Quarterly Journal of Political Science, 2008, 3: 1–29

Communitarian versus Universalistic Norms*

Jonathan Bendor[1] and Dilip Mookherjee[2]

[1]*Graduate School of Business, Stanford University, Stanford, CA 94305, USA*
[2]*Department of Economics, Boston University, Boston, MA 02215, USA*

ABSTRACT

The celebration of communitarianism by political philosophers (Sandel, M. 1982 *Liberalism and the Limits of Justice.* Cambridge: Cambridge University Press) has apparently been extended to strategic analyses of ascriptively attuned norms (Fearon, J. and D. Laitin. 1996. "Explaining Interethnic Cooperation." *American Political Science Review* 90: 715–735) — an intriguing development, given game theory's individualistic premises. We believe, however, that game theory offers little comfort to prescriptive theories of communitarian rules: a hardheaded strategic analysis supports the Enlightenment view that such norms tend to be Pareto inefficient or distributionally unjust. This survey uses a specific criterion — supporting cooperation as a Nash equilibrium — to compare communitarian norms, which turn on people's ascriptive identities, to universalistic ones, which focus on people's actions. We show that universalistic rules are better at stabilizing cooperation in a broad class of circumstances. Moreover, communitarian norms hurt minorities the most, and the advantages of universalism become more pronounced the more ascriptively fragmented a society is or the smaller is the minority group.

Though they have had well-known disagreements, most communitarians and many identity theorists share a focus on ascriptive characteristics: race, ethnicity, gender, and (especially for communitarians) family. For example, Michael Sandel, one of the best-known communitarians, wrote that for people "bound by a sense of community. . .community describes. . .not a relationship they choose (as in a voluntary association) but an

* We would like to thank Russell Hardin, Nannerl Keohane, Robert Keohane, Keisuke Nakao, and an anonymous referee for their helpful comments.

MS submitted 22 June 2007; *final version received 8 February 2008*
ISSN 1554-0626; DOI 10.1561/100.00007028

attachment they discover" (1982, p. 152) — i.e., Tonnies' Gemeinschaft (1887). And though some identity theorists (e.g., Young 1990, pp. 232–236) have criticized communitarians for emphasizing certain ascriptive groups — families, villages and other "communities of place" — they celebrate other kinds of ascriptive identities such as gender. Further, both types of scholars criticize the individualism and "abstract universalism" of liberal political theory (Young 1990, p. 228).

Some of these criticisms were based on serious misunderstandings of Enlightenment liberalism. In particular, advocates of universalism have never advocated eliminating communities or the tight interpersonal bonds nourished in them. As Stephen Holmes (1994) put it, "the much reviled "Enlightenment project" did not aim impossibly to extirpate particular attachments from the repertoire of human interaction. [Liberals] strove to dilute or relativize group identity just enough to increase the chances for peaceful coexistence and mutually beneficial cooperation. . .Liberals are not blind to loyalty, therefore, but instead assume that loyalty is sometimes good and sometimes bad, depending largely on the way conflicting affliations and affinities are handled politically. While liberals are not anticommunitarian in a militant sense, they reasonably refuse to apotheosize loyalty as the source of all meaning and the highest human good, insisting that a sharp distinction be made between group identifications to be encouraged and factionalisms and xenephobias to be discouraged" (1994, pp. 601–602). Relatedly, some empirically oriented scholars (e.g., Horowitz 1985) continue to stress the dark side of ascriptively defined identities such as ethnicity, with xenophobia and violence leading the list of negative effects.

Nevertheless, although Enlightenment universalism has been vigorously defended on both normative (Okin 1989; Hardin 1995; Barry 2001; Benhabib 2002) and empirical grounds, liberal political theory no longer enjoys the overwhelming intellectual dominance it once had. Now, most aspects of the theory are contested.

Most social scientists and philosophers recognize that Hobbes' problem of order does not vanish with the wave of a magic wand labeled "community." Hence, the revived interest in Gemeinschaft has been complemented by more attention to how such systems regulate themselves. A major control mechanism in communities is norms: informal rules backed by sanctions.[1] Eventually, rational choice theorists began to explore norms (e.g., Calvert 1995); a few even began to examine communitarian rules (e.g., Bendor and Mookherjee 1990; Hardin 1995; Wintrobe 1995; Fearon and Laitin 1996; Bowles and Gintis 2004). Although some of these writers (prominently Hardin) have criticized communitarianism sharply, others seem more positive. For example, both Bendor and Mookherjee (1990) and Fearon and Laitin (1996) can be read as saying that norms with ascriptively based sanctions can effectively regulate interethnic conflict. Bowles and Gintis (2004) argue that informational problems make communitarian codes useful.[2]

---

[1] This is the standard sociological definition of norms (e.g., Homans 1950, p. 123).

[2] However, rational choice modelers who say positive things about communitarian norms usually do so via theories of the *second best*: given certain real-world constraints, ascriptively based norms may be efficient. (E.g., in Fearon and Laitin's model intra-ethnic sanctions for inter-ethnic crimes are desirable when A's cannot ascertain which B cheated their kinsman. Given this informational constraint, the A's recourse is to punish the B's collectively; a spiral of interethnic violence could

When rational choice theorists, with their natural inclinations toward individualism and associated liberal theories, begin to see benefits of communitarian norms, one wonders whether the celebration has gone too far.

Indeed, so we think. In this survey we seek to redress the balance by arguing that the benefits of communitarian norms have been exaggerated.[3] We argue that communitarian norms are often normatively deficient, even when the primary part of a communitarian rule prescribes cooperating with Outsiders (people outside one's group) as well as with Insiders (people in one's group). In this latter respect they do not differ from the realistic version of universalism, which — as Holmes suggested — also allows for more cooperation within groups than between them. The key difference is the nature of sanctions levied on transgressors of the cooperative norm. Universal norms require punishments to be applied uniformly; the ascriptive features of deviants or victims are considered irrelevant. In contrast, sanctions in communitarian norms do differentiate on such characteristics. (For example, the equilibrium studied by Fearon and Laitin exhibits xenophobic behavior: members of the victim's group punish a deviator who is an Outsider; members of the wrong-doer's group do nothing.)

We show that such communitarian norms are generally less effective in upholding cooperation than universalistic ones: in a significant set of circumstances such norms fail to support inter-group cooperation when universalistic rules could work.[4] Further, we show that communitarian norms hurt minorities the most: the smaller the minority, the greater the harm caused by these norms. Thus more than Pareto efficiency is at stake: distributive justice is also involved.

The survey is organized as follows. Section "THE CONCEPT OF NORMS" briefly examines the fundamental concept of norms. Section "DURKHEIMIAN MECHANICAL SOLIDARITY AND SUPERFLUOUS THIRD-PARTY SANCTIONS" establishes that when interpersonal relations are homogeneous, *all* norm-like strategies — universalistic or communitarian — are in a significant sense superfluous: for stabilizing an important class of outcomes, these strategies add nothing above and beyond what can be sustained by dyadic ones alone (Theorem 1). Because this result covers all third-party sanction strategies, it identifies conditions under which universalistic norms are no better than communitarian ones. Section "DURKHEIMIAN ORGANIC SOLIDARITY" then establishes our central positive result: given natural heterogeneities in interpersonal relations, universalistic norms sustain cooperation as Nash equilibria more effectively than particularistic ones do (Proposition 1).

---

ensue.) In contrast, many communitarians and identity theorists seem to be saying — though not with this language — that certain ascriptively-attuned rules are part of a *first*-best system. We side with the game theoretic perspective and, more importantly, with the Enlightenment view that communitarian norms are typically either Pareto deficient or unjust.

[3] Henceforth we shall use "communitarian norms" as shorthand for "norms based on membership in ascriptively defined groups." For the sake of variety we sometimes follow sociological usage and call such norms "particularistic." ("Communitarian" is more common in political science; "particularistic," in sociology.)

[4] We will show that this conclusion is consistent with the empirical regularity that cooperation within ascriptively defined groups (e.g., ethnicities) is more common and more stable than is cooperation between such groups.

Section "COMMUNITARIAN NORMS HURT MINORITIES THE MOST" analyzes the impact of communitarian norms on minority groups. Proposition 2 establishes that such norms hurt minorities the most. Further, the advantages of universalism become more pronounced the more lopsided are the sizes of the majority and minority groups (Proposition 3) or the more ascriptively fragmented a society is (Proposition 4). Section "IMPERFECT MONITORING" extends our central result to contexts of fragmentary information and of imperfect monitoring: provided that the error rate is not too great, it still pays to have everyone who has relevant information punish deviants (Proposition 5). Section "CONCLUSIONS" concludes.

Because tractability pressures intensify as the analysis proceeds, we impose increasing amounts of structure as we move from section to section.

THE CONCEPT OF NORMS

Before presenting our formal model, we must unpack a key concept: norms. For a major concept in the social sciences, ordinary language and academic definitions exhibit a surprising degree of consistency: in both, a norm is a rule of action that prescribes or proscribes certain actions.[5] Sociological definitions usually add that the rule is informal rather than formal, thus distinguishing norms from laws, and that norms are backed by sanctions.[6] We hew to all of these parts of the sociological notion: in this survey, norms are informal social rules backed by sanctions.

Under the sociological conception, norms may enhance conformity to a code for three reasons.[7] First, humans — even young ones (Kochanska and Thompson 1997) — sometimes internalize norms.[8] Internalization produces what Ellickson (1991) calls first-party control of behavior: the more a decision maker believes that, e.g., the norm of honesty is a legitimate rule, the less likely s/he is to lie. Second, people hurt by the violation of a norm may punish the deviant — Ellickson's second-party control. Third, norms are usually backed by third-party sanctions: Susan punishes Bill if Bill has cheated Joe, even if Bill cooperated with Susan.[9]

---

[5] See Gibbs (1981, pp. 7–9) for an extensive review of sociological definitions.

[6] For example, in the Webster's *Ninth Collegiate Dictionary*'s first definition of "norm" ("a principle of right action binding upon the members of a group and serving to guide, control or regulate proper and acceptable behavior"), sanctions are implicit in the statement that behavior is controlled and regulated. Homans' definition makes sanctions explicit.

[7] The following is taken from Bendor and Mookherjee (1990).

[8] Most developmental psychologists now agree that children do not passively "absorb" parental socialization; they actively reconstruct what adults try to teach them. Nevertheless, specialists also agree that at the end of the day children do often internalize parental values. On both points see Kuczynski and Grusec (1997).

[9] Empirically, all three aspects of norms can occur simultaneously, e.g., a person may feel morally compelled to tell the truth, those victimized by a lie may retaliate, and third parties may punish the miscreant. Because ascertaining the strengths of these different controls is difficult, scholars often study the effects of one kind of control in isolation, holding the other two constant experimentally or analytically.

Second-party control is a weak indicator of a norm because it does not rely on social ties; it can arise even in situations characterized only by bilateral relationships. In contrast, third party sanctions are a quintessentially social dimension of norms: people who are not parties to an original interaction become involved because social networks connect them to the original players and because community-wide codes of conduct, germane to the original interaction, were violated.[10]

We use the important distinction between second- and third-party controls in our analysis. Game-theoretically, player i's strategy is dyadic, or uses bilateral sanctions, if its behavior toward any other player j is based only on the history of play between i and j. The strategy of i is *norm-like*, or uses third-party sanctions, if at some time i's conduct toward j is based on j's behavior toward some third player k.

DURKHEIMIAN MECHANICAL SOLIDARITY AND SUPERFLUOUS THIRD-PARTY SANCTIONS

In this section, we focus on a group where all bilateral relationships are homogeneous in all strategically relevant ways: payoffs, information, and frequency of interaction. Ascriptive differences may exist, but if so, they do not affect any payoff-relevant attributes.

Consider, for example, an ideal-typical village in which networks are maximally dense — every person encounters everyone else in every period — and all pairs play the same game of bilateral exchanges. Through gossip and observation everybody knows everyone else's business: what happened in all past interactions is common knowledge to everyone. Clearly, this community is tightly-knit. Nevertheless, in this prototypical village we show below that *norms do not matter* in an important sense: whenever the strongest such norm can uphold adherence to prescribed symmetric behavior as a Nash equilibrium, *so can some dyadic strategy*. Equivalently, if no dyadic strategy can ensure that people do what they are supposed to, *then neither can any norm*. So norms are useless in this tightly knit village.[11]

To see why these properties hold, consider the interactions among three members of the village, i, j, and k. Suppose that in period t player i is thinking about cheating j. Suppose further that no dyadic strategy packs enough punch to deter i: even if j

---

[10] Although our models do not directly represent internalization, they do lend themselves to such an interpretation. If i sanctions j for cheating k, even though j did not harm i, one can well believe that i, adhering to fair dealing as a code of conduct, is morally outraged by j's behavior and so punishes j (Fehr and Fischbacher 2004). However, as Ellickson's categorization makes clear, the standard sociological conception of norms involves multiple, redundant controls. If socialization worked perfectly then the other two controls would be superfluous. However, socialization is never perfect — to believe otherwise is to succumb to the "overly socialized" view of humans (Wrong 1961) — so second- and third-party controls also matter.

[11] Because they use third-party sanctions, norm-like strategies are more complex than dyadic ones. So if using more complex strategies imposes a cost or if simpler strategies are lexicographically preferred then norms would be Pareto dominated in our ideal-typical village. To see the implications of taking strategic complexity into account, see, e.g., Banks and Sundaram (1989) or Binmore and Samuelson (1992).

retaliates by punishing i in all subsequent encounters, cheating j today would still be i's best move. How would i's thinking be affected if he knew that he would be punished not only by j but also by a third party, k? Because this doubles the maximal punishment that j alone can unleash, it seems obvious that this norm-like rule would restrain i in some settings in which dyadic sanctions fail. But this reasoning ignores two key related facts. First, it does not analyze i's optimal deviation in period t. Second, it overlooks the fact that i's relations to j and to k are identical. Hence, if it is optimal for i to cheat j today, given that only j will retaliate, then it must be best for i to cheat k as well, anticipating k's bilateral response. But then k, as a third party trying to enforce cooperation between i and j, has nothing with which to threaten i, who is contemplating defecting against k anyway for dyadic reasons. Thus, if bilateral sanctions cannot support cooperation here then neither can any norm-like strategies. (For further details on this example, see Bendor and Mookherjee (1990).)

It is well-understood that informal methods of social control — e.g., maintaining cooperation in iterated PDs by strategies of conditional cooperation — generally work better when networks are dense and monitoring is informative. (See Ellickson (1991) for a good summary of this perspective.) One might call these village-attributes. However, many scholars also maintain that one type of informal control, norms, are especially useful in village-like communities (e.g., Ostrom 1990, pp. 88, 89; Ellickson 1991, pp. 167–169, 177–182). *These two points are easily confused with each other*. As the preceding example suggests, they are logically unrelated. Our first result, Theorem 1, generalizes this example. The next section shows that what enables norms to be more efficacious than dyadic strategies is *relational heterogeneities* — Durkheim's organic solidarity — not the village attributes at all.[12] The latter strengthens *all* kinds of informal social controls, including dyadic ones.

These ideas generalize considerably. Consider a finite set of players, $N = \{1, \ldots, n\}$, with $n \geq 3$ to allow for third-party sanctions. As in the above example, all pairwise relations in this community are the same. Formally, we say that the (i, j) relation and the (i, k) relation are *homogeneous* if two conditions hold for all t: (1) i and j meet in period t if and only if i and k also meet in t; (2) i and j play the same game in t as i and k play in t. In addition, we shall assume payoff separability throughout this survey: a player's total payoff in any period is simply the sum of her payoffs from all her different bilateral relationships.

Although the notion of two relations being identical is rather intuitive, formalizing an intuitive idea usually introduces some conceptual nuances. Given that game-theoretic analyses of norms usually assume a conventional setting — the agents are playing a standard repeated game — understanding what relational homogeneity does *not* presume is important. For several reasons it does not presuppose a standard repeated game. First, people may be playing a finite game, with a commonly known ending date. Second, the nature of interactions could change over time. Imagine a cohort of people, all the same

[12] "Mechanical solidarity [i.e., where everyone and their ties are alike] does not bind men together with the same strength as does the division of labour" (Durkheim 1997 [1893], p. 123). The division of labor creates relational heterogeneities.

age, who together go through childhood, adulthood and old age. All relations are the same at any one date; all change the same way as people age. Finally, relational homogeneity does not require that interaction probabilities are stationary: meetings could be based on a nonstationary calendar. (For example, people meet at the start of the full moon and never otherwise.) Thus the notion of relational homogeneity is not confined to standard repeated games; it can be applied to a much wider set of environments.

Whenever players meet they play a game that belongs to a class of oneshot games, G. (The temporal sequence of games is exogenously fixed and is common knowledge.) Because norms include sanctions for deviations, G is defined by the kind of punishments that are available. Throughout this paper we focus on games with *simple punishment* by requiring that every game in G has an action, a_p, with the following properties: it minmaxes the other player and it is a best response to itself.[13]

Requiring that each game in G has this kind of sanction simplifies our analysis greatly: it implies that irrevocably deploying a_p after someone deviates is the maximal credible punishment (Abreu 1988) in any subgame-perfect Nash equilibrium. Hence we can easily characterize the set of (subgame-perfect) Nash equilibrium outcomes: those that can be sustained by the threat of reverting to this punishment path following any deviation from the prescribed pattern of behavior.

We assume that every game in G is symmetric: the same actions are available to both players and reversing actions reverses payoffs. The set of actions in a particular game, say τ, is $A(\tau)$; every action-set is finite.[14] Let $v_{1\tau}(a_x(\tau), a_y(\tau))$ be player 1's payoff when she does $a_x(\tau)$ and player 2 uses $a_y(\tau)$ in game τ. Similarly, $v_{2\tau}(a_x(\tau), a_y(\tau))$ is player 2's payoff from this pair of actions. Because the game is symmetric, $v_{1\tau}(a_x(\tau), a_x(\tau)) = v_{2\tau}(a_x(\tau), a_x(\tau))$ in any game in G, so we suppress players' subscripts when both players select the same action. Players' utilities are additive across their pairwise games and over time, and future payoffs are discounted by a common parameter $\delta \in (0, 1)$.

Theorem 1 *Suppose that all bilateral relationships are homogeneous and every oneshot game they play belongs to G. Then any sequence of symmetric outcomes can be supported by a subgame perfect Nash equilibrium by any norm-like strategy if and only if it can be supported by some dyadic strategy.*[15]

Theorem 1 is not confined to standard repeated games or to stationary norms. (A natural example of a nonstationary norm: one kind of behavior is prescribed for the young and another for the old.) But to get a sense for the result's content let us focus briefly on stationary norms used in a conventional repeated game. People in this context can adhere

---

[13] Thus G includes the prisoners' dilemma-type games, but many others as well, e.g., Stag Hunt, in which mutual cooperation is a Nash equilibrium in the oneshot game. Such coordination games can represent interactions in cohesive communities in which altruism or norm-internalization transform objective, PD-like payoffs into Stag Hunt-like utilities.

[14] Although most of our results extend to games with compact action sets, such an extension provides little insight.

[15] Note the restriction to symmetric outcomes, where both members of a given bilateral relationship choose the same action. See Bendor and Mookherjee (1990) and Section "DURKHEIMIAN ORGANIC SOLIDARITY" for further discussion of this restriction.

to a stationary social rule with a very simple structure. For instance, by coding members of a community as being in either Good or Bad Standing (a sociologically intuitive notion), a Grim Trigger-type norm could be defined as follows. (1) In every period every player must be categorized as being either in Good or Bad Standing. (2) Everyone is in Good Standing initially. (3) A player is in Good Standing in $t + 1$ if and only if in t she was in Good Standing and played the socially prescribed action, $a^\star$, with every partner who was in Good Standing at that time. (4) Play the punishment action, a_p, toward any partner who is in Bad Standing.[16] Thus, this norm-like strategy requires that everyone in the community keep track of each other's social standing. Given a person's standing, the prescribed action follows immediately. In this world, reputation, in the colloquial sense, is everything.

Note that the prescribed outcome of $(a^\star, a^\star)$ need not be Pareto optimal. (Consider rat races in organizations in which working long hours is inefficient but normatively prescribed.) Of course, if the payoff to the prescribed outcome is less than that of mutual punishment, then that $a^\star$ cannot be supported as a Nash equilibrium. Even then, however, bilateral and norm-like strategies are equivalent: neither works if $v(a^\star, a^\star) < v(a_p, a_p)$.

Although the example that opened this section presumed dense interaction — people encounter each other in every period — Theorem 1 does not require this temporal density or any particular interaction rate at all. Thus, it shows that what is central is not the cohesiveness of the community, as reflected by, e.g., interaction rates, but the *homogeneity of ties*. Suppose, for example, that interaction is not stationary; instead, people are more likely to meet on weekdays than weekends. The result still holds because whenever i meets j — no matter how irregular the calendar of interaction — she also meets all other players.

Theorem 1 identifies circumstances in which universalistic rules are not superior to particularistic ones: they are equally useless. This property extends to some settings where bilateral relationships are not homogeneous. Suppose there are two ascriptive groups, A and B. All pairs play repeated games. Intragroup and intergroup pairs play different games: specifically, (A, B) pairs play a Prisoner's Dilemma, while intragroup pairs play a Stag Hunt.[17] Between-group relations — all (A, B) pairs — are homogeneous. Then the following properties can be established. (1) Some norms are useful in supporting intergroup cooperation: they can support cooperation with Outsiders as a subgame-perfect Nash equilibrium when the former cannot. For instance, Insiders can punish someone who cheats an Outsider. The difference between inter-group and intra-group ties can allow players to want to deviate in the latter but not in the former.

---

[16] Whereas evolutionary stability requires that norms have a complete metanorm structure (Bendor and Swistak 2001), subgame perfection and the existence of a punishment action that is a best reply to itself do not entail this type of structure. If k knows that j is going to punish him forever with a_p, k's best response is to do likewise, i.e., mutual punishment is a subgame-perfect Nash equilibrium. Hence enforcement is credible on a bilateral basis; metanorms — i sanction j if j does not punish the deviant k — are unnecessary.

[17] Stag Hunt's key property is that mutual cooperation (as well as mutual defection) is a Nash equilibrium in the one-shot game.

The sanction imposed by one's own group can thus effectively deter the cheating of Outsiders. (2) The best communitarian norm is just as effective as the best universalistic rule in supporting intergroup cooperation. In this example, only the threat of sanctions imposed by comrades can credibly inter-group transgressions; Outsiders' sanctions are impotent because one does not cooperate with people from other groups in the first place.

Relational homogeneity of all inter-group matches is driving property (2). If A_i is thinking about cheating B_j then optimal deviation entails cheating all other B's as well, since their ties to A_i are the same as his relation to B_j. Hence a within-group norm — only A_i's kin punish him for his misconduct — is maximally effective; the sanctions of B_j's kin add nothing. Hence, no universalistic rule is better than the best particularistic one.

A key pattern is emerging: third-party sanctions matter when they do not suffer from correlated failure. A_i's cheating B_j is not linked to his cheating other A's, but it is perfectly correlated with his hurting other B's. So A_i's ties to other A's are useful strategic back-ups to the (A_i, B_j) pair, but his ties to the other B's are not. This suggests that in a fully differentiated community — all ties are heteogeneous and do not suffer from correlated failure — universalistic norms will out-perform communitarian ones. We examine this idea next.

DURKHEIMIAN ORGANIC SOLIDARITY

We now examine a community where relationships are differentiated in two ways. The first arises from a difference in matching patterns: every pair does not meet in every period. Instead, any given players meets exactly one other in any given period. The exact pairing is decided by a random draw. We shall assume that any given pair of players can meet with postive probability in every period, independently and identically across periods. We also assume that a given pair play the same stage-game whenever they meet. Having completed play in period t, the probability that the population will be re-matched and play in period $t + 1$ is a fixed probability $\delta \in (0, 1)$. With the complementary probability of $1 - \delta$ play ends. Following common usage, we call this *social matching*, as distinct from the round-robin play examined in the previous section.

With social matching, relationships are differentiated *ex post* even if they are identical *ex ante*: only some of these relationships result in an actual meeting in any given period, while others do not. As will become evident shortly, this kind of heterogeneity suffices to nullify the result of Theorem 1. Cooperation is harder to sustain by the threat of bilateral sanctions alone: with these, the deviant may not expect to re-encounter the victim soon after the infraction. In contrast, with third party sanctions the wrong-doer could be punished by everyone she meets in all subsequent dates.

We also allow a second source of differentiation: the population is broken up into ascriptively defined groups. Relations with Outsiders and with Insiders can differ *ex ante* as well as *ex post*: they may involve different pairwise games and different matching-probabilities. There are $m\,(1 < m \leq n)$ ascriptively defined groups, labeled $\{A_1, \ldots, A_m\}$.

(For convenience, if there are only two groups then they are labeled A and B.) We assume that the groups partition the society: each person belongs to exactly one group.[18] Hence $\sum_{r=1}^{m} n_r = n$, where n_r is the number of people in group A_r. If there are only two groups then B_j denotes player j and indicates that he belongs to group B, and analogously for A_i.[19]

We have defined homogeneous ties in an ideal-typical way. Hence, heterogeneity is a residual category: relations that are not homogeneous must be heterogeneous. Hence, once we have stepped off the well-defined island of homogeneity we are in a more complicated world. Accordingly, we impose more structure on (1) how players encounter each other and (2) the games they play. We assume that a given pair whenever matched with one another at any date play the same stage-game.

It would be simpler to assume that all intergroup pairs are the same in all respects (i.e., payoff and frequency of interaction), and all intragroup relationships are the same. But as this is inessential for the results of this section, we do not make this assumption here. What *is* essential is that the model represent intergroup cooperation as problematic in some fundamental way. We describe this in more detail below.

Regarding the stage games, we restrict attention to two nested subsets of G. Intragroup pairs play stage games that belong to G'. In addition to the simple punishment feature shared by all games in G, games in G' have a uniquely efficient cooperative action a_c, distinct from a_p, such that $2v(a_c, a_c) \geq v(a_x, a_y) + v(a_y, a_x)$ for all x, y, with equality holding if and only if $x = y = c$. Hence, in these games the meaning of "cooperating" is clear: it means playing a_c.[20] (Anything other than (a_c, a_c) we call a noncooperative outcome.) Cooperation may or may not be a Nash equilibrium of the one-shot game: we impose no restriction in this regard.

Intergroup pairs play stage games that belong to a subset of G', called G'', where cooperation is *not* a Nash equilibrium. Thus, in addition to the two properties defining G', games in G'' add a third: there is a unique best response to a_c, but that response, called a_d for "defection", is not a_c.[21] Thus intergroup pairs meet each other in situations where cooperation is *collectively valuable* but *strategically problematic*. In contrast, intragroup stage games need not belong to G'': ascriptive groups might be so cohesive — due

---

[18] We do not model group-formation or how different ascriptive labels are activated. (Hence we take no stand on the primordialism-versus-constructivism debate. For a lucid discussion of this controversy, see Laitin (1998), Chapter 1.) However, we do examine the effects of an exogenous change in the number of politically active ascriptive groups.

[19] Group membership need not involve any objective differences other than the observable feature (e.g., skin color) that enables ascriptive labeling. Game theorists have known for some time that ascriptively-oriented strategies can be based on such purely nominal differences. See Axelrod (1984, p. 147) for an early analysis.

[20] If this condition did not hold then the socially optimal pattern could be to alternate, as in the Battle of the Sexes or in PDs in which $2R < T + S$. As is well-known, ascriptive labels can be normatively desirable in such situations: they can aid coordination.

[21] The punishment action a_p need not be the best response to a_c. It is the best response in the ordinary PD, where "punishment" and "defection" are the same, but in some games in G'' they are distinct actions.

perhaps to altruism or because norms prescribing cooperation with Insiders are well-internalized — that intragroup cooperation could be a Nash equilibrium in *one-shot* encounters (e.g., the Stag Hunt game).[22]

Thus, any game in G'' exhibits a gap between collective optimality and individual rationality: the socially optimal outcome of (a_c, a_c) does not constitute a Nash equilibrium outcome in intergroup relations. However, the feasibility of punishment — using a_p via either dyadic or norm-like sanctions — creates the possibility that cooperation can be stabilized by stick-and-carrot strategies of conditional cooperation. We focus on the sustainability of cooperation in across-group relationships. The presumption is that cooperation within groups can be sustained via either altruism or the high frequency of interactions; the essential question concerns the effectiveness of different kinds of norms for sustaining cooperation *between* groups.

For convenience we use the payoff notation from the standard binary-choice Prisoner's Dilemma to label intergroup payoffs associated with cooperation and defection. Thus let $v(a_c, a_c) \equiv R$, $v(a_d, a_c) \equiv T$, $v(a_c, a_d) \equiv S$, and $v(a_p, a_p) \equiv P$. The above assumptions imply that $T > R > \max(P, S)$ and $2R > T + S$ for any intergroup game. A player's *ex ante* payoffs for the entire game is the discounted sum of her expected pairwise payoffs.

We also continue to assume that monitoring of intergroup breaches is perfect. Section "IMPERFECT MONITORING" shows that our results are robust against monitoring imperfections, as long as they are not too big.

The result in this section, Proposition 1, establishes sufficient conditions for the superiority of universalistic over particularistic norms in supporting intergroup cooperation.[23] Any number of ascriptive groups is allowed; the groups may be of arbitrary sizes. Further, no significant sociometric assumptions are imposed. For example, the proposition allows for cliques: part of a group could encounter each other more often than they meet anyone outside the clique. Or there could be brokers who interact often with people in many groups. More importantly, Proposition 1 holds *even when intragroup dyadic ties are stronger than intergroup ones and hence within-group policing is better than between-group policing at sustaining intergroup cooperation*. (We do not mention these parametric advantages of intragroup ties in the proposition itself, as we prefer to state it in greater generality. The discussion after the result, however, will make clear that it does indeed hold when intragroup ties are stronger than intergroup ones.)

A *universalistic* norm is defined as one in which sanctions do not condition on ascriptive features: any deviation by any player in any intergroup relationship invites the same response from all other players, regardless of group membership. In the best universalistic norm, any deviation is followed by perpetual reversion to minmax actions thereafter, by every other player in the game. In *communitarian* norms, sanctions are ascriptively

---

[22] Quite a few scholars have argued that hominid preferences evolved so that intragroup cooperation was a Nash equilibrium even in finite encounters. See Fehr and Fischbacher (2003) for an overview of the argument and a literature survey.

[23] The appendix actually proves a stronger result: universalistic norms are always at least as good as communitarian ones in supporting cooperation in *both* intergroup *and* intragroup pairs. Since this survey concentrates on the former, the proposition in the survey confines itself to that subset.

differentiated: the sanctions of people in one's own group differ (in terms of their payoff consequences for the sanctioned players) from those of other groups. This implies that communitarian norms punish intergroup deviations less vigorously than does the best universalistic norm: some other player, either from one's own group or some other, will respond with a lighter sanction. An example of such a norm is where only members of the victim's group sanction the deviator; the latter's compatriots are passive.

Proposition 1 *Suppose the game involves social matching and* $m(1 < m \leq n)$ *ascriptive groups. Each intragroup pair when matched play a stage game which belongs to* G'*; each intergroup pair when matched play a stage game which belongs to* G''*. Then*

(i) *cooperation between people in different groups can be sustained as a subgame perfect Nash equilibrium if and only if it can be sustained by the best universalistic norm;*
(ii) *every communitarian norm has a higher threshold for cooperation (a higher* δ^**) in every intergroup pair than does the best universalistic norm.*

Part (i) tells us that if no universalistic norm can support intergroup cooperation as Nash behavior, then it cannot be done — not by dyadic sanctions and not any particularistic norm. Part (ii) implies that universalism can work even when the best communitarian norm falls apart.[24]

Part of the explanation for these effects is simple (almost misleadingly so): the bigger the stick, the stronger the deterrent. And no strategy or norm wields a bigger stick than the best universalistic norm, wherein everyone in a community is supposed to punish miscreants. But this simple logic is only part of the story. By itself it would incorrectly imply that the best universalistic norm is *always* better at supporting cooperation. But Theorem 1 tells us that this is false: if relations are homogeneous then adding more punishers is futile.

Thus Proposition 1 requires the vital premise that social ties are not homogeneous. They are heterogeneous simply because people meet at different times: if A_i plays B_j today, he is not playing B_k. This temporal difference by itself makes norms — especially universalistic ones — efficacious. It suffices because a *strong universalistic norm makes the social network denser than it otherwise would be*. Given social matching, the best universalistic norm ensures that a deviant will be punished in *every* period. In contrast, under any particularistic norm a deviant will sometimes encounter people who will not punish him. So *de facto*, the social network is denser under the former than the latter.

Although Proposition 1 did not say that within-group cooperation must be easier than between-group cooperation, it is general enough to allow for that possibility. We underscore this point by describing a simple example of this result. Suppose that a society is divided into two equal-size ethnic groups, A's and B's. For simplicity, assume that all within-group stage games are the same and all between-group games are the same; each

[24] Players described by Proposition 1 have heterogeneous relations. Therefore, one can easily extend part (ii) to cover dyadic strategies: all of them also have a higher threshold for cooperation in every intergroup pair than does the best universalistic norm. Indeed, these heterogeneous relations also enable many communitarian norms to be more effective than dyadic strategies; given heterogeneity, norms in general matter.

is a binary-choice PD. Matching probabilities are also simple: all intragroup matches are equally likely; all between-group matchups are equally likely. However, we load the dice in favor of within-group cooperation by assuming that two properties hold in this example. First, the stage-game payoff to mutual cooperation is higher in intragroup matches than intergroup ones. The cause of this difference could be objective (the groups live apart and travel makes cooperation costly) or subjective (altruism declines with social distance) or both. Second, within-group pairs form more often than do between-group ones.

These two assumptions ensure that dyadic strategies of conditional cooperation, such as Grim Trigger or Tit for Tat, work better in within-group pairings than between-group ones for several reasons. (1) The temptation to cheat an Insider is less than the temptation to cheat an Outsider.[25] (2) The threat of retaliation — withdrawing cooperation — is more potent in within-group pairs.[26] (3) Social networks within groups are tighter than between-group ties — a person is more likely to run into a specific Insider than a specific Outsider — which intensifies punishment.

Proposition 1's assumptions hold in this example. Hence *the best universalistic norm is more effective at supporting cooperation than is the best communitarian norm*. The reason is simple: the greater potency of within-group controls does not imply that Outsiders' sanctions are worthless. A_i's anticipation that he will be punished by everyone, Insider and Outsider alike, for cheating B_j today reflects greater deterrent power than being punished only by peers. Accordingly, a norm with universalistic sanctions can support intergroup cooperation when a particularistic norm with restricted punishments cannot.[27]

Proposition 1 implies that the relative merits of universalism and particularism depend on the value of δ, the discount factor. Define an *intermediate* range of δ (for a particular intergroup pair) as follows. If δ is close enough to one then the Folk Theorem tells us that even dyadic sanctions can uphold cooperation. Hence both universalistic and communitarian norms will also do the job. If δ is too low then even universal sanctions will not work. If, however, δ is intermediate — in-between the high and low ranges — then the best universalistic norm will support cooperation but the best particularistic norm will not. (Proposition 1 implies that such an intermediate range must exist for every intergroup pair.) Clearly, it is this intermediate range of δ that is empirically

---

[25] To see this, let R_w denote the payoff to mutual cooperation in the within-group stage game, and R_b the payoff to between-group cooperation. Since $R_w > R_b$ but the other payoffs are the same, the temptation to cheat an Outsider, $T - R_b$, exceeds the temptation to cheat an Insider, $T - R_w$.

[26] For example, a trigger strategy's threat is to substitute a punishment phase for mutual cooperation. Because mutual cooperation is more valuable within groups than between them, ending it hurts more, so the intragroup per period penalty, $R_w - P$, exceeds that of between-group pairs, $R_b - P$.

[27] The empirical regularity of more cooperation within than between groups is consistent with the survey's discussion: if stage-game payoffs and social networks make the former easier to support, we should observe more within-group cooperation than between-group. Thus, for certain ranges of δ cooperation across groups cannot be sustained by *any* norm, while some norms will support it within groups. (E.g., suppose a community has evolved within-group norms but not universalistic ones. Given the advantages of within-group cooperation, δ has a range where only within-group cooperation is a Nash equilibrium outcome.)

interesting, for there we get discriminating predictions about the effects of different norms. Specifically, if δ is intermediate for *any* intergroup pair, then the best symmetric equilibrium supported by universalistic norms Pareto dominates all symmetric equilibria supported by communitarian norms.

COMMUNITARIAN NORMS HURT MINORITIES THE MOST

The previous section showed that for intermediate values of the discount factor, universalism Pareto-dominates communitarian norms. In this section, we examine distributive implications. We argue that failing to use universalistic norms hurts small groups the most. So the consequences of communitarian norms are inequitable as well as inefficient.

The next result focuses on what can be achieved by the best strategy — those involving the strongest feasible punishments — in two empirically important classes of particularistic norms: within-group punishments (the cheater's kin punish the miscreant) and between-group punishments (the victim's kin punish). We restrict attention to outcomes based on symmetric actions within each bilateral intergroup relationship.[28]

Because the analysis in this section is more complicated, we employ several simplifying assumptions on stage games and intergroup matching probabilities.

A1: All between-group games are identical and all within-group games are identical.

A2: All between-group pairings are equally likely to form, in every period.

Even with these simplifying assumptions, we can represent the substantively significant situation of cohesive intragroup behavior and problematic intergroup relations, e.g., between-group pairs form less often than intragroup matches. (A2 puts no restrictions on network patterns within groups.) We label the groups in descending order of size: $n_1 \geq \cdots \geq n_m$. In what follows, a *within-group norm* (resp. *between-group norm*) is one in which a deviation with someone from a different group is sanctioned by all members of one's own group (resp. by all members of the victim's group). Recall that the best within- or between-group norm deploys a maximal sanction — perpetual reversion to the minmax action thereafter. To examine how much cooperation can be sustained by such norms, we can restrict attention to the best version of either norm.

---

[28] Despite its natural appeal as a fair outcome of a symmetric relationship, this restriction is significant in terms of payoffs that can be achieved. Bendor and Mookherjee (1990) showed that allowing asymmetric levels of cooperation in intergroup relationships would generally Pareto-dominate the maximal supportable symmetric payoff. However, because incentives for cooperative behavior are harder to provide to minority members, in these equilibria members of majority groups provide more cooperation to minority members than vice versa. Such patterns of cooperation are rarely observed, presumably because they conflict with notions of fair exchange. For instance, in the standard 2×2 Prisoners Dilemma, a big-group player would allow, as equilibrium behavior, a small-group partner to sucker him repeatedly. This reverse communitarianism may be enlightened but it is probably uncommon (Brown 1991).

Proposition 2 *Suppose A1, A2 and the hypotheses of Proposition 1 hold; further,* $m \geq 3$ *and* $n_2 > n_m$. *Then there exist thresholds* $\underline{\delta}$ *and* $\overline{\delta}$, *with* $0 < \underline{\delta} < \overline{\delta} < 1$, *such that the following hold.*

(i) *For any fixed* $\delta \in (\underline{\delta}, \overline{\delta})$, *in any symmetric equilibrium of either the best within-group or the best between-group norm the groups are partitioned into two subsets: one contains the k biggest groups (*$1 < k < m$*); the other, the* $m - k$ *smallest ones. People in the big groups cooperate with each other but those in small groups do not cooperate with any Outsiders.*

(ii) *Under the best within- or between-group norm, the number of groups that can sustain between-group cooperation is weakly increasing in* $\delta \in (\underline{\delta}, \overline{\delta})$, *and strictly so if* δ *rises sufficiently.*

(iii) *The best universalistic norm can support cooperation between people in all groups for any* δ *in* $(\underline{\delta}, \overline{\delta})$.

Thus, for intermediate values of δ, universalism allows intergroup cooperation to flourish generally, while under communitarian norms members of small groups can cooperate only with their peers: trapped within their groups, their exchanges are confined to small networks. This is harmful, given the following weak restriction on payoffs.

Corollary 1 *Suppose the hypotheses of Proposition 2 hold; further, within-group cooperation delivers higher payoffs than does any symmetric noncooperative between-group outcome. If* $\delta \in (\underline{\delta}, \overline{\delta})$ *then the following properties hold in the symmetric equilibrium with cooperation under either the best within-group or the best between-group norm:*

(i) *The expected payoffs of everyone in the large groups (*A_1 *through* A_k*) exceed those of anyone in the small ones.*

(ii) *The expected payoffs of people in the small groups are strictly increasing in the size of their groups.*

Thus for intermediate ranges of δ small groups are hurt the most by particularistic codes. Under the conditions of Corollary 1, members of these groups have the most to gain from a transition from communitarianism to universalism.

Changes in Group Composition

Now we examine how changes in group size affect the viability of intergroup cooperation and the ensuing welfare impacts on minorities. First, we continue the analysis of size asymmetries. Proposition 2 and its corollary showed that given a fixed set of ascriptively defined groups, the smaller groups can support less cooperation and suffer accordingly. We now study how changes in size asymmetries affect cooperation and welfare. To keep the comparative statics clean, the number of groups and the total population are held constant. Proposition 3 and its corollary will show that as the size distribution of groups becomes increasingly lopsided, the problems that communitarian norms impose on minorities intensify.

To keep the analysis simple we consider a society with two groups, A and B, with A the bigger of the two. Further, we continue to presume A1 and A2. However, these

assumptions as stated are not adapted to comparative static analysis: they are silent on whether exogenous changes can affect the stage game or matching probabilities. Hence, we introduce extensions of A1 and A2 that address this issue. The extension of the stage-game assumption is simple, but because interaction probabilities are interdependent — they must sum to one — this extension is a bit more complex.[29]

A1′: All between-group stage games are identical and all within-group games are identical. Further, neither type of game is affected by any parametric change.

A2′: (a) Before a parametric change all between-group pairings are equally likely; after the change all such matches are equally likely. (b) A parametric change may affect the probability of such encounters, subject to the constraint that the chance that a player meets an Outsider is increasing in the number of Outsiders.

A2(b) implies, e.g., that a B is more likely to run into an A after the parametric change, when there are more A's, than before — a rather weak constraint.

Proposition 3 *Assume that A1′, A2′, and the hypotheses of Proposition 1 hold, with* $n_A > n_B$. *Suppose there is an exogenous shift in group-composition: some people shift from minority group B to the majority group A. Then the critical* δ*-thresholds of the best within- and between-group norms rise monotonically, equaling their dyadic values when* n_B *equals one.*

The explanation is straightforward. If policing is only by one's own group and constant cooperation is required, then the smaller group's size is the binding constraint on the viability of cooperation. Hence eventually the intragroup communitarian norm can no longer uphold intergroup cooperation because there are too few B's to force themselves to cooperate with A's. The between-group norm experiences the same effect for the opposite reason: once there are enough A's, they will start cheating B's because there are too few of the latter to deter the A's. As usual, the universalistic norm is unaffected.

Now consider the welfare effect of this change in identity demographics. If the majority group grows, there is one obvious negative effect on minority member payoffs since there are fewer people in their group with whom to cooperate. This results whenever cooperating with members of one's own group generates higher payoffs than cooperating with Outsiders. The next result shows that minority players can be hurt (and never benefit) as group-sizes become more lopsided *even when within- and between-group cooperation are equally valuable*.

In the statement of the results, A's include people who kept this identity throughout as well as individuals who changed from B to A. Similarly, B's include people who maintain this identity throughout as well as anyone who changed from A to B. (Proposition 3 and its corollary allows for the possibility that no majority group members start passing as minority; the welfare results show why few people would want to make such transitions.)

---

[29] More specifically, whereas we assume that the parametric shift does not affect stage games, this cannot hold for matching probabilities. Some of these must change as identities change: since the number of A's rises, if intergroup pairings occur with the same probability as before then B's must become less likely to encounter peers.

Corollary 2 *Suppose the hypotheses of Proposition 3 hold, with* $n_A > n_B > 1$. *Further, within-group cooperation is more valuable than is any symmetric noncooperative between-group outcome. If there are still some B's left after the identity-shifts have occurred, then for either the best within-group norm or the best between-group norm the following conclusions hold.*

(i) *Increased lopsidedness never makes B's (old or new) better off.*
(ii) *If* δ *is in* $(\underline{\delta}, \overline{\delta})$, *where* $0 < \underline{\delta} < \overline{\delta} < 1$, *then the parametric change makes B's worse off.*
(iii) *The change never yields a Pareto-improvement: if it benefits any A, old or new, then it harms the B's.*

Under the above assumptions, the infeasibility of intergroup cooperation hurts the minority more than the majority because the former necessarily have fewer peers. Hence, the smaller the minority the more damaging is the absence of intergroup cooperation. And since by Proposition 3 this collapse occurs for bigger ranges of δ as the groups becomes more lopsided, for the minority everything goes bad at the same time as their group shrinks: supporting cooperation via either communitarian regime becomes more difficult and the collapse has more dire effects.

Thus Proposition 3 and its corollary are normatively significant: they tell us that *the minority group has more to gain from a transition from particularistic to universalistic codes*, and the smaller it is the more it has to gain.

Further, if identity were endogenous — e.g., people could sometimes, perhaps at a cost, change their identity — then in certain parametric environments the minority group will disappear in equilibrium.[30]

Ascriptive Fragmentation

With the rise of identity politics in the seventies, some social commentators remarked that the United States was becoming more ascriptively fragmented. In this subsection we will argue that a society's degree of ascriptive fragmentation differentially affects the abilities of universalistic and communitarian norms to sustain cooperation: *as the number of (active) ascriptive groups increases, the bigger the gap between the effectiveness of these types of norms*.

By "the number of (active) ascriptive groups" we mean a specific, though empirically important, type of identity-change: those which cause every citizen to lose some peers and not gain any new ones. (Hence i's peers *ex post* are a proper subset of her *ex ante* ones, for all i.) For example, suppose that initially a society is divided into blacks and whites. Then gender is politically activated in addition, dividing the society into white females, white males and so on. In this parametric change, everyone loses peers.

In order to focus on the joint effects of increased fragmentation and particularistic norms, we abstract from other impacts that are probably correlated with changes in

[30] See Laitin (1998) for a related analysis of an "assimilation equilibrium". On the choice-theoretic foundations of identity more generally, see Calvert (2002).

identity: we assume that the rise in the number of active ascriptive groups affects neither the chance that any two specific players meet nor the stage game they play. Thus, the result shows that increased ascriptive fragmentation impairs cooperation *even when this change does not make pairwise relationships less valuable*. Hence, the effects that Proposition 4 uncovers must be due to the isolated causal mechanism: how communitarian norms implement more fragmented ascriptive identities.

The result allows for completely heterogeneous matching probabilities and stage games.

Proposition 4 *Suppose the hypotheses of Proposition 1 hold. Initially there are at least two groups; the smallest group has at least three members. Let the society become more ascriptively fragmented in the sense stipulated above; this change affects neither stage games nor matching probabilities. Then the following hold, for either the best within- or between-group norm.*

(i) *For any two players i and j who are in different groups ex ante, there exist threshold discount factors $\underline{\delta}_{ij}, \overline{\delta}_{ij}$ such that if δ is in $(\underline{\delta}_{ij}, \overline{\delta}_{ij})$ then cooperation between i and j is a subgame-perfect Nash equilibrium outcome ex ante but not ex post.*

(ii) *Suppose in addition that no ex post group is a singleton, intragroup stage games belong to G'' and within-group cooperation is regulated by the best within-group communitarian norm. Then part (i)'s conclusion holds for every pair of players.*

Thus, under part (ii)'s mild restriction on the parametric change, increased ascriptive fragmentation makes *all* pairwise cooperation, with Insiders as well as Outsiders, more difficult. This happens because greater ascriptive fragmentation, mediated by the best particularistic norms, makes social networks sparser, *de facto*; hence, deviants encounter punishment less often.[31] Therefore, insisting on being punished only by one's own kind is probably inconsistent with stable intergroup cooperation in pluralistic societies.[32]

In contrast, fragmentation has no effect on the universalistic norm, since it ignores ascriptive characteristics. Consequently, the more ascriptively fragmented the society, the bigger is the difference in the effectiveness of universalistic versus the above communitarian regimes.

The increased fragility of the best particularistic norms in the face of greater ascriptive fragmentation has sharp welfare effects.

---

[31] As the formulation of the proposition suggests, this argument's validity depends on comparing the best communitarian norms to themselves, before and after the parametric change. It need not hold for suboptimal communitarian norms. Consider, e.g., a community initially divided into racial groups. Within-group norms regulate interracial cooperation. Suppose, however, the prevailing norm says that a deviator should be punished only by someone of his or her own gender (which is clearly suboptimal). Then, if groups become finer, so that white females are in one group and white males another, cooperation is unaffected.

[32] There *is* an ascriptively-based norm that does not suffer from this effect: if i cheats an Outsider then everyone from *uninvolved* tribes punish i. An increase in m helps this norm, and at the limit ($m = n$) it converges to universalism. This might also reflect ideas of impartiality, for it is a particularistic and decentralized version of a specialized universalistic institution (police and courts): instead of an impartial judicial system, A_1's are supposed to be impartial about A_2's cheating A_3's and so on.

Corollary 3 *If all the hypotheses of Proposition 4 hold then the following conclusions obtain for either kind of communitarian norm.*

(i) *Increased ascriptive fragmentation never makes anyone better off.*
(ii) *It makes some people worse off if δ is intermediate for at least one pair.*

The criterion of δ-intermediacy identified in part (ii) is very weak: we only need one pair, say i and j, for whom δ is intermediate. That is, δ is high enough so that before the parametric change i and j can cooperate but low enough so that after the change they cannot. If there are many pairs playing different stage games then the union of these intermediate δ-intervals will be large, whence the condition is easily satisfied.

IMPERFECT MONITORING

Thus far we have assumed that people flawlessly observe everyone's play. But if ethnic contact has a spatial aspect — e.g., intragroup encounters are nearby, intergroup ones are distant — then two problems might attend the monitoring of the latter. First, knowledge of intergroup play might be more fragmentary: an A might be less likely to observe an (A, B) encounter than an (A, A) one. Second, it might be more error-prone: greater distance, spatial or social, makes mistakes more likely.

The problem of fragmentary knowledge can be addressed swiftly. The main implication is that it entails a redefinition of universalism. Because enforcing a norm requires knowing about violations, the universalistic rule should now be understood as, "*if* you have observed cheating, *then* punish the cheater." (If A_i cheats B_j but B_k does not observe this, then B_k's conduct toward A_i is unaffected.)[33]

Particularistic norms require that enforcement be based partly on group identity. Hence, they typically impose weaker punishments than does the universalistic rule described above, which requires everyone who observes a deviation to sanction the wrongdoer. Thus under communitarian norms the deviation may go unpunished and for some configurations of group identities it will not, in fact, be punished, whereas it would have been sanctioned under a universalistic code. Hence our essential point — that universalism is superior to particularism — remains intact, provided that it is common knowledge between B_k and A_i if/when the former observes the latter cheating someone else.

Thus fragmentary knowledge *per se* causes no serious problems for universalism: its injunctions must merely be conditioned on the possession of relevant information.

---

[33] Similarly, fragmentary knowledge about specific aspects of between-group interaction (e.g., an Outsider's identity) does not pose insuperable obstacles to universalism. Suppose, following Fearon–Laitin, that A_i knows that he interacted with a B but may not know which B it is, and the same problem plagues bystander A's. Yet, so long as the A's have *some* chance of ascertaining the B's identity, then it is straightforward to show that a universalistic rule of the form, "if a player shirked *and* you know his identity, then punish him..." is better at supporting cooperation than is the communitarian within-group norm. Thus, the claim that universalism adds nothing is not robust: it holds if *and only if* the probability of discerning an Outsider's identity is exactly zero.

(Not coincidentally, this helps to define "legitimate witness" in a universalistic legal code.) The second problem — a potential enforcer has *incorrect* information — is thornier. We follow Fearon–Laitin in positing that a player can tremble in between-group encounters and fail to cooperate despite intending to do so; further, such mistakes are more likely between groups than within them.[34] (In order to focus on the implications for intergroup cooperation, we assume that within-group transactions are error-free.)

The expected value of mutual cooperation in across-group encounters could now be less than that of within-group (even if the stage game payoffs are the same in the two kinds of encounters) for two linked reasons. First, the higher tremble probabilities mean that intergroup pairs cooperate with each other less than do intragroup pairs. Second, cooperation is the socially desirable outcome in games in G'. Nevertheless, including between-group sanctions as part of an overall norm can strengthen cooperation by increasing the punishment for defection, if trembles do not occur too often. Our next result shows that our tremble-free results are not knife-edge: in particular, Proposition 1's central message holds as long as errors are not too common.[35]

We now informally describe the set-up with trembles; for full details see Bendor and Mookherjee (2001). We focus on finite trigger strategy versions of particularistic and universalistic norms.[36] Given slightly imperfect monitoring, these strategies generate payoffs close to those produced by these norms under perfect monitoring. The basic technical result is that a necessary and sufficient condition for the existence of such equilibria (if errors are sufficiently infrequent) is that the corresponding grim trigger strategy

---

[34] We also use Fearon–Laitin's assumption that players defect without error. Relaxing this assumption makes the algebra messier without producing new insights.

[35] Of course, if mistakes occur very often then no guarantees can be given; indeed, cooperation may then not be a Nash equilibrium outcome. (Similar problems can arise if other common knowledge assumptions are relaxed.) These problems make cooperation hard to sustain by *any* kind of informal social control.

[36] However, because the results of Abreu (1988) do not apply to contexts of imperfect information, there is no *a priori* reason to restrict attention to (grim) trigger strategies here. As a referee pointed out, this may give rise to the following concern: might there exist some non-trigger-strategy version of the ascriptive norm which payoff-dominates any trigger strategy version of that norm, as well as any universalistic one?

We think this is unlikely: the continuity result described below indicates that the difference between trigger strategy versions of these respective norms will probably hold when this restriction is dropped. For instance, suppose that the discount factor is in the intermediate range. Then with perfect public monitoring, our preceding analysis has shown that the highest universalistic norm payoff exceeds the highest ascriptive norm payoff. The continuity result below shows that if the monitoring imperfection is sufficiently small, there is a finite trigger strategy version of the universalistic norm that produces a payoff arbitrarily close to that of the best universalistic rule when monitoring is perfect. Further, standard game theoretic arguments imply that the payoff set associated with (any version of) ascriptive norms will be upper-semi-continuous with respect to the extent of imperfect monitoring. Hence the highest ascriptive norm payoff (across *all* versions of these rules, not just finite trigger strategies) with small enough imperfections in monitoring will at most be arbitrarily close to the highest ascriptive norm payoff with perfect monitoring. Since the latter is less than the highest universalistic norm payoff with perfect monitoring, it follows that the difference between the two sets of norms would hold even when we do not restrict attention to their finite trigger strategy versions.

equilibria exist in the limit case of perfect monitoring.[37] This result yields a substantively important implication: if discount rates are in the crucial intermediate range — under perfect monitoring the best universalistic norm can support cooperation while no particularistic rule can — then this asymmetry continues to hold when monitoring is imperfect, provided mistakes are not too common.

To save space, the basic continuity result is established for only one regime: ascriptively based norms with between-group sanctions. The other cases are analogous. We now informally describe the set-up for this case.

Heavy punishments are desirable in an error-free world: they can enforce cooperation for lower values of δ yet never be used. However, it is well-known that if monitoring errors can occur, then heavy punishments may reduce expected payoffs.[38] Accordingly, when trembles can arise it is important, as theorists have long recognized, to consider using finite trigger strategies, and we shall do so here. We therefore analyze a class of trigger strategies where punishment lasts K periods. (If K is infinite then we recover the Grim Trigger, which is allowed as a limiting case.)

We define finite trigger strategies in a conventional way, modified slightly for the present context. Players are deemed to be in either *good standing* or *bad standing*. Everyone begins the game in good standing. If player A_i is in good standing at the start of period t then he remains in that status so long as he cooperates with a partner who is also in good standing. Suppose, however, A_i defects against B_j, who is in good standing. Further, there are enough A's and B's in good standing in this period. Then A_i's deviation puts him in bad standing for the next K periods, when he will be punished by every B he meets. Once the punishment phase has ended the player returns unconditionally to good standing.

---

[37] This result differs from the Folk Theorem in the theory of repeated games (see, e.g., Fudenberg and Maskin (1986), or Horner and Olszewski (2007)) in several ways. The Folk Theorem concerns the set of all possible sequential equilibrium payoffs for discount factors close to one. In contrast, we fix a discount factor and ask how the payoff sets across two specific classes of equilibrium strategies compare.

Recent versions of the Folk Theorem address the problem of private monitoring. In such contexts the history of past plays is not common knowledge across players: players obtain different signals of past plays. We deal instead with the problem of imperfect (public) monitoring: all players observe a common noisy signal of past actions, which is common knowledge amongst them. Our result does, however, relate in some ways to the principal result of Horner and Olszewski (2007): for high discount factors, they show that the set of equilibrium payoffs is continuous with respect to small degrees of imperfection in the monitoring (regardless of the privateness of the monitoring). Our result is that for any given discount factor, the set of equilibrium payoffs within each class (universalistic or ascriptive trigger strategy) is continuous with respect to small degrees of imperfection in public (but not private) monitoring.

[38] While avoiding overkill is good, it does not imply abandoning universalism. Suppose, e.g., that for trigger strategies the shortest punishment period that makes cooperation equilibrium behavior, given the values of δ and the payoffs, is two. Then, given social matching, an efficient universalistic rule would say that the first two people — *any* two people with the requisite information and retaliatory capacity — who meet a deviant should punish him. Universalism does not require that everyone sanction the deviant forever; that is merely one kind of universalistic strategy which happens to be useful in proving results.

Whenever A_i is in good standing, then B's who are also in good standing cooperate with A_i as long as enough members of *both* groups are in good standing. If the number of people in good standing falls below this threshold in either group then everyone defects in intergroup matches.

If trembles cannot occur and δ is sufficiently high — so, i.e., the Grim Trigger can support cooperation as Nash outcomes — then many finite triggers are observationally equivalent to each other and to the Grim strategy: all members of this class (with sufficiently long punishment phases) support cooperation as equilibrium behavior, and one observes constant cooperation. But if the tremble probability, denoted by ϵ, is strictly positive then mistakes will happen, thus triggering punishment phases. Hence, bouts of mutual punishment will occur even in "cooperative" equilibria.

Now consider a norm with between-group punishments for between-group cheating. Suppose A's and B's play the same finite trigger strategy. Suppose A_i and B_j, both in good standing, meet in period t. B_j cooperates; due to a tremble, A_i defects. Then A_i falls into bad standing for the next K periods, and B's who meet A_i in that time-span are supposed to sanction him. Because this norm is particularistic, A's do not punish A_i for his transgression against B_j.

This norm is a valuable resource — social capital (Coleman 1990), if one likes that metaphor. However, though in a tremble-free world the resource is never depleted, in a world with errors it waxes and wanes: the number of people in good standing rises and falls. These fluctuations can disrupt intergroup conduct. Suppose, e.g., that few A's are in good standing in period t. People in bad standing defect in intergroup encounters. Hence, few A's are available to discipline deviant B's; in the short-run most A's will defect regardless of how B's treat A's. But then the B's are inadequately policed by third-party sanctions in period t, so they will not cooperate with A's. Anticipating this, the A's will not cooperate either. Hence the stochastic depletion of a critical social resource — the number of people in good standing in *either* group — can shatter intergroup cooperation. Moreover, because trembles happen with positive probability and independently across players, breakdowns will occur.

Encounters between A's and B's thus exhibit the following patterns. Either macro-relations between the two groups are in a good phase (enough A's and B's are in good standing) or a bad one. In a good macro-phase, when two people in good standing meet, both try to cooperate and succeed with probability $(1 - \epsilon)^2$. If either is in bad standing then both defect. Thus even in good macro-phases one will observe isolated breakdowns in intergroup cooperation, either because an agent trembles and defects despite good intentions or because at least one player is in bad standing and so both defect intentionally. In a bad macro-phase one observes a *community-wide* breakdown in intergroup cooperation: all A's and B's defect against each other. Thus the community, as well as individuals, moves probabilistically between good and bad states.

This stochastic process is strongly influenced by the tremble probability ϵ — transparently so for individuals: a person playing an Outsider falls from grace with exactly that tremble probability. And because whether the overall community is in a good macro-phase depends on how many people in each group are in good standing, the chance that the community falls into a bad macro-phase also depends on ϵ. Further, since the

probability of micro-movements is continuous in ϵ and the macro-transitions are functions of the micro ones, the chance of macro-movements is also continuous in ϵ. Finally, since payoffs are functions of individual- and system-states, these too are continuous in ϵ. This is the intuitive basis for the following result.

Proposition 5 *Assume that the hypotheses of Proposition 4 hold.*

(i) *Suppose discount rates are in the intermediate range, so that with perfect monitoring some universalistic norm would support symmetric intergroup cooperation while no particularistic norm would. Then the same holds for sufficiently small amounts of imperfect monitoring.*

(ii) *Suppose the discount rates are either all in the high range or all in the low range, so that the best norm of each type support the same degree of intergroup cooperation (in terms of actions and payoffs). Then the same is approximately true in terms of payoffs and exactly true in terms of actions with sufficiently small amounts of imperfect monitoring.*

CONCLUSIONS

It is important to note that the models in this survey examine norms that impose sanctions only on the people who deviate from those social rules. Empirically, however, repressive regimes often use collective punishment, e.g., a miscreant's relatives are also punished. Such rules violate Enlightenment criteria: to punish a wrong-doer's innocent kin is to base sanctions on people's particularistic ties rather than on their actions. Unfortunately, this brutal method might be effective. Subjects are more likely to obey a ruler if they realize that deviation could lead to the deaths of loved ones. A similar point may hold for the horizontal relations of ascriptive groups as well as the vertical ones of regimes and subjects (Nakao 2006).[39] However, though punishing the innocent can be effective, it is obviously unjust. This violation of universalism is, we hope, beyond the pale.

Overall, then, our results indicate that game theory offers little comfort to prescriptive theories of communitarian norms. Universalistic rules are generally better at sustaining cooperation between groups as stable (equilibrium) behavior, and when communitarian norms are superior they are clearly unjust. Thus, classical liberalism's emphasis on formal equality — like cases should be treated alike; codes should be based on people's actions — stands up well when viewed through the lens of noncooperative game theory. Indeed, emphasizing identity politics can perversely end up hurting most the minorities that such an emphasis was intended to support.

---

[39] Suppose that player A_i cares about the well-being of other A's: her utility is increasing in their payoffs. A_i knows that the B's follow a communitarian norm that involves punishing the innocent. Hence, A_i realizes that if she cheats B_j then the B's will punish not only A_i but also all other A's. This will be an effective deterrent: as long as social ties are appropriately heterogeneous, the B's communitarian norm of collective punishment will induce A_i to cooperate when an individualistic sanction would not. Moreover, norm-enforcement could be enhanced via collective punishment, e.g., B-players threaten to punish all A's unless the A's find and sanction the A who cheated a B (Nakao 2006).

Empirically, of course, ascriptive identity can correlate with what liberal theory considers fundamental variables. For example, social networks often have ethnic contours, so intragroup encounters are easier to monitor than between-group ones.

Normatively, however, we believe that ascriptive identity should be regarded as merely a proxy for more basic properties: here, who has accurate information about deviations from social codes and who can enforce those codes. These fundamental considerations pertain to information and actions. These in principle are universalistic; only contingent social facts link them to ethnicity or other ascriptive features. At most, communitarian rules are second-best: desirable only when certain practical constraints bind.[40]

Taking dynamics into account makes this view matter still more. Stereotypes can be self-fulfilling. Consequently, if we shrug helplessly at current ascriptive constraints (e.g., all A's look alike to B's, so the personal identity of miscreant A's is unknown to B's) they are more likely to endure. Insisting on universalistic principles — what should matter, in courts or in informal systems of code-enforcement, is not who one is but what one does — might help destabilize these social traps.

Our game-theoretic analysis of the merits of universalism is consistent with the classical position in political sociology that proclaimed the superiority of such norms over particularistic ones. However, whereas we only analyzed equilibria, sociologists such as Tonnies were also interested in dynamics: will societies move from communitarian norms to universalistic ones? The present survey cannot speak to this question.

But work on the evolution of norms has shown that in many strategic settings efficient norms enjoy an evolutionary advantage — larger basins of attraction (Bendor and Swistak 2001) — over inefficient ones. Since universalistic norms are often more efficient than particularistic ones, this is promising news for universalism.

APPENDIX

Proof of Theorem 1

Since norm-like strategies can have features (i.e., third-party sanctions) lacked by dyadic ones, sufficiency must hold: if a dyadic strategy can support a sequence of symmetric outcomes as Nash equilibrium outcomes, then there must be a norm-like strategy that also does so. Hence, we focus on proving necessity.

Consider a sequence of symmetric outcomes,
$\{(a^*(g_{t_1}), a^*(g_{t_1})), (a^*(g_{t_2}), a^*(g_{t_2})), \ldots, \}$, where $a^*(g_{t_k})$ denotes the prescribed action at date t_k, in the stage game of that date. Suppose that no dyadic strategy upholds this sequence of outcomes as a subgame perfect Nash equilibrium. So there is a date $t_\bullet$ such

[40] For related reasons we suspect that descriptive theories that try to explain particularistic norms purely as efficient solutions to strategic problems will also encounter difficulties. This kind of neo-functionalist explanation must confront the fact that ascriptively based rules are very common, often prevailing even when there are good reasons for believing that universalistic codes would be superior. There is room for non-rational factors in descriptive theories of norms, especially those pertaining to group identity, as social psychologists (e.g., Brewer and Brown 1998) have long suggested.

that in the play between i and j, player i is better off deviating at $t_\bullet$ even though j will punish him in all succeeding periods (if there are any) with the maximal punishment available to j in each of those periods.

Since all ties are homogeneous, it must be optimal for i to deviate with *all* her partners at date $t_\bullet$, even if each partner will respond by punishing i maximally.

Hence collectively, i's partners minmax i in all encounters after $t_\bullet$, i.e., collectively they impose the maximal feasible punishment in all periods after i's optimal deviation in $t_\bullet$. But since no strategy, norm-like or otherwise, can do more than to inflict this maximal sanction on a deviator, it follows that no norm-like strategy can make the optimal deviation unprofitable for i in $t_\bullet$. ■

Proof of Proposition 1

(i) Sufficiency is obvious. Regarding necessity, note that given perfect monitoring and the properties of punishment actions in stage-games in G, it is easily established that maximal credible punishment is provided by the Grim Trigger (GT) strategy. Hence the best symmetric equilibrium of any norm involves everyone in the appropriately defined group using GT against a deviant.

Given social matching, player i meets only one other player j at any given date and i will not defect at that date if and only if j's retaliation plus the norm's punishment (weakly) exceeds i's temptation:

$$T_{i,j} - R_{i,j} \leq \frac{\delta}{1-\delta} \sum_{k=1}^{m} p_{i,k}(R_{i,k} - P_{i,k}), \tag{A.1}$$

where payoffs are subscripted to denote pair-specific stage games and $p_{i,k}$ is the probability that i meets k in any period. Since all stage games belong to G', $R_{i,k} - P_{i,k} > 0$ for all i and k; since all pairs form with positive probability, $p_{i,k} > 0$. So dropping people from the defined group of punishers must decrease the RHS of (A.1) while leaving the LHS unchanged. Hence, if the universalistic GT's punishment does not exceed the temptation then nothing does, which establishes necessity.

(ii) A norm's critical δ-threshold, δ^*, is determined by making Eq. (A.1) an equality and solving for δ. Since $p_{i,k}(R_{i,k} - P_{i,k}) > 0$ for all i and k, adding people to the defined group of punishers must increase the RHS of (A.1). Thus, given that the RHS is increasing in δ, restoring equality in (A.1) entails reducing δ^*, if one adds new punishers. Since $m < n - 2$ for any particularistic norm that is not *de facto* universalistic, new punishers could in fact be added, and the result follows. ■

Proof of Proposition 2

It is convenient to prove part (iii) first. We know from Proposition 1 that the best universalistic norm has a lower threshold for cooperation (a lower δ^*) than does any communitarian norm. Call this $\underline{\delta}$. (Of course, even $\underline{\delta}$ exceeds zero, given the short-run temptation to defect, per Equation (A.1).) At the other end of the spectrum, even purely dyadic responses can enforce cooperation if δ is sufficiently close to one. Call this threshold $\overline{\delta}$. By the proof of part (ii) of Proposition 1, the threshold for cooperation is

decreasing in the number of punishers; among other things, this means that $\underline{\delta} < \overline{\delta}$. So we have $0 < \underline{\delta} < \overline{\delta} < 1$. Since $\underline{\delta}$ is the cooperation-threshold for the best universalistic norm, cooperation in all pairs is a subgame-perfect Nash outcome for any $\delta > \underline{\delta}$, so the same is true trivially for any $\delta \in (\underline{\delta}, \overline{\delta})$, thus establishing part (iii).

Parts (i) and (ii) follow from Equation (A.1) and the fact that in any intergroup match it is the smaller group's size that determines cooperation's viability, for both types of particularistic norms. To see this, consider the best between-group norm: if someone from group A_r cheats a member of A_s, with $n_r > n_s$, then all A_s's punish the deviant in every opportunity thereafter. Hence an A_r will not cheat iff $T_b - R_b \leq n_s p \frac{\delta}{1-\delta}(R_b - P_b)$, where p is the probability of encountering a specific Outsider, T_b is the between-group payoff for exploiting a cooperating Outsider and so forth. An A_s will not cheat iff $T_b - R_b \leq n_r p \frac{\delta}{1-\delta}(R_b - P_b)$. Since the temptation to cheat is the same but an A_s faces a bigger deterrent ($n_r > n_s$), it follows that the critical δ-threshold is higher for a would-be cheater from the larger group. Under a within-group norm, would-be cheaters from the *smaller* group have the higher δ^*, for the same reason: there are fewer punishers. So for both norms, δ^* is determined by the smaller group's size. Since the proof of Proposition 1 showed that δ^* is decreasing in the number of punishers, we get the following ordering of threshold discount factors for cooperation across different pairs of groups: $\delta^*(A_q, A_r) < \delta^*(A_q, A_s) = \delta^*(A_r, A_s)$, where $n_q \geq n_r > n_s$. This rank-ordering implies the conclusions of parts (i) and (ii). ■

Proof of Corollary 1

(i) Compare the expected per period payoff, in the best symmetric equilibrium of either communitarian norm, for someone from a big group, say $A_{i,r}$ (player i in group r), to that of someone from a small group, say $A_{j,s}$. Between-group play gives $A_{i,r}$ an edge: she cooperates (plays a_c) with people from other big groups whereas $A_{j,s}$ does not cooperate with any Outsider (and all between-group stage games are the same, with R_b the highest symmetric payoff). The only other difference in their expected per period payoffs is that $A_{i,r}$ is more likely to meet Insiders than is $A_{j,s}$, since all between-group pairs are equally likely to form and $A_{j,s}$ has more of them. Hence, since the payoff from cooperating with Insiders exceeds payoffs produced by any symmetric noncooperative outcome in games with Outsiders, the person from the big group must do better, in expectation, than does the person in the small group.

(ii) No one in small groups cooperates with Outsiders and (in the best symmetric equilibrium) cooperates with Insiders. Hence, the only difference in their expected per payoffs arises from the size of their own groups. If j's group is bigger than l's, then j is more likely to wind up cooperating with his partner. So, since such cooperation with Insiders gives higher payoffs than does any symmetric noncooperative outcome with Outsiders, the conclusion follows. ■

Proof of Proposition 3

Consider first the within-group norm. If anyone cheats in an intergroup match, he or she is punished indefinitely by all of his or her peers. Since A's outnumber B's and A2′

holds, any A is more likely to encounter a peer, either before or after the parametric change, than is a B likely to run into a B-peer. Since all within-group games are the same, by Equation (A.1) the δ-threshold is determined by the size of the B-group.

Now consider the between-group norm. Under this norm, a cheating A is punished thereafter by all B's; a deviating B, by all A's. Again, since A's outnumber B's and all between-group games are the same, the former's sanction-threat outweighs the latter's, so once again the δ-threshold is determined by the number of B's.

Therefore, the feasibility of intergroup cooperation under either communitarian norm is determined by the size of the smaller group. So as n_B falls, B-enforcers are lost and δ^* must rise by Equation (A.1). Once $n_B = 1$, the solitary B has no peers to help him punish a deviant A under the between-group norm or to discipline himself under the within-group norm. Hence δ^* becomes what can be supported by dyadic punishments alone. ■

Proof of Corollary 2

(i) Either the parametric change alters the *ex ante* equilibrium or it does not. If it does not then payoffs are unaffected so B's are not better off. Alternatively, suppose it does change the equilibrium. Proposition 3 implies that if the best symmetric within-group communitarian norm can support intergroup cooperation as a Nash equilibrium *ex post* then it must have been able to do so *ex ante*; the same property holds for the best symmetric between-group norm. Hence, if the parametric change alters the equilibrium, then *ex post* A's and B's do not cooperate with each other. Since the stage game is in G', the (a_c, a_c) outcome Pareto-dominates all other symmetric ones, so this effect of the collapse of cooperation must reduce B-payoffs. Further, after the B-group has shrunk, the remaining B's are now relating as Outsiders to people who had previously been Insiders but who now pass as A's. Finally, the corollary assumes that the payoff from cooperating with Insiders exceeds those of any symmetric noncooperative outcome with Outsiders. This too lowers B-payoffs.

(ii) If δ is in $(\underline{\delta}, \overline{\delta})$ then by Proposition 3 the (a_c, a_c) between-group equilibrium is destroyed. As it can be replaced only by an equilibrium with symmetric noncooperation, B-payoffs fall, by the reasoning of (i), above.

(iii) If the increase in lopsidedness does not alter the *ex ante* equilibrium then A-payoffs do not change. So if A's become better off then the equilibrium *has* changed, and part (i) showed that this hurts the B's. ■

Proof of Proposition 4

(i) By the definition of "increased ascriptive fragmentation," if i and j are in different groups (called for convenience A and B, respectively) *ex ante* then so they remain *ex post*. Consider first the best within-group norm. Since the punishment action minmaxes the deviant and is a best reply to itself, in the best within-group norm everyone else in A punishes i indefinitely if i cheats j; the parallel norm operates in B if j cheats. In the best between-group norm, everyone in B punishes i for deviating; similarly, all A's sanction j if the latter deviates.

In either case, increased fragmentation means that *ex post* i and j belong to groups that are proper subsets of A and B, respectively. Hence at least one feasible punisher is lost in each group. By Equation (A.1) and the proof of Proposition 1, this loss means that the critical threshold on δ for cooperation between players i and j must rise. This implies the existence of the required $(\underline{\delta}_{ij}, \overline{\delta}_{ij})$ interval.

(ii) Since (i) dealt with circumstances in which i and j started out in different groups, here we need only consider those in which they are initially in the same group. There are two subcases. In case 2a they wind up in different groups; in 2b, they stay in the same one.

Case 2a. *Ex ante*, the best within-group norm for regulating the relationship between i and j requires everyone in the group (call it A) to punish i, if he cheats, or j, if she does. In this case the two players wind up in different groups, called $A(i)$ and $A(j)$, respectively. By the definition of increased fragmentation both $A(i)$ and $A(j)$ are proper subsets of A. Hence *ex post*, the best within-group norm for regulating the i–j relationship has lost punishers, relative to the *ex ante* situation, whence Equation (A.1) and the proof of Proposition 1 imply that the critical δ-threshold must rise. The same logic applies to the best *ex post* between-group norm.

Case 2b. Here i and j stay in the same group. Given the parametric change, this group contracts. Hence, given that the best within-group norm for regulating the i–j relationship has lost enforcers, the conclusion obtains as in case 2a. ■

Proof of Corollary 3

Consider an arbitrary player i. If δ is not in any $(\underline{\delta}_{ij}, \overline{\delta}_{ij})$ interval, for any partner j of i, then i maintains *ex post* the same equilibrium-relationship with every partner that she had *ex ante*. Hence her payoffs do not change.

Alternatively, suppose there is at least one partner j such that $\delta \in (\underline{\delta}_{ij}, \overline{\delta}_{ij})$. Then Proposition 4 implies that after the parametric change i and j can no longer support cooperation as subgame-perfect Nash equilibrium outcomes. Since all stage games are in G', the new *ex post* equilibrium is Pareto-inferior to the cooperative outcome sustained earlier.

Hence, any such player i is made worse off by increased fragmentation. ■

Because the proof of Proposition 5 is rather technical, it has been omitted. It has been provided to the referees and can be obtained from the authors upon request.

REFERENCES

Abreu, D. 1988. "On the Theory of Infinitely Repeated Games." *Econometrica* 56: 383–396.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Banks, J., and R. Sundaram. 1989. "Repeated Games, Finite Automata, and Complexity." *Games and Economic Behavior* 2: 97–117.

Barry, B. 2001. *Culture and Equality: An Egalitarian Critique of Multiculturalism*. Cambridge, MA: Harvard University Press.

Benhabib, S. 2002. *The Claims of Culture: Equality and Diversity in the Global Era*. Princeton: Princeton University Press.
Bendor, J., and D. Mookherjee. 1990. "Norms, Third-Party Sanctions, and Cooperation." *Journal of Law, Economics, and Organization* 6: 33–63.
Bendor, J., and D. Mookherjee. 2001. "Regulating Intergroup Conflict: Ascriptive versus Universalistic Norms." Unpublished manuscript, Stanford University.
Bendor, J., and P. Swistak. 2001. "The Evolution of Norms." *American Journal of Sociology* 106: 1493–1545.
Binmore, K., and L. Samuelson. 1992. "Evolutionary Stability in Repeated Games Played by Finite Automata." *Journal of Economic Theory* 57: 278–305.
Bowles, S., and H. Gintis. 2004. "Persistent Parochialism: Trust and Exclusion in Ethnic Networks." *Journal of Economic Behavior and Organization* 55: 1–23.
Brewer, M., and R. Brown. 1998. "Intergroup Relations." In *The Handbook of Social Psychology*, eds. Daniel Gilbert, Susan Fiske, and Gardner Lindzey, Boston: McGraw-Hill.
Brown, D. 1991. *Human Universals*. Philadelphia: Temple University Press.
Calvert, R. 1995. "Rational Actors, Equilibrium, and Social Institutions." In *Explaining Social Institutions*, eds. Jack Knight and Itai Sened, Ann Arbor: University of Michigan Press.
Calvert, R. 2002. "Identity, Expression, and Rational-Choice Theory." In *Political Science: The State of the Discipline*, eds. Ira Katznelson and Helen Milner, New York: W. W. Norton.
Coleman, J. 1990. *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.
Durkheim, E. [1893] 1997. *The Division of Labor in Society*. New York: Free Press.
Ellickson, R. 1991. *Order without Law.* Cambridge: Harvard University Press.
Fearon, J., and D. Laitin. 1996. "Explaining Interethnic Cooperation." *American Political Science Review* 90: 715–135.
Fehr, E., and U. Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425: 785–791.
Fehr, E., and U. Fischbacher. 2004. "Social Norms and Human Cooperation." *Trends in Cognitive Sciences* 8: 185–190.
Fudenberg, D., and E. Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54: 533–554.
Gibbs, J. 1981. *Norms, Deviance, and Social Control*. New York: Elsevier North Holland.
Hardin, R. 1995. *One for All: The Logic of Group Conflict*. Princeton: Princeton University Press.
Holmes, S. 1994. "Liberalism for a World of Ethnic Passions and Decaying States." *Social Research* 61: 599–610.
Homans, G. 1950. *The Human Group*. New York: Harcourt, Brace.
Horner, J., and W. Olszewski. 2007. "The Folk Theorem for Games with Private Almost-Perfect Monitoring." *Econometrica* 74: 1499–1544.
Horowitz, D. 1985. *Ethnic Groups in Conflict*. Berkeley: University of California Press.
Kochanska, G., and R. Thompson. 1997. "The Emergence and Development of Conscience in Toddlerhood and Early Childhood." In *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory*, eds. L. Kuczynski and J. Grusec, New York: Wiley.
Kuczynski, L., and J. Grusec. 1997. "Future Directions for a Theory of Parental Socialization." In *Parenting and Children's Internalization of Values*, eds. Kuczynski and Grusec, New York: Wiley.
Laitin, D. 1998. *Identity in Formation*. Ithaca: Cornell University Press.
Nakao, K. 2006. "The Construction of Social Orders and Interethnic Cooperation." SUNY Stony Brook Game Theory Conference.
Okin, S. 1989. *Justice, Gender, and the Family*. New York: Basic.
Ostrom, E. 1990. *Governing the Commons*. Cambridge: Cambridge University Press.
Sandel, M. 1982. *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.
Tonnies, F. [1887] 1963. *Community and Society*. New York: Harper and Row.
Wintrobe, R. 1995. "Some Economics of Ethnic Capital Formation and Conflict." In *Nationalism and Rationality*, eds. A. Breton, G. Galeotti, P. Salmon and R. Wintrobe, Cambridge: Cambridge University Press.
Wrong, D. 1961. "The Oversocialized Conception of Man in Modern Sociology." *American Sociological Review* 26: 184–193.
Young, I. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.