Stability and fracture of social groups

Soheil Eshghi, Grace-Rose Williams, Gualtiero B. Colombo, Liam D. Turner,
David G. Rand, Roger M. Whitaker, Leandros Tassiulas

***Abstract*—In this paper, we present a mathematical model for the mutation of social groups. Group mutability has been studied in multiple domains, with insights generated on significant factors at differing scales. Mathematical modeling enables the simultaneous study of such phenomena, understanding interactions and generating hypotheses for experiments. In particular, we focus on group fracture, where individuals leave groups of which they are members. For example, this can be due to perceived differences with other group members due to norm related conflict (such as extreme actions by some members). Our aim is to consider simple mathematical models incorporating a selection of social and psychological theory which describes these phenomena as a way to understand their interplay, and describe the trade-offs and challenges.**

I. Introduction

The existence, persistence, and stability of groups is an emerging topic of study in differing contexts (see [1] for a summary). Key problems in the field include the evolution of costly cooperation [2], [3] and the persistence of groups, even in the face of (possibly negative) externalities [4]. The effect of norms as factors that internalize externalities and affect decision-making has been mathematically formalized by Coleman [5], albeit for rather simple settings. Furthermore, the importance of an individual's identity and the resulting knock-on effects on their decision-making have been studied in the economic and social psychology literature [6]. The integration of the effects of these different phenomena on group behavior and dynamics, across individual to group scales, will help generate testable hypotheses for real-world settings and provide actionable insights into the drivers of group actions.

Theoretical studies on group behavior have either directly focused on simple two-player ultimatum and prisoner's dilemma games [7], [8], [9], [10], [11], [12], [13], or bespoke games relevant to particular scenarios (e.g., group conflict [14], [15], [16]). Recently, Kranton [17] provided a roadmap for the integration of more complex models to create an identity-based meta-model for group behavior. We examine this modeling approach, as well as the social psychology-inspired factors that should be considered in such models that aim to examine group stability in an earlier work [18]. In this paper, we integrate micro-economic models that have been developed to describe group persistence-relevant phenomena to develop a mathematical theory of group fracture and stability.

Quantifying the stability of a group and relating it to the motivations and perceptions of individuals and the norms of the group help us to identify stable groups and their most vulnerable/reluctant group members, and serve as a basis for reasoning about the interactions of the aforementioned phenomena. We categorize group persistence-related phenomena based on their social dependence: interdependent phenomena are dependent on the choices and actions of other group members, while independent phenomena are related to the psychology and perception of the individual irrespective of other group members.

In [18], we explicitly describe how norms and information externalities (e.g., through social comparison [19]) are some such inter-dependent phenomena, while identity [6][1]/self-categorization [20] and individual characteristics/ability are some such independent phenomena.

A. Research Question

In particular, we focus on the effect of norm-related conflict and informational externalities in the fracture of a group. For example, this can capture the case where empirical expectations of individuals conflict with normative expectations, leading to the supremacy of empirical expectations and the changing of normative expectations [21]. Conflicts can also manifest when a specific social norm puts a significant strain on a particular individual (conflicting with a fairness norm), or when the burden it places on an individual is significant enough to convince them to leave the group (and to risk the associated negative consequences). Finally, the cost of group membership is not always explicit: decision-making can be affected by the information gleaned by an individual (e.g., through signaling [1] or through

SE and LT are with the Yale Institute for Network Science (YINS) and Electrical Engineering Department, Yale University, New Haven, USA, E-mail: {soheil.eshghi, leandros.tassiulas}@yale.edu. GRW is with the Defence Science and Technology Laboratory (Dstl), Porton Down, UK, E-mail: grwilliams1@dstl.gov.uk. GBC, LDT, and RMW are with the School of Computer Science & Informatics, Cardiff University, Cardiff, UK, E-mail: {ColomboG, TurnerL9, whitakerrm}@cardiff.ac.uk. DGR is with the Psychology Department & Economics Department & Yale School of Management (YSOM), Yale University, New Haven, USA, E-mail: david.rand@yale.edu.

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

[1] In [6], identity also depends on the actions of others (and in our parlance would be considered inter-dependent). In this paper, we use "identity" to only represent the independent part of the individual's utility, capturing the effect of others' actions on an individual's decision-making through our model of "externalities".

observational learning [22], [23], [24]). We seek to quantify insights about the relationship of these factors to the stability of social groups and to provide testable hypotheses and actionable predictions about the fracture of social groups.

B. Contribution

In this paper, we propose a new mathematical model that captures the tensions between individuals and the groups to which they belong and how this influences the stability or fracture of the group. This paper elucidates the high-level linkage between mathematical theory and social phenomena through which group stability can be assessed by way of representations of utility. We formally quantify a group-member's relative attachment to the group, which can be used to predict the least satisfied members of a group and thus the most likely to leave. This is especially important to characterize for internally-stable extremist groups, where tactically targeting particular individuals with incentives to leave the group (e.g., monetary incentives, information campaigns) so as to create division within their structure, and where such knowledge can remove the need for costly and dangerous tactical missions.

It should be noted that while we borrow liberally from the underlying assumptions of many of the works we integrate (as will be stated), our modeling approach is distinct, especially, from the game-theory literature which considers repeated interactions modeled as simple games. Our focus is on the inter-play of many different phenomena at different scales. Such a meta-model of group stability will, for the first time, allow the simultaneous exploitation of diverse insights from social psychology, sociology, and economics, as well as facilitating the understanding of their interplay.

II. Mathematical model

In this work, we develop models for individual utilities that incorporate the group-based phenomena described above. These utilities will have three broad parts:

1) *Intrinsic/Personal*: A simple cost-benefit calculation that determines the effort the individual expends in group-related activities. The calculation of actions is complicated by the diverse abilities and aptitudes of individuals.

2) *Externalities*: The actions chosen by other group members affect the perceptions of the individual and their choice of effort/action. This externality modifies the utility of an individual as well as possibly changing their chosen action. This is complicated by the limited observations of other group members and the difficulty in inferring the reason behind their actions.

3) *Group-based effects*: While the previous two parts of the utility are related to individual interactions, there are group-based effects that manifest on longer time-scales, e.g. a psychological utility due to attachment to groups (related to the salience of the group), and a representative utility related to norm-based interactions with other group members,[2] which may decrease the utility of a single individual to the benefit of other group members. More precisely, this latter utility models normative expectations by the individual.

These utilities form a special case of the identity-based utility proposed by Akerlof and Kranton [6], where the utility of an individual j in a group is modeled as:

$$U_j = U_j(\mathbf{a}_j, \mathbf{a}_{-j}, \mathbf{I}_j(\mathbf{a}_j, \mathbf{a}_{-j}; \mathbf{c}_j, \epsilon_j, \mathbf{P})),$$

where a_j represents the actions chosen by individual j, $\mathbf{a}_{-j}$ represents the actions chosen by others, and I_j is person j's identity/self-image, which other than actions, depends on the assigned social categories $\mathbf{c}_j$ and the match between the individual's characteristics ϵ_j and the group ideals $\mathbf{P}$. In our parlance, the terms with $\mathbf{a}_j$ and ϵ_j would represent intrinsic utility, the terms with $\mathbf{a}_{-j}$ would primarily represent externalities, while the $\mathbf{c}_j$-dependent terms would represent group-based effects.

To capture the mapping of different scales of group-relevant phenomena into these outlined types of utility, we break our analysis into the three broad categories outlined by Kranton [17], which are dubbed the *short*, *medium*, and *long run*. In her framework, individuals choose actions in the *short run* taking expectations, norms, identities, and categories to be fixed. In the *medium run*, individuals can take some actions to modify their empirical and normative expectations (to resolve conflict) or their relative attachment to groups (categorization). In the *long run*, nothing is fixed.

In the short run, we identify social comparison [19] as one of the externality-causing phenomena and use the mathematical framework set up by Clark and Oswald [26] to capture its effect on decision-making.[3] Social comparison is a mechanism through which individuals compare their opinions and actions with others to gain a better and possibly accurate self-evaluation [19]. This internal mechanism both affects individual decision-making and, indirectly, the decision-making of others [26], [27], [28]. Accordingly, the aggregation of information about others is hypothesized to affect the utility, and thus the decision-making, of individuals in the short-run.

We also incorporate possible differences in abilities among individuals. Social comparison among intrinsically similar individuals may lead to the straightforward adoption of successful behavior [29]. However, with heterogeneity in abilities, observing the actions of others is not necessarily informative of their effort (i.e., their strategy). Thus, the effect of social comparison is a comparison of observable actions/behaviors, or rewards, with other group members. One of the interesting results of the framework in [26] is that it has been shown to implicitly model both convention-following and contrarian characteristics in individuals.

In the medium run, we consider empirical and normative expectations [21] and group norms [30], and identity [31], [32] and their effect on group members.

[2]This is related to the social exchange theory-posited view of social interactions [25].

[3]Nothing in our analysis of the medium and long-run prevents the incorporation of other phenomena in the short-run model.

Normative expectations act as a belief about expected behavior of the individual, while empirical expectations act as an expectation of future behavior by others. When they are in conflict, the conditional preference property of group norms may make an individual less likely to follow them, adopting the empirical norms instead. From the theoretic perspective, norms are considered to act as correlating devices of a correlated equilibrium [33]. We aggregate norm-based interactions/exchanges among group-members (alternately, consider the correlated equilibrium that is played) to see their effect on individuals in the medium-run.[4] A complicating factor for norm-based consideration may also possibly be present in the medium-run: individuals are inclined to reject unfair norms (*inequity aversion*) in some instances [35].

In the medium-run, we also quantify how important the group is to an individual's self-concept[5] (i.e., their self-categorization), as well as their perception of the threats they perceive from the group as possible punishments for norm deviation. The effects, though different in origin, manifest as a group friction/stickiness in aggregate: they measure how relatively willing an individual is to suffer onerous norms in the group.

The questions investigated in this time horizon are whether a norm is self-consistent (e.g., will empirical expectations match normative expectations) and whether it is unfair [35]. If the answer to any of these questions is no, then one can predict that the normative expectations will evolve in the long-run to resolve these conflicts (in other words, the description of the norm is not stable). This view of norms and their evolution is inspired by Bicchieri [21].

The primary question under investigation in the *long run* is whether a given group is stable under the evolved (and thus self-consistent) normative expectations of the medium-run. For the purpose of this study, we define stability to mean that no member of the group would be incentivized to leave the group. We use the model of rational agents with clear preferences used in economics in this definition. If it is indeed stable, and no member will disassociate with the group of their own volition, it is instructive to understand which member is the most vulnerable group member to target with an incentive to facilitate their leaving of the group.

After making the case for a model for group stability that considers social comparison and group norms, in the next subsections we describe the constituent parts of the mathematical model and its underpinnings from an analytic perspective.

A. Short Run

We now present the additive social comparison model courtesy of Clark and Oswald [26] for the short run. In this model, individuals choose how much effort to put into an action (that is related to the purpose of the group) given their ability in that task, which is a personal, unobservable, and heterogeneous trait, as well as a subjective social comparison. In this model, an individual's private ability/fitness to perform tasks related to that goal is captured by their type $\theta \in [\underline{\theta}, \overline{\theta}]$. We assume this parameter does not change in the time-scale of consideration. An individual considers their type in choosing their effort which leads to their (observable) action $a \geq 0$. Thus, there is a trade-off for each individual, between the rewards related to taking an action, and the cost of the effort it requires. This internal trade-off is further complicated by the psychological effects of social comparison.

In this setting, an individual's short-term utility function is:

$$w(\theta, a) := (1 - s)u(a) - c(a, \theta) + sv(a - a^*),$$

where:

- $u(a)$ is the benefit of observable action $a \geq 0$ to an individual.[6] In this model, outcomes are related to actions and not to (private) abilities. This term can, for example, model the likelihood of success of the action.
- $c(\cdot, \cdot)$ is the ability-dependent cost of effort, and for each type θ is an increasing convex function of action a. For a constant a, and given that θ represents ability, the function is decreasing in θ (i.e., more capable people can achieve the same outcome with less effort).
- $v(\cdot)$ is the subjective effect of comparison with some measure (e.g., mean) of other group-member actions has on the individual. Clark and Oswald [26] focus on the case where individuals gain satisfaction from surpassing the actions of others (so downward comparison has a positive effect and upward comparison has a positive effect), but the framework can easily be extended to consider the opposite case.
- a^* is what the individual perceives to be the representative "group action". We assume, in line with the original model, that this representative action is the population mean action. This variable couples the actions of individuals in the short-run and is the source of information-related externalities.
- $s \in [0, 1]$ is a variable that tunes the relative importance of the objective and subjective (comparative) utilities to the decision-maker. If s is set to 1, the utility has no comparative element, and the model devolves into a classical economic model with objective costs and benefits.

Clark and Oswald show that in this model, if $v(\cdot)$ is concave, an increase in a^* will lead to an increase in the action a chosen by the individual, while if it is concave, the individual will decrease their effort (and thus their action) in response to an increase in a^*. Thus, this model implicitly captures behavioral types of individuals (conventional vs contrarian) in addition to capturing ability types.

[4]This is also to account for so-called generalized exchange [34], where reciprocity in interactions is not direct in every exchange, and happens at the population level over many interactions

[5]Equivalent to Akerlof and Kranton [6]'s $\mathbf{c}_j$.

[6]For simplicity, we assume scalar actions. The approach can easily be extended to encapsulate vectors of actions.

In these settings, the individual's problem to find their best action is:

$$y(\theta) := \arg\max_{a \geq 0} (1-s)u(a) - c(a,\theta) + sv(a - a^*).$$

The utility that an individual derives from comparison against their reference group depends on a characterization of the group's actions, e.g. via the mean observed action a^*. To estimate this value accurately, individuals must accumulate information. Therefore, this representative action may not always align with the true population mean. However, as an individual encounters more and more people, the empirical mean observed action will converge to the mean of the distribution. If $f(\theta)$ is the probability density function of individuals across types in terms of their ability, then $a^* = \mathbb{E}_\theta\{y(\theta)\} = \int_{\underline{\theta}}^{\overline{\theta}} y(\theta) df(\theta)$. Thus, the short-run utility $u_{sr}(\theta)$ of each individual is:

$$\begin{aligned} u_{sr}(\theta) :=& (1-s)u(y(\theta)) - c(y(\theta),\theta) \\ &+ sv(y(\theta) - \mathbb{E}_\theta\{y(\theta)\}). \end{aligned} \quad (1)$$

In this time-frame, group norms and identities can be assumed to be fixed.

B. Medium run

In the medium run, we focus on group effects. We assume that the short run dynamics have reached an equilibrium, such that the perception of a^* by group members has converged to the real population average, and individuals receive a utility of $u_{sr}(\theta)$ from the short-run dynamics.

The group has a salience to an individual that is captured through a parameter $\gamma \in [\underline{\gamma}, \overline{\gamma}]$. This parameter can also model how important a group is to a person's self-concept, or, inversely, how difficult a group is to leave (i.e., what adverse consequences or punishments would result from such an action). In effect, this acts as a "friction" term that keeps group members inside the group. For example, for a minimal group, γ would be small, while it would be large for a group that is especially important to an individual's self-concept. We assume that this parameter is fixed in the short and medium run. Thus, we can assume that there is a probability density function $f(\theta,\gamma)$ over the set of (θ,γ) pairs which describes the population. Note that this allows there to be a possible correlation between ability and salience of the group to the individual.

Each individual, knowing their private ability and the salience of the group to them, takes the action they believe other group members expect someone in their situation to take (normative expectations). We quantify the preferences of the individuals over these norm-related actions through a function, $b_n(\theta,\gamma)$, that considers the net effect of these actions on the individual. Positive $b_n(\theta,\gamma)$ denotes the case where the individual expects to derive additional utility from norm-based interactions with other group members, while negative $b_n(\theta,\gamma)$ denotes the case where the individual expects to have to contribute to other group members (i.e., take on a burden) due to the norm. For example, some norms may involve some group members helping other in-group members. In these circumstances, adhering to normative expectations may not just place no burden on the individual being helped, they might also decrease the effort they need to exert in completing the task. On the other hand, the additional burden on the helpers/donors may make group membership less desirable to them.

Note that we assume the individual has knowledge about the whole function of $b_n(\cdot,\cdot)$ (i.e., normative expectations of group members), but acts only according to what the function specifies for someone with their attributes.

Each individual will also have expectations about norm-related behavior from other group-members. These expectations will be empirical [21], and will align with their observed behavior. We quantify the preference of an individual over these empirical actions through the function $b_e(\theta,\gamma)$, which signifies the understanding of the individual about the actions taken by other group members given their private information (translated to the same scale as $b_n(\theta,\gamma)$).

There may arise a case where an individual's empirical expectations conflict with their normative ones. This will be the case when the behavior they observe from others is incompatible with the behavior they deem the others expect from them. Under these conditions, the normative expectation will be amended over time to be compatible with the empirical expectation [21]. Since we are not explicitly concerned with the process under which these changes happen,[7] we consider the end-result, where we have a convergence of b_n and b_e to a compatible functional description of the norm, $b(\theta,\gamma)$.[8] This would be the shared norm that is enforced by the group in the medium run, possibly through sanctions [36]. While there is significant work in understanding sanctioning decisions and their effect on norms, we do not consider them explicitly in this framework.[9]

For this norm to be self-compatible, it must be possible for each individual in the group to perform the action which they believe the norm prescribes for them, and for other group members to be able to sustain the expectations of an individual. In norms governing resources that are concrete (having the same value to individuals performing and receiving the action) and non-particularistic (with value that is irrespective of the identities of the actors), e.g., money and goods, as defined in Foa's resource theory of social exchange [37], this condition results in the following necessary condition for a self-sustaining norm:[10]

[7] i.e., convergence to a correlated equilibrium.

[8] This may require the pre-supposition that individuals know the correct distribution of $f(\theta,\gamma)$, as θ and γ are private variables. Understanding the informational conditions that are necessary for this convergence is beyond the scope of this paper.

[9] In the long-run, we discuss the fact that leaving the group would result in the loss of the benefit gained from group membership (codified in our model through γ). This loss can capture the effect of ostracism and sanctions for not abiding by norms.

[10] Equation (2) may apply in a more complex way for norms governing resources that do not fall within these categories.

$$\int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\gamma}}^{\overline{\gamma}} b(\theta,\gamma) f(\theta,\gamma) d\gamma d\theta \leq 0. \tag{2}$$

One could also add other constraints on to the group norm that align with psychological and social constraints. For example, it has been argued that a complementary measure of fairness of a norm is required to capture effects such as inequity aversion [38]. One could, in principle, include other mathematical constraints on the set of norms under consideration in to capture such observations. One such constraint could penalize norms that place large expectations, or even give large benefits, to a subset of people [39].

$$\int_{\underline{\theta}}^{\overline{\theta}} \int_{\underline{\gamma}}^{\overline{\gamma}} \big(b(\theta,\gamma)\big)^w f(\theta,\gamma) d\gamma d\theta \leq F, \tag{3}$$

where F is a fixed upper-bound and $w > 1$ is a parameter that represents how sensitive the fairness constraint is to placing higher burdens on individuals, with higher w indicating more concern for the fairness of the norm. One can also formulate versions of (3) that account for fairness within specific subgroups of the social group.

Putting norm-related and group-identity related factors together, the medium-run utility $u_{mr}(\theta,\gamma)$ of each individual is:

$$u_{mr}(\theta,\gamma) := \gamma + b(\theta,\gamma). \tag{4}$$

C. Long run

In the medium run, we assumed that individuals have "learned" the norms of the group (e.g., perceived norms of group members and collective norms have converged). In the long-run, given that even the group itself is not considered fixed, we focus on the stability of the group. In particular, we discuss each individual's choice of whether to remain in the group and to be bound by its norms, or to leave the group and risk sanctions by group members. The preference is codified through a comparison of the individual's utility within the group and outside the group[11].

Leaving the group would also modify the expected short-run and medium-run utility of the individual, as in the short-run, outside the group, the individual will be deprived of the feedback provided by observing the actions of group-members. This would translate to a change in an individual's utility function, and therefore their chosen action.

$$y^o(\theta) := \arg\max_{a \geq 0} u(a) - c(a,\theta).$$

Thus, the expected short-run utility of an individual outside the group can be calculated from:

$$u^o_{sr}(\theta) := u(y^o(\theta)) - c(y^o(\theta),\theta)). \tag{5}$$

[11]This approach is similar to the "comparison level for alternative" in social exchange theory (defined by Thibaut and Kelley [40]), which is technically "the lowest level of relational rewards a person is willing to accept given available rewards from alternative relationships or being alone" [41].

The medium-run effects we described are both related to group membership, and will have no effect once the individual leaves the group[12].

1) Long run group stability with no fairness norms: Individuals make the decision to leave the group or to stay by considering their preference over these choices, as captured by a comparison of total in-group utility with that they could expect to sustain outside the group in the short and medium run:

$$u_{mr}(\theta,\gamma) + u_{sr}(\theta) \geq u^o_{mr}(\theta,\gamma) + u^o_{sr}(\theta), \tag{6}$$

where $u_{sr}(\theta)$ and $u_{mr}(\theta,\gamma)$ (respectively $u^o_{sr}(\theta)$ and $u^o_{mr}(\theta,\gamma)$) are an individual's expected short-run and medium-run utility inside (outside) the group. The difference between the two sides of the inequality represents the starkness of the difference between the choices for (θ,γ)-individuals. One can think of this difference as representing the additional encouragement (in the form of outside incentives) the individual would need to be convinced to leave the group. For example, this can capture the rewards offered to members of terrorist groups to facilitate their de-radicalization [42]).

Without loss of generality, henceforth we will only consider a finite number of (θ,γ) pairs and a related probability mass function $P(\theta,\gamma)$ so as to have cleaner definitions.

We now provide a mathematical definition for long-run *stability* $\mathcal{S}$ of a group: $\mathcal{S}$ is the minimum additional incentive that has to be offered to group members to cause one of them to leave in the long run:

$$\begin{aligned} \mathcal{S} := \min \quad & k \\ \text{s.t. } & u_{mr}(\theta,\gamma) + u_{sr}(\theta) - u^o_{mr}(\theta,\gamma) - u^o_{sr}(\theta) \leq k \ \exists_{\theta,\gamma}. \end{aligned} \tag{7}$$

Notice that if (6) does not hold for an individual in the group, then by default $\mathcal{S} < 0$. This means that such an individual will leave the group of their own volition and thus the group is unstable. In these cases, one can use this framework to study how many individuals would have to leave the group to make it stable, potentially via computational simulations.

Also note that (7) is equivalent to:

$$\begin{aligned} \mathcal{S} = \max \quad & k \\ \text{s.t. } & u_{mr}(\theta,\gamma) + u_{sr}(\theta) - u^o_{mr}(\theta,\gamma) - u^o_{sr}(\theta) \geq k \ \forall_{\theta,\gamma}. \end{aligned} \tag{8}$$

This is useful because the constraints of the maximum formulation fit the typical constrained optimization framework.

III. Stability-maximizing norms

In the long run, we seek to study what type of shared norms maximize the stability of a groups. In this study, we are not investigating how these norms are generated nor the exact mechanisms by which they are maintained, but only on their effect on the mutability of the group. Note that characterizing stability-maximizing norms allows the quantification of the most stable group that can exist with

[12]This is due to the way we have encoded γ. One can equally plausibly define γ to emphasize the relative dis-utility of being in the out-group.

any possible norm, and allows the design of interventions that would be successful without knowledge of the specific norm. We start with the case that is not constrained by (3):

$$\arg\max_{b(\cdot,\cdot)} \quad \mathcal{S}(\theta,\gamma) \tag{9}$$
$$\text{s.t.} \quad \sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b(\theta,\gamma) P(\theta,\gamma) \leq 0.$$

Thus, the problem of identifying the most stable type of group norm is equivalent to finding a solution to the following problem:

$$\arg\max_{b(\cdot,\cdot)} \quad \max_{k\geq 0} \quad k \tag{10}$$
$$\text{s.t.} \quad b(\theta,\gamma) \geq u^o_{sr}(\theta) - u_{sr}(\theta) - \gamma + k \quad \forall_{\theta,\gamma} P(\theta,\gamma) > 0,$$
$$\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b(\theta,\gamma) P(\theta,\gamma) \leq 0.$$

For simplicity, we define $\Delta(\theta) := -u^o_{sr}(\theta) + u_{sr}(\theta)$, which is the short-run utility difference of group membership. Note that (10) becomes:

$$\arg\max_{b(\cdot,\cdot)} \quad \max_{k\geq 0} \quad k \tag{11}$$
$$\text{s.t.} \quad b(\theta,\gamma) - k \geq -\Delta(\theta) - \gamma \quad \forall_{\theta,\gamma} P(\theta,\gamma) > 0,$$
$$\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b(\theta,\gamma) P(\theta,\gamma) \leq 0.$$

We now explicitly show the closed-form solution for the most stable norm $b^*(\theta,\gamma)$ and most stable group $\mathcal{S}^*(\theta,\gamma)$.

Theorem 1. *For the inequity-unconstrained case, and for all (θ,γ) such that $P(\theta,\gamma) > 0$:*

$$b^*(\theta,\gamma) = -(\gamma - \mathbb{E}_\gamma\{\gamma\}) - (\Delta(\theta) - \mathbb{E}_\theta\{\Delta(\theta)\}).$$

Furthermore:

$$\mathcal{S}^*(\theta,\gamma) = \mathbb{E}_\theta\{\Delta(\theta)\} + \mathbb{E}_\gamma\{\gamma\}$$

This means that the most stable norm asks of each individual to contribute to the group in direct proportion to the relative salience of the group to the individual and the relative positive informational externalities they gain by virtue of group membership. Interestingly, this is easily compatible with real-world cases where norms persist even though they impose more onerous tasks on less capable individuals, as stability is related to relative informational externalities and not to innate ability.

This theorem also implies that in the inequity-unconstrained case, any outside offer of above $\mathbb{E}_\theta\{\Delta(\theta)\} + \mathbb{E}_\gamma\{\gamma\}$ (i.e., the mean informational externalities of group membership plus mean salience of the group to its members) for members to leave the group will be successful irrespective of the particular norm of the group.[13]

[13]i.e., which correlated equilibrium is being played.

Finally, this helps show how conditions under which individuals suffer negative short-run group-membership externalities (e.g., through social comparison) from group membership affect the stability of a group, and why an individual may choose to remain within the group under such conditions: 1) high salience of the group to the individual, 2) threats of punishment for leaving the group, and 3) compensation from other group members through the prevalent social norms.

A. Proof of Theorem 1

We prove the theorem by putting together the results of two observations, both proved in the appendix. Assume $b^*(\cdot,\cdot)$ and k^* are the optimal solutions to (11). We first define $R := \{(\theta,\gamma) \in [\underline{\theta},\overline{\theta}] * [\underline{\gamma},\overline{\gamma}] | b^*(\theta,\gamma) - k^* = -\Delta(\theta) - \gamma\}$, the set of all (θ,γ) pairs for which the individual group membership inequality constraint holds with equality for the optimal solution.

Observation 1. $P\{R\} = 1$.

Thus, we have for all (θ,γ) such that $P(\theta,\gamma) > 0$:

$$b^*(\theta,\gamma) - k^* = -\Delta(\theta) - \gamma. \tag{12}$$

We also show that the self-sustaining inequality is tight for the most stable norm:

Observation 2. $\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b^*(\theta,\gamma) P(\theta,\gamma) = 0$.

Replacing (12) into the statement of Observation 2, we have:

$$\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b^*(\theta,\gamma) P(\theta,\gamma) = \sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} (k^* - \Delta(\theta) - \gamma) P(\theta,\gamma)$$
$$= k^* - \mathbb{E}_\theta\{\Delta(\theta)\} - \mathbb{E}_\gamma\{\gamma\} = 0.$$

Therefore, $k^* = \mathbb{E}_\theta\{\Delta(\theta)\} + \mathbb{E}_\gamma\{\gamma\}$, and so for all such (θ,γ), $b^*(\theta,\gamma) = -(\gamma - \mathbb{E}_\gamma\{\gamma\}) - (\Delta(\theta) - \mathbb{E}_\theta\{\Delta(\theta)\})$.

IV. Conclusions, Challenges, and Next Steps

In this paper, we provide a novel mathematical model for social group stability that quantifies the relative strength of a group by measuring the motivation of its most vulnerable member. We show that the most stable group norms are those that place burdens on individuals in accordance with the relative salience of the group to the individual (as compared to their peers) and the relative benefit they gain from informational externalities provided by the group. We also show that these informational externalities need not be in alignment with the abilities of group members, leading to seemingly paradoxical cases where stable norms ask more of less able group members.

The insights derived from the framework presented, which are gleaned from integrating models at different scales, are difficult to hypothesize in experimental settings due to the complexity of the interactions among the numerous social and psychological processes. Understanding the interplay between these processes mathematically allows us to posit

hypotheses about such complex interactions that can be amenable to experimentation.

In future work, we will seek to understand the effects of inequity aversion on the stability of norms, as well as explicitly characterizing the effects of contrarian and conformist behavior types on group stability. Analyzing the process by which unstable groups can stabilize by shedding members is another interesting question for study.

APPENDIX

A. Proof of Observation 1

Proof: By contradiction. Assume $0 < P(R) < 1$. This means[14] that there exists (θ_1, γ_1) such that $P(\theta_1, \gamma_1) = \epsilon > 0$ and $b^*(\theta_1, \gamma_1) - k^* + \Delta(\theta_1) + \gamma_1 = \omega > 0$. Now, consider the norm $b'(\cdot, \cdot)$ where $b'(\theta, \gamma) = b^*(\theta, \gamma) + \frac{\omega m}{(1-\epsilon)}$ for $(\theta, \gamma) \neq (\theta_1, \gamma_1)$ such that $P(\theta, \gamma) > 0$, $b'(\theta_1, \gamma_1) = b^*(\theta_1, \gamma_1) - \frac{\omega m}{\epsilon}$, and where $m = \min_{\{(\theta,\gamma):P(\theta,\gamma)>0\}} P(\theta, \gamma)$. In this case,

$$
\begin{aligned}
&\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b'(\theta,\gamma)P(\theta,\gamma) \\
&:= \sum_{(\theta,\gamma)\neq(\theta_1,\gamma_1)} b'(\theta,\gamma)P(\theta,\gamma) + b'(\theta_1,\gamma_1)P(\theta_1,\gamma_1) \\
&= \sum_{(\theta,\gamma)\neq(\theta_1,\gamma_1)} \big(b^*(\theta,\gamma) + \frac{\omega m}{(1-\epsilon)}\big)P(\theta,\gamma) \\
&\qquad + \big(b^*(\theta_1,\gamma_1) - \frac{\omega m}{\epsilon}\big)P(\theta_1,\gamma_1) \\
&= \sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b^*(\theta,\gamma)P(\theta,\gamma) \leq 0,
\end{aligned}
$$

so $b'(\cdot, \cdot)$ fulfills (2) (is a self-sustaining norm). Furthermore, this new norm satisfies the constraints in (8) (stated for k^*) with strict inequalities for all (θ, γ) that have positive probability (as $\omega - \frac{\omega m}{\epsilon} = \omega(\frac{\epsilon-\omega}{\epsilon}) > 0$). Therefore, there exists a real value $\delta > 0$ such that the equalities will also hold with strict inequality for $k^* + \delta > k^*$, which is in contradiction with the optimality of k^*. ∎

B. Proof of Observation 2

Proof: By contradiction.

Assume $\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b^*(\theta, \gamma)P(\theta, \gamma) = -\tau < 0$. As in the proof of Observation 1, we define a secondary norm that for each (θ, γ) such that $P(\theta, \gamma)$, $b''(\theta, \gamma) := b^*(\theta, \gamma) + \frac{\tau}{P(\theta,\gamma)}$. In this case, $\sum_{\theta=\underline{\theta}}^{\overline{\theta}} \sum_{\gamma=\underline{\gamma}}^{\overline{\gamma}} b''(\theta, \gamma)P(\theta, \gamma) = 0$ (the new norm fulfills (2)). Furthermore, this new norm satisfies the constraints in (8), stated for $k^* + \frac{\tau}{M} > k^*$, where $M := \max_{(\theta,\gamma)} P(\theta, \gamma)$, which is a contradiction with the optimality of k^*. ∎

[14] Given that we are have a discrete, finite-valued distribution.

REFERENCES

[1] W. J. Wildman and R. Sosis, "Stability of groups with costly beliefs and practices," *Journal of Artificial Societies and Social Simulation*, vol. 14, no. 3, p. 6, 2011.

[2] H. Gintis, E. A. Smith, and S. Bowles, "Costly signaling and cooperation," *Journal of theoretical biology*, vol. 213, no. 1, pp. 103–119, 2001.

[3] A. Bear and D. G. Rand, "Intuition, deliberation, and the evolution of cooperation," *Proceedings of the National Academy of Sciences*, vol. 113, no. 4, pp. 936–941, 2016.

[4] J. S. Coleman, "Free riders and zealots: The role of social networks," *Sociological Theory*, vol. 6, no. 1, pp. 52–57, 1988.

[5] ——, *Foundations of social theory*. The Belknap Press of Harvard University, 1990.

[6] G. A. Akerlof and R. E. Kranton, "Economics and identity," *The Quarterly Journal of Economics*, vol. 115, no. 3, pp. 715–753, 2000.

[7] S. M. Allen, G. Colombo, and R. M. Whitaker, "Cooperation through self-similar social networks," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 5, no. 1, p. 4, 2010.

[8] D. Nettle and R. I. Dunbar, "Social markers and the evolution of reciprocal exchange," *Current Anthropology*, vol. 38, no. 1, pp. 93–99, 1997.

[9] R. M. Axelrod, *The evolution of cooperation*. Basic books, 2006.

[10] D. Hales and B. Edmonds, "Applying a socially inspired technique (tags) to improve cooperation in p2p networks," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 35, no. 3, pp. 385–395, 2005.

[11] M. A. Nowak, "Five rules for the evolution of cooperation," *science*, vol. 314, no. 5805, pp. 1560–1563, 2006.

[12] K. Gray, D. G. Rand, E. Ert, K. Lewis, S. Hershman, and M. I. Norton, "The emergence of us and them in 80 lines of code: Modeling group genesis in homogeneous populations," *Psychological science*, vol. 25, no. 4, pp. 982–990, 2014.

[13] F. Fu, C. E. Tarnita, N. A. Christakis, L. Wang, D. G. Rand, and M. A. Nowak, "Evolution of in-group favoritism," *Scientific reports*, vol. 2, p. 460, 2012.

[14] D. Verma, G. Pearson, D. Felmee, A. Verma, and R. M. Whitaker, "A generative model for predicting terrorist incidents," in *SPIE Defense + Security Symposium: Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VIII*, 2017.

[15] R. M. Whitaker, L. D. Turner, G. Colombo, D. Verma, D. Felmee, and G. Pearson, "Intra-group tension under inter-group conflict: a generative model using group social norms and identity," in *Proceedings of the 8th International Conference on Applied Human Factors and Ergonomics*, 2017.

[16] A. B. Naugle and M. L. Bernard, "Using computational modeling to examine shifts towards extremist behaviors in european diaspora communities," in *Advances in Cross-Cultural Decision Making*. Springer, 2017, pp. 321–332.

[17] R. E. Kranton, "Identity economics 2016: Where do social distinctions and norms come from?" *The American Economic Review*, vol. 106, no. 5, pp. 405–409, 2016.

[18] S. Eshghi, G. R. Williams, G. B. Colombo, L. Turner, D. G. Rand, R. M. Whitaker, and L. Tassiulas, "Mathematical models for social group behavior," in *Workshop on Distributed Analytics InfraStructure and Algorithms for Multi-Organization Federations (DAIS)*, 2017.

[19] L. Festinger, "A theory of social comparison processes," *Human relations*, vol. 7, no. 2, pp. 117–140, 1954.

[20] J. C. Turner and P. J. Oakes, "Self-categorization theory and social influence." 1989.

[21] C. Bicchieri, *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, 2016.

[22] A. V. Banerjee, "A simple model of herd behavior," *The Quarterly Journal of Economics*, vol. 107, no. 3, pp. 797–817, 1992.

[23] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.

[24] L. Smith and P. Sørensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, no. 2, pp. 371–398, 2000.

[25] G. C. Homans, "Social behavior as exchange," *American journal of sociology*, vol. 63, no. 6, pp. 597–606, 1958.

[26] A. E. Clark and A. J. Oswald, "Comparison-concave utility and following behaviour in social and economic settings," *Journal of Public Economics*, vol. 70, no. 1, pp. 133–155, 1998.

[27] G. Roels and X. Su, "Optimal design of social comparison effects: Setting reference groups and reference points," *Management Science*, vol. 60, no. 3, pp. 606–627, 2013.

[28] R. M. Whitaker, G. B. Colombo, S. M. Allen, and R. I. Dunbar, "A dominant social comparison heuristic unites alternative mechanisms for the evolution of indirect reciprocity," *Scientific Reports*, vol. 6, 2016.

[29] D. Friedman, "On economic applications of evolutionary game theory," *Journal of Evolutionary Economics*, vol. 8, no. 1, pp. 15–43, 1998.

[30] D. C. Feldman, "The development and enforcement of group norms," *Academy of management review*, vol. 9, no. 1, pp. 47–53, 1984.

[31] J. C. Turner, "Social categorization and the self-concept: A social cognitive theory of group behavior," *Advances in group processes*, vol. 2, pp. 77–122, 1985.

[32] J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell, *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, 1987.

[33] H. Gintis, "Social norms as choreography," *politics, philosophy & economics*, vol. 9, no. 3, pp. 251–264, 2010.

[34] T. Yamagishi and K. S. Cook, "Generalized exchange and social dilemmas," *Social Psychology Quarterly*, pp. 235–248, 1993.

[35] I. Castelli, D. Massaro, C. Bicchieri, A. Chavez, and A. Marchetti, "Fairness norms and theory of mind in an ultimatum game: judgments, offers, and decisions in school-aged children," *PloS one*, vol. 9, no. 8, p. e105024, 2014.

[36] E. Fehr and U. Fischbacher, "Social norms and human cooperation," *Trends in cognitive sciences*, vol. 8, no. 4, pp. 185–190, 2004.

[37] E. B. Foa and U. G. Foa, "Resource theory," in *Social exchange*. Springer, 1980, pp. 77–94.

[38] E. Fehr and K. M. Schmidt, "A theory of fairness, competition, and cooperation," *The quarterly journal of economics*, vol. 114, no. 3, pp. 817–868, 1999.

[39] J. S. Adams, "Inequity in social exchange," *Advances in experimental social psychology*, vol. 2, pp. 267–299, 1965.

[40] H. H. Kelley and J. W. Thibaut, *Interpersonal relations: A theory of interdependence*. John Wiley & Sons, 1978.

[41] M. E. Roloff, *Interpersonal communication: The social exchange approach*. Sage Publications, Inc, 1981, vol. 6.

[42] M. Crenshaw, "Theories of terrorism: Instrumental and organizational approaches," *The Journal of strategic studies*, vol. 10, no. 4, pp. 13–31, 1987.