10

The Analytics of Human Sociality

> The whole earth had one language. Men said, "Come, let us build ourselves a city, and a tower with its top in the heavens." The Lord said, "Behold, they are one people, and they have all one language. Nothing will now be impossible for them. Let us go down and confuse their language." The Lord scattered them over the face of the earth, and they ceased building the city.
>
> Genesis 11:1

> An economic transaction is a solved political problem. Economics has gained the title of Queen of the Social Sciences by choosing solved political problems as its domain.
>
> Abba Lerner

10.1 Explaining Cooperation: An Overview

It is often said that sociology deals with cooperation and economics deals with competition. Game theory, however shows that cooperation and competition are neither distinct nor antithetical. Cooperation involves aligning the beliefs and incentives of agents with distinct interests, competition among groups requires cooperation within these groups, and competition among individuals may be mutually beneficial.

A major goal of economic theory is to show the plausibility of wide-scale cooperation among self-regarding individuals. In an earlier period, this endeavor centered on the Walrasian model of general market equilibrium, culminating in the celebrated Fundamental Theorem of Welfare Economics (Arrow and Debreu 1954, Debreu 1959, Arrow xxand Hahn 1971). However, the theorem's key assumption that market exchange can be enforced at zero cost to the exchanging parties is often violated (Arrow 1971, Bowles and Gintis 1993, Gintis 2002, Bowles 2004).

The game theory revolution replaced reliance on exogenous enforcement with repeated game models in which punishment of defectors by cooperators secures cooperation among self-regarding individuals. Indeed, when a game $\mathcal{G}$ is repeated an indefinite number of times by the same players, many of the anomalies associated with finitely repeated games (§5.1,5.7,4.11) dis-

appear. Moreover, Nash equilibria of the repeated game arise that are not Nash equilibria of $\mathcal{G}$. The exact nature of these equilibria is the subject of the Folk Theorem (§10.3), which shows that when individuals are Bayesian rational, self-regarding, have sufficiently long time-horizons, and there is adequate public information concerning who obeyed the rules and who did not, efficient social cooperation can be achieved in a wide variety of cases.

The Folk Theorem requires that a defection on the part of a player carry a signal that is conveyed to other players. We say a signal is *public* if all players receive the same signal. We say the signal is *perfect* if it accurately reports whether or not the player in question defected. The first general Folk Theorem that does not rely on incredible threats was proved by Fudenberg and Maskin (1986) for the case of perfect public signals (§10.3).

We say a signal is *imperfect* if it sometimes mis-reports whether or not the player in question defected. An imperfect public signal reports the same information to all players, but it is at times inaccurate. The Folk Theorem was extended to imperfect public signals by Fudenberg, Levine and Maskin (1994), as will be analyzed in §10.4.

If different players receive different signals, or some receive no signal at all, we say the signal is *private*. The case of private signals has proved much more daunting than that of public signals, but Folk Theorems for private but near-public signals (i.e., where there is an arbitrarily small deviation ϵ from public signals) have been developed by several game theorists, including Sekiguchi (1997), Piccione (2002), Ely and Välimäki (2002), Bhaskar and Obara (2002), Hörner and Olszewski (2006), and Mailath and Morris (2006). It is difficult to assess how critical the informational requirements of these Folk Theorems are, because generally the theorem is proved for "sufficiently small ϵ," with no discussion of the actual order of magnitude involved.

The question of the signal quality required for efficient cooperation to obtain is especially critical when the size of the game is considered. Generally, the Folk Theorem does not even mention the number of players, but in most situations, in real life, the larger the number of players participating in a cooperative endeavor, the lower the average quality of the cooperation vs. defection signal, because generally a player only observes a small number of others with a high degree of accuracy, however large the group involved. We explore this issue in §10.4, which illustrates the problem by applying the Fudenberg et al. (1994) framework to the Public Goods Game (§3.9) which in many respects is representative of contexts for cooperation in humans.

10.2 Bob and Alice Redux

	C	D
C	5,5	-3,8
D	8,-3	0,0

Suppose Bob and Alice play the Prisoner's Dilemma shown on the right. In the one-shot game there is only one Nash equilibrium, in which both parties defect. However, suppose the same players play the game at times $t = 0, 1, 2, \ldots$ This is then a new game, called a *repeated game*, in which the payoff to each is the sum of the payoffs over all periods, weighted by a *discount factor* δ, with $0 < \delta < 1$. We call the game played in each period the *stage game* of a *repeated game* in which at each period the players can condition their moves on the complete history of the previous stages. A strategy that dictates cooperating until a certain event occurs and then following a different strategy, involving defecting and perhaps otherwise harming one's partner, for the rest of the game is called a *trigger strategy*.

Note that we have exactly the same analysis if we assume that players do not discount the future, but in each period the probability that the game continues at least one more period is δ. In general, we can think of δ as some combination of discount factor and probability of game continuation.

We show that the cooperative solution (5,5) can be achieved as a subgame perfect Nash equilibrium of the repeated game if δ is sufficiently close to unity and each player uses the trigger strategy of cooperating as long as the other player cooperates, and defecting forever if the other player defects on one round. To see this, consider a repeated game that pays 1 now and in each future period to a certain player, and the discount factor is δ. Let x be the value of the game to the player. The player receives 1 now and then gets to play exactly the same game in the next period. Because the value of the game in the next period is x, its present value is δx. Thus $x = 1 + \delta x$, so $x = 1/(1-\delta)$.

Now suppose both agents play the trigger strategy. Then, the payoff to each is $5/(1-\delta)$. Suppose a player uses another strategy. This must involve cooperating for a number (possibly zero) of periods, then defecting forever; for once the player defects, his opponent will defect forever, the best response to which is to defect forever. Consider the game from the time t at which the first player defects. We can call this $t = 0$ without loss of generality. A player who defects receives 8 immediately and nothing thereafter. Thus the cooperate strategy is Nash if and only if $5/(1-\delta) \geq 8$, or $\delta \geq 3/8$. When δ satisfies this inequality, the pair of trigger strategies

is also subgame perfect, because the situation in which both parties defect forever is Nash subgame perfect.

This gives us an elegant solution to the problem, but in fact there are lots of other subgame perfect Nash equilibria to this game. For instance, Bob and Alice can trade off defecting on each other as follows. Consider the following trigger strategy for Alice: alternate $C, D, C, \ldots$ as long as Bob alternates $D, C, D, \ldots$. If Bob deviates from this pattern, defect forever. Suppose Bob plays the complementary strategy: alternate $D, C, D, \ldots$ as long as Alice alternates $C, D, C, \ldots$. If Alice deviates from this pattern, defect forever. These two strategies form a subgame perfect Nash equilibrium for δ sufficiently close to unity.

To see this, note that the payoffs are now $-3, 8, -3, 8, \ldots$ for Alice and $8, -3, 8, -3, \ldots$ for Bob. Let x be the payoffs to Alice. Alice gets -3 today, 8 in the next period and then gets to play the game all over again starting two periods from today. Thus, $x = -3 + 8\delta + \delta^2 x$. Solving this, we get $x = (8\delta - 3)/(1 - \delta^2)$. The alternative is for Alice to defect at some point, the most advantageous time being when it is her turn to get -3. She then gets zero in that and all future periods. Thus, cooperating is Nash if and only if $x \geq 0$, which is equivalent to $8\delta - 3 \geq 0$, or $\delta \geq 3/8$.

For an example of a very unequal equilibrium, suppose Bob and Alice agree that Bob will play $C, D, D, C, D, D \ldots$ and Alice will defect whenever Bob supposed to cooperate, and vice-versa. Let v_B be the value of the game to Bob when it is his turn to cooperate, provided he follows his strategy and Alice follows hers. Then, we have

$$v_B = -3 + 8\delta + 8\delta^2 + v_B\delta^3,$$

which we can solve, getting $v_B = (8\delta^2 + 8\delta - 3)/(1 - \delta^3)$. The value to Bob of defecting is 8 now and zero forever after. Hence, the minimum discount factor such that Bob will cooperate is the solution to the equation $v_B = 8$, which gives $\delta \approx 0.66$. Now let v_A be the value of the game to Alice when it is her first turn to cooperate, assuming both she and Bob follows their agreed strategies. Then we have

$$v_A = -3 - 3\delta + 8\delta^2 + v_A\delta^3,$$

which give $v_A = (8\delta^2 - 3\delta - 3)/(1 - \delta^3)$. The value to Alice of defecting rather than cooperating when it is her first turn to do so is then given by $v_A = 8$, which we can solve for δ, getting $\delta \approx 0.94$. With this discount factor, the value of the game to Alice is 8, but $v_B \approx 72.47$, so Bob gains more than nine times as much as Alice.

10.3 The Folk Theorem

The *Folk Theorem* is so called because no one knows who first thought of it—it is just part of the "folklore" of game theory. We shall first present a stripped-down analysis of the Folk Theorem with an example and provide a more complete discussion in the next section.

Consider the stage game in §10.2. There is a subgame perfect Nash equilibrium in which each player gets zero. Moreover, neither player can be forced to receive a negative payoff in the repeated game based on this stage game, because zero can be assured simply by playing D. Also, any point in the region OEABCF in Fig. 10.1 could be attained in the stage game, assuming the players could agree on a mixed strategy for each. To see this, note that if Bob uses C with probability α and Alice uses C with probability β, then the expected payoff to the pair is $(8\beta - 3\alpha, 8\alpha - 3\beta)$, which traces out every point in the quadrilateral OEABCF for $\alpha, \beta \in [0, 1]$. Only the points in OABC are superior to the universal defect equilibrium (0,0), however.

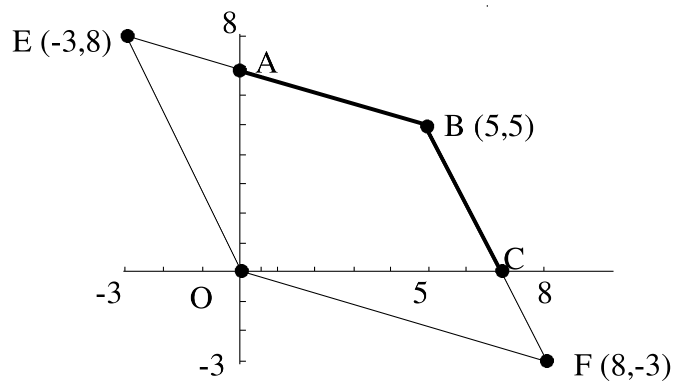

Figure 10.1. The Folk Theorem: any point in the region OABC can be sustained as the average per-period payoff the subgame perfect Nash equilibrium of the repeated game based on the stage game in §10.2.

Consider the repeated game $\mathcal{R}$ based on the stage game $\mathcal{G}$ of §10.2. The Folk Theorem says that under the appropriate conditions concerning the cooperate/defect signal available to players, any point in the region OABC can be sustained as the average per-period payoff of a subgame perfect

Nash equilibrium of $\mathcal{R}$, provided the discount factors of the players are sufficiently near unity.

More formally, consider an n-player game with finite strategy sets S_i for $i = 1, \ldots, n$, so the set of strategy profiles for the game is $S = \prod_{i=1}^{n} S_i$. The payoff for player i is $\pi_i(s)$, where $s \in S$. For any $s \in S$ we write s_{-i} for the vector obtained by dropping the ith component of s, and for any $i = 1, \ldots, n$ we write $(s_i, s_{-i}) = s$. For a given player j, suppose the other players choose strategies m^j_{-j} such that j's best response m^j_j gives j the lowest possible payoff in the game. We call the resulting strategy profile m^j the *maximum punishment payoff* for j. Then, $\pi^*_j = \pi_j(m^j)$ is j's payoff when everyone else "gangs up on him." We call

$$\pi^* = (\pi^*_1, \ldots, \pi^*_n), \tag{10.1}$$

the *minimax point* of the game. Now define

$$\Pi = \{(\pi_1(s), \ldots, \pi_n(s)) | s \in S, \pi_i(s) \geq \pi^*_i, i = 1, \ldots, n\},$$

so Π is the set of strategy profiles in the stage game with payoffs at least as good as the maximum punishment payoff for each player.

This construction describes a stage game $\mathcal{G}$ for a repeated game $\mathcal{R}$ with discount factor δ, common to all the agents. If $\mathcal{G}$ is played in periods $t = 0, 1, 2, \ldots$, and if the sequence of strategy profiles used by the players is $s(1), s(2), \ldots$, then the payoff to player j is

$$\tilde{\pi}_j = \sum_{t=0}^{\infty} \delta^t \pi_j(s(t)).$$

Let us assume that information is *public* and *perfect*, so that when a player deviates from some agreed-upon action in some period, a signal to this effect is transmitted with probability one to the other players. If players can use mixed strategies, then any point in Π can be attained as payoffs to $\mathcal{R}$ by each player using the same mixed strategy in each period. However, it is not clear how a signal indicating deviation from a strictly mixed strategy should be interpreted. The simplest assumption guaranteeing the existence of such a signal is that there is a *public randomizing device* that can be seen by all players and that players use to decide which pure strategy to use, given that they have agreed to use a particular mixed strategy. Suppose, for instance, the randomizing device is a circular disc with a pointer that can

be spun by a flick of the finger. Then, a player could mark off a number of regions around the perimeter of the disc, the area of each being proportional to the probability of using each pure strategy in a given mixed strategy to be used by that player. In each period, each player flicks his pointer and chooses the appropriate pure strategy, this behavior is recorded accurately by the signaling device, and the result is transmitted to all players.

With these definitions, we have the following, where for $\pi \in \Pi$, $\sigma_i(\pi) \in \Delta S_i$ is a mixed strategy for player i such that $\pi_i(\sigma_1, \ldots, \sigma_n) = \pi_i$:

THEOREM 10.1 Folk Theorem. *Suppose players have an available public randomizing device and the signal indicating cooperation or defection of each player is public and perfect. Then, for any $\pi = (\pi_1, \ldots, \pi_n) \in \Pi$, if δ is sufficiently close to unity, there is a Nash equilibrium of the repeated game such that π_j is j's payoff for $j = 1, \ldots, n$ in each period. The equilibrium is effected by each player i using $\sigma_i(\pi)$ as long as no player has been signaled as having defected, and playing the minimax strategy m_i^j in all future periods after player j is first detected defecting.*

The idea behind this theorem is straightforward. For any such $\pi \in \Pi$, each player j uses the strategy $\sigma_j(\pi)$ that gives payoffs π in each period, provided the other players do likewise. If one player deviates, however, all other players play the strategies that impose the maximum punishment payoff on j forever. Because $\pi_j \geq \pi_j^*$, player j cannot gain from deviating from $\sigma_j(\pi)$, so the profile of strategies is a Nash equilibrium.

Of course, unless the strategy profile $(m_1^j, \ldots, m_n^j)$ is a Nash equilibrium for each $j = 1, \ldots, n$, the threat to minimax even once, let alone forever, is not a credible threat. However, we do have the following:

THEOREM 10.2 The Folk Theorem with Subgame Perfection. *Suppose $y = (y_1, \ldots, y_n)$ is the vector of payoffs in a Nash equilibrium of the underlying one-shot game, and $\pi \in \Pi$ with $\pi_i \geq y_i$ for $i = 1, \ldots, n$. Then, if δ is sufficiently close to unity, there is a subgame perfect Nash equilibrium of the repeated game such that π_j is j's payoff for $j = 1, \ldots, n$ in each period.*

To see this, note that for any such $\pi \in \Pi$, each player j uses the strategy s_j that gives payoffs π in each period, provided the other players do likewise. If one player deviates, however, all players play the strategies that give payoff vector y forever.

10.4 The Folk Theorem with Imperfect Public Information

An important model due to Fudenberg et al. (1994) extends the Folk Theorem to many situations in which there is public imperfect signaling. Although their model does not discuss the n-player Public Goods Game, we shall here show that this game does satisfy the conditions for applying their theorem.

We shall see that the apparent power of the Folk Theorem in this case comes from letting the discount factor δ go to one *last*, in the sense that for any desired level of cooperation (by which we mean the level of *intended*, rather than *realized* cooperation), for any group size n and for any error rate ϵ, there is a δ sufficiently near unity that this level of cooperation can be realized. However, given δ, the level of cooperation may be quite low when n and ϵ are relatively small. Throughout this section, we shall assume that the signal imperfection takes the form of players defecting by accident with probability ϵ and hence failing to provide the benefit b to the group, although they expend the cost c.

The Fudenberg, Levine, and Maskin stage game consists of players $i = 1, \ldots, n$, each with a finite set of pure actions $a_1, \ldots, a_{m_i} \in A_i$. A vector $a \in A \equiv \prod_{j=1}^{n} A_i$ is called a pure action *profile*. For every profile $a \in A$ there is a probability distribution $y|a$ over the m possible public signals Y. Player i's payoff, $r_i(a_i, y)$, depends only on his own action and the resulting public signal. If $\pi(y|a)$ is the probability of $y \in Y$ given profile $a \in A$, i's expected payoff from a is given by

$$g_i(a) = \sum_{y \in Y} \pi(y|a) r_i(a_i, y).$$

Mixed actions and profiles, as well as their payoffs are defined in the usual way, and denoted by greek letters, so α is a mixed action profile, and $\pi(y|\alpha)$ is the probability distribution generated by mixed action α.

Note that in the case of a simple Public Goods Game, in which each player can cooperate by producing b for the other players at a personal cost c, each action set consists of the two elements {C,D}. We will assume that players choose only pure strategies. It is then convenient to represent the choice of C by 1 and D by 0. Let A be the set of strings of n zeros and ones, representing the possible pure strategy profiles of the n players, the kth entry representing the choice of the kth player. Let $\tau(a)$ be the number of ones in $a \in A$, and write a_i for the ith entry in $a \in A$. For any

$a \in A$, the random variable $y \in Y$ represents the imperfect public information concerning $a \in A$. We assume defections are signaled correctly, but intended cooperation fails and appears as defection with probability $\epsilon > 0$. Let $\pi(y|a)$ be the probability that signal $y \in A$ is received by players when the actual strategy profile is $a \in A$. Clearly, if $y_i > a_i$ for some i, then $\pi(y|a) = 0$. Otherwise

$$\pi(y|a) = \epsilon^{\tau(a)-\tau(y)}(1-\epsilon)^{\tau(y)} \qquad \text{for } \tau(y) \leq \tau(a). \tag{10.2}$$

The payoff to player i who chooses a_i and receives signal y is given by $r_i(a_i, y|a) = b\tau(y)(1-\epsilon) - a_i c$. The expected payoff to player i is just

$$g_i(a) = \sum_{y \in Y} \pi(y|a) r_i(a_i, y) = b\tau(a)(1-\epsilon) - a_i c. \tag{10.3}$$

Moving to the repeated game, we assume in each period $t = 0, 1, \ldots,$ the stage game is played with public outcome $y^t \in Y$. The sequence $\{y^0, \ldots, y^t\}$ is thus the *public history* of the game through time t, and we assume that the strategy profile $\{\sigma^t\}$ played at time t depends only on this public history (Fudenberg, Levine, and Maskin show that allowing agents to condition their play on their previous private profiles does not add any additional equilibrium payoffs). We call a profile $\{\sigma^t\}$ of public strategies a *perfect public equilibrium* if, for any period t and any public history up to period t, the strategy profile specified for the rest of the game is a Nash equilibrium from that point on. Thus, a public perfect equilibrium is subgame perfect Nash equilibrium implemented by public strategy profiles. The payoff to player i is then the discounted sum of the payoffs from each of the stage games.

The *minimax* payoff for player i is largest payoff i can attain if all the other players collude to choose strategy profiles that minimize i's maximum payoff—see equation 10.1. In the Public Goods Game, the minimax payoff is zero for each player, because the worst the other players can do is universally defect, in which case i's best action is to defect himself, giving payoff zero. Let V^* be the convex hull of stage game payoffs that dominate the minimax payoff for each player. A player who intends to cooperate and pays the cost c (which is not seen by the other players) can fail to produce the benefit b (which is seen by the other players) with probability $\epsilon > 0$. In the two-player case, V^* is the quadrilateral ABCD in Figure 10.2, where $b^* = b(1-\epsilon) - c$ is the expected payoff to a player if everyone cooperates.

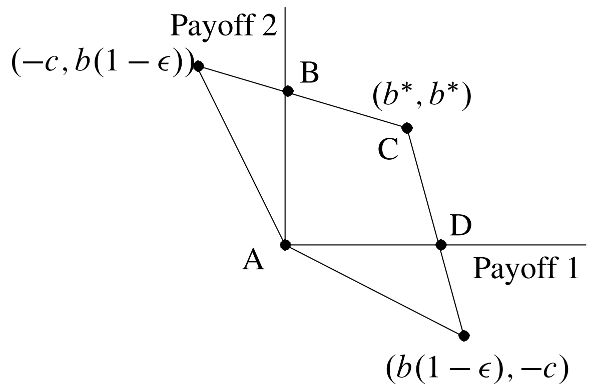

Figure 10.2. Two-player Public Goods Game

The Folk Theorem (Theorem 6.4, p. 1025 in Fudenberg, Levine, and Maskin) is then as follows.[1] We say $W \subset V^*$ is *smooth* if W is closed and convex, has a nonempty interior, and such that each boundary point $v \in W$ has a unique tangent hyperplane P_v that varies continuously with v (e.g., a closed ball with center interior to V^*). Then if $W \subset V^*$ is smooth, there is a $\underline{\delta} < 1$ such that for all δ satisfying $\underline{\delta} \leq \delta < 1$, each point in W corresponds to a strict perfect public equilibrium with discount factor δ, in which a pure action profile is played in each period. In particular, we can choose W to have a boundary as close as we might desire to $\mathbf{v}^* \equiv (b^*, \ldots b^*)$, in which case the full cooperation payoff can be approximated as closely as desired.

The only condition of the theorem that must be verified in the case of the Public Goods Game is that the full cooperation payoff $\mathbf{v}^* = \{b^*, \ldots, b^*\}$ is on the boundary of an open set of payoffs in $\mathbf{R}^n$, assuming players can used mixed strategies. Suppose player i cooperates with probability x_i, so the payoff to player i is $v_i = \pi_i - cx_i$, where

$$\pi_i = b \sum_{j=1}^{n} x_j - x_i.$$

If J is the Jacobian of the transformation $x \to v$, it is straightforward to show that

$$\det[J] = (-1)^{n+1}(b-c)\left(\frac{b}{n-1} + c\right)^{n-1},$$

[1] I am suppressing two conditions on the signal y that are either satisfied trivially or irrelevant in the case of a Public Goods Game.

which is non-zero, proving the transformation is not singular.

The method of recursive dynamic programming used to prove this theorem in fact offers an equilibrium construction algorithm, or rather, a collection of such algorithms. Given a set $W \subset V^*$, a discount factor δ, and a strategy profile α, we say α is *enforceable* with respect to W and δ if there is a payoff vector $v \in \mathbf{R}^n$ and a *continuation function* $w: Y \to W$ such that for all i,

$$v_i = (1-\delta)g_i(a_i, \alpha_{-i}) + \delta \sum_{y \in Y} \pi(y|a_i, \alpha_{-i}) w_i(y) \quad \text{for all } a_i \text{ with } \alpha_i(a_i) > 0, \tag{10.4}$$

$$v_i \geq (1-\delta)g_i(a_i, \alpha_{-i}) + \delta \sum_{y \in Y} \pi(y|a_i, \alpha_{-i}) w_i(y) \quad \text{for all } a_i \text{ with } \alpha_i(a_i) = 0. \tag{10.5}$$

We interpret the continuation function as follows. If signal $y \in Y$ is observed (the same signal will be observed by all, by assumption), each player switches to a strategy profile in the repeated game that gives player i the long-run average payoff $w_i(y)$. We thus say that $\{w(y)_{y \in Y}\}$ *enforces* α with respect to v and δ, and that the payoff v is *decomposable* with respect to α, W, and δ. To render this interpretation valid, we must show that $W \subseteq E(\delta)$, where $E(\delta)$ is the set of average payoff vectors that correspond to equilibria when the discount factor is δ.

Equations (10.4) and (10.5) can be used to construct an equilibrium. First, we can assume that the equations in (10.4) and (10.5) are satisfied as equalities. There are then two equations for $|Y| = 2^n$ unknowns $\{w_i(y)\}$ for each player i. To reduce the underdetermination of the equations, we shall seek only pure strategies that are symmetrical in the players, so no player can condition his behavior on having a particular index i. In this case, that $w_i(y)$ depends only on whether or not i signaled cooperate, and the number of other players who signaled cooperate. This reduces the number of strategies for a player at this point from 2^n to $2(n-1)$. In the interests maximizing efficiency, we assume that in the first period all players cooperate, and as long as y indicates universal cooperation, players continue to play all cooperate.

To minimize the amount of punishment meted out in the case of observed defections while satisfying (10.4) and (10.5), we first assume that if more than one agent signals defect, all continue to cooperate. If there is a single defection, this is punished by all players defecting an amount that just