.e., $N(S) = \cup_{z \in S} N_z$. We will also write $(S_P)^c := \Lambda_P \backslash S_P$ and $(S_R)^c := \Lambda_R \backslash S_R$.

To study general interaction structures for two-population games, we extend the concept of *cohesiveness*, proposed by Morris (2000), to bipartite networks. Cohesiveness captures how many interactions two sets of players, each from a different population, share with each other.

Definition 1 (q-cohesiveness)**.** *We say that $S_P \subset \Lambda_P$ is q-cohesive with $T_R \subset \Lambda_R$ if*

$$\min_{z \in T_R} \frac{|N_z \cap S_P|}{|N_z|} \geq q.$$

In words, a set of poor villages S_P in a population is q-cohesive with another set T_R from the rich population if every rich agent has at least q proportion of its interaction with S_P. Note that it is not symmetric: T_R can have many of its members interact with S_P while members of S_P interact with many members of T_R as well as many agents outside of T_R.

We can use *q-cohesion* to obtain a simple sufficient condition for stability of the U. We simply require that i) every poor village set is sufficiently cohesive with its neighboring rich so that a stable cluster (described in (10)) can emerge and an expanded cluster can stabilize and ii) every set of poor villages neighboring rich agents are sufficiently cohesive with the *other* poor villages, so that the stable cluster can propagate throughout the network.

Definition 2. *We say that a bipartite graph is (q_P, q_R)-cohesive if*
(i) Every S_P is q_R-cohesive with $N(S_P)$.
(ii) Every S_R is q_P-cohesive with $N(S_R)$.

We can then show the following theorem:

Theorem 1. *Suppose that the bipartite graph is $(\frac{b}{b+a_P}, \frac{b}{b+a_R})$-cohesive. Then there exists ι^* and η^* such that for all $\iota > \iota^*$ and $\eta > \eta^*$, $\mathcal{E}_U$ is stochastically stable.*

Proof. This follows from Theorem 2 below. □

While (q_P, q_R)-cohesiveness is intuitively appealing, it is too strong, and does not admit many of the cases we have already studied above. First, for the complete network ($|\Lambda_P| = |\Lambda_R| = N$) which yields the random matching model in Section 2.3, every singleton $\{x\}$ is

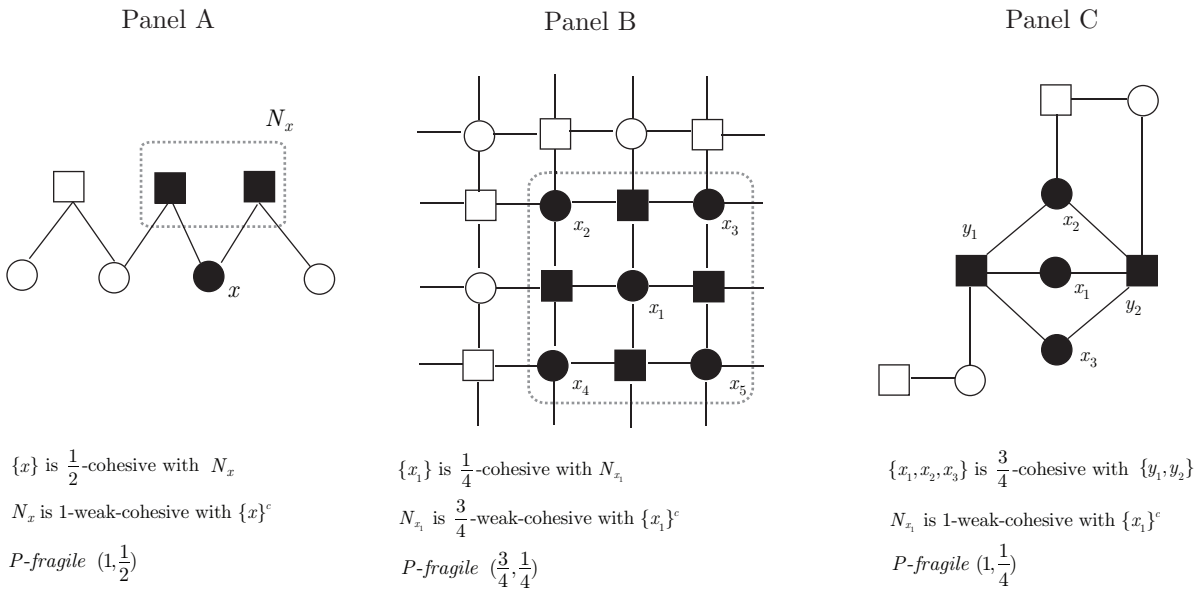

Figure 3: **Illustration of Definitions 1, 3, and 2.** We call the bipartite network in Panel B a 2-*d* network. Squares indicate Rich agents, circles indicate clusters of size η Poor agents. Colored indicates U strategy played.

$\frac{1}{N}$-cohesive with N_x ($q_R = \frac{1}{N}$ in Definition 2 (i)). Thus when N is large for the complete network, the condition in Theorem 1 is difficult to be satisfied. Second, for the 1-d network in Figure 2, every $\{x\}$ is $\frac{1}{2}$-cohesive with N_x ($q_P = \frac{1}{2}$ in Definition 2 (ii)). Since the poor prefer E to U, $\frac{b}{b+a_P} > \frac{1}{2}$ and thus a sparse network like the 1-d network fails to satisfy the condition in Theorem (1). Thus, we introduce a slightly weaker condition than cohesiveness which also induces a fast propagation mechanism described in (12).

Definition 3 (Bipartite weak-cohesive)**.** *We say that a $S_R \subset \Lambda_R$ is q-weak-cohesive with $T_P \subset \Lambda_P$ if*

$$\max_{z \in T_P} \frac{|N_z \cap S_R| + 1}{|N_z|} \geq q. \tag{14}$$

If $T_P = \varnothing$, we set $q = 1$. Obviously, if S_R is q-cohesive with T_P, then S_R is q-weak-cohesive with T_P. In words, S_R is *q-weak-cohesive* with T_P if there exists a poor village (z) in the poor village set, T_P, whose has at least $q - \frac{1}{|N_z|}$ proportion of interactions with S_R. Thus, when all rich agents in S_R play strategy, say, U, there exists a poor village z which has at lease $q|N_z| - 1$ of rich neighbors playing strategy U. If one rich agent in the neighborhood of z (outside of S_R) switches from E to U and if q is greater than the threshold fraction, derived from payoffs, inducing U to be the best responses of the poor agent at z, U indeed

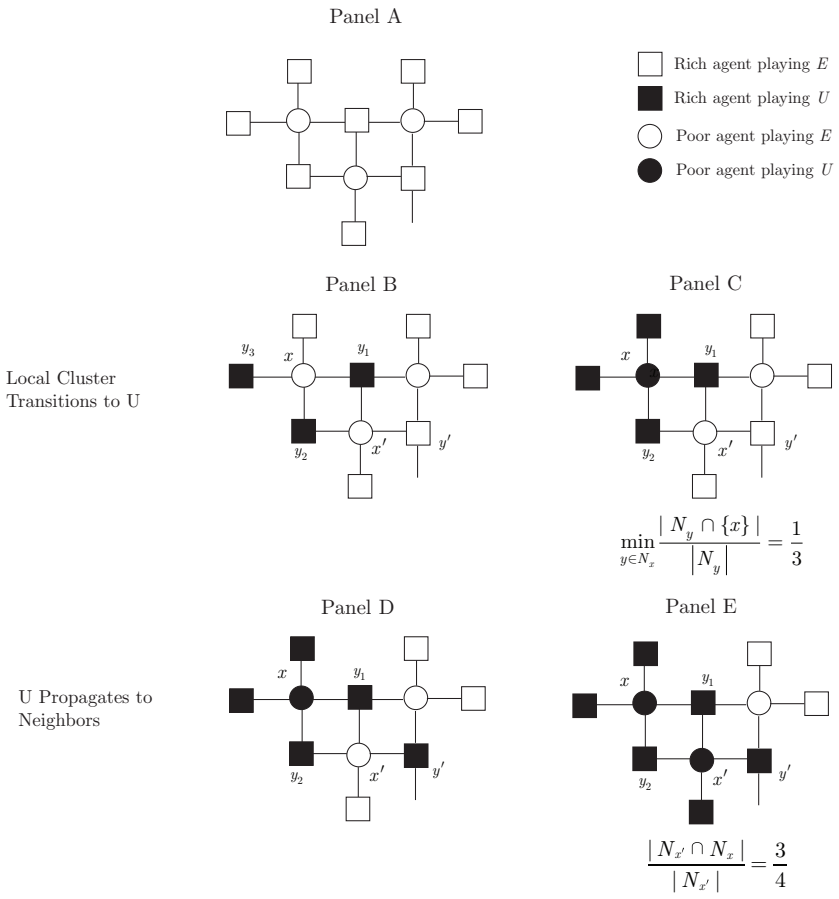

Figure 4: **Illustration of Definition 2.** From the E-convention $\mathcal{E}_E$ in Panel A, three rich agents, y_1, y_2, y_3, idiosyncratically play U, leading to Panel B. Then, $\{x\}$ being $\frac{1}{3}$-cohesive with N_x ensures that the state of the cluster playing U agents becomes stable in Panel C. Then the rich agent at y' idiosyncratically plays U in Panel D, and N_x being $\frac{3}{4}$-weak-cohesive with $\{x_1\}^c$ ensures that the poor village at x' will best respond with U, leading to Panel E.

becomes the best response of the poor agent at z. This process guarantees the propagation mechanism. Note that since we are interested in the fast propagation and convergence, Definition 3 relies on the propagation mechanism induced by a single agent (hence 1 in the numerator of (14)). However, Definition 3 can be relaxed to study other (possibly slower) propagation mechanisms by more than one agents, by changing the number 1 to 2 or 3 etc. accordingly. We provide some examples of Definition 3 in Figure 3.

In Figure 4, we explain, in details, how to determine the cost of transitioning between conventions using cohesion and weak-cohesion. Suppose that the initial convention is $\mathcal{E}_E$ (Panel A), and that three rich agents playing U in the neighborhood of a poor village ensures that the poor village's best response is U. Now suppose that the three rich agents at y_1, y_2,

and y_3 play U idiosyncratically at the same time. Then whether the resulting state induced by the deviant plays of the rich agents is stable depends on the number of neighbors of those deviant rich agents (see Panel C in Figure 4). In this example, a minimal stable set consists of one poor village and its neighbors, $(\{x\}, N_x)$. If $\{x\}$ is q_R-cohesive with N_x, it means that the rich agents in the neighborhood of the poor village x have at least q_R proportion of interactions with the poor village. Thus if $a_R q_R \geq b(1-q_R)$ (or $q_R \geq \frac{b}{a_R+b}$) holds, playing U is the best response for the remaining rich agents in N_x. Thus all agents in $(\{x\}, N_x)$ playing U constitutes a stable, autonomous, cluster of agents playing U (see Panel C). In the 1-d network in Figure 2 (or the 2-d network in Figure 4), if $\{x\}$ is q_R-cohesive with N_x, then every poor set S'_P containing $\{x\}$ is again q_R-cohesive with $N(S'_P)$. Definition 4 below requires the existence of a set S_P such that every S'_P containing S_P is again q_R-cohesive with $N(S'_P)$, which ensures that $(S'_P, N(S'_P))$ can be a stable cluster.

Next, consider an idiosyncratic play of U by a rich agent neighboring this cluster of U agents (see y' in Panel D in Figure 4). If N_x is q_P-weak-cohesive with $\{x\}^c = \Lambda_P \backslash \{x\}$, there exists $x' \neq x \in \Lambda_P$ such that the poor village at x' neighboring the cluster faces at least $q_P|N_{x'}| - 1$ number of the rich agents playing U. If $a_P q_P \geq b(1-q_P)$ (or $q_P \geq \frac{b}{a_P+b}$) and a single idiosyncratic strategy of U is played by a rich agent in $N_{x'}$ (that is, y' in Panel D), then playing U is the best response for the poor agent. Thus, this condition, together with the condition for cohesivness, ensures a transition to a new stable state with a greater number of agents playing U. In Definition 4, we requires that every $N(S_P)$ is q_P-weak cohesive with $(S_P)^c$, where we recall that $(S_P)^c := \Lambda_P \backslash S_P$.

Definition 4. *We say that a bipartite graph is P-fragile* (q_P, q_R) *if*
(i) For some S_P*, every* S'_P *containing* S_P *is* q_R*-cohesive with* $N(S'_P)$*.*
(ii) Every $N(S_P)$ *is* q_P*-weak-cohesive with* $(S_P)^c$*.*

The definition of *P-fragile* is somewhat more complicated than Definition 2, however the following two lemmas give sufficient conditions for *P-fragililty* that are readily applicable to various network structures.

Lemma 2 (Sparse networks)**.** *Let* D *be the maximum degree of the poor village and every set neighboring a poor village set is* q_P*-weak cohesive with the complement of the poor village.*

Then the bipartite graph Λ *is P-fragile* $(q_P, \frac{1}{D})$.

Proof. We first check Definition 4 (i). Let $S_P \subset \Lambda_P$ and $z \in N(S_P)$. Then for some $x \in S_P$, $z \in N_x$. Thus $x \in N_z$ and $|N_z \cap S_P| \geq 1$. Then for any $z \in N(S_P)$, we have

$$\frac{|N_z \cap S_P|}{|N_z|} \geq \frac{1}{|N_z|} \geq \frac{1}{D}.$$

Thus every S_P is $\frac{1}{D}$-cohesive with $N(S_P)$ which implies Definition 4 (i). Also by the assumption, any $N(S_P)$ is q_P-weak cohesive with $(S_P)^c$. □

Note that for the 1-d bipartite network in Figure 2, every rich agents set neighboring S faces a poor village site (x) outside S which has one interaction with the rich agent set (out of total 2 interactions). Thus

$$\frac{|N_x \cap N(S_P)| + 1}{|N_x|} = \frac{1+1}{2}$$

and $N(S_P)$ is 1-weak-cohesive with $(S_P)^c$ and thus the 1-d bipartite graph is *P-fragile*$(1, \frac{1}{2})$. Similarly, for the 2-d bipartite network in Panel B in Figure 3, every rich agents set neighboring S faces a poor village site (x) outside S which has 2 interactions with the rich agent set (out of total 4 interactions).Thus

$$\frac{|N_x \cap N(S_P)| + 1}{|N_x|} = \frac{2+1}{4}$$

which shows that the 2-d bipartite network is *P-fragile* $(\frac{3}{4}, \frac{1}{4})$. Thus Lemma 2 includes the 1-d bipartite network in Figure 2 and 2-d bipartite network in Panel B in Figure 3 as special cases. In addition, Lemma 3 covers the complete network (hence random matching interactions in Section 2.3).

Lemma 3 (Dense networks)**.** *Suppose that there are* $|\Lambda_P| = |\Lambda_R| = N$ *and each pair of sites in a population shares at least* $N - k$ *neighbors for some* $0 \leq k < N$ *(i.e.,* $|N_x \cap N_{x'}| \geq N - k$ *for all* $x, x' \in \Lambda_P$ *and* $|N_y \cap N_{y'}| \geq N - k$ *for all* $y, y' \in \Lambda_R$*). Then the bipartite graph,* Λ*, is P-fragile* $(\frac{N-k}{N}, \frac{N-k}{N})$.

Proof. Choose $S_P = \Lambda_P$ in Definition 4 (i). Then for all $y \in N(S_P) = \Lambda_R$,

$$\frac{|N_y \cap S_P|}{|N_y|} = \frac{|N_y \cap \Lambda_P|}{|N_y|} \geq \frac{N-k}{N}$$

and hence Λ_P is $\frac{N-k}{N}$-cohesive with $N(\Lambda_P)$. Thus, Definition 4 (ii) is satisfied. Also let $N(S_P)$ such that $S_P \neq \Lambda_P$ be given. Choose $z \notin S_P$. Then again by the assumption,

$$\frac{|N_z \cap N(S_P)| + 1}{|N_z|} \geq \frac{N-k+1}{N} \geq \frac{N-k}{N}$$

Thus every $N(S_P)$ is $\frac{N-k}{N}$ weak-cohesive with S_P. □

The condition for *P-fragility* combines the two aforementioned distinctive features of network structure, which make the P population vulnerable to idiosyncratic play by the R population. Following Young (2011), who uses very similar conditions to characterize diffusion of efficient social innovations in a symmetric coordination game, we can call the first condition *Autonomy*: essentially every set of P players is in a neighborhood of sufficiently many R players such that a small amount of idiosyncratic play of the R players can generate a local stable cluster of U play. The rich share enough interactions with the poor that they best respond with U to the poor, who are induced to best-respond with U to idiosyncratic play of U by a small subset of rich. The second condition can be called *Contagion*: each P player is in a neighborhood of R players that share neighbors with sufficiently many other R players. Thus our theorem is in the spirit of Young (2011), but adapted to the asymmetric bipartite networks with intentionality, where it yields a sufficient condition for even *inefficient* conventions to be stochastically stable in the long-run.

Theorem 2. *Suppose that the bipartite graph is P-fragile* $(\frac{b}{b+a_P}, \frac{b}{b+a_R})$. *Then there exists* ι^* *and* η^* *such that for all* $\iota > \iota^*$ *and* $\eta > \eta^*$, $\mathcal{E}_U$ *is stochastically stable.*

Proof. See Appendix A and the precise expressions for ι^* and η^*. □

Theorem 2 identifies sufficient conditions for $\mathcal{E}_U$ to be stochastically stable. This is because in a networked model, there are many intermediate stable states between conventions, unlike in the uniform interaction case. Thus we are only able to bound the transition times

between the two conventions, preventing us from obtaining a sharper characterization. We also prove a partial converse of Theorem 2 in Proposition 5 in Appendix A.1, which provides similar sufficient conditions for convention $\mathcal{E}_E$ to be stochastically stable. As with *P-fragility* guaranteeing vulnerability of the poor to the idiosyncratic play of the rich, so that convention U is easily accessible, this $R-fragility$ condition guarantees that the network is sufficiently vulnerable to idiosyncratic play of the poor so that E is easily accessible.

The proof of Theorem 2 is presented in Appendix A. In the proof, we first show that (i) the cohesive sets give the upper bound of the cost of transition from $\mathcal{E}_E$ (Lemma 4) and (ii) the weak-cohesive sets, together with cohesive sets, ensure that the cost of a transition from each intermediate stable state ($\mathcal{E}_M$) is always 1 (Lemma 5). Then we show that (iii) a condition similar to (11) is sufficient for the stochastic stability of the U-convention $\mathcal{E}_U$.

4 Information Structure and Equilibrium Selection

Because it is sampling noise that drives institutional transitions, classes whose play is not completely observed by members of the other class have an advantage similar to that conferred by larger or smaller η. In this section we show that our model can be re-interpreted so that even if the interactions are uniform and population sizes are identical ($\eta = 1$), we can consider each class observing only a sample of the other class (see Figure 5). Then players best-respond not to the average play of the whole opposing population, but rather a subsample. Then, if the sample sizes differ between the two classes, the class with a larger sample will have an advantage, even if the class sizes are identical. Correspondingly, having information on what only a small share of the other class is doing will heighten the responsiveness to small amounts of idiosyncratic play, increasing the likelihood of a transition to a disfavored convention when play is intentional.

Allowing populations to respond to only a sample of their neighbor's play breaks the symmetry implicit in the model interaction structure above. If agent P plays a best-response to a neighborhood containing agent R, then in our model thus far this implies that agent P is also in the neighborhood whose play agent R best responds to. Our model thus far also implies that populations can automatically infer the state of play from the payoff they

received. But if population P agents only sample a number κ_P of their neighbor's play, and population R agents only sample a number κ_R of their neighbors play, then the payoff-relevant population is no longer identical to the behavior-relevant population.

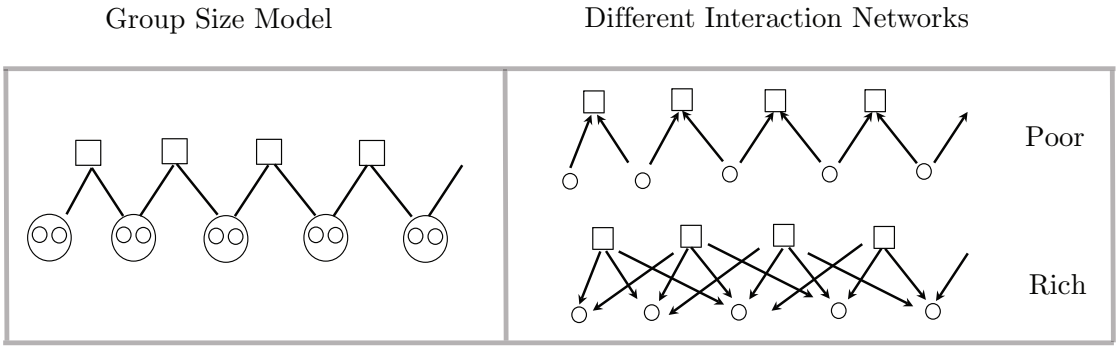

Figure 5: **An analogy between the group size model and an asymmetric structure model.** The image on the left shows the $1-d$ interaction structure with $\eta = 2$. The image on the right shows an equivalent information structure with $\eta = 1$ and $\kappa = \frac{\kappa_R}{\kappa_P} = 2$, where the arrows are directed from players who are sampling to the players whose behavior is being sampled.

The stability of unequal gender norms provides an illustration. Male-female interactions may be differentially structured in traditional patriarchies, despite numerical parity and similar interaction structures. However, men who can publicly circulate and fraternize with other men have an informational advantage vis-a-vis women confined to domestic roles and family networks. This may have a similar effect as men being less numerous in a random matching model of the kind in the last section. By indirectly observing the behavior of many women, male strategies will not be altered in response to idiosyncratic play of a few women, while women may be induced to abandon a favored convention by the idiosyncratic play of just a few men. This may explain the well-documented persistence of unequal gender norms, such as women working outside the home (Alesina et al., 2011), patrilocality, and son-preference (Jayachandran, 2015).

Beyond gender, we also have in mind rural societies. Pre-capitalist agrarian institutions (Gellner, 1983) entailed very different information sets between classes. Elite upper classes communicated readily amongst themselves and therefore had information about the recent play of a large segment of the less well-off class. The geographical, cultural and linguistic isolation of the Ps, by contrast, militated against information sharing beyond ones' local

community.

The advantage enjoyed by the Rs is not that a given R-patron may engage the P-clients of other Rs. Rather, by drawing information from a larger sample of Ps, the R's less noisy signal of the distribution of play reduces the likelihood that their myopic best response will overreact to the chance occurrence of a high level of idiosyncratic play among their particular P-clients.

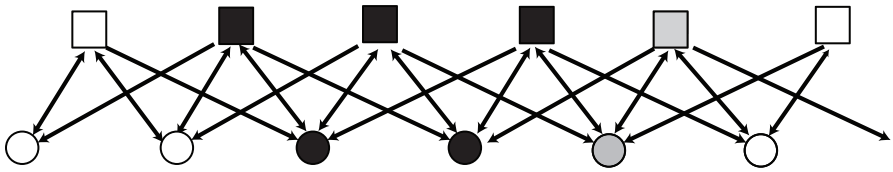

Figure 6: **Asymmetric Network: an autonomous state and propagation.** Note that $\lceil 2\frac{b}{b+a_P} \rceil = 2$ and we assume that $\kappa = \frac{\kappa_R}{\kappa_P} = \frac{4}{2} = 2$ and $\lceil 2\kappa\frac{b}{b+a_R} \rceil = 2$. The sites in black constitute the stable cluster of poor and rich agents, hence an autonomous state. The sites in grey yields the propagation.

This idea can be readily captured in a re-interpreted version of our model. Fix $\eta = 1$ and consider again the 1-d bipartite network in Section 3. To simplify analysis, suppose that $\iota = \infty$, but the analysis to follow can be readily extended to arbitrary ι. Now we suppose that each poor site is occupied by one poor agent and each poor observes the behavior of κ_P rich agents while each rich agents observes the behavior of κ_R poor agents. Further let $\kappa = \frac{\kappa_R}{\kappa_P}$ be a parameter capturing the relative scope of vision of the rich agents. In Figure 7, A rich agent observes 2κ poor neighbors, while a poor agent observes the behavior of 2 rich neighbors. To explain how convention U becomes stochastically stable, we consider an escape from convention U first. Observe that at convention U, only the poor agents idiosyncratically play contract E. Thus the minimum cost of escaping convention U can be similarly computed as in (11) and is given by

$$\min_{\mathcal{E}'} C(\mathcal{E}_U, \mathcal{E}') \geq \left\lceil 2\kappa \frac{a_R}{a_R + b} \right\rceil \tag{15}$$

That is, to induce a rich agent to best respond with E, at least $\frac{a_R}{a_R+b}$ proportion of the 2κ poor agents neighboring the rich agent need to idiosyncratically play strategy E.

Next we consider an escape from convention E. Also recall that, when $\iota = \infty$, at conven-

tion E only rich agents idiosyncratically play. Thus, the minimum number of idiosyncratic plays by the rich agents to escape from convention U needs to ensure that the number of poor agents best-responding to the rich agents' idiosyncratic plays again is enough to support the idiosyncratic play U of the rich agents as their best responses. Note that $\lceil 2\frac{b}{b+a_p} \rceil = 2$ rich agents in the neighborhood of a poor agent ensures U to be the best response of the poor agent and that $\lceil 2\kappa\frac{b}{b+a_R} \rceil$ poor agents in the neighborhood of a rich agent ensures U to be the best response of the rich agent. Thus, the minimum number of the rich agents' idiosyncratic play must be at least $\lceil 2\kappa\frac{b}{b+a_R} \rceil + 1$ (see Figure 6 where $\lceil 2\frac{b}{b+a_P} \rceil = 2$ and $\lceil 2\kappa\frac{b}{b+a_R} \rceil = 2$). Thus, we find that

$$\min_{\mathcal{E}'} C(\mathcal{E}_E, \mathcal{E}') \leq \left\lceil 2\kappa \frac{b}{a_R + b} \right\rceil + 1$$

Finally, from the stable cluster consisting of U playing $\left\lceil 2\kappa\frac{b}{a_R+b} \right\rceil + 1$ rich agents and U playing $\left\lceil 2\kappa\frac{b}{a_R+b} \right\rceil$ poor agents, if one neighboring (E-playing) rich agent plays U-strategy idiosyncratically, then the dynamics move to the new stable cluster with a new U playing rich agent and a new U playing poor agent (see sites in grey in Figure 6). Thus similarly to Theorem 2, we obtain the following result:

Proposition 3. *Suppose that $\iota = \infty$ and consider the network in Figure 6, where $\kappa \geq 2$. Then, we have*

$$\min_{\mathcal{E}'} C(\mathcal{E}_U, \mathcal{E}') \geq \left\lceil 2\kappa \frac{a_R}{a_R + b} \right\rceil \text{ and } \min_{\mathcal{E}'} C(\mathcal{E}_E, \mathcal{E}') \leq \left\lceil 2\kappa \frac{b}{a_R + b} \right\rceil + 1 \tag{16}$$

and there exists κ^* such that for all $\kappa \geq \kappa^*$, U is a stochastically stable state.

Proof. See Appendix C. □

Changing gender norms as a result of female labor force participation may result from the increased availability of information about male behavior (lower κ or higher κ_{Female} relative to κ_{Male}), as women could increasingly share information about their husbands when working outside the household. In the pre-industrial context, Gellner argues that the geographic, industrial, and occupational mobility characteristic of modern labor markets (coupled with the spread of literacy and greater ease of communication) made workers less responsive to the demands of local patrons, as they had access to information about conditions elsewhere

via shared national culture. In our U.S. South example below, ongoing mechanization of the agricultural economy, lowered transportation costs, the spread of black newspapers, and increased urbanization may have played a similar role, allowing poor blacks to see a much larger swath of white behavior, and therefore no longer reacting purely to the idiosyncratic behavior of the local whites. This is independent of any increased market competition: rather it comes through the availability of information about the norms prevalent in other places, measured as an increase in κ.

A similar informational role might be played by leaders or prominent activists, as in Acemoglu and Jackson (2014), who model leaders as agents whose play is visible to all future agents. In our context, leaders could be modelled as agents whose behavior is always sampled by the other players. The idiosyncratic play of only a few leaders in one population could induce best-responses by the all players in the other population. We leave this extension to future work.

5 Historical "Bottom up" Transitions Among Conventions

We now go through a number of well-known historical examples of changes in conventions, motivating our modifications of the standard evolutionary model and contrasting them with the conventional top down approaches. They have in common the following: First, a transition was induced by challenges to a long standing unequal status quo convention by at best loosely coordinated members of the disadvantaged group. Second, these challenges altered the best response of the advantaged group, allowing for the decentralized stabilization of an alternative convention. Any legislation or other formal recognition and enforcement of the new convention came only after the *de facto* decentralized change had occured. Third, transitions followed structural transformations, either in population size (changes in η) or, qualitatively, changes in interaction structure away from *P-fragile* networks.

5.1 The Demise of Serfdom In England

Consistent with the centralized top down approach, the emancipation of Russia's serfs by Tsar Alexander II in 1863 was a deliberate choice to implement a new set of institutions resulting from bargaining within Russia's elite (Blum, 1971). In contrast, the demise of English serfdom was not the result of explicit bargaining among political parties implemented by a central authority.[5] Serfdom was never formally repealed by law England, and indeed all legal labor restrictions were arguably not abolished until the 1875 repeal of Master and Servant criminal fines (Naidu and Yuchtman (2013)). But agricultural serfdom had, in practice, disappeared centuries earlier.

The practice of serfdom in England centered on customary villeinage, whereby serfs (villeins) were generally tied to lords on inherited customary contracts, and had labor obligations (e.g. tallage, the land tax) as well as tax obligations like merchet (dues for marriage of a daughter), heriot (death taxes), childwite (fine for illegitimate pregnancy) and chevage (head taxes paid to the lord). Beyond the economic claims, serfdom came with a distinct inequality in status, with the chevage being "psychologically burdensome,signifying,...the yoke of servitude." (Bailey, 2014, pp. 46). As with sharecropping practices or norms of racial segregation in the U.S. the transition looked like "a bewildering variety of practices ... a mosaic of variable bargains" even within the same locality of upon the same seignioral estate." (Bailey, 2014, pp. 23).

Between 1350 and 1450 this entire system disappeared after centuries of peristence, replaced by shorter leases (either copyhold or leasehold) where rents were fixed in cash, status was no longer inherited, and no feudal dues were collected. The most immediate candidate explanation during this period is undoubtedly the 1349 Black Death, which lowered population by up to 50% in some areas.

North and Thomas (1971) provide the economic interpretation of the change. Essentially

[5]Other countries are less clear-cut, but an arguably similar process was at play in France. Protracted agrarian conflict culminated in the 1789 peasant rebellions, forcing local lords to abandon many of their feudal privileges well before any legislation was passed:"Peasant uprisings kept rural France on the legislative agenda and drowned out the tendencies to silence on seigneurial rights that characterized much of the nobility" (Markoff, 1996, pp. 509). The abolition of seigniorial dues by the Estates General in 1789 confirmed the new order, it did not introduce it. Instead, a series of uncoordinated actions by dispersed peasants, each taking the grievances of the entire group as their own, induced the aristocratic class to change the terms of agricultural labor.

the fall in the labor-land ratio increased the value of labor and competition among manors for scarce villeins induced a change in customs. North and Thomas describe a process very consistent with our model, but our model makes the interpretation of the Black Death shock more nuanced. The Black Death changed population shares such that idiosyncratic play by the serfs could rapidly induce a best-response from the lords, as well as changing the relative payoffs from different institutional arrangements. Indeed, the timing of the decline of serfdom occurs a full generation after the Black Death, largely during a period of falling real wages, and with considerable local variation (Bailey, 2014). North and Thomas (1971) point out that were it simple changes in scarcity alone, it would be difficult to account for the simultaneous change of so many customs together, rather than a simple change in the terms of various obligations. It required a change in tenant-landlord conventions, rather than simply the terms of the contract.[6] In addition, the mechanism is not simply manorial competition for labor: there was a wave of peasant unrest in the decades after the Black Death, culminating in the Great Peasant Revolt of 1381, ranging from conspiracies to not pay merchet to physical attacks on lords and the destruction of land records. The bargaining was more collective than individual, as whole groups of villeins simultaneously offered different terms: In Holywell-cum-Needingworth, there were large strikes in 1353 and 1386, along with 191 cases of individual refusal to perform labor services between 1353 and 1403 (compared to the 21 such instances between 1288 and 1339) (Bailey, 2014). In 1379 Essex county, "the tenants [collectively] offered their lord 40 [shillings] to set fixed monetary sums for rents and services." (Poos, 1991, pp. 247). There is also evidence of a "seigneural reaction" where lords attempted to squeeze more servile dues out of their tenants, as well as using powers granted to them by the 1351 Statute of Labourers, as our model would suggest. Nonetheless, by the mid-15th century, the panoply of feudal norms was extinct, with little change in formal law. Beyond the changes in economic conventions, the cultural norms regulating the interactions between peasants and lords also changed. For example, Bailey (2014) notes that the language used in manor records to describe relationships with villeins was upgraded from "bondage"

[6] As North and Thomas (1971, pp. 799) write "When a change of parameters offers potential gains from establishing new secondary institutional arrangements, these may not be directly realizable simply because they run counter to the basic rules of society." While North and Thomas stress common law changes, recent research shows that the changes in customs predated the legal changes by at least a century Bailey (2014).

and "villeinage" to more dignified modes of address. The historian E.B. Fryde (Fryde, 1996, pp. 6) writes:

> throughout the 1380s and long beyond them...the servile velleins refused with ever increasing persistence to accept the implications of serfdom, ... In this atmosphere of frequent local disorder and of continuous tension between lords and tenants, the direct exploitation of domanial estates would largely disappear from England in the fifty years after the [1381] Great Revolt.

Proposition 1 qualitatively captures this transition. Suppose the serfdom convention is U and the leasehold convention is E. The Black Death likely both made E risk-dominant and also lowered η. In an intentional deviation environment (where leaseholders do not ask to be villeins and lords do not spontaneously offer serfs leaseholdings) then the equilibrium could have stayed at U if all that had changed was the payoffs. But because η also fell, the transition from U to E was quite rapid: with relatively few peasants, it only took a small number of peasants demanding new terms to induce lords to concede their traditional privileges. Serf migration also increased in the years following the Black Death: in the language of our model, the bipartite peasant-lord network was less likely to be $P - fragile$, further facilitating a transition from U to E. The dynamics of this transition are captured by our approach: the heterogeneous, piecemeal, occasionally reversed, and gradual transition, well after the initial shock, is how a transition from an unequal to equal convention would occur in the network extension of our model studied in Section 3.

5.2 South Africa

South Africa's transition to democracy provides another reason to extend the standard evolutionary model, and a contrast with the top down approach. Acemoglu and Robinson (2006, pp. 13) write that "the basic structure of apartheid was unaltered" until "De Klerk concluded that the best hope for his people was to negotiate a settlement from a position of strength". For Acemoglu and Robinson, South Africa's new institutions were introduced as the result of the formal constitutional negotiations beginning in 1990. Consistent with their view that economic institutions will only change after the political institutions change, owing to commitment failures, they conclude that the change in economic institutions resulted from

the introduction of a new political system. However, our reading of the historical evidence is that fundamental changes in economic practices and hence *de facto* economic institutions predate De Klerk's rise to prominence in the National Party, and are more plausibly seen as the cause of the subsequent political transition, rather than its consequence.

Rather than simply a set of laws, crucial dimensions of South African apartheid can be modeled as a convention regulating relations between employers and black African workers. Formal apartheid labor regulations, such as pass-laws and Master and Servant laws were an important component of Apartheid, but they buttressed and formalized a large set of informal racial practices not explicitly legislated.

Having existed de facto throughout most of South Africa's recorded history, the apartheid system was formalized in the early 20th century and strengthened in the aftermath of World War II. For white business owners, the convention might be expressed as follows: Offer only low wages for menial work to blacks. For black workers the convention was: Offer one's labor at low wages, do not demand access to skilled employment. These actions represented mutual best responses: As long as (almost) all white employers adhered to their side of the convention, the black workers' best response was to adhere to their aspect of the convention, and conversely.

A wave of strikes beginning in the 1970s (particularly the 1973 Durban strike) and peaking in the 1980s together with the refusal of Soweto students to attend classes taught in Afrikaans and the ensuing 1976 uprising signaled large scale rejection of the apartheid. While these protests were not entirely spontaneous, of course, the African National Congress leadership were in exile or prison at the time and had limited capacity to coordinate or direct these bottom up challenges to the status quo.

Consistent with this, Mariotti (2012) shows that African occupational segregation was falling well before the formal democratization. Indeed, her data shows that the relative share of black production workers in manufacturing sharply grows between 1974 and 1979, consistent with a response to the unrest (and prior to the enactment of 1979 labor market reforms). Many business leaders concluded that adherence to the apartheid convention was no longer a best response, leading them independently to alter their labor relations, raising real wages and promoting black workers. One Anglo-American executive commented:

"...in the business community we were extremely concerned about the long-run ability to do business..." (Wood, 2003, pp. 171). By 1981, the CEO of Barlow Rand found himself describing the new convention: "[he] said that Black trade unions were a fact of life" (Wood, 2003, pp. 137).

The business community eventually did come together to develop a political strategy for managing the transition, but only well after significant dimensions of apartheid had already unravelled. Starting in the mid 1980s, the Anglo-American Corporation developed new policies for 'managing political uncertainty' and to address worker grievances, even granting workers a half day off to celebrate the Soweto uprising. In September 1985, Anglo American's Gavin Relly led several business leaders on a clandestine "trek" to a secret place near Lusaka to seek common ground with African National Congress leaders in exile, on both political and labor market institutions. In 1986 the Federated Chamber of Industries issued a business charter with this explanation: "the business community has accepted that far reaching political reforms have to [be] introduced to normalize the environment in which they do business." The centralized bargaining over political institutions only took place well-after the decentralized transition in economic conventions was underway.

Note the following about this process: As with our serfdom example above, the concession of best-responding businesses to rejection of apartheid by black workers occurred well before the political transition. Second, the process of transition was extremely abrupt, bringing to an end in less than a decade de facto class and race relations that had endured for decades. Third, while trade unions, 'civics' (community organizations), and other groups were involved in the rent strikes, student stay aways, and strikes against employers, the piecemeal and *de facto* transition away from apartheid was substantially decentralized and only loosely coordinated prior to the 1990 unbanning of the ANC.

5.3 Racial Desegregation in the Southern U.S

The desegregation of the U.S. South in the 1960s also followed a logic close to our model. Racial segregation in the U.S. South was of course sustained by laws, but had an important component of informal convention. And during the 1950 and early 1960s it was well on the way to unravelling on the ground before desegregation was recognized in law. Protests,

intentional violations of status quo norms, played an important, if understudied, role in *de facto* desegregation. Our task is not to explain the complex process that produced the Civil Rights movement and its successes, but rather to show that informal racial conventions were destabilized by a large number of relatively small, intentional, shocks generated by activists and participating citizens.

The status-quo unequal convention was that blacks and whites would interact in segregated public spaces, while an alternative, more equal convention was integrated public interaction. Jim Crow etiquette included hierarchical modes of address, where blacks referred to whites in positions of authority as "Boss", and their children as "Massa", but white employers and customers referred to black workers as "Boy" or "Uncle" or "Old Man" (regardless of age). A simple, concrete example could be of bus seats: the unequal convention is that black riders would sit in the back of the bus and as more whites boarded the bus would give up their seat to white riders. The equal convention is that whoever arrives first retains the seat, but many others could be constructed.

The desegregation of the U.S. South in the 1960s shows the importance of intentional deviations that, prior to formal recognition by government, change a status-quo convention. The eruption of boycotts, sit-ins, pickets and other challenges to racial norms (Figure 7) eventually won the passage of the Civil Rights Act of 1964 (McAdam, 1990). Before the formal legal changes, however, the protests immediately forced whites in some localities to change their behavior. Wright reports that "[Dallas] civic leaders responded to picketing by arranging for blacks to be served at forty-nine downtown restaurants on July 26, 1961, followed by removal of white-only signs." Figure 8 shows the extent of de facto desegregation in the South prior to the first federal legislation in 1964. Segregation had been a convention, conformism to which was initially an individual best response for both whites and blacks. But when large numbers of blacks rejected the convention, this altered the best response calculations of whites, inducing them to abandon racial norms of exclusion.

The importance of the civil rights movement was not just in winning legislation from the federal government, but also in changing norms in the South. Wright (2013, pp. 67) writes "when the demonstrators showed their persistence by returning for new rounds of protest, business and civic leaders in many cities were ready to acquiesce". As can be seen

from Figure 7, black students and citizens began protesting the Jim Crow institutions of the south beginning in the 1950s, with little success or momentum. However, as McAdam (1983) notes, beginning in 1960, a wave of protest took over in the South. Louis Oberdorfer, assistant attorney general in the early 1960s, wrote in his report to the federal government that "after demonstrations by Negroes, positive, concrete steps have been taken towards the meeting of their just demands by desegregating movie theaters and removing racial signs from restrooms.". White citizens in over 100 cities voluntarily desegregated lunch counters within a year of the first sit in (Wright, 2013).

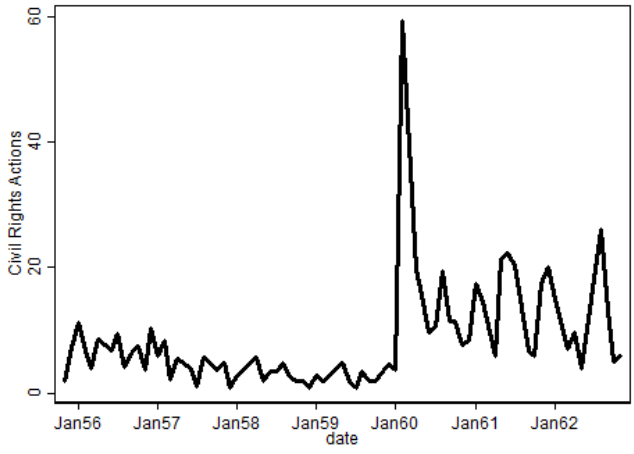

Figure 7: **Civil rights movement actions.** Data from McAdam (1983).

Data from Oberdorfer's reports (Oberdorfer, 1963) documents that the diffusion of these norms was uneven, with some cities desegregating earlier than others. The subsequent pattern of desegregation showed the "local uniformity" and "global heterogeneity" that characterize models of local updating Young and Burke (2001).[7] Oberdorfer continued his description of the changes in the South prior to 1965 with:

> "As desegregation has so rapidly spread, a curious patchwork pattern of desegregated establishments has been created. In some cities movie theatres are desegregated while in others, everything is except theaters. Lunch-counters are the

[7] Young and Burke (2001, pp. 560) describe similar patterns in their evolutionary account of sharecropping contracts in Illinois "There are regional "patches" where contractual terms are nearly uniform, separated by boundaries where contractual norms jump substantially from one set of terms to another."

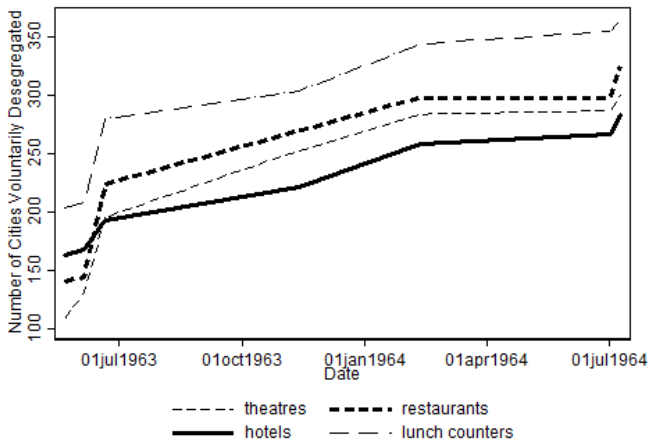

Figure 8: **Voluntary Desegregations in Southern cities prior to July 2, 1964 Civil Rights Act.** Data from reports to Attorney General (Oberdorfer, 1963).

> most desegregated facility; yet restaurants are the least. In a number of towns, four wall movie houses desegregated without incident yet drive-in theaters in these same towns remain segregated. There are places where motels are desegregated, hotels are not; in others, hotels are but motels are not. In some motels or hotels, Negroes can rent rooms but can not use the restaurant; in some others, they can eat but not sleep. Some hotels or motels open their doors of their restaurants to "transient" Negroes, but not to those who are residents of the area,"

Our model asks what features of the Jim Crow equilibrium made it stable for 100 years after the end of slavery, what made it vulnerable to mass norm violation, as well as accounting for its sudden yet patchwork erosion across venues even within a town. The model echoes forces that other scholars (Chong, 1991) have suggested as contributing factors: organization by churches, unions, and the NAACP, which generated intentional deviations from Jim Crow conventions, the smaller population of blacks, particularly following the waves of the Great Migration in the preceding 50 years, the increased urbanization and industrialization of the South, as well as the expansion of national media, changed the interaction structure between blacks and whites from the previous rural economy, making blacks less susceptible to idiosyncratic local white behavior. This occured in part by expanding the once "parochial" scope of vision of rural African-Americans who moved to the city (McAdam, 1990). Importantly, as

Wright (2013) argues, market and technological changes made it so that many whites stood to gain from de-segregation (decreasing a_R), which plays an important role in determining which conventions persist in the long-run.

Theorem 2 gives conditions that could have made segregation a durable convention. The salience of race guaranteed a high degree of intentionality ι, as both whites and blacks were unlikely to experiment with conventions that went against that group interest. In addition, blacks were both a large share of the population and not highly mobilized for much of the post-Reconstruction period, and this resulted in a high η. Finally, the population interaction structure made blacks very vulnerable to idiosyncratic play by whites, as many different black communities had to deal with only a few landowning and money-lending whites. Idiosyncratic play by a few whites could induce many black communities to acquiesce to segregation, and this made transitioning to the segregated strategy easier for all the other whites with whom these communities interacted.

The Southern interaction structure was fragile for blacks, contributing to the stability of the Jim Crow equilibrium. Geographic and social segregation ensured that rural black communities relied on whites as a bridge to information or trade with the larger economy, for example landowners or storeowners, who in turn interacted with many black communities. The transition from Reconstruction to Jim Crow, while certainly enshired in law, also entailed a decentralized change in racial conventions. Thus a small amount of idiosyncratic U behavior by the whites (e.g. demanding deferential behavior of blacks, such as relinquishing seats) could induce a set of blacks to respond by acquiescing. The local whites who shared a large proportion of their cross-race interactions with that set of blacks will then find it optimal to play U also, generating a local stable pocket of segregation that was vulnerable to further idiosyncratic behavior of whites. Thus Jim Crow could diffuse even in the absence of government statute.

There are many historical examples of communities that held out against the propagation of racial norms in the pre-civil rights U.S. South (Dittmer, 1994). For example, Mound Bayou, Mississippi, was a town in the Mississippi Delta founded by ex-slaves in 1887, and close to 100% black. Notably, even the plantation owners in the town were black, the town had no racial codes, and even white visitors complied with the desegregated convention.

It became a springboard the civil rights movement starting in the early 1950s, holding annual rallies, housing civil rights leaders from out of the Delta, and producing many key boycott organizers.While economic benefits, violence, and government fiat were undoubtedly important in changing the payoffs to various strategies, our model shows how these payoff changes combine with *P-fragility* in the interaction structure and intentional idiosyncratic play to stabilize unequal conventions.

Theorem 2 also shows conditions that may have increased the likelihood of a transition from the segregated convention to the unsegregated one. A fall in a_R, for example, which would be a fall in the benefits to whites of segregation, owing to farm mechanization, lack of external investment or federal sanctions, would lower the thresholds ι^*, η^* as well as increasing the set of networks with the required level of *P-fragility*. All these forces would increase the likelihood of a transition, holding ι and η constant. NAACP organizing and black church consciousness raising could have increased ι, while the earlier Great Migration and civil rights mobilization would have both lowered η (recall that η captures both the (inverse) rate of idiosyncratic play as well as the relative population size) (Hornbeck and Naidu, 2014; McAdam, 1983).

The transition paths generated in the model in this section look like the actual historical transition paths in the U.S. South, where, starting from the U convention, small clusters playing E formed and gradually diffused.

In our model, Mound Bayou mentioned above was an early intermediate state consisting of a small cohesive cluster of E playing agents in a larger population of agents playing U. But this cluster interacted with changing payoffs and networks during the civil rights era: it made it easier for adjacent poor agents, such as the Mound Bayou insurance agent Medgar Evers, to idiosyncratically play E, for example boycotting segregated gas stations in the 1950s, propagating the E convention to other white populations. While there are many differences between our stylized model and the dynamics of historical desegregation, the percolation of new, egalitarian racial codes across the South prior to the Civil Rights Act in response to idiosyncratic deviations of black agents is a feature of our model that other models do not capture as naturally.

We think this pattern is not unique to the civil rights movement or changing norms

of racial segregation, but instead is a common feature of rapid cultural changes that upset unequal conventions (as in the serfdom and apartheid examples above). Deliberate violations of norms induce a small location to change equilibrium, and this new equilibrium percolates through the whole population depending on the pattern of interactions.

6 Conclusion

Unequal social norms, such as racial, gender, and labor market conventions are present throughout history, and transitions between them often the outcome of intentional deviations from the status quo. The standard model of equilibrium selection in evolutionary games provides guidance for the emergence and persistence of conventions in single-population coordination game environments where idiosyncratic play is likely to be undirected. But an evolutionary model with directed idiosyncratic behavior of the type we have presented may be more appropriate for selecting among conventions with distributional consequences, where it is likely that individuals are influenced by organizations and ideologies that account for group interests in alternative equilibria.

We have shown that enriching the standard evolutionary model with intentional idiosyncratic play, differential group size and explicit network structure yields predictions in the contract game consistent with a wider range of historical experiences: when intentionality (measured by ι) is high, transitions will be driven by those who stand to gain from the transition, larger and less mobilized populations will be disfavored, and lack of information among the less well off will result in unequal conventions being selected.

The enhanced evolutionary model also provides a framework for studying influences on the political economy of inequality stressed by historians and social scientists. Included is the privileged access to information that elite groups often enjoy and the way that ideology, organization, and leadership may affect the frequency and intentionality of deviations from a status quo unequal convention.

In Appendix D, we explore an extension of our model that endogenizes population structure, generating a relationship between intergenerational mobility and cross-sectional inequality. In the extension, we show that there is a barrier to upward mobility out of the dis-

advantaged class sufficient to stabilize an unequal convention. This occurs because restricted mobility enlarges the poorer class sufficiently to make a high level of inequality stochastically stable, even if the convention implementing it is not risk-dominant. Our model's evolutionary dynamic thus yields a simple version of the "Great Gatsby Curve" (Corak, 2012), where low intergenerational mobility is associated with high cross-sectional inequality.

We have provided a number of historical examples of deliberate deviations from the status quo inducing a transition to a more egalitarian convention. In our examples, centralized collective action at the population level played little role, and *de jure* changes in laws were likewise not important. While we leave the formulation of precise empirical tests to future work, one distinct feature of our theory is the formulation of explicit dynamics of change (and attempted, unsuccesful, change), something that existing non-evolutionary models have a difficult time capturing. Empirical work examining dynamic patterns of changes in social norms (for example the linguistic changes studied in Naidu et al. (2017)) is likely to provide an informative test of our model against, say, the standard evolutionary model. Finally, while we have deliberately abstracted from the role of formal laws and sanctions in modelling the evolution of unequal social norms, it is not that we think these are unimportant. Extending the model to allow richer interactions between formal laws and informal social norms is an important direction for future work.

References

Acemoglu, D. and M. O. Jackson (2014). History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies 82*(2), 423–456.

Acemoglu, D. and J. Robinson (2006). *Economic Origins of Dictatorship and Democracy*. Cambridge University Press.

Acemoglu, D. and J. Robinson (2008). Persistence of power, elites and institutions. *The American economic review 98*(1), 267–293.

Alesina, A., P. Giuliano, and N. Nunn (2011). On the origins of gender roles: Women and the plough. Technical report, National Bureau of Economic Research.

Bailey, M. (2014). *The decline of serfdom in late medieval England: from bondage to freedom*. Boydell & Brewer Ltd.

Becker, G. S. (1981). *A Treatise on the Family*. Harvard University Press.

Bergin, J. and B. Lipman (1996). Evolution with state-dependent mutations. *Econometrica*, 943–956.

Bewley, T. F. (1999). *Why wages don't fall during a recession*. Harvard University Press.

Binmore, K., L. Samuelson, and P. Young (2003). Equilibrium selection in bargaining models. *Games and economic behavior 45*(2), 296–328.

Bisin, A. and T. Verdier (2000). "beyond the melting pot": Cultural transmission, marriage, and the evolution of ethnic and religious traits. *The Quarterly Journal of Economics 115*(3), 955–988.

Blum, J. (1971). *Lord and Peasant in Russia*. Princeton University Press.

Borgerhoff Mulder, M., S. Bowles, T. Hertz, A. Bell, J. Beise, G. Clark, I. Fazzio, M. Gurven, K. Hill, P. L. Hooper, et al. (2009). Intergenerational wealth transmission and the dynamics of inequality in small-scale societies. *science 326*(5953), 682–688.

Bowles, S. (2004). *Microeconomics: Behavior, Institutions, and Evolution*. Princeton: Princeton University Press.

Brown, R. and A. Gilman (1960). The pronouns of power and solidarity. style in language, ed. by thomas a. sebeok, 253-76.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics 129*(4), 1553–1623.

Chong, D. (1991). *Collective Action and the Civil Rights Movement*. University of Chicago Press.

Clyne, M., C. Norrby, and J. Warren (2009). *Language and human relations: Styles of address in contemporary language*. Cambridge University Press.

Corak, M. (2012). Inequality from generation to generation: The united states in comparison. In R. Rycroft (Ed.), *The Economics of Inequality, Poverty, and Discrimination in the 21st Century*. ABC-CLIO.

Dittmer, J. (1994). *Local people: The struggle for civil rights in Mississippi*, Volume 82. University of Illinois Press.

Edgerton, R. B. (1992). *Sick societies*. Simon and Schuster.

Ellison, G. (1993). Learning, local interaction, and coordination. *Econometrica 61*(5), 1047–1071.

Ellison, G. (2000, January). Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. *Review of Economic Studies 67*(1), 17–45.

Foster, D. and H. P. Young (1990). Stochastic evolutionary game dynamics. *Theoretical Population Biology 38*(2), 219–232.

Fryde, E. (1996). *Peasants and Landlords in later Medieval England, C. 1380-C. 1525*. Stroud, Gloucestershire: Alan Sutton Publishing.

Gellner, E. (1983). *Nations and nationalism*. New perspectives on the past. Ithaca: Cornell University Press.

Goody, J. R., J. Thirsk, and E. P. Thompson (1976). Family and inheritance. *Past and Present Publications (USA)*.

Gulati, M. and R. E. Scott (2012). *The three and a half minute transaction: Boilerplate and the limits of contract design*. University of Chicago Press.

Hanes, C. (1993). The development of nominal wage rigidity in the late 19th century. *The American Economic Review*, 732–756.

Hornbeck, R. and S. Naidu (2014). When the levee breaks: Black migration and economic development in the american south. *The American Economic Review 104*(3), 963–990.

Hwang, S.-H., W. Lim, P. Neary, and J. Newton (2018). Conventional contracts, intentional behavior and logit choice. Working paper.

Jackson, M. O. and A. Watts (2002). On the formation of interaction networks in social coordination games. *Games and Economic Behavior 41*(2), 265–291.

Jayachandran, S. (2015). The roots of gender inequality in developing countries. *Annual Reviews of Economics 7*(1), 63–88.

Kandori, M. G., G. Mailath, and R. Rob (1993). Learning, mutation, and long run equilibria in games. *Econometrica 61*, 29–56.

Kreindler, G. and H. P. Young (2011). Fast convergence in evolutionary equilibrium selection. *Economics Series Working Papers*.

Lim, W. and P. Neary (2016). An experimental investigation of stochastic adjustment dynamics. Working paper.

Mariotti, M. (2012). Labour markets during apartheid in south africa. *The Economic History Review 65*(3), 1100–1122.

Markoff, J. (1996). *The Abolition of Feudalism: Peasants, Lords, and Legislators in the French Revolution*. Pennsylvania State Univ Pr.

Mäs, M. and H. H. Nax (2016). A behavioral study of "noise" in coordination games. *Journal of Economic Theory 162*, 195–208.

McAdam, D. (1983). Tactical innovation and the pace of insurgency. *American Sociological Review*, 735–754.

McAdam, D. (1986). Recruitment to high-risk activism: The case of freedom summer. *American Journal of Sociology*, 64–90.

McAdam, D. (1990). *Political Process and the Development of Black Insurgency, 1930-1970*. University of Chicago Press.

Morris, S. (2000). Contagion. *Review of Economic Studies 67*, 57–78.

Naidu, S., S.-H. Hwang, and S. Bowles (2010). Evolutionary bargaining with intentional idiosyncratic play. *Economics Letters 109*(1), 31–33.

Naidu, S., S.-H. Hwang, and S. Bowles (2017). The evolution of unequal linguistic conventions. *American Economic Review Papers and Proceedings* (5), 31–35.

Naidu, S. and N. Yuchtman (2013). Coercive contract enforcement: Law and the labor market in nineteenth century industrial britain. *The American Economic Review 103*(1), 107–144.

North, D. C. and R. P. Thomas (1971). The rise and fall of the manorial system: A theoretical model. *The Journal of Economic History 31*(04), 777–803.

Nunn, N. (2014). Historical development. *Handbook of Economic Growth 2*, 347–402.

Oberdorfer, L. (1963). United states government memorandum re: Report of desegregation activity.

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Parsons, T. (1964). Evolutionary universals in society. *American Sociological Review 3*(29), 339–57.

Poos, L. R. (1991). *A rural society after the Black Death: Essex 1350-1525*. Number 18. Cambridge University Press.

Scott, J. C. (1985). *Weapons of the weak: Everyday forms of peasant resistance*. yale university Press.

Shiryaev, A. (2000). *Probability*.

Silbey, S. S. (2010). J. locke, op. cit.: Invocations of law on snowy streets. *J. Comp. L. 5*, 66.

van Damme, E. and J. W. Weibull (2002, October). Evolution in games with endogenous mistake probabilities. *Journal of Economic Theory 106*(2), 296–315.

Wood, E. (2003). *Insurgent Collective Action and Civil War In El Salvador*. Cambridge: Cambridge University Press.

Wright, G. (2013). *Sharing the Prize*. Harvard University Press.

Young, H. P. (1993a, January). The evolution of conventions. *Econometrica 61*(1), 57–84.

Young, H. P. (1993b). The evolution of conventions. *Econometrica 61*(1), 57–84.

Young, H. P. (1998a, October). Conventional contracts. *Review of Economic Studies 65*(4), 773–92.

Young, H. P. (1998b). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press.

Young, H. P. (2011). The dynamics of social innovation. *Proceedings of the National Academy of Sciences 108*, 21285–21291.

Young, H. P. and M. Burke (2001). Competition and custom in economic contracts: A case study of illinois agriculture. *American Economic Review 91*(3), 559–573.

A Appendix (Not For Publication)

To economize on notation, throughout the Appendix we let

$$\zeta_P := \frac{b}{b+a_P}, \quad \zeta_R := \frac{b}{b+a_R}$$

i.e., ζ_P is the minimum fraction of rich agents in the poor village which induces the poor agents switch their best responses from E to U. ζ_R can be interpreted similarly. Observe that from our assumption in Section 2.1, $a_P < b < a_R$, we have $\frac{1}{2} < \zeta_P < 1$ and $0 < \zeta_R < \frac{1}{2}$.

We begin by re-stating and proving Proposition 1.

Proposition 4. *There exist $\iota^* > 1, \bar{\eta} > 1$ and $\eta^* > 1$ such that:*
(i) Suppose that E is risk-dominant. Then U is stochastically stable if and only if $\iota > \iota^*$ and $\eta > \eta^*$.
(ii) Suppose that U is risk-dominant. Then E is stochastically stable if and only if $\iota < \iota^*$ and $\eta > \bar{\eta}$.

Proof. For the convenience of notations, let $\alpha := 1 - \zeta_R$ and $\beta := \zeta_P$. Then, $\alpha > \frac{1}{2}$ and $\beta > \frac{1}{2}$ hold. We let $\hat{c}(U,E) = \min\{\eta\alpha, \iota(1-\beta)\}$ and $\hat{c}(E,U) = \min\{\beta, \iota\eta(1-\alpha)\}$. We first prove (i). Notice that E is risk-dominant if and only if $\alpha < \beta$. We let

$$\iota^* := \frac{\beta}{1-\beta} \qquad \eta^* := \frac{\beta}{\alpha}$$

Then, if $\iota > \iota^*$, then $\iota(1-\beta) > \beta$ and if $\eta > \eta^*$, then $\eta\alpha > \beta$. Thus, U is stochastically stable and "if part" follows. To show "only if part", we show that if $\iota < \iota^*$ or $\eta < \eta^*$, then E is stochastically stable (when $\iota = \iota^*$ and $\eta = \eta^*$, the similar argument shows that both U and E are stochastically stable and hence the desired result follows). Suppose that $\iota < \iota^*$ or $\eta < \eta^*$. We divide cases.
Case 1: $\eta\alpha < \iota(1-\beta)$ and $\beta < \iota\eta(1-\alpha)$. In this case, we have $\hat{c}(U,E) = \eta\alpha, \hat{c}(E,U) = \beta$. If $\eta < \eta^*$, then $\eta\alpha < \beta$, thus E is stochastically stable. If $\iota < \iota^*$, then $\eta\alpha < \iota(1-\beta) < \beta$ and again $\eta\alpha < \beta$, thus E is stochastically stable.
Case 2: $\eta\alpha > \iota(1-\beta)$ and $\beta < \iota\eta(1-\alpha)$. Then, $\hat{c}(U,E) = \iota(1-\beta), \hat{c}(E,U) = \beta$. Thus if $\iota < \iota^*$, then E is stochastically stable. If $\eta < \eta^*$, then $\iota(1-\beta) < \eta\alpha < \beta$. Thus again, E is stochastically stable.
Case 3: $\eta\alpha > \iota(1-\beta)$ and $\beta > \iota\eta(1-\alpha)$. Then, $\hat{c}(U,E) = \iota(1-\beta), \hat{c}(E,U) = \iota\eta(1-\alpha)$. Since $\eta \geq 1$, $\iota\eta(1-\alpha) > \iota(1-\beta)$ and thus E is stochastically stable.
Case 4: $\eta\alpha < \iota(1-\beta)$ and $\beta > \iota\eta(1-\alpha)$. In this case, we have $\alpha < \iota(1-\beta)$ and $\beta > \iota(1-\alpha)$. Thus we have $\frac{\beta}{1-\alpha} > \frac{\alpha}{1-\beta}$ which implies that $\beta - \alpha > (\beta-\alpha)(\beta+\alpha)$ and since $\beta > \alpha$, this implies $\alpha + \beta < 1$ which is a contradiction. Thus Case 4 cannot occur. When $\eta\alpha = \iota(1-\beta)$ or $\eta = \eta^*$, the similar arguments as above hold and thus we obtain the desired result.

We next prove (ii). Then U is risk-dominant if and only if $\alpha > \beta$. We let

$$\iota^* := \frac{\beta}{1-\beta} \qquad \bar{\eta} := \frac{1-\beta}{1-\alpha}.$$

Then suppose that $\iota < \iota^*$ and $\eta > \bar{\eta}$. Then $\iota\eta(1-\alpha) > \iota(1-\beta)$ and $\beta > \iota(1-\beta)$. Thus E is

stochastically stable and "if part" follows. To show "only if part", we show that if $\iota > \iota^*$ or $\eta < \bar{\eta}$, then U is stochastically stable (again when $\iota = \iota^*$ and $\eta = \eta^*$, the similar argument shows that both U and E are stochastically stable and hence the desired result follows). Suppose that $\iota > \iota^*$ or $\eta < \bar{\eta}$. We divide cases.
Case 1: $\eta\alpha < \iota(1-\beta)$ and $\beta < \iota\eta(1-\alpha)$. In this case, we have $\hat{c}(U,E) = \eta\alpha, \hat{c}(E,U) = \beta$. Since $\alpha > \beta$, $\eta\alpha > \beta$. Thus U is stochastically stable.
Case 2: $\eta\alpha > \iota(1-\beta)$ and $\beta < \iota\eta(1-\alpha)$. Then, $\hat{c}(U,E) = \iota(1-\beta), \hat{c}(E,U) = \beta$. Thus if $\iota > \iota^*$, then U is stochastically stable. If $\eta < \bar{\eta}$, then $\beta < \iota\eta(1-\alpha) < \iota(1-\beta)$. Thus again, U is stochastically stable.
Case 3: $\eta\alpha > \iota(1-\beta)$ and $\beta > \iota\eta(1-\alpha)$. Then, $\hat{c}(U,E) = \iota(1-\beta), \hat{c}(E,U) = \iota\eta(1-\alpha)$. If $\iota > \iota^*$, then $\iota(1-\beta) > \beta > \iota\eta(1-\alpha)$. Thus U is stochastically stable. If $\eta < \bar{\eta}$, $\iota(1-\beta) > \iota\eta(1-\alpha)$. Thus U is stochastically stable.
Case 4: $\eta\alpha < \iota(1-\beta)$ and $\beta > \iota\eta(1-\alpha)$. In this case, we have $\hat{c}(U,E) = \eta\alpha, \hat{c}(E,U) = \iota\eta(1-\alpha)$. Since $\eta\alpha > \beta > \iota\eta(1-\alpha)$, U is stochastically stable.

□

A.1 Cohesiveness and Stochastic Stability of the E Convention

In this Appendix we define *R-fragility* analogously to *P-fragility* in the text, and provide conditions on $\Lambda = (\Lambda_R, \Lambda_P)$ under which the E convention is stochastically stable.

Definition 5. *We say that a bipartite graph is R-fragile* (q_P, q_R) *if*
(i) For some S_R*, every* S'_R *containing* S_R *is* q_P*-cohesive with* $N(S'_R)$*.*
(ii) Every $N(S_R)$ *is* q_R*-weak-cohesive with* $(S_R)^c$*.*

Note that, compared with Definition 2, q_P in (i) and q_R in (ii) are switched. This is because the network condition (i) stabilizes the deviant play of the poor agents, hence need to be compared with the payoff condition stabilizing the deviant play (E) of the poor agents ($1-\zeta_P$) and the network condition (ii) stabilizes the reacting play (E) of the rich agents, hence need to be compared with the payoff condition stabilizing the reacting play of the rich agents ($1-\zeta_R$).
We then have the following characterization for the equal convention.

Proposition 5. *Suppose that the bipartite graph is R-fragile* $(1-\zeta_P, 1-\zeta_R)$*. Suppose also that*

$$\frac{\min_{x\in\Lambda_P}|N_x|}{\min_{y\in\Lambda_R}|N_y|} \geq \frac{1-\zeta_R}{\zeta_P}. \tag{17}$$

and $\eta = 1$*. There exists* $\iota^* < 1$ *such that for all* $\iota < \iota^*$*,* E *is stochastically stable.*

Proof. See the appendix in B. □

The additional restriction on minimum vertex degree is necessary in the case of the E convention because we have imposed a lower bound of 1 on η. We also have the following corollary.

Corollary 1. *Suppose that the bipartite graph is R-fragile* $(1-\zeta_P, 1-\zeta_R)$ *and* $\eta = 1$ *and* $\min_{x\in\Lambda_P} |N_x| \geq \min_{y\in\Lambda_R} |N_y|$ *Then there exist* $\iota^* < 1$ *such that for all* $\iota < \iota^*$*, if* E *is risk-dominant, then* E *is stochastically stable.*

Proof. The condition that E is risk-dominant if and only if $b^2 > a_R a_P$ if and only $1 > \frac{1-\zeta_P}{\zeta_R}$. Thus the result immediately follows from Proposition 5. □

A.2 Proofs

Before we prove Theorem 2, we show Lemma 1.

Proof of Lemma 1. If $\iota < \infty$, then the chain can reach any state with a positive probability, hence the chain is irreducible. Suppose that $\iota = \infty$. In this case, the poor agents idiosyncratically play E only, while the rich agents idiosyncratically play U only. We show that the chain has only one recurrent class. First, since the chain is finite, there exists at least one recurrent class. Recall that $\mathcal{E}_U$ and $\mathcal{E}_E$ be the states where every agent uses U and E, respectively. Then by the intentional idiosyncratic behaviors, $\mathcal{E}_U$ and $\mathcal{E}_E$ communicate and belong to the same recurrent class, denoted, $\mathcal{R}$. Let $\sigma \notin \mathcal{R}$. Then again by the intentional idiosyncratic behaviors of either poor agents or rich agents, the chain can reach either $\mathcal{E}_E$ or $\mathcal{E}_U$ with a positive probability. Since $\sigma \notin \mathcal{R}$, this shows that σ is transient. Thus there exists a unique recurrent class, $\mathcal{R}$, containing $\mathcal{E}_E$ and $\mathcal{E}_U$. Since the chain has a unique (positive) recurrent class, the standard results on Markov chains show that there exists a unique invariant measure (see, e.g., Shiryaev (2000)). □

A.3 Proof of Theorem 2

We prove Theorem 2 using a series of lemmas. First, we define some notations. Recall that

$$\sigma(x) = (\sigma_1(x), \sigma_2(x), \cdots, \sigma_\eta(x)) \text{ for all } x \in \Lambda_P$$

and

$$\sigma = (\{\sigma(x)\}_{x\in\Lambda_P}, \{\sigma(y)\}_{y\in\Lambda_R})$$

Let $\beta_x(\sigma)$ for $x \in \Lambda_P$ be the best responses of the agents at x at σ and $\beta_y(\sigma)$ for $y \in \Lambda_R$ be the best response of the agent at y at σ. We let $\beta : \Xi \to \Xi$ be the map in which $\beta(\sigma)$ is the state obtained by making all the agents to play the best responses at σ: i.e.,

$$\beta(\sigma) = (\{\beta_x(\sigma)\}_{x\in\Lambda_P}, \{\beta_y(\sigma)\}_{y\in\Lambda_R})$$

Then we say that $\mathcal{E}$ is a stable state if $\beta(\mathcal{E}) = \mathcal{E}$. We denote by $\mathcal{E}_E$ and $\mathcal{E}_U$ the stable states where all agents play E and U, respectively (see Panel A in Figure 4). We also denote by $\mathcal{E}_M$ be the stable state in which some agents play E and other agents play U. Finally, we let $\#(\sigma)$ be the number of sites, whether poor or rich, at which agents play strategy U:

$$\#(\sigma) = \sum_{x\in\Lambda_P} \mathbf{1}\{\sigma(x) = U\} + \sum_{y\in\Lambda_R} \mathbf{1}\{\sigma(y) = U\}$$

where $\mathbf{1}_A = 1$ if A is true and $\mathbf{1}_A = 0$, otherwise.

Using the definition of q_P and q_R given above, the condition of theorem 2 is that the bipartite graph is *P-fragile* (q_P, q_R). Then, the following two conditions hold:

A1 There exists a S_P and there exists q_R such that for all $S'_P \supset S_P$

$$\min_{y \in N(S'_P)} \frac{|N_y \cap S'_P|}{|N_y|} \geq q_R$$

A2 For every S_P, there exists $x' \notin S_P$ such that

$$\frac{|N_{x'} \cap N(S_P)| + 1}{|N_{x'}|} \geq q_P$$

In the following proof, we will use the following facts:

If $\lceil |N_x| \zeta_P \rceil$ of the rich agents in the neighborhood of the poor village at x play U, then $\beta_x(\sigma) = U$ (18)

If $\lceil \eta |N_y| \zeta_R \rceil$ of the poor agents in the neighborhood of the rich site at y play U, then $\beta_y(\sigma) = U$ (19)

Let S_P be the set satisfies **A1**. We let $S_R^0(S_P)$ be the minimum size set of rich agents which ensures that each poor in S_P has at least $\lceil |N_x| q_P \rceil$ interactions with the rich agents in the set. That is, the set, $S_R^0(S_P) \subset N(S_P)$ is the minimum size set which is q_P-cohesive with S_P:

$$|S_R^0(S_P)| := \min\{|S_R| : \min_{x \in S_P} \frac{|N_x \cap S_R|}{|N_x|} \geq q_P\} \tag{20}$$

where S_P is given by **A1**. Indeed, for $x \in S_P$, $|N_x \cap S_R| \geq |N_x| q_P$ implies $|N_x \cap S| \geq \lceil |N_x| q_P \rceil$ (since $|N_x \cap S|$ is an integer). Thus, each poor x in S_P has at $\lceil |N_x| q_P \rceil$ interaction with the rich agents in $S_R^0(S_P)$.

Lemma 4. *Suppose that **A1** holds.*

$$\min_{\mathcal{E}' \neq \mathcal{E}_E} C(\mathcal{E}_E, \mathcal{E}') \leq |S_R^0(S_P)|.$$

Proof. We will construct a path γ from $\mathcal{E}_E$ to some stable state $\mathcal{E}'$. Let $S_P^* := S_P$ be the set given by **A1**. Suppose that rich agents in $S_R^* := S_R^0(S_P)$(in (20)) idiosyncratically play U. We denote by σ' the new state obtained by these switches (Panel B in Figure 4). We let $\mathcal{E}' = \beta^l(\sigma')$ for some $l \geq 2$ such that $\beta^{l+1}(\sigma') = \beta^l(\sigma')$(Panel C in Figure 4). Here observe that from σ' to $\beta(\sigma')$ the poor village agents change their best responses, while from $\beta(\sigma')$ to $\beta^2(\sigma')$ the rich agents change their best responses and so on. Since each $x \in S_P^*$ interacts with $\lceil |N_x| q_P \rceil$ number of U playing rich agents,

$$\beta_x(\sigma') = U \text{ for all } x \in S_P^*. \tag{21}$$

which implies that

$$\beta_x^2(\sigma') = U \text{ for all } x \in S_P^*. \tag{22}$$

Now let $y \in N(S_P^*)$. Then from A1,

$$\frac{|N_y \cap S_P^*|}{|N_y|} \geq q_P \iff \eta|N_y \cap S_P^*| \geq \lceil \eta|N_y|q_P \rceil \tag{23}$$

since $\eta|N_y \cap S_P^*|$ is an integer. Then from $\beta_x(\sigma') = U$ for all $x \in S_P^*$, there are at least $\eta|N_y \cap S_P^*|$ number of the poor village agents in the neighbor of y playing U at the state, $\beta(\sigma')$; thus $\beta_y(\beta(\sigma')) = U$. Thus we have

$$\beta_y^2(\sigma') = U \text{ for all } y \in N(S_P^*). \tag{24}$$

Thus, (22) and (24) show that

$$\beta_x^l(\sigma') = U \text{ for all } x \in S_P^*, \;\; \beta_y^l(\sigma') = U \text{ for all } y \in N(S_P^*)$$

which shows that $\mathcal{E}' \neq \mathcal{E}_E$. Now we let

$$\gamma : \mathcal{E}_E \to \sigma' \to \beta(\sigma') \to \beta^l(\sigma') = \mathcal{E}'$$

Then $c(\gamma) = |S_R^0(S_P^*)|$, where $c(\gamma)$ is the cost of a path γ and we obtain the desired result. □

Next, we show the following lemma.

Lemma 5. *Suppose that **A1** and **A2** hold. Then we have*

$$C(\mathcal{E}_M, \mathcal{E}') = 1 \text{ for some } \mathcal{E}' \text{ such that } \#(\mathcal{E}') > \#(\mathcal{E}_M)$$

Proof. Let $\mathcal{E}_M$ be given. Then by the definition of $\mathcal{E}_M$ and from Lemma 4, there exists $S_P^* \times S_R^* := S_P^* \times N(S_P^*)$ such that for all $x \in S_P^*$, $\sigma(x) = U$, for all $x' \notin S_P^*$, $\sigma(x') = E$, and for all $y \in S_R^* := N(S_P^*)$, $\sigma(y) = U$. Then from A2, there exists $x' \notin S_P^*$ such that

$$|N_{x'} \cap N(S_P^*)| \geq \lceil |N_{x'}|q_P \rceil - 1$$

Since $x' \notin S_P^*$, $\sigma(x') = E$. Then we chooses $y \in N_{x'}$ such that $\sigma(y) = E$. (such y exists, otherwise for all $y \in N_{x'}$, $\sigma(y) = U$ implies $\sigma(x) = U$, a contradiction.) Let σ' be the state induced by one idiosyncratic play of the rich agent at y' from E to U and let $\mathcal{E}' = \beta^l(\sigma')$ for some $l \geq 2$ such that $\beta^l(\sigma') = \beta^{l+1}(\sigma')$. Then out of $|N_{x'}|$ neighbors of x', there are (at least) $|N_{x'} \cap N(S_P^*)|$ number of U playing rich agents (since all agents in $N(S_P^*)$ play U). Thus one idiosyncratic play (from E to U) by the rich agent at $y' \in N_{x'}$ induces a change in the poor agents' best responses at x' from E to U, since $|N_{x'} \cap N(S_P^*)| + 1 \geq \lceil |N_{x'}|q_P \rceil$ (Panels D and E in Figure 4): i.e., $\beta_{x'}(\sigma') = U$ which implies

$$\beta_{x'}^2(\sigma') = U \tag{25}$$

Let $S_P = S_P^* \cup \{x'\}$, $N(S_P) = S_R^* \cup \{N_{x'}\}$. Then from **A2** and (23), S_P is again q_P-cohesive

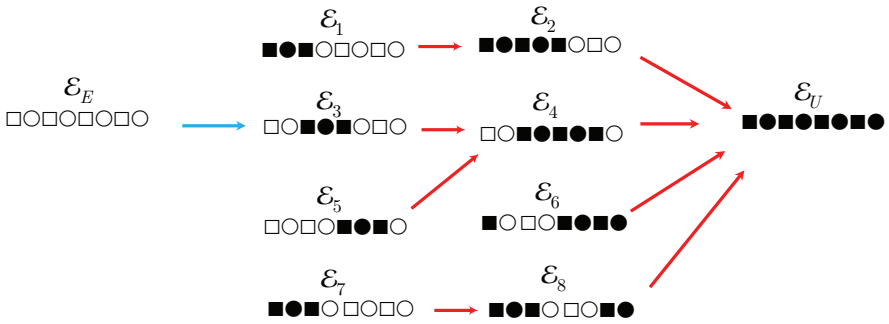

Figure 9: **A minimal tree**. Here we present a minimal tree for the 1-d bipartite graph in Figure 2.

with $N(S_P)$ and we find that

$$\beta_y^2(\sigma') = U \text{ for all } y' \in N(S_P). \tag{26}$$

Thus (25) and (26) show that

$$\beta_x^l(\sigma') = U \text{ for all } x \in S_P^* \cup \{x'\}, \;\; \beta_y^l(\sigma') = U \text{ for all } y \in N(S_P^* \cup \{x'\}),$$

which shows again $\mathcal{E}' \neq \mathcal{E}_M$. Then,

$$\gamma : \mathcal{E}_M \to \sigma' \to \beta(\sigma') \to \beta^l(\sigma') = \mathcal{E}'$$

and observe that $\#(\mathcal{E}') > \#(\mathcal{E}_M)$. Thus, we obtain the desired result. □

Lemma 6. *We have*

$$\min_{\mathcal{E}' \neq \mathcal{E}_U} C(\mathcal{E}_U, \mathcal{E}') \geq \min_{x \in \Lambda_P} \iota \lceil |N_x|(1-\zeta_P) \rceil \wedge \min_{y \in \Lambda_R} \lceil \eta |N_y|(1-\zeta_R) \rceil$$

where $a \wedge b := \min\{a, b\}$.

Proof. Observe that to escape $\mathcal{E}_U$, either the rich agents or the poor agents play idiosyncratically which induce the other population agents to switch their best responses. By the definitions of q_P and q_R, we see that $\min_{x \in \Lambda_P} \lceil |N_x|(1-q_P) \rceil$ is the minimum number of the rich agents to induce changes in the best responses of the poor population and the cost of such idiosyncratic plays is given by $\min_{x \in \Lambda_P} \iota \lceil |N_x|(1-q_P) \rceil$. Similarly, the minimum number of the rich agents to induce the changes in the best responses of the poor population is given by $\min_{y \in \Lambda_R} \lceil \eta |N_y|(1-q_R) \rceil$, which give us the desired results. □

We have the following lemma.

Lemma 7. *Suppose that*

$$\textit{For all } \mathcal{E}_M, C(\mathcal{E}_M, \mathcal{E}') = 1 \textit{ for some } \mathcal{E}' \textit{ such that } \#(\mathcal{E}') > \#(\mathcal{E}_M) \textit{ and } \min_{\mathcal{E}' \neq \mathcal{E}_U} C(\mathcal{E}_U, \mathcal{E}') > \min_{\mathcal{E}' \neq \mathcal{E}_E} C(\mathcal{E}_E, \mathcal{E}').$$

Then $\mathcal{E}_U$ *is stochastically stable.*

Proof. The proof uses Proposition 2 in Binmore et al. (2003). Under the hypothesis, we construct a minimal tree whose root is at $\mathcal{E}_U$ (see Figure 9). First consider the state $\mathcal{E}_E$. We find $\mathcal{E}''$ such that $C(\mathcal{E}_E, \mathcal{E}'') = \min_{\mathcal{E}'} C(\mathcal{E}_E, \mathcal{E}')$. We connect $\mathcal{E}_E$ to $\mathcal{E}''$ with an edge. If $\mathcal{E}'' = \mathcal{E}_U$, then we stop. Otherwise applying the fact that $C(\mathcal{E}_M, \mathcal{E}') = 1$ for some $\mathcal{E}'$ successively, we obtain a sequence of edges leading to $\mathcal{E}_U$. In this way, we can construct a sequence of edges from $\mathcal{E}_E$ to $\mathcal{E}_U$. Next, choose $\mathcal{E}_M$ such that $\#(\mathcal{E}_M)$ is smallest. If $\mathcal{E}_M$ already has an edge to $\mathcal{E}'_M$ such that $\#(\mathcal{E}'_M) > \#(\mathcal{E}_M)$, then we choose other $\mathcal{E}_M$. Again using the fact that $C(\mathcal{E}_M, \mathcal{E}') = 1$ for some $\mathcal{E}'$, we can find a sequence of edges leading to $\mathcal{E}_U$. In this way, we define sequences of edges from all the stable set $\mathcal{E}_M$ such that $\#(\mathcal{E}_M)$ is smallest. Next we move on to $\mathcal{E}_M$ such that $\#(\mathcal{E})$ is second smallest. By proceeding in this way, we can construct a $\mathcal{E}_U$ rooted tree (see Figure 9). Then it is easy to see that the constructed tree is a naive minimization tree and hence the desired result follows. □

Now we are ready to prove our main theorem, Theorem 2, which we state again for convenience of readers.

Theorem 1. *Suppose that the bipartite graph is P-fragile* (q_P, q_R)*. Then there exists* ι^* *and* η^* *such that for all* $\iota > \iota^*$ *and* $\eta > \eta^*$*, U is stochastically stable. More precisely,* ι^* *and* η^* *are given by*

$$\iota^* := \frac{|S_R^0(S_P)|}{\min_{x\in\Lambda_P}\lceil |N_x|(1-\zeta_P)\rceil} \text{ and } \eta^* := \frac{|S_R^0(S_P)|}{\min_{y\in\Lambda_R}|N_y|(1-\zeta_R)}$$

where $|S_R^0(S_P)|$ *is defined in* (20)*.*

Proof. Let $\iota > \iota^*$ and $\eta > \eta^*$. Then if we show that

$$\min_{x\in\Lambda_P} \iota \lceil |N_x|(1-q_P)\rceil \wedge \min_{y\in\Lambda_R} \lceil \eta |N_y|(1-q_R)\rceil > |S_R^0(S_P)|$$

Lemmas 4,5,6 and 7 show that U is stochastically stable. First, $\min_{x\in\Lambda_P} \iota \lceil |N_x|(1-q_P)\rceil > |S_R^0(S_P)|$ follows from our choice of ι^*. Next, we show that $\min_{y\in\Lambda_R} \lceil \eta|N_y|(1-q_R)\rceil > |S_R^0(S_P)|$. Observe that $\lceil \eta|N_y|(1-q_R)\rceil \geq \eta|N_y|(1-q_R)$. Thus we obtain $\min_{y\in\Lambda_R}\lceil \eta|N_y|(1-q_R)\rceil \geq \min_{y\in\Lambda_R} \eta|N_y|(1-q_R)$. Thus we find

$$\min_{y\in\Lambda_R}\lceil \eta|N_y|(1-q_R)\rceil \geq \min_{y\in\Lambda_R} \eta|N_y|(1-q_R) > \min_{y\in\Lambda_R} \eta^*|N_y|(1-q_R) > |S_R^0(S_P)|$$

which is the desired result and $\mathcal{E}_U$ is stochastically stable. □

B Proof of Proposition 5

Proof. We show the following three:
(i) Since the bipartite graph is *R-fragile* $(1-\zeta_P, 1-\zeta_R)$ for all $y \in \Lambda_R$, $\{x\}$ is q_P-cohesive with N_y. Thus for all $y \in \Lambda_R$,

$$\min_{x\in N_y} \frac{|N_x \cap \{y\}|}{|N_x|} \geq 1-\zeta_P$$

Thus, we have for all $x \in \Lambda_P$, $1 > |N_x|(1-\zeta_P)$ and hence $1 \geq \max_{x\in\Lambda_P}\lceil |N_x|(1-\zeta_P)\rceil$. Thus one rich playing E ensures that all his neighboring poor agents play E as best responses (autonomous condition).
(ii) Again since the bipartite graph is *R-fragile* $(1-\zeta_P, 1-\zeta_R)$, for every S_R such that

$$\frac{|N_{y'} \cap N(S_R)| + 1}{|N_{y'}|} \geq q_R$$

This implies that for every $y \in \Lambda_R$, there exists $y' \in \Lambda_R$ such that

$$|N_y \cap N_{y'}| \geq \lceil |N_{y'}|(1-\zeta_R)\rceil - 1$$

which ensures the propagation condition.
(iii) First recall that $\zeta_R < 1/2$ and $\zeta_P > 1/2$. Also if (17) holds, then $\min_{x\in\Lambda_P}\lceil |N_x|\zeta_P\rceil > \min_{y\in\Lambda_R}\lceil |N_y|(1-\zeta_R)\rceil$. Thus we have $\iota \min_{x\in\Lambda_P}\lceil |N_x|\zeta_P\rceil > \iota \min_{y\in\Lambda_R}\lceil |N_y|(1-\zeta_R)\rceil$. We choose $\iota^* < 1$ such that for all $\iota < \iota^*$, $\min_{y\in\Lambda_R}\lceil |N_y|\zeta_R\rceil > \iota \min_{y\in\Lambda_R}\lceil |N_y|(1-\zeta_R)\rceil$. Then we have

$$\iota \min_{x\in\Lambda_P}\lceil |N_x|\zeta_P\rceil \wedge \min_{y\in\Lambda_R}\lceil |N_y|\zeta_R\rceil > \iota \min_{y\in\Lambda_R}\lceil |N_y|(1-\zeta_R)\rceil$$

for all $\iota < \iota^*$. Then by following the same steps as in the proof of Theorem 1, we obtain the desired result. □

C Proof of Proposition 3

Proof. Since $\iota = \infty$, from $\mathcal{E}_U$ only the poor agents play idiosyncratically and from this we obtain

$$\min_{\mathcal{E}'} C(\mathcal{E}_U, \mathcal{E}') \geq \left\lceil 2\kappa \frac{a_R}{a_R + b} \right\rceil$$

Next observe that $\lceil 2\frac{b}{b+a_p}\rceil = 2$ rich agents in the neighborhood of a poor agent ensures U to be the best response of the poor agent and that $\lceil 2\kappa\frac{b}{b+a_R}\rceil$ poor agents in the neighborhood of a rich agent ensures U to be the best response of the rich agent. Also notice that from our assumption that $a_P < b$, $\left\lceil 2\frac{b}{b+a_P}\right\rceil = 2$. To estimate $\min_{\mathcal{E}'} C(\mathcal{E}_E, \mathcal{E}')$, first observe that to escape E, $\left\lceil 2\frac{b}{b+a_P}\right\rceil = 2$ number of R agents need to play idiosyncratically. Then $\left\lceil 2\frac{b}{b+a_P}\right\rceil - 1 = 1$ number of P agents who are surrounded by R agents best respond with U. For each R agent who has idiosyncratically played U to retain U as a best response, at least $\left\lceil 2\kappa\frac{b}{a_R+b}\right\rceil$ out of 2κ neighboring P agents need to play U. Thus, to induce $\left\lceil 2\kappa\frac{b}{b+a_R}\right\rceil$ P agents to use U, $\left\lceil 2\kappa\frac{b}{b+a_R}\right\rceil + 1$ surrounding R agents use U, since $\left\lceil 2\frac{b}{b+a_P}\right\rceil = 2$ number of R agents induces the neighboring one poor agent to play U. We show that a state where a cluster of (consecutively located) $\left\lceil 2\kappa\frac{b}{b+a_R}\right\rceil + 1$ R agents and neighboring $\left\lceil 2\kappa\frac{b}{b+a_R}\right\rceil$ P agents play U and all other agents play E is a stable state (see the sites in black in Figure 6). Thus,

we find that

$$\min_{\mathcal{E}'} C(\mathcal{E}_E, \mathcal{E}') \leq \left\lceil 2\kappa \frac{b}{a_R + b} \right\rceil + 1.$$

Next we suppose that $\mathcal{E}'$ be a stable state where some agents play U and other agents play E. Then it must contains the cluster of sites found previously (see again the sites in black in Figure 6). Then there exists one poor agent playing E with one neighboring rich agent playing U and the other neighboring rich agent playing E (if not, we lead to a contradiction) (see the sites in grey in Figure 6). Then by the idiosyncratic play of the rich agent from E to U, the rich agent induces the poor to best respond with U. We let $\mathcal{E}''$ be the resulting state. Also, since $\mathcal{E}'$ is stable, the U playing rich agent faces $\left\lceil 2\kappa \frac{b}{a_R+b} \right\rceil$ poor agents playing U and thus the newly switched rich agent's best response is also U (see Figure 6). Thus $\mathcal{E}''$ is stable. This shows that $C(\mathcal{E}', \mathcal{E}'') = 1$ for some $\mathcal{E}''$ such that $\#(\mathcal{E}'') > \#(\mathcal{E}')$ and by the same argument of the proof of Theorem 2, we have the desired result. □

D Endogeneous Demographics

In this Appendix section, we endogenize the relative size of the two classes, η, and sketch an extension of our model that incorporates an intergenerational mobility dynamic to the convention selection process, following Bisin and Verdier (2000). Here, the equilibrium relative size η^* of the P and R classes depends on the wealth of the two classes which, because a rich individual may interact with more than one poor individual, depends on the relative size of the classes. We focus on the uniform random matching case from Section 2.1 with $\iota = \infty$, as it is more historically relevant when selecting between conventions with large differences in payoffs for each group. In addition, to keep things tractable, we make a simplifying assumption that the time scale of the population adjustment is sufficiently slow relative to the transition times between conventions. This means that to each relative population η, we can associate a stochastically stable convention.

Recall that Proposition 1 (ii) shows that when $\iota = \infty$, U is stochastically stable if and only if

$$\left\lceil \eta N^R \frac{a_R}{a_R + b} \right\rceil > \left\lceil N^R \frac{b}{b + a_P} \right\rceil$$

Then, given the relative size of populations, η, we are able to determine the stochastic stable states. More precisely, let $\Phi^*(\eta)$ be the stochastically stable state given relative population size η:

$$\Phi^*(\eta) = \begin{cases} U \text{ if } \eta \geq \eta^{*,\infty}(\rho, \theta) \\ E \text{ if } \eta < \eta^{*,\infty}(\rho, \theta) \end{cases} \tag{27}$$

Without loss of generality where $\eta = \eta^{*,\infty}(\rho, \theta)$ (from equation (5) in Section 2.1), we assume U is stochastically stable, ensuring $\Phi^*(\eta)$ is unique. In the resulting model, the stochastically stable contract and the relative sizes of the two classes will be jointly determined.

In the intergenerational mobility dynamic, we suppose that each generation randomly matches into mating pairs, has 2 offspring, and then dies. We assume that when parents

belong to the same class, the two offspring retain the parents' class membership and only the cross class couple's offspring can change the class membership. We would like to capture the fact that the probability of a child becoming a member of the R class is increasing in parents' joint wealth. To do this we posit a minimum parental wealth barrier to upward class mobility, which could arise because class membership requires that one undertake a project with a minimum size, for example, inheriting capital goods sufficient to employ an economically viable team of workers or the amount needed to acquire the educational credentials and social connections necessary to be an elite member.

We will define wealth as the cumulative payoff across all interactions within a period. Rather than modelling the complex distribution of individual wealth, we take a reduced-form approach and define the per-capita wealth of each class as the product of the number of members of the other class and the average payoff from each interaction. Thus, the wealth of a R agent ($y_R = y_R(\eta)$) is the payoff from the game multiplied by the number of P agents with whom the R interacts (which on average is equal to η), while the wealth of a P agent (y_P) is simply the payoff from the game since each P agent interacts only once with the R.

$$(y_P, y_R(\eta)) := \begin{cases} (bN^R, b\eta N^R) & \text{if } \Phi^*(\eta) = E \\ (a_P N^R, a_R \eta N^R) & \text{if } \Phi^*(\eta) = U. \end{cases} \tag{28}$$

Since the cross class couple will be formed by one class agent's matching with the other class agent and thus the joint wealth of the cross class couple is

$$y_c(\eta) := y_P + y_R(\eta). \tag{29}$$

To capture the relationship between parental wealth and class mobility explained above, we suppose that there is a $\bar{y}$ called the baseline wealth barrier. We then suppose that inter-class mobility occurs when cross-class couples have enough wealth to send their children into the R class. Concretely, a cross-class couple gives birth to two R children, resulting in a net increase of 1 R if $y_c(\eta) > \bar{y}$ and gives birth to two P children, resulting in a net increase of 1 C if $y_c(\eta) < \bar{y}$. Cross-class couples with $y = \bar{y}$ have 1 R and 1 P child, keeping η constant. Formally, we consider the following dynamic for η_T, the ratio of the number of P agents to the number of R agents:

$$\underbrace{\frac{1}{1+\eta_{T+\Delta T}}}_{\substack{\text{Future fraction} \\ \text{of R}}} = \underbrace{\frac{1}{1+\eta_T}}_{\substack{\text{Current fraction} \\ \text{of R}}} + \underbrace{\frac{1}{1+\eta_T}\frac{\eta_T}{1+\eta_T}}_{\substack{\text{Probability of forming} \\ \text{Cross-class couples}}} \underbrace{\frac{1}{(1+\eta_T)N^R} sgn(y_c(\eta_T) - \bar{y})}_{\text{increase/decrease in the fraction of R}} \tag{30}$$

In equation (30), $\frac{1}{1+\eta}$ and $\frac{\eta}{1+\eta}$ are the fractions of the rich and poor agents in the whole population consisting of the rich and poor agents, respectively. Thus, the future fraction of R is given by the sum of the current fraction of R and a change in the fraction of R multiplied by the frequency of forming cross-class couples. Here, the change is either an increase in the R's fraction ($+\frac{1}{(1+\eta_T)N^R}$) if the cross class wealth is greater than the wealth barrier ($y_c(\eta_T) > \bar{y}$) or a decrease in the R's fraction ($-\frac{1}{(1+\eta_T)N^R}$) if the cross class wealth

is less than the wealth barrier ($y_c(\eta_T) < \bar{y}$). By rearranging (30), we find that

$$\eta_{T+\Delta T} > \eta_T \text{ if and only if } \bar{y} > y_c(\eta_T) \tag{31}$$

thus the relative size of the poor population increases if and only if the cross class couple wealth is less than the wealth barrier.

We use capital T to denote time in the dynamic in (30) to account for the difference in the time scales between the dynamic governing η and the convention selection dynamic. As explained earlier, we assume that T occurs on a much slower time scale than t (the time for the convention selection dynamic), so that we only have to consider the stochastically stable conventions. In other words, ΔT is sufficiently large so that the convention selection dynamic is always at a stochastically stable convention, and that the stochastically stable state is achieved within a ΔT increment. This simplification rules out possible intermediate states where agents are playing different strategies. We leave the full formal analysis of the coupled two time scale stochastic dynamical system to future work.

Using (27) and (29), we find that:

$$y_c(\eta_T) = \begin{cases} bN^R + b\eta_T N^R & \text{if } \eta < \eta^{*,\infty}(\rho,\theta) \\ a_R N^R + a_P \eta_T N^R & \text{if } \eta \geq \eta^{*,\infty}(\rho,\theta) \end{cases} \tag{32}$$

Combining (32) with the $y_c(\eta_T) = \bar{y}$ condition for steady-state η implied by (31), and abstracting from any discrete time issues, we get

Proposition 6. *Suppose that $\frac{\bar{y}}{N^R} > 2b$ so that $\eta > 1$ is in steady-state. Then we have:*
(i) There exists $\eta^{**}(\bar{y}) \in \mathbb{R}$ defined by $y_c(\eta^{**}(\bar{y})) = \bar{y}$ such that η^{**} is stable with respect to the dynamic in (30).
(ii) As $\bar{y}$ increases, $\eta^{**}(\bar{y})$ increases.
(iii) When $\bar{y}$ is sufficiently large, then convention U is stochastically stable.

Proof. This easily follows from setting $y_c(\eta)$ in equation (32) equal to $\bar{y}$. □

Figure 10 shows the mechanics of Proposition 6 clearly. In Panel A, $\bar{y}$ is low and thus intersects $y_c(\eta)$ at a low η^{**}. Since this $\eta^{**} < \eta^{*,\infty}$, defined in equation (5), the corresponding stochastically stable convention is the E convention. In Panel B, $\bar{y}$ is high and intersects $y_c(\eta)$ at a high η^{**}, which is greater than $\eta^{*,\infty}$ and thus the U contract is stochastically stable.

In our model, high barriers to mobility make the poor class larger, making a sufficient fraction engaging in collective action less likely, and thus unequal contracts will persist for longer. This occurs because large classes require many more deviant players to induce the other side to change behavior, and unequal contracts make it harder for cross-class couples to send their children into the wealthier class. Evidence for this correlation is abundant across a wide variety of historical contexts: from pre-industrial populations (Borgerhoff Mulder et al., 2009) to the modern day (Corak, 2012; Chetty et al., 2014) "Great Gatsby Curve", suggesting that it is general phenomenon unrelated to particular technological or political settings, and therefore possibly illuminated by a relatively abstract evolutionary model such as ours.

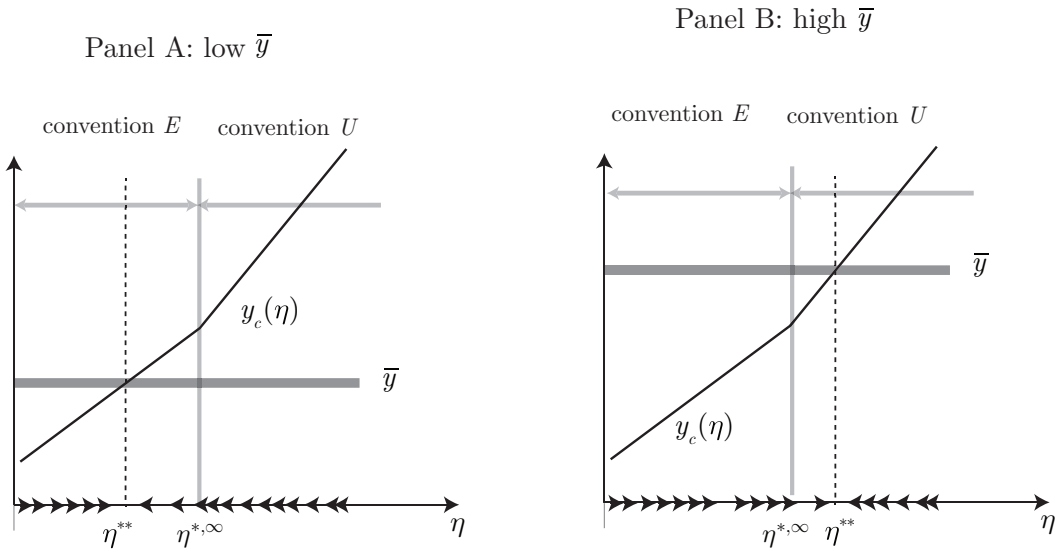

Figure 10: **Illustration of Proposition 6** . Panel A shows low $\bar{y}$, low η^{**} and E as the stochastically stable convention. Panel B shows high $\bar{y}$, high η^{**} and U as the stochastically stable convention. We assume that $2b < a_R + a_P$.

The American South case also highlights the role of endogenous class sizes. Barriers to intergenerational mobility facing non-white populations were very high, as skin color was an obvious inherited obstacle to mobility. This restricted the size of the wealthy population, and created a large population of poor workers, making sufficient social unrest difficult to generate and enabling the segregation convention to persist for quite some time, despite being very unequal.