Individual Variation and Population Structure in Complex Social Systems: Dynamics and Consequences

Mari Kawakatsu

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Program in
Applied and Computational Mathematics
Advisers: Simon A. Levin and Naomi E. Leonard

September 2022

Abstract

From ant colonies to human societies, social groups exhibit remarkable collective behavior. Central to the functioning of these self-organizing systems are individual heterogeneity and population structure. This dissertation explores how the interplay between these features influences collective dynamics, and vice versa, in complex social systems.

Chapter 1 investigates patterns of behavioral specialization in heterogeneous groups, using a dynamical model based on behavioral response thresholds. Testing the model predictions against experimental data from ant colonies reveals that the simplest form of response thresholds can capture the full range of observed social organization, but only if we consider variation in previously overlooked behavioral parameters.

Chapter 2 probes the role of opinion diversity in the coupled dynamics of interindividual cooperation and political polarization. Using a cultural evolution model grounded in evolutionary game theory, we show that increasing interest diversity can improve individual and collective outcomes. But partisanship reduces the dimensionality of opinion space via self-sorting along party lines, potentially yielding greater in-group cooperation at the cost of heightened polarization—an emergent tension between the individual and the collective.

Chapter 3 studies the consequences of stereotyping, or generalizing beliefs about social groups, for cooperation. Using a game-theoretic model of indirect reciprocity, we identify conditions under which stereotype use spreads via social imitation. While stereotyping behavior can boost cooperation in some scenarios, group structure in information availability gives rise to in-group favoritism, potentially resulting in an asymmetric improvement in cooperation levels, with individuals cooperating more on average but preferentially with their in-group.

Chapter 4 examines mechanisms underlying the emergence of enduring hierarchies. Using an adaptive network model, we prove that feedback between social prestige and individual-level decision-making alone can lead to stratification among otherwise equal competitors. Fitting the model to empirical data, such as hiring patterns among mathematicians, reveals that observed social systems may be near the critical threshold between egalitarianism and hierarchy.

Complex social systems—from social media platforms to democracies—shape how we organize our societies. A greater understanding of the connections among diversity, social structure, and collective dynamics may enable us to become better stewards of these systems.

Acknowledgments

I am deeply indebted to my advisers, Simon Levin and Naomi Leonard, for their mentorship. Simon has been a consistent source of wise guidance, with not only his encyclopedic knowledge but also his big heart and sense of humor. Naomi's enthusiasm for creative research has propelled me to explore the art of marrying mathematics with collective phenomena. From the start, Simon and Naomi encouraged me to pursue and refine my own research interests. This process, though highly nonlinear, has helped me become a better thinker and scientist. Thank you for believing in my abilities and allowing me the freedom to follow my curiosity.

As an unofficial co-adviser, Corina Tarnita welcomed me into her lab as if I were her own student. My research and career have benefited tremendously from her generous investment, for which I am grateful. Thank you also for showing me how to give and receive honest, constructive feedback.

I thank Dan Rubenstein for serving as an examiner for my dissertation defense and Howard Stone for serving as an examiner for my preliminary exam.

The talented scientists of the Levin, Leonard, and Tarnita labs, past and present, provided intellectually stimulating environments in which to learn and work. I thank them for sharing their enthusiasm and offering generous feedback over the years. I also thank Sandi Milburn, Ksenia Rodionova, and my officemates for making Eno Hall a welcoming home base. I am especially grateful to Sandi for her assistance with many administrative matters, her excellent birthday-party-planning skills, and her thoughtfulness in all her interactions. Special thanks go to Carole Levin, whose contagious joy always brings much warmth to the Levin lab.

This dissertation is the fruit of a team effort. I thank my insightful, creative, and generous (and often humorous) collaborators, especially Phil Chodrow, Nicole Eikmeier, Taylor Kessinger, Daniel Kronauer, Dan Larremore, Yph Lelkes, Sebastián Michel-Mata, Josh

Plotkin, Chris Tokita, and Yuko Ulrich. They provided much valuable input and feedback along the way. I am particularly grateful to Josh, my future post-doctoral mentor, for his encouragement and advice over the last few years.

Teaching has also been an important part of my graduate training. I was fortunate to be an Assistant in Instruction with four excellent teachers, Howard, Simon, Corina, and Jon Fickenscher. I thank my co-AIs in 2020, Luojun Yang and Merlijn Staps, whose dedication, humor, and companionship made pandemic-time teaching enjoyable and rewarding. I am also grateful to the McGraw Center staff for supporting me as a Graduate Teaching Fellow.

My graduate studies would have been impossible without the supportive staff in the Program in Applied and Computational Mathematics. I thank Victoria Beltra, Tina Dwyer, Lisa Giblin, Gina Holland, Audrey Mainzer, and Bernadeta Wysocka for ensuring that I and all other PACM students had everything we needed to complete our degrees.

I gratefully acknowledge support from the National Science Foundation through Grant DMS-1514606, the Army Research Office through Grant W911NF-18-1-0325, and Princeton University through the First Year Fellowship in Natural Sciences and Engineering and summer and travel funding from PACM. I also had the opportunity to participate in several scientific meetings in Princeton and Berlin as part of the Cooperation and Collective Cognition Network (CoCCoN), thanks to a Strategic Grant by Princeton University and the Humboldt University of Berlin.

Serving as a Resident Graduate Student helped me feel at home at Princeton. To the First College staff—including Anne Caswell-Klein, Sachiko Datta, Sue Giranda, Laurie Hebditch, AnneMarie Luijendijk, Garrett Meggs, Dianne Spatafore, and Johanna Rossi Wagner—and my fellow RGSs, thank you for the opportunity to be part of a team so invested in building community and supporting undergraduates. I have tremendous respect for the star RCAs and PAAs I worked with, especially Hifsa Chaudhry, Eliot Chen,

Hamza Hashem, Thomas Hontz, Kanishkh Kanodia, Lane Marsh, Natalie Nagorski, and Angie Sheehan. Advising first-years while handling life at Princeton is no small feat, let alone during a pandemic. I also want to thank my four years' worth of zees for enriching my experience with their energy, talent, and ideas. I am excited to see what they will do in the coming years.

I have been fortunate to sing with the Princeton University Glee Club under the leadership of its director extraordinaire, Gabriel Crouch, for all five years of graduate school. Among my favorite memories from Princeton is performing choral masterworks ranging from Handel's *Dixit Dominus* and Fauré's *Requiem* to Britten's *War Requiem* and Talbot's *Path of Miracles*. Joining voices with others is a beautiful and uplifting experience unlike any other.

Graduate school has also been a time of personal growth. I thank members of the Princeton Christian Fellowship, Stone Hill Church, and the Episcopal Church at Princeton, communities that helped me stay grounded and grow in my faith. I am especially grateful to Debbie Boyce and Fr. Allen Wakabayashi for their wisdom and support.

Finally, I thank my friends and family who supported me on this journey. To Bárbara Cruvinel Santiago, Sierra Janik, Skye Jerpbak, and Ralitsa Racheva, thank you for always being there for me. To my grandfather, Yoshihiro Ishibashi, a physicist, thank you for showing me from a young age that math puzzles can be fun. My doctorate won't be in physics, but I carry its spirit that math can explain (at least some of) the world around us. To Jonathan, thank you for always making me smile. Lastly, to my parents, Michiko and Michio, and my brother, Hiro, thank you for being my biggest cheerleaders, no matter where in the world I am. I am who I am today because of you.

To my family.

Contents

List of Tables

List of Figures

Introduction

Social systems across the tree of life exhibit remarkable collective behavior. Colonies of ants, termites, and other social insects harmoniously perform complex tasks, such as nest building, nursing, and navigation (Gordon, 2010; Gordon and Schwengel, 1999; Perna and Theraulaz, 2017; Seeley, 2010). Birds and fish display mesmerizing patterns, from starling murmuration in the evening sky to predator avoidance among schooling fish (Sumpter, 2006). Humans are no exception: we navigate crowded crossings with ease, vote to achieve distributed decision making, and form an extensive economy through individual transactions.

These systems typically operate without a leader or central control. For instance, no single fish plans the spatial arrangement of every individual in a group; instead, individuals adjust their positions, speeds, and headings to those of their neighbors, organizing themselves into what we recognize as a school. Such a process—in which local interactions among the components, often following simple rules, give rise to collective order that transcends the individual—is called self-organization.

Many self-organizing systems are also adaptive. Individuals comprising a system can adjust their behavior based on experience or the external environment. This can occur either within the lifetime of an individual entity—e.g., birds become more efficient navigators with experience (Pettit et al., 2013)—or across generations—e.g., species evolve for better survival in changing environments.

This dissertation explores social systems across scales that self-organize and adapt. In complexity theory, such systems are called *complex systems* or *complex adaptive systems*. While definitions of these terms vary slightly (Gell-Mann, 1994; Holland, 2006; Levin, 2002), Herbert Simon captured the general idea in his classic 1962 essay: complex (adaptive) systems are "made up of a large number of parts that interact in a non-simple way" and have the property that "the whole is more than the sum of the parts...in the...sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole" (Simon, 1962, p. 468).

Fascination with complex social systems is hardly new, and decades-long research has identified their key features. One is individual heterogeneity. Diversity among components has long been thought to make systems robust to perturbations (Levin, 2002). Recent empirical studies have revealed that behavioral variation can also facilitate collective living in animal groups, from synchronized movements of fish schools (Couzin et al., 2011, 2005; Jolles et al., 2017) and bird flocks (Aplin et al., 2014) to group coordination among primates (King and Sueur, 2011) (see Jolles et al. (2020) for a review). Fewer theoretical studies explicitly account for group composition (Jolles et al., 2020), but those that do show that even small differences among individuals can have large effects on system-level behavior (Couzin et al., 2011, 2005, 2002; Leonard et al., 2012).

Population structure also plays a fundamental role in complex social systems. Early work on complex systems identified modularity—the division of a whole into groups or clusters with more interconnections within than between them—as a critical factor that protects systems from disturbances (Levin, 2002; May, 1972; Simon, 1962). More recent work has focused on the effects of interaction patterns on contagion processes, such as information transfer in networks of honeybees (Naug, 2008) and fish (Rosenthal et al., 2015) or disease transmission in ants (Stroeymeyt et al., 2014) and, of course, humans. In

social evolution, population structure can alter whether and how much cooperation evolves (Chu, 2021; Nowak et al., 2010; Ohtsuki et al., 2006; Santos et al., 2006). Network structure can also facilitate (Bizyaeva et al., 2022; Gray et al., 2018) or skew (Stewart et al., 2019) collective decision-making. And even without an exhaustive survey of the literature, we only need to look around to understand the enduring preoccupation with social organization: groups and group identities are central to our social lives.

Despite the long-standing interest in inter-individual variation and social structure, we still lack a systematic understanding of how these factors shape collective dynamics. This stems partly from an empirical challenge: it is difficult to measure the effects of heterogeneity or interaction patterns because they are often hard to control experimentally. Theoretically, capturing these variables in a tractable manner is not simple: analyzing heterogeneous, structured populations typically requires mathematical tools different from those used for homogeneous, well-mixed populations (although significant advances have been made for the former). Moreover, the interplay among composition, structure, and collective behavior can vary across phenomena in subtle and unpredictable ways.

At the interface of mathematical, biological, and social sciences, this dissertation explores how individual heterogeneity and population structure influence collective dynamics, and vice versa, in complex social systems. The four research chapters address this overarching theme from different angles:

- *Self-organization*: How do different axes of behavioral variation drive the dynamics of task allocation in colonies of a social insect? (Chapter 1)
- *Multi-level dynamics*: How does partisanship mediate the coupled dynamics of individual-level cooperation and collective-level polarization among individuals with diverse political interests? (Chapter 2)

- *Social information*: What roles do moral reputations based on individual actions and stereotyped reputations based on group affiliations play in promoting norm-based cooperation? (Chapter 3)
- *Emergence*: How do stable, global-level rankings emerge from noisy, individual-level endorsements in networked populations? (Chapter 4)

This dissertation focuses on theory: to address these questions, I develop and analyze theoretical frameworks using computational and mathematical tools. Where possible, I have collaborated with researchers in experimental biology and network science to explore connections between theory and data.

Here I provide a summary of each chapter:

Self-organization: behavioral specialization

Social groups across taxa divide tasks among specialized individuals. Such division of labor is considered key to the ecological success of many social groups, especially social insect colonies (Robinson, 1992). Theory suggests that variation in behavioral response thresholds can generate specialized behavior among otherwise identical workers (Beshers and Fewell, 2001; Bonabeau et al., 1996a). However, few studies have considered behavioral variation along axes other than thresholds (Jeanson and Weidenmüller, 2014).

In Chapter 1, I collaborated with experimental biologists to investigate the dynamics of behavioral specialization in heterogeneous groups of social insects. The clonal raider ants in our study exhibit different behavioral types based on their genotype, morphology, or age. Previous work has invoked complex mechanisms, such as social learning (Alem et al., 2016; van de Waal et al., 2013) or information transfer (Berdahl et al., 2013; Rosenthal et al., 2015), to explain behavioral patterns in heterogeneous

groups. However, our analysis shows that simple behavioral rules suffice for specialization: a simple, dynamical model based on fixed thresholds—the simplest form of response thresholds—alone can recapitulate all experimentally observed patterns of social organization, but only if we allow for inter-individual differences in parameters other than thresholds.

These parameters, such as task efficiency or larval demand for food, remain under-explored in the literature on social insects and on behavioral specialization more broadly. Our study thus underscores the need for collective behavior research to consider diverse sources of heterogeneity. It also identifies larvae as important regulators of worker specialization, opening up novel avenues for further investigation into colony regulation. Overall, our study demonstrates that simple models can advance our understanding of collective behavior even in groups with complex compositions.

Multi-level dynamics: political polarization

Shifting the focus to human societies, Chapter 2 explores the problem of political polarization. In his essay, *Federalist No. 10*, James Madison argued that the then-nascent republic could mitigate the dangers of political sectarianism by fostering a diversity of political interests (Madison, 1787). But although Americans today care about many more political issues than 75 years ago, polarization plagues the United States.

Motivated by this paradox, Chapter 2 explores how individual-level interactions in a multidimensional issue space can shape collective-level polarization. With collaborators in theoretical biology and political science, I develop a model of cultural evolution grounded in evolutionary game theory (Nowak, 2006; Tarnita et al., 2009a), which couples cooperative behavior and opinion dynamics. Our analysis shows that societal cohesion should increase with increasing diversity of issues, confirming Madison's intuition. However, partisanship complicates the picture. Under extreme levels of

partisanship, individuals are less willing to learn from political opposites. This reduces the effective dimensionality of opinion space via self-sorting along party lines. We find that the resulting political tribalism leads to high levels of within-ideology cooperation at the expense of between-ideology polarization. In other words, tribal instincts in the political arena, as harmful as they might be to collective cohesion, could pay off at the individual level.

However, we only find this tug-of-war between the individual and the collective when individuals learn primarily from their peers. This provides a silver lining: we might be able to escape the worst perils of ideological tribalism if we occasionally explore new issues independently. Our findings emphasize the need to study polarization in a coupled, multi-level context.

Social information: stereotypes

In Chapter 3, we explore another distinctive feature of human societies: systems of social norms and reputations (Tomasello and Vaish, 2013). Norm-based reputations promote cooperation via indirect reciprocity: altruistic behavior, typically considered a moral good, may improve individuals' standings, making them more likely to receive help from others (Nowak and Sigmund, 2005). But cognitive constraints may limit access to individual-level reputations, particularly because interactions in modern societies often involve strangers whose information may be difficult to obtain. As a result, people may resort to stereotypes—generalized reputations based on individuals' group affiliations.

Chapter 3 investigates the effects of stereotype use and its evolution on cooperation. We develop a game-theoretic model of indirect reciprocity in which individuals are assigned both individual reputations based on their actions and stereotyped reputations based on their group memberships. Our analysis shows that the use of stereotypes can spread via social imitation when access to individual reputations is costly, reputation

assessments and strategy execution are error-prone, or people's actions are judged privately. But stereotyping may not always be bad for collective welfare. Despite their diminished precision relative to individualized information, stereotypes can promote cooperation if shared more widely than individual reputations, thus facilitating greater agreement among individuals about their peers' standings.

However, an important subtlety arises when reputation information is shared only within each group. Such group-wise monitoring gives rise to in-group favoritism: the structure of information sharing leads individuals to view in-group members more favorably than out-group members. This emergent phenomenon can result in an asymmetric improvement in cooperation levels, with individuals cooperating more on average but preferentially with their in-group. Our findings highlight the importance of group structures in indirect reciprocity, a topic that remains under-explored.

Emergence: stable hierarchies

Chapter 4 turns to the origins of individual variation and population structure, focusing on emergent social hierarchies. Systems across scales exhibit enduring hierarchies—stable sets of relative rankings among individuals—including those governing prestige in academia (Clauset et al., 2015) and dominance in animal groups (Hobson and DeDeo, 2015). These hierarchies play a critical role in social life: social rank can shape who gets hired by universities or attacked by conspecifics. But by what general mechanisms do persistent hierarchies arise from individual interactions? In particular, can hierarchies form even without intrinsic differences among competitors?

Developed in collaboration with network scientists, Chapter 4 examines how individual endorsements give rise to prestige-based hierarchies. We introduce an adaptive-network model in which endorsements (links) occur based on the utilities of individuals (nodes). The utility function considered preferences for those high in rank

(prestige) and those close in rank (proximity). Under different notions of rank, we prove the existence of a critical transition between egalitarianism and persistent hierarchy that depends only on prestige preference.

Strikingly, when fit to data on several real-world systems (e.g., employment flows among math PhDs), our model estimated that they were near the critical threshold. This finding offers hope: small reductions in prestige preference could unset entrenched hierarchies in real systems. But it also urges caution: feedback between social prestige and individual-level decision-making alone can lead to stratification among otherwise equal competitors. Observed differences in social rank may be "the product of accident, not worth" (DeDeo and Hobson, 2021).

Co-author contributions and prior publications

While a common theme unites the chapters in this dissertation, each of Chapters 1–4 represents a self-contained manuscript, either published (Chapters 1, 2 and 4) or to be submitted soon for publication (Chapter 3). The Notes section at the start of each chapter details publication status and prior scholarly presentations based on the material in that chapter. Chapters 1–4 contain the main texts for the manuscripts; supplementary analyses, tables, and figures are in Appendices A–D.

I am a primary author for all chapters in this dissertation. I share primary authorship with Yuko Ulrich in Chapter 1; with Sebastián Michel-Mata in Chapter 3; and with Philip Chodrow and Nicole Eikmeier in Chapter 4. The contributions of all co-authors are described in the Notes sections.

Other work related to this dissertation

In addition to Chapters 1–4, I co-authored a paper on the dynamics of cooperation and nonlinear opinion dynamics in multi-agent systems (Park et al., 2022):

> Shinkyu Park, Anastasia Bizyaeva, Mari Kawakatsu, Alessio Franci, Naomi Ehrich Leonard. Tuning cooperative behavior in games with nonlinear opinion dynamics. *IEEE Control Systems Letters*, 6:2030–2035 (2022). doi:10.1109/LCSYS.2021.3138725

Although not part of this dissertation, this work complements Chapter 2: whereas Chapter 2 describes a model grounded in evolutionary game theory, in which both opinions and behavior evolve via social *imitation*, Park et al. (2022) builds on a dynamical-systems model of opinion dynamics, in which opinions and behavior update via social *influence*. My hope is that the multiplicity of methods will help us better understand the intersection of opinion dynamics and cooperation.

Thank you for reading.

Chapter 1

Response thresholds alone cannot explain empirical patterns of division of labor in social insects

1.1 Notes

This chapter is adapted from:

> Yuko Ulrich*, Mari Kawakatsu*, Christopher K. Tokita, Vikram Chandra, Jonathan Saragosti, Corina E. Tarnita**, Daniel J. C. Kronauer**. Response thresholds alone cannot explain empirical patterns of division of labor in social insects. *PLOS Biology*, 19(6):e3001269 (2021). doi:10.1371/journal.pbio.3001269

Author contributions. Y. Ulrich and I contributed equally to this study as co-first authors, and C. E. Tarnita and D. J. C. Kronauer as co-senior authors. Y. Ulrich, C. K. Tokita, J. Saragosti, C. E. Tarnita, D. J. C. Kronauer, and I developed the study. Y. Ulrich and D. J. C. Kronauer designed the experiments, and Y. Ulrich, J. Saragosti, and V. Chandra performed the experiments. C. K. Tokita, C. E. Tarnita, and I developed the theoretical approach. C. K. Tokita and I performed the simulations, and C. K. Tokita,

C. E. Tarnita, and I analyzed the simulation results. I performed analytical calculations with input from C. E. Tarnita. Y. Ulrich, C. K. Tokita, C. E. Tarnita, D. J. C. Kronauer, and I drafted the paper, and all authors provided comments.

Prior presentations. I have given poster presentations on this work at the following conferences:

- Fields Workshop on Mathematical Ecology, Queen's University (June 2019).
- Society for Mathematical Biology 2019 Annual Meeting, Montreal, QC (July 2019).
- Association for Women in Mathematics Graduate Student Poster Session, Joint Mathematics Meetings 2020, Denver, CO (January 2020).

Acknowledgments. We thank A. Gal for advice on analyses and O. Feinerman and M. Liu for contributions to the tracking algorithms. Research reported in this chapter was supported by grants from the Faculty Scholars Program of the Howard Hughes Medical Institute, the Pew Biomedical Scholars Program, and the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM127007 (D. J. C. Kronauer); Swiss National Science Foundation Advanced Postdoc Mobility (P300P3-147900) and Ambizione (PZ00P3_168066) fellowships, and a Rockefeller University Women & Science fellowship (Y. Ulrich); Army Research Office Grant W911NF-18-1-0325 (M. Kawakatsu); National Science Foundation Graduate Research Fellowship no. DGE1656466 (C. K. Tokita); and a Henry and Marie-Josée Kravis Postdoctoral Fellowship (J. Saragosti).

1.2 Abstract

The effects of heterogeneity in group composition remain a major hurdle to our understanding of collective behavior across disciplines. In social insects, division of

labor (DOL) is an emergent, colony-level trait thought to depend on colony composition. Theoretically, behavioral response threshold models have most commonly been employed to investigate the impact of heterogeneity on DOL. However, empirical studies that systematically test their predictions are lacking because they require control over colony composition and the ability to monitor individual behavior in groups, both of which are challenging. Here, we employ automated behavioral tracking in 120 colonies of the clonal raider ant with unparalleled control over genetic, morphological, and demographic composition. We find that each of these sources of variation in colony composition generates a distinct pattern of behavioral organization, ranging from the amplification to the dampening of inherent behavioral differences in heterogeneous colonies. Furthermore, larvae modulate interactions between adults, exacerbating the apparent complexity. Models based on threshold variation alone only partially recapitulate these empirical patterns. However, by incorporating the potential for variability in task efficiency among adults and task demand among larvae, we account for all the observed phenomena. Our findings highlight the significance of previously overlooked parameters pertaining to both larvae and workers, allow the formulation of theoretical predictions for increasing colony complexity, and suggest new avenues of empirical study.

1.3 Introduction

The study of collective behavior and self-organization is an active area of research across fields, from animal movement (Sumpter, 2006) to robotics (Werfel et al., 2014), from tissue engineering (Cohen et al., 2014) to public health (Nadell et al., 2013), and from voting (Stewart et al., 2019) to conservation (Westley et al., 2018). Despite considerable theoretical and empirical advances, however, our understanding remains limited by a poor grasp on the impacts of heterogeneity in group composition on collective

organization. This limitation stems from the difficulty in precisely controlling the sources of heterogeneity and rigorously and comprehensively measuring their impacts experimentally. This empirical challenge, in turn, has hindered the systematic testing and refining of the conceptual and theoretical frameworks employed to investigate the mechanisms underlying the collective dynamics.

The colonies of social insects are striking examples of highly integrated, complex biological systems that can self-regulate without centralized control (Gordon, 1996). Consequently, social insects have emerged as powerful systems to study collective behavior and social dynamics, both experimentally and theoretically (Brahma et al., 2018; Greenwald et al., 2018; Huang and Robinson, 1996; Khuong et al., 2016; Seeley et al., 2012). An emergent, colony-level trait that has long been thought to depend on colony composition (e.g., in age, genotype, or morphology) is division of labor (DOL), the nonrandom interindividual variation in task performance among members of a social group that is consistent over time (Beshers and Fewell, 2001; Robinson, 1992). However, few experimental studies have comprehensively measured this dependence because the inherent complexity of social insect colonies usually renders their composition intractable: a typical social insect colony consists of one or more queens, dozens to thousands of workers of different (and often unknown) age, genotype, and morphology, and various brood development stages. This difficulty in controlling and replicating colony composition has hampered attempts to systematically test and refine the theoretical framework for collective organization in insect societies. Consequently, we have a limited understanding of how colony composition affects individual behavior and the emergent DOL, which, in turn, limits our understanding of the evolution of collective organization (Duarte et al., 2011).

While several proximate mechanisms have been proposed to explain DOL in social insects (see Beshers and Fewell (2001) for a review), the "vast majority of studies on the

impact of variability on colony behaviour have so far focused on the distribution of individual response thresholds and how this distribution affects the collective response behaviour" (Jeanson and Weidenmüller (2014), p. 679). In this framework, colony members are assumed to differ in their response thresholds, i.e., in their propensity to respond to task-specific stimuli indicating the group-level demand for a given task (Bonabeau et al., 1996b, 1998; Gautrais et al., 2002; Huber, 1814; Myerscough and Oldroyd, 2004; Page and Mitchell, 1998; Waibel et al., 2006). Individuals with lower thresholds perform the corresponding task more readily than individuals with higher thresholds. Stimulus intensity, in turn, decreases with the number, efficiency, and time investment of individuals performing the task. With this negative feedback loop, response thresholds offer a simple mechanism for both robust and flexible allocation of individuals to tasks (Beshers and Fewell, 2001). While refinements of response threshold models have included a self-reinforcement mechanism, whereby thresholds are modulated through experience such that individuals become more likely to perform a task that they have already performed (Beshers and Fewell, 2001; Theraulaz et al., 1998), DOL can emerge in the absence of threshold reinforcement so long as individuals differ in their response thresholds. Indeed, the simplest version of the model, which only assumes intrinsic (i.e., fixed) variation in individual thresholds, has been successful in recapitulating certain empirically observed patterns of DOL (Brahma et al., 2018; Fewell and Jr, 1999; Gordon, 1989; Holbrook et al., 2013; Jeanson and Fewell, 2008; Jeanson et al., 2007; Pankiw and Page, 2000; Ulrich et al., 2018).

Empirically, worker behavior in social insect colonies often correlates with individual traits (Jeanson and Weidenmüller, 2014). For example, within a colony, workers of different age (Hinze and Leuthold, 1999; Naug and Gadagkar, 1998; Seeley, 1982; Tripet and Nonacs, 2004), experience (Ravary et al., 2007), genotype (Fewell and Page, 1993) (e.g., patrilines (Eyer et al., 2012; Frumhoff and Baker, 1988) or matrilines (Blatrix et al., 2000)), or morphology (e.g., size (Blanchard et al., 2000; Eyer et al., 2012; Kwapich et al.,

2018; Spaethe and Weidenmüller, 2002; Wetterer, 1999) can vary in their propensity to engage in tasks such as foraging, nursing, or nest construction. Such behavioral variation is often attributed to the developmental or genetic modulation of response thresholds. However, empirical evidence suggests that response thresholds are only one of several axes of possible individual variation. For example, workers can also vary in the efficiency with which they perform tasks (Kay and Rissing, 2005; Mertl and Traniello, 2009; Wilson, 1980) or in the average time spent performing a given task (Weidenmüller, 2004). These empirical findings suggest that previously under-explored parameters may vary depending on developmental or genetic factors and may play a role in colony organization. This possibility has led to recent calls for a diversity of parameters to be considered when investigating the relationship between colony composition and DOL (Jeanson, 2019; Jeanson and Weidenmüller, 2014; Weidenmüller et al., 2019).

Here, we combine theoretical modeling with behavioral tracking experiments in the clonal raider ant, *Ooceraea biroi*, to both assess the explanatory power of existing behavioral response threshold models and explore other axes of individual variation. The unique biology of this species affords unparalleled control over the main axes of colony composition that are thought to affect individual- and group-level behavior in social insects: genotype, age, and morphology. Specifically, colonies of clonal raider ants are naturally queenless and exclusively composed of workers that all reproduce asexually and synchronously, so that all adults within a colony are genetically almost identical and emerge in discrete age cohorts. Furthermore, individuals show variation in ovariole number that is associated with body size and other morphological features (Teseo et al., 2014), making it possible to approximately sort individuals into "regular workers" (2 to 3 ovarioles) and "intercastes" (4 to 6 ovarioles) based on their size (Teseo et al., 2014). Intercastes typically represent a small fraction (3.7% to 6.3% (Ravary and Jaisson, 2004)) of individuals in unmanipulated colonies, but colonies with higher fractions of intercastes (50% or more) do occur occasionally and are functional (Teseo

et al., 2014). Conveniently, workers of different clonal genotypes, age cohorts, and morphologies can be mixed to create functional chimeric experimental colonies (Teseo et al., 2014). Additionally, colony behavior is controlled by larvae (Ravary et al., 2006; Ulrich et al., 2016, 2018), which solicit food and care from the workers and induce them to forage. This means that colony-level task demand can be standardized or manipulated across colonies by controlling larvae number or, potentially, genotype. Finally, while colonies collected in the field contain between approximately a dozen and several hundred workers (Ravary and Jaisson, 2002; Trible et al., 2020; Tsuji and Yamauchi, 1995), smaller colonies of approximately 10 workers have high fitness and show complex collective behavior (e.g., group raiding (Chandra et al., 2021), stable DOL, and phasic reproduction (Ulrich et al., 2018)) in the laboratory. Taking advantage of these features, we quantify individual and collective behavior of *O. biroi* in response to precise, independent manipulations of colony genetic, morphological, and demographic composition, as is uniquely possible in this system.

1.4 Results and discussion

1.4.1 Theoretical model

We adopt the simplest and most commonly employed formulation of the response threshold model, which assumes that individual thresholds do not change over time (Bonabeau et al., 1996b). We consider a colony of n individuals, N_X of which are of type X and $N_Y = n - N_X$ are of type Y. Types X and Y represent any pair of the experimentally manipulated subcolony compositions (i.e., genotypes A and B, Young and Old, or Regular Workers and Intercastes). The colony must perform m tasks; for consistency with the experimental approach (see below), we assume that there are 2 tasks ($m = 2$). At a given time step, an individual can be either performing one of the m tasks (active) or not performing any (inactive). The task state of individual i at time t is given by the

binary variable $x_{ij,t}$: if individual i is active and performing task j at time t, then $x_{ij,t} = 1$ and $x_{ij',t} = 0$ for all $j' \neq j$; if individual i is inactive and resting, then $x_{ij,t} = 0$ for all j.

Each task j has an associated stimulus $s_{j,t}$, signaling the group-level demand for that task. The stimulus for a task changes depending on the rate at which the demand increases (e.g., the demand for foraging increases due to increased hunger in the colony), the efficiency with which workers perform the task (e.g., more efficient foragers decrease hunger faster), and the number of individuals performing the task. Mathematically, the stimulus $s_{j,t}$ is governed by Eq. (1.1):

$$s_{j,t+1} = s_{j,t} + \delta_j - \frac{\alpha_j^X n_{j,t}^X + \alpha_j^Y n_{j,t}^Y}{n}\,, \tag{1.1}$$

where δ_j is the task-specific demand rate, taken to be constant over time; α_j^X (respectively, α_j^Y) is the task-specific performance efficiency (i.e., the rate with which an individual decreases stimulus intensity by performing the corresponding task) of type X (respectively, type Y) individuals; and and are the numbers of type X and Y individuals performing task j at time t, respectively. We assume that individuals $i = 1, \ldots, N_X$ are of type X and individuals $i = N_X + 1, \ldots, n$ are of type Y.

Each individual i is assumed to have an internal threshold for each task j, θ_{ij}, drawn at time $t = 0$ from a normal distribution with mean μ_j and normalized standard deviation σ_j (i.e., expressed as a fraction of the corresponding mean μ_j). Thus, an individual can, and typically does, have different thresholds for different tasks. Although thresholds may change over the individuals' lifetime (Robinson, 1987), they are assumed to be fixed over the timescale of the experiments and, consequently, over the simulation runs. We refer to μ_j as the mean task threshold (or mean threshold) and to σ_j as the threshold variance for task j; each can be type and/or task specific (i.e., μ_j^X, μ_j^Y, σ_j^X, σ_j^Y).

At each time step, inactive individuals assess the m task stimuli in a random sequence until they either begin performing a task or have encountered all stimuli without landing on a task. For each encountered stimulus, individual i evaluates whether to perform the task by comparing the stimulus level to its internal threshold. Specifically, given stimulus $s_{j,t}$ and internal threshold θ_{ij}, individual i commits to performing task j with probability

$$P_{ij} = \frac{s_{ij}^{\eta}}{s_{ij}^{\eta} + \theta_{ij}^{\eta}}, \tag{1.2}$$

where parameter η governs the steepness of this response threshold function. The larger the value of η, the more deterministic the behavior; in the limit $\eta \to \infty$, the response function becomes a step function. Active individuals spontaneously quit their task with a constant quit probability τ. Active individuals can neither evaluate stimuli nor switch tasks without first quitting their current task.

Each agent-based simulation began with both stimuli set to 0 (i.e., $s_{j,t} = 0$ for $j = 1, 2$) and lasted $T = 10,000$ time steps (see Table A.2 for parameter settings).

1.4.2 Baseline model predictions

To establish baseline predictions for ant colonies with different compositions, we use the simplest implementation of this model, which assumes that ant types differ only in mean response threshold (Bonabeau et al., 1996b). We simulated colonies that were either homogeneous (pure), with a single type of ant, or heterogeneous (mixed), with two types in equal proportions. The individual thresholds for each type were drawn from a normal distribution with the type-specific mean ($\mu_j^X = \mu^X$ or $\mu_j^Y = \mu^Y$). All other model parameterstask performance efficiency, demand rate, and threshold variancewere constant across types. Thus, the only source of heterogeneity in pure colonies was the distribution of individual response thresholds, while in mixed colonies that

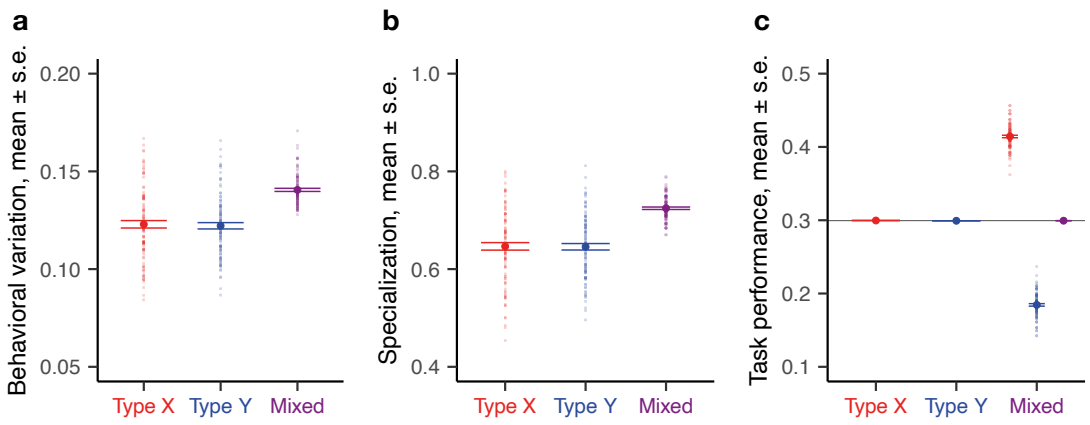

Figure 1.1: Baseline theoretical predictions. Division of labor (DOL, measured by colony-level behavioral variation (a), colony-level specialization (b)) and task performance frequency for a single task (c) shown as a function of colony composition. Opaque circles represent individual replicate colonies ($n = 100$ replicates for each composition); solid circles represent the average value across replicates; horizontal lines represent s.e.; and the horizontal gray line (c) represents the average of the pure colonies (first 2 columns). Types X and Y differ in mean threshold: $\mu^X = 10$, $\mu^Y = 20$; all other parameters are identical across types (see Table A.2).

heterogeneity was compounded by differences in the means of the type-specific distributions.

To quantify individual behavior, we computed each individual's task performance frequency for each task, defined as the fraction of time that an individual spent performing a given task. For example, if an ant spent 2,000 time steps performing task 1 (e.g., foraging), 4,000 performing task 2 (e.g., nursing), and the remaining 4,000 being inactive in a simulation of 10,000 time steps, then it had a task 1 performance frequency of 0.2 and a task 2 performance frequency of 0.4. To quantify the mean behavior of ants in a given colony for a given task, we then averaged the individual task performance frequencies for that task across all individuals in that colony. In a mixed colony, we also quantified the type-specific mean behavior for a given task by taking the average across all individuals of a given type in the colony instead. To quantify DOL, we measured two colony-level properties: behavioral variation, defined as the standard deviation of task

performance frequency across all individuals in a colony; and specialization, defined as the mean correlation in individual task performance frequencies across time, measured as the Spearman rank correlation on consecutive windows of 200 time steps. Thus, specialization measures how consistent ants in a colony are in their task performance relative to each other.

In pure colonies, there is a single normal distribution of individual thresholds for a given task. In contrast, mixed colonies have a bimodal distribution of thresholds for each task, with the thresholds of the two types clustered around the different modes. This wider distribution of thresholds resulted in both greater behavioral variation (because individuals from the lower end of the distribution for a task are more sensitive to the stimulus for that task, they tended to perform that task more often than those from the higher end) and greater colony-level specialization (those performing a task in a given time step are likely to be from the lower end of the distribution and therefore also likely to be performing that task in a future time step) relative to pure colonies, resulting in more pronounced DOL (Fig. 1.1A and B). However, all colonies, irrespective of their composition, had the same mean behavior (Fig. 1.1C). This is because while colonies may differ in how they allocate workers to tasks–within mixed colonies, the two ant types differed in their mean task performance because the type with the lower average threshold for a given task took up that task more often than the other type–they must perform the same amount of work overall to satisfy a given demand. Thus, on average, colony members spent the same fraction of time performing each task across pure and mixed colonies.

In summary, the simple model predicted that (P1) mixed colonies would exhibit higher overall DOL but that (P2) all colonies would have the same mean behavior (Fig. 1.1C), although (P3) the two types would diverge behaviorally in mixed colonies

(Fig. 1.1C). The same predictions held if, instead of differences in the means of the response thresholds, we assumed differences in the variances (Fig. A.1).

1.4.3 Empirical tests of the theoretical predictions

We then tested these theoretical predictions in experimental colonies that were either pure or 1:1 mixes of clonal raider ants that differed in one of 3 factors thought to influence DOL: genotype (A vs. B (Teseo et al., 2014; Ulrich et al., 2018)), age (around 3-month-old old ants vs. 1-month-old young ants; the life span of workers in this species is around 1 year), and size (large intercastes vs. smaller regular workers (Teseo et al., 2014)) (Table A.1). Colonies contained larvae of the same genotype as the workers; in the case of genotype effects, the experiment was performed twice, once with larvae of each genotype (see *Supplementary methods*). We analyzed individual behavior in 120 experimental colonies using automated tracking (Ulrich et al., 2018).

Because work in insect societies is spatially organized (e.g., foraging and waste disposal occur away from the nest, whereas nursing only occurs at the nest), individual spatial distribution can be used as a proxy for individual behavioral roles (Crall et al., 2018; Mersch et al., 2013; Pamminger et al., 2014; Sendova-Franks and Franks, 1995). Here, the spatial distribution of each ant was measured as the two-dimensional root-mean-square deviation (r.m.s.d.) of its spatial coordinates:

$$\text{r.m.s.d.} = \sqrt{\frac{\sum_i \left((x_i - \bar{x})^2 + (y_i - \bar{y})^2 \right)}{F}} \tag{1.3}$$

where x_i and y_i are the coordinates of the focal ant in frame i, $\bar{x}$ and $\bar{y}$ are the coordinates of the center of mass of the focal ant's overall spatial distribution, and F is the number of frames in which the focal ant was detected. As previously shown (Ulrich et al., 2018), the r.m.s.d. of an ant captures its tendency to leave the nest: workers that spend a lot of time

at the nest with the brood (e.g., nursing the larvae) and little time performing extranidal tasks (e.g., foraging or waste disposal) have low r.m.s.d. values, whereas workers that spend more time away from the brood have higher r.m.s.d. values (Fig. 1.2A). Consequently, the mean r.m.s.d. of a colony reflects its collective foraging activity, as shown by the fact that r.m.s.d. increases in response to experimentally inflated nutritional demand (Ulrich et al., 2018). We therefore use the r.m.s.d. as a proxy for the propensity to perform tasks away from the nest (e.g., foraging) rather than at the nest (e.g., nursing) (Ulrich et al., 2018). Analogously to the simulations, we quantified the mean behavior of a given ant type as the average r.m.s.d. of all ants of that type in a colony; similarly, to quantify colony-level DOL, we computed behavioral variation as the standard deviation across r.m.s.d. values of all ants in a colony and specialization as the mean correlation in individual r.m.s.d. across time, measured as the Spearman rank correlation on consecutive days in the experiment (Ulrich et al., 2018) (see *Supplementary methods*).

Colonies with different compositions often differed in mean behavior (Fig. 1.2B–D), inconsistent with prediction (P2). For instance, pure colonies of genotype A on average spent more time at the nest than pure colonies of genotype B (Fig. 1.2B: B_{pure} vs. A_{pure}, LME post hoc tests: $z = 7.75$, $p = 3.64 * 10^{-14}$; Fig. 1.2C: B_{pure} vs. A_{pure}: $z = 7.45$, $p = 2.80 * 10^{-13}$). Similarly, colonies of young workers spent more time at the nest than colonies of old workers (Fig. 1.2D: Old_{pure} vs. $Young_{pure}$: $z = -6.05$, $p = 4.39 * 10^{-09}$). That ants of different genotype (Blatrix et al., 2000; Eyer et al., 2012; Frumhoff and Baker, 1988) and age (Biedermann and Taborsky, 2011; Seeley, 1982; Tripet and Nonacs, 2004) differ in their task performance is consistent with observations in other social insects. However, such behavioral differences are theoretically only predicted to emerge (and have empirically mostly been documented) within mixed colonies (as also observed here: Fig. 1.2B: B_{mixed} vs. A_{mixed}: $z = 4.61$, $p = 8.06 * 10^{-06}$, Fig. 1.2C: B_{mixed} vs. A_{mixed}: $z = 7.68$, $p = 6.57 * 10^{-14}$, Fig. 1.2D: Old_{mixed} vs. $Young_{mixed}$: $z = -13.31$, $p < 2 * 10^{-16}$),

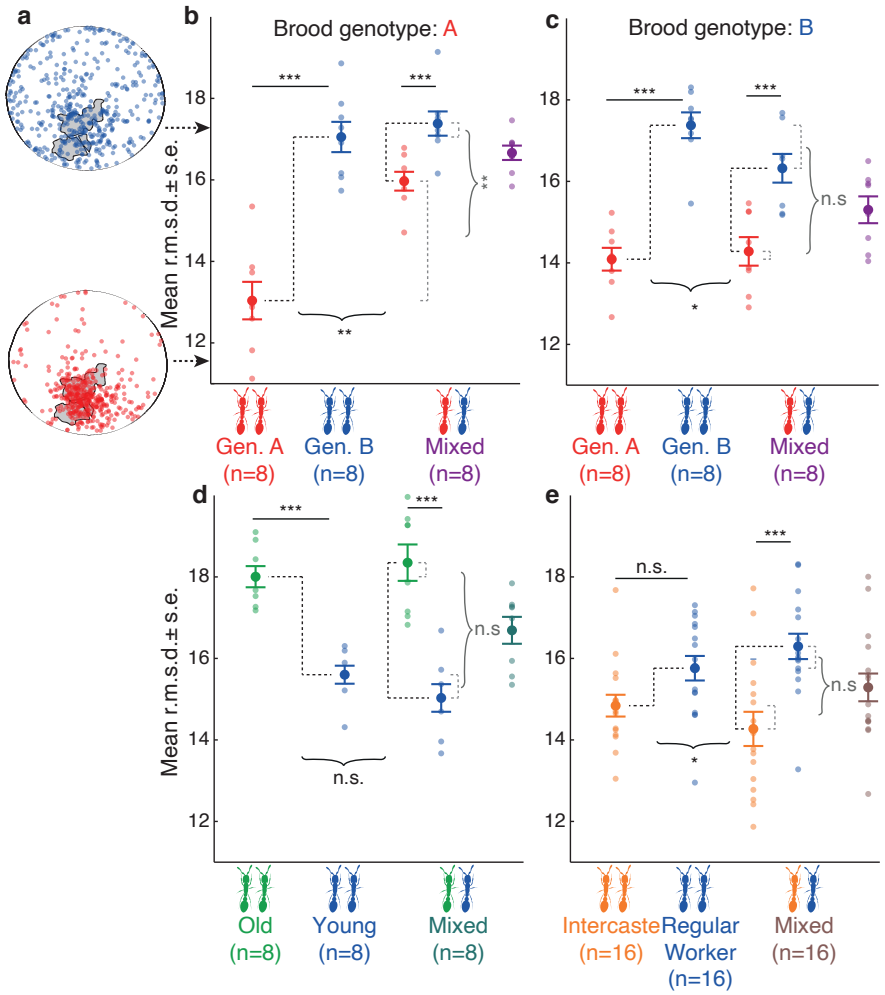

Figure 1.2: Behavior as a function of colony composition. (a) Spatial distributions of 2 ants with high (blue; genotype B) and low (red; genotype A) activity outside the nest. Arrows point to corresponding r.m.s.d. values. Gray areas represent the position of the larvae. (b–e) Mean behavior (mean r.m.s.d.) as a function of colony composition. Opaque circles represent mean behavior across individuals in replicate colonies or subcolonies. Solid circles represent average behavior across replicate colonies or subcolonies. For mixed colonies, data are shown both as type-specific and colony-level mean behavior (in average color). Sample sizes indicate the number of replicate colonies. Black curly brackets represent the effect of mixing on behavioral differences between types. (b) Behavioral convergence in genetically mixed colonies with A brood. $B_{pure} - A_{pure}$ vs. $B_{mixed} - A_{mixed}$: t test: $t = 3.86, p = 0.002$. (c) Behavioral convergence in genetically mixed colonies with B brood. $B_{pure} - A_{pure}$ vs. $B_{mixed} - A_{mixed}$: $t = 2.63, p = 0.025$. (d) No effect of mixing in demographically mixed colonies. $Old_{pure} - Young_{pure}$ vs. $Old_{mixed} - Young_{mixed}$, $t = -1.50, p = 0.157$. (e) Behavioral divergence in morphologically mixed colonies. Regular $Worker_{pure} - Intercaste_{pure}$ vs. Regular $Workers_{mixed} - Intercaste_{mixed}$: $t = -2.44, p = 0.022$. n.s., nonsignificant; $^{*}, p < 0.05$; $^{**}, p < 0.01$; $^{***}, p < 0.001$; r.m.s.d., root-mean-square deviation.

and not across pure colonies (Fig. 1.1C). Moreover, while the simple model predicted behavioral divergence between types in mixed colonies relative to pure colonies (P3), experiments produced all possible outcomes. Most surprisingly, mixing different genotypes resulted in behavioral convergence (see definitions in *Materials and methods*), whereby genotypes behaved more similarly in mixed colonies than in separation (i.e., across pure colonies) (Fig. 1.2B and C). In contrast, mixing different age cohorts had no detectable effect on mean behavior (henceforth no effect) (Fig. 1.2D). Only mixing regular workers and intercastes produced behavioral divergence as predicted by the simple model: intercastes spent more time at the nest than regular workers in mixed colonies (Fig. 1.2E: Regular $\text{Worker}_{\text{mixed}}$ vs. $\text{Intercaste}_{\text{mixed}}$: $z = 8.95$, $p < 2 * 10^{-16}$) but not in pure ones (Fig. 1.2E: Regular $\text{Worker}_{\text{pure}}$ vs. $\text{Intercaste}_{\text{pure}}$: $z = 2.14$, $p = 0.098$). Because intercastes have more ovarioles than regular workers, this behavioral difference is consistent with observations that reproductive potential often negatively correlates with the propensity to forage (Bernadou et al., 2018; Ito and Higashi, 1991; Ravary and Jaisson, 2004).

While in most cases mixing had symmetric effects on behavior—i.e., the behavior of both types was equally affected (Fig. 1.2D: $|\text{Young}_{\text{mixed}} - \text{Young}_{\text{pure}}|$ vs. $|\text{Old}_{\text{mixed}} - \text{Old}_{\text{pure}}|$: $t = 0.94$, $p = 0.365$; Fig. 1.2E: $|\text{Regular Worker}_{\text{mixed}} - \text{Regular Worker}_{\text{pure}}|$ vs. $|\text{Intercaste}_{\text{mixed}} - \text{Intercaste}_{\text{pure}}|$: $t = -0.68$, $p = 0.501$; see *Supplementary methods*)—we found that asymmetric effects are also possible: in genetically mixed colonies with A larvae, mixing affected the behavior of A workers more than that of B workers, manifesting in asymmetric behavioral convergence (Fig. 1.2B: $|\text{A}_{\text{mixed}} - \text{A}_{\text{pure}}|$ vs. $|\text{B}_{\text{mixed}} - \text{B}_{\text{pure}}|$ t test: $t = -3.86$, $p = 0.002$). Such an asymmetry was not apparent in the presence of B larvae, however (Fig. 1.2C: $|\text{A}_{\text{mixed}} - \text{A}_{\text{pure}}|$ vs. $|\text{B}_{\text{mixed}} - \text{B}_{\text{pure}}|$: $t = 0.53$, $p = 0.607$).

Consistent with these behavioral patterns, mixed colonies had overall higher DOL than pure colonies in the age and morphology experiments, in line with the baseline model prediction (P1) (Figs. A.2 and A.3). However, this trend was weakened (i.e., half of the pairwise comparisons were not significant) in the genotype experiments by the emergent behavioral convergence (Figs. A.2 and A.3), so that mixed colonies did not systematically have higher DOL than pure colonies in all experiments, violating (P1).

Taken together, our experimental results revealed a greater diversity of behavioral patterns than predicted by the simple model: colonies differed in mean behavior, thus violating (P2) (Fig. 1.2); the direction and magnitude of behavioral changes in mixed colonies depended on the specific source of workforce heterogeneity, thus violating (P3) (Fig. 1.2); and consequently, DOL was not necessarily higher in mixed than in pure colonies, thus violating (P1) (Figs. A.2 and A.3). Thus, heterogeneity in response thresholds alone was insufficient to explain our observations. This discrepancy prompted us to consider other biologically realistic sources of heterogeneity in the model.

1.4.4 An expanded model of DOL

Previous work revealed that the developmental trajectory of *O. biroi* larvae—i.e., the size of the resulting adults—depends on nonlinear interactions between the larval genotype and the genotype of the caregiving adults (Teseo et al., 2014). This finding suggests (i) that larvae of different genotypes signal different levels of demand, e.g., for food or care; and (ii) that workers of different genotypes differ in their response to a given level of larval demand, possibly via differences in their response thresholds or in the efficiency with which they perform the corresponding task. Indeed, when we added differences in task performance efficiency and in larval-induced task demand to the simple model (with between-type differences in response thresholds, i.e., consistent variation in threshold across types), we were able to qualitatively recapitulate the phenomena

observed in genotype-mixing experiments (Fig. 1.3A and B). Differences in task performance efficiency were, in fact, sufficient to robustly produce both colonies with different mean behaviors and behavioral convergence in mixed colonies, where the more efficient ants compensated for the less efficient ones by spending more time performing the task than they did in pure colonies. By affecting how much more the efficient ants needed to work in the mixed colonies, the differences in larval-induced task demand determined the asymmetry of the convergence. In particular, when task demand was so high that the less efficient type could not keep up with the demand on their own, we recovered the experimental pattern in Fig. 1.2B (see Fig. 1.3A; see also stimulus dynamics in Fig. A.4 and *Supplementary analyses*). While the simulations assumed, for simplicity, that the two tasks had the same level of demand, the analytical calculations suggest that varying demand across tasks would produce patterns qualitatively identical to either Fig. 1.3A or B, depending on the demand levels (see *Supplementary analyses*).

Exploring the efficiency-threshold parameter space broadly recapitulated not only the behavioral convergence observed in the genotype experiments, but also the divergence and no effect patterns observed in the morphology and age experiments, respectively (Fig. 1.3C). The emergent pattern in mixed colonies depended on the interplay between differences in efficiency, which increased behavioral similarity, and differences in threshold, which decreased similarity. Manipulating genotypic composition corresponded to regions of the parameter space with relatively strong effects of differences in efficiency and relatively weak effects of differences in threshold (Fig. 1.2A and B, Fig. 1.3A and B). Manipulating morphological composition corresponded to regions where differences in threshold had a relatively stronger effect (Fig. 1.2E and Fig. 1.3D). Finally, manipulating age composition corresponded to an intermediate scenario in which the 2 effects balanced each other out (Fig. 1.2D and Fig. 1.3E). Consistent with the experiments, DOL was higher in mixed colonies than in pure colonies when threshold effects were at least as strong as efficiency effects—i.e., in

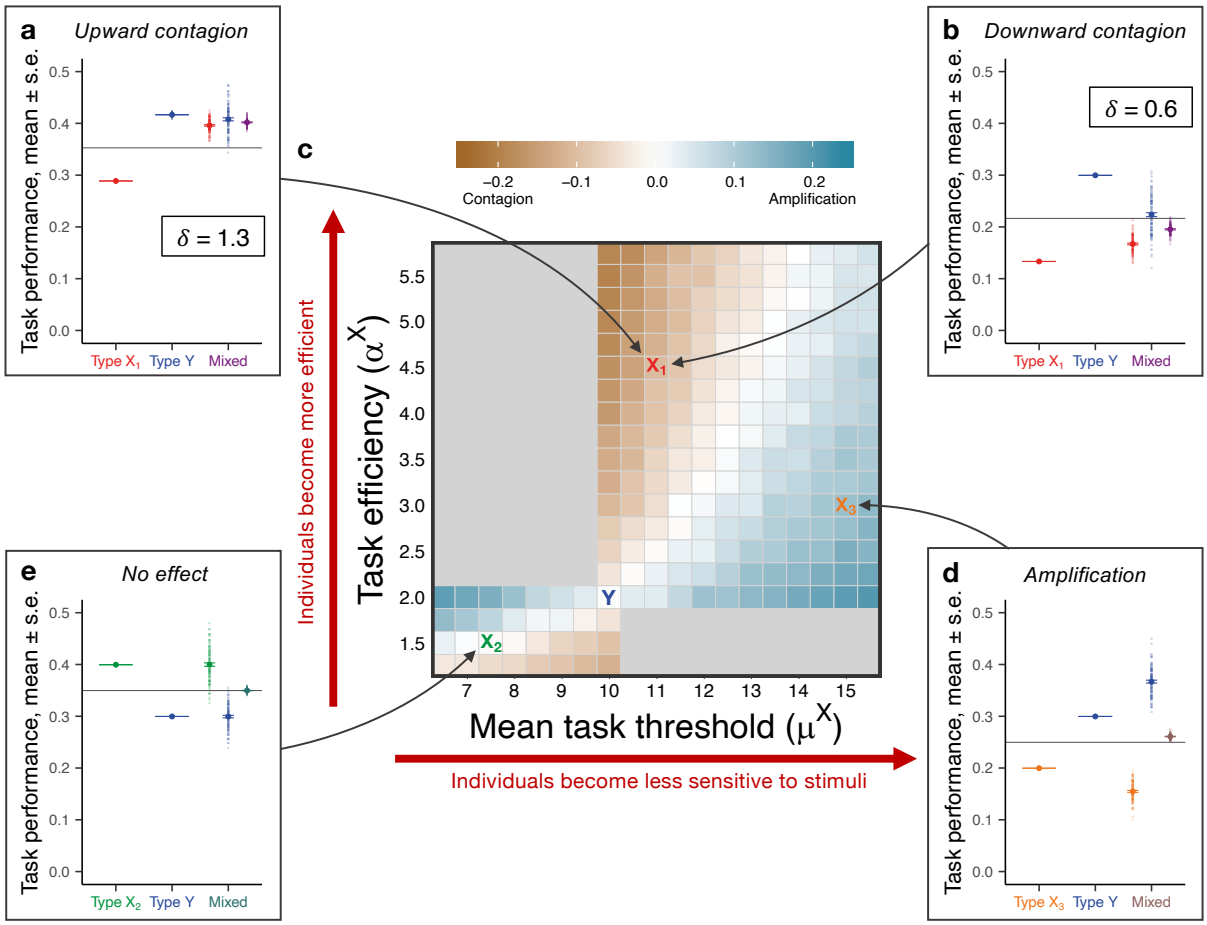

Figure 1.3: Theoretical predictions of the expanded model. (a, b, d, and e) Task performance frequency for a single task as a function of colony composition. Opaque circles represent replicate colonies ($n = 100$ replicates per composition); solid circles represent the average across replicates; horizontal bars represent s.e.; and horizontal gray lines represent the average of the pure colonies (first two columns). Identical colors indicate ants of the same type; in particular, type Y ants are the same across all panels ($\alpha^Y = 2$, $\mu^Y = 10$). (a and b) Differences in both task efficiency and mean threshold ($\alpha^{X_1} = 4.5$, $\mu^{X_1} = 11$) capture asymmetric behavioral convergence, with directionality determined by the demand rate: (a) upward ($\delta = 1.3$) and (b) downward ($\delta = 0.6$). (d and e) Differences in both task efficiency and mean threshold capture both (d) behavioral divergence ($\alpha^{X_3} = 3$, $\mu^{X_3} = 15$) and (e) a lack of effects from mixing ($\alpha^{X_2} = 1.5$, $\mu^{X_2} = 7.5$). (c) Change in relative task performance between mixed and pure colonies (measured as $(Y_m - X_m) - (Y_p - X_p)$) as a function of type X's efficiency and mean threshold ($n = 50$ replicates per parameter combination). Types X_1, X_2, X_3, and Y correspond to those in a, b, d, and e. Blue gradient indicates behavioral divergence ($Y_m - X_m > Y_p - X_p$); brown gradient indicates convergence ($Y_m - X_m < Y_p - X_p$); and light gray indicates regions with behavioral patterns falling outside our definitions (see *Materials and methods*). See Table A.2 for other parameter values.

areas of behavioral divergence or no effectbut not when threshold effects were weaker—i.e., in areas of behavioral convergence (Figs. A.5 and A.6).

1.5 Conclusions

In most social insect colonies, all factors studied here (worker genotype, age, morphology, and larval genotype) influence behavior simultaneously and in largely intractable ways. However, the unique biology of *O. biroi* allows us to break this complexity down experimentally and study each effect independently, thereby providing insight into the basic organizing principles of behavior in social groups. Our finding that the magnitude and direction of effects on DOL depend on the specific factor being manipulated underscores the importance of considering and controlling the various sources of heterogeneity that naturally act in social groups in order to study the different (and possibly opposing) effects that they have on collective organization. Moreover, our work also underscores the importance of considering factors beyond the usual suspects (e.g., age and morphology): while larval cues (Mas and Kölliker, 2008) are known to affect worker physiology (Maisonnasse et al., 2009; Oldroyd et al., 2001) and behavior (Pankiw et al., 1998; Ulrich et al., 2016), our results highlight larvae as important players in the actual regulation of DOL between workers, something that has rarely been considered. And, on longer timescales, our findings suggest the need to consider a broader array of factors when investigating the evolution of DOL (Duarte et al., 2011).

The integrated empirical and theoretical analysis reveals that models based on threshold variation alone fail to recapitulate the diverse outcomes observed in heterogeneous colonies. However, consistent with recent calls to expand theoretical investigations to other sources of heterogeneity (Jeanson, 2019; Jeanson and Weidenmüller, 2014; Weidenmüller et al., 2019), incorporating differences in

larval-induced demand and in worker task performance efficiency, two parameters that, like response thresholds (Detrain and Pasteels, 1991; Pankiw and Page, 2000; Pankiw and Page Jr., 1999; Robinson, 1992), are known to vary in nature (Dornhaus, 2008; Kaptein et al., 2005; Kay and Rissing, 2005; Mertl and Traniello, 2009; Wilson, 1980), allowed us to recapitulate all empirically observed behavioral patterns. Importantly, the expanded threshold model could recapitulate these patterns using only simple individual behavioral rules and without invoking social interactions. For example, behavioral convergencea phenomenon that intuitively appears to rely on direct social interactionscould emerge without invoking complex social processes, such as social learning (Alem et al., 2016; van de Waal et al., 2013) or direct information transfer between group members (Berdahl et al., 2013; Rosenthal et al., 2015). Although the theoretical treatment can only suggest candidate mechanisms, it is reassuring that the observed behaviors are robust and generic, i.e., the parameter values chosen to illustrate the versatility of the model are representative of large regions of parameter space. Nevertheless, rigorous empirical quantifications of thresholds, efficiency, and demand for realistic task-stimulus pairs—which have only rarely been attempted (Detrain and Pasteels, 1991; Merling et al., 2020; Weidenmüller, 2004) and remain very challenging—are a critical next step toward bridging the gap between theory and empirical observations.

While we focused on the simplest model that could recapitulate our empirical results, we recognize that DOL can be influenced by an even broader set of parameters, whose roles deserve further empirical and theoretical work. For example, experience and social interactions (Fewell and Bertram, 1999; Jeanson et al., 2007; Pacala et al., 1996; Tokita and Tarnita, 2020) might dynamically change individual thresholds (Ravary et al., 2007) and/or task efficiency (O'Donnell and Jeanne, 1992; Tripet and Nonacs, 2004) over time, potentially modulating the effects observed here. It will be important to consider such effects in future theoretical extensions. At the same time, this simple model can

nevertheless be used to make rich testable predictions for colonies with increasingly complex composition. A first attempt using different ratios of ant types led to a striking range of patterns even among the four parameter combinations in Fig. 1.3A–D: the model predicts that behavior can change linearly or nonlinearly as a function of colony composition depending on the between-type differences in mean threshold (Fig. A.7, *Supplementary analyses*). In other words, despite one type of ant being more efficient than the other in all cases considered, replacing an individual of the former with one of the latter led to proportional, greater-than-proportional, or less-than-proportional changes in task performance. Testing these predictions empirically will accelerate the productive crosstalk between theory and experiments.

Our findings add to the growing literature on the role of individual heterogeneity in the collective behavior of complex biological (e.g., schools of fish, neurons in a brain, pathogen strains sharing a host, etc.) and artificial (e.g., robot swarms, synthetic microbial communities, etc.) systems. Much like colonies of the clonal raider ant, these systems exhibit patterns that can be interpreted as behavioral convergence (Berdahl et al., 2013; Broly and Deneubourg, 2015; Centola, 2018; Christakis and Fowler, 2007; van de Waal et al., 2013), divergence (Bettenworth et al., 2019), and nonlinear effects of mixing on group-level phenotypes (Buttery et al., 2010; Kaushik et al., 2006; Pande and Velicer, 2018). In turn, these patterns affect important processes such as collective decision-making (Stewart et al., 2019), the transmission and evolution of disease (Bell et al., 2006; Read and Taylor, 2001), and the evolution of cooperative behavior (Diggle et al., 2007; Strassmann et al., 2000). While different variants of threshold-based models have been employed to study several of these systems (Dodds and Watts, 2005; Hopfield, 1982; Melke et al., 2010; Ward et al., 2008), we still lack a unified theoretical framework to understand the consequences of individual differences on collective dynamics (Jolles et al., 2020). Thus, a comparative approach to the study of the basic

organizing principles of heterogeneous systems across scales constitutes an important next step toward understanding the behavior of complex biological systems.

1.6 Materials and methods

For details of the experiments designed and performed by my collaborators (V. Chandra, D. J. C. Kronauer, J. Saragosti, Y. Ulrich), see *Supplementary methods*.

1.6.1 Definitions of behavioral patterns

We use the following definitions to characterize the qualitative outcomes of mixing individuals with different behavioral tendencies on individual behavior. Let X_k and Y_k denote the mean behavior of ant types X and Y, respectively, in pure ($k = p$) or mixed colonies ($k = m$). We assume that $Y_p > X_p$ and $Y_m > X_m$, to reflect our observation that the type with higher r.m.s.d. in pure colonies always also had higher r.m.s.d. in mixed colonies. Given this assumption, mixing could, in principle, result in one of the following patterns:

1. *No effect* of mixing on individual behavior: The mean behavioral difference between types across pure colonies is the same as the mean behavioral difference between types within mixed colonies, so that $Y_p - X_p = Y_m - X_m$;
2. *Behavioral convergence*: Individuals of different types are behaviorally more similar on average to each other when mixed, so that $Y_p - X_p > Y_m - X_m$; or
3. *Behavioral divergence*: Individuals of different types are behaviorally more different on average from each other when mixed, so that $Y_p - X_p < Y_m - X_m$.

Chapter 2

Interindividual cooperation mediated by partisanship complicates Madison's cure for "mischiefs of faction"

2.1 Notes

This chapter is adapted from:

> Mari Kawakatsu, Yphtach Lelkes, Simon A. Levin, Corina E. Tarnita. Interindividual cooperation mediated by partisanship complicates Madison's cure for "mischiefs of faction." *Proceedings of the National Academy of Sciences*, 118(50):e2102148118 (2021). doi:10.1073/pnas.2102148118

Author contributions. Y. Lelkes, S. A. Levin, C. E. Tarnita, and I designed this study and developed the theoretical framework. I performed computational simulations and analytical calculations with input from C. E. Tarnita. Y. Lelkes, C. E. Tarnita, and I drafted the paper, and all authors provided comments.

Prior presentations. I have given talks on this work at the following conferences, meetings, and seminars:

- Princeton-Humboldt Cooperation and Collective Cognition Network Meeting, Humboldt University of Berlin (July 2019).
- Political Polarization Workshop, Princeton University (August 2019).
- Social Decisions Workshop, University of Houston (October 2019).
- Theoretical Ecology Lab Tea, Princeton University (November 2019).
- Collective Information Processing Workshop, Humboldt University of Berlin (online; March 2020).
- Dialogues in Complexity II: Political Polarization, Arizona State University (online; August 2020).
- PNAS Political Polarization Conference (online; January 2021).
- Society for Industrial and Applied Mathematics Conference on Applications of Dynamical Systems (online; May 2021).
- Networks 2021 (online; July 2021).
- American Mathematical Society Eastern Sectional Meeting (online; March 2022).
- Complex/Dynamical Systems Seminar, Department of Applied Mathematics, University of Colorado Boulder (online; April 2022).

Acknowledgments. We thank Joshua Plotkin, Samuel Wang, and Bernard Grofman for helpful discussions and feedback. M. Kawakatsu acknowledges support from the Army Research Office Grant W911NF-18-1-0325. S. A. Levin thanks The College of Liberal Arts and Sciences at Arizona State University for its support.

2.2 Abstract

Political theorists have long argued that enlarging the political sphere to include a greater diversity of interests would cure the ills of factions in a pluralistic society. While the scope of politics has expanded dramatically over the past 75 years, polarization is markedly worse. Motivated by this paradox, we take a bottom-up approach to explore how partisan individual-level dynamics in a diverse (multidimensional) issue space can shape collective-level factionalization via an emergent dimensionality reduction. We extend a model of cultural evolution grounded in evolutionary game theory, in which individuals accumulate benefits through pairwise interactions and imitate (or learn) the strategies of successful others. The degree of partisanship determines the likelihood of learning from individuals of the opposite party. This approach captures the coupling between individual behavior, partisan-mediated opinion dynamics, and an interaction network that changes endogenously according to the evolving interests of individuals. We find that while expanding the diversity of interests can indeed improve both individual and collective outcomes, increasingly high partisan bias promotes a reduction in issue dimensionality via party-based assortment that leads to increasing polarization. When party bias becomes extreme, it also boosts interindividual cooperation, thereby further entrenching extreme polarization and creating a tug-of-war between individual cooperation and societal cohesion. These dangers of extreme partisanship are highest when individuals' interests and opinions are heavily shaped by peers and there is little independent exploration. Overall, our findings highlight the urgency to study polarization in a coupled, multilevel context.

2.3 Introduction

Two hundred and twenty-six years ago, George Washington, in his farewell address, predicted that factions—or monolithic parties—would yield precisely the political sectarianism that the United States now experiences. As party sectarianism has increased, democratic norms have eroded, and the United States seems to be at a breaking point. However, a decade prior to Washington's speech, James Madison argued that the "mischiefs of faction" could be prevented by expanding the sphere of politics: in a society with diverse interests, no faction could act as a monolith and agendas could be pursued only by negotiating across differences and forming alliances toward shared goals.

The scope of politics has dramatically increased over the past 75 years. Potentially driven by increases in educational attainment, the nationalization of politics, and changes to the information environment (Chaffee and Wilson, 1977; Green and Hobolt, 2008), the number of issues people care about and consider within the realm of national politics has markedly increased (Edy and Meirick, 2018; Jennings et al., 2011; McCombs and Zhu, 1995). Despite this trend, and the consequent expectation that an abundance of issues will improve the collective cohesion by decreasing the likelihood of monoliths, polarization is markedly worse.

A potential explanation for this paradox is the decreasing dimensionality of the issue space. In other words, although the number of issues may have increased, individuals' opinions on these issues might be so strongly correlated with their political ideology that, in effect, there are only one or two issue dimensions (Taagepera and Grofman, 1985; Treier and Hillygus, 2009). While some papers have argued that the decreasing dimensionality of issue attitudes (DellaPosta, 2020; Webster and Abramowitz, 2017) is at the core of current political tensions, any demonstrated relationship between

dimensionality reduction and polarization has been merely correlational. In fact, some have argued that "[a]lthough polarization and the reduction in dimensionality tend to coincide, there is no necessary logical connection between the two trends" (Barber and Lelkes, 2015, p. 42).

Here we propose a bottom-up mechanism that might offer a resolution for the paradox of polarization in the face of rising issue diversity. In particular, we focus on individual-level interactions that are influenced by issue stances, coupled with social learning that is mediated by partisan bias. The issues individuals care about (political or otherwise) and the stances they take on these issues have become both increasingly visible to others (e.g., via social media) and strong determinants of individual behaviors (Settle, 2018): how trustful, forgiving, or helpful we are—even in quotidian, pairwise interactions with neighbors, colleagues, friends, or strangers (Carlin and Love, 2013, 2018; Iyengar and Westwood, 2015; Rand et al., 2009)—can hinge on our respective views on a variety of issues, from preferred sports teams to art tastes (Billig and Tajfel, 1973) to gun control or to favored political candidates [even in a primary election (Rand et al., 2009)]. Simultaneously, the stronger the perceived partisan bias, the less likely it is that individuals leaning toward one end of the political spectrum will embrace issues or opinions held by those at the opposite end (e.g., mask wearing in the COVID-19 pandemic) (Allcott et al., 2020; Guilbeault et al., 2018; Milosh et al., 2020).

We propose that the interplay between individual-level behavior on the one hand and the degree of partisanship on the other hand mediates the effect of issue dimensionality both on individual-level dynamics and on emergent collective-level factioning. To investigate this proposition, we extend an evolutionary game theoretic (Hofbauer et al., 1998; Nowak, 2006) model of cultural evolution (Tarnita et al., 2009a) that allows the coevolution of individual states and social networks (Castellano et al., 2009): individuals imitate others—i.e., adopt their interests, opinions, and

strategies—depending on their relative success in a pairwise donation game (also known as a simplified Prisoner's Dilemma). Our choice of game is motivated by previous behavioral studies that have used similar pairwise games, such as the dictator game or the trust game, to measure cooperation between individuals with different political or other attitudes (Carlin and Love, 2013, 2018; Iyengar and Westwood, 2015; Rand et al., 2009). However, our framework is sufficiently versatile to allow multiplayer interactions, such as public goods games, or even multilevel interactions, in which individuals can not only cooperate with peers but also contribute to their party.

Finally, but importantly, we assume that the imitation process is influenced by political affiliations and partisan bias, a mesolevel societal organization—intermediate between the individual and the collective—that governs the extent of a politically mediated reduction in issue dimensionality (Levendusky, 2010). Because we focus on the United States, where third parties have minimal influence (Goff and Lee, 2019), we model a two-party system (L, R) with individuals distributed equally between the parties. We also ignore unaffiliated independents since a majority of independents admit to leaning Democrat or Republican and act much like their partisan counterparts, at least in their voting behavior (Keith et al., 1992). However, because independents may perceive partisan bias differently in their day-to-day pairwise interactions, future work should extend this model to consider an independent class.

2.4 Model description

2.4.1 Population

Building on Tarnita et al. (2009a), we consider a population of N individuals distributed over M potentially overlapping groups, each representing a political issue of interest (e.g., climate change, gun control; Fig. 2.1A). A priori, we do not assume any relationship

among the issues; i.e., we assume that all M issues are independent, so that M gives the dimensionality (or diversity) of the issue space. Individuals can care about (or have an interest in) any non-zero number of issues. Individual i cares about issue k if she takes either a liberal ($h_{ik} = -1$) or a conservative ($h_{ik} = +1$) position on it; future extensions could explore different strengths of interest by allowing values along a scale (e.g., 1–7). We say that she does not care about issue k if i takes a neutral position ($h_{ik} = 0$) on it. We define the opinion vector of i as $\mathbf{h}_i = [h_{i1}, h_{i2}, \cdots, h_{iM}] \in \{-1, 0, 1\}^M$ and the corresponding issue interest vector as $\overline{\mathbf{h}}_i = [|h_{i1}|, \cdots, |h_{iM}|] \in \{0, 1\}^M$, where $|h_{i1}| = 1$ if i cares about issue k and 0 otherwise. For simplicity, we assume that every individual cares about exactly $K \leq M$ issues, but the set of K issues can differ among individuals.

Individuals also have political affiliations, but their opinions on issues are not necessarily perfectly correlated with their political label. In other words, someone who identifies as a member of a left-leaning party can hold right-leaning opinions and vice versa (e.g., an American might identify as a Democrat based on stances on economic and racial issues but still oppose the party on some social issues). The strength of correlation between one's party label and opinions is subject to model dynamics as described below in *Imitation dynamics*.

2.4.2 Pairwise interactions

In our model, an interaction takes the form of a one-shot pairwise donation game. In a game, the donor must choose whether to cooperate with the recipient. A cooperating donor ('cooperator', C) incurs a cost c to provide a benefit b to the recipient; a defecting donor ('defector', D) incurs no cost and provides no benefit to the recipient.

Interactions are entirely determined by issues and are not influenced by party affiliation. Specifically, individuals i and j (independent of their party labels) interact if and only if there is at least one issue that they are both interested in, and they interact as

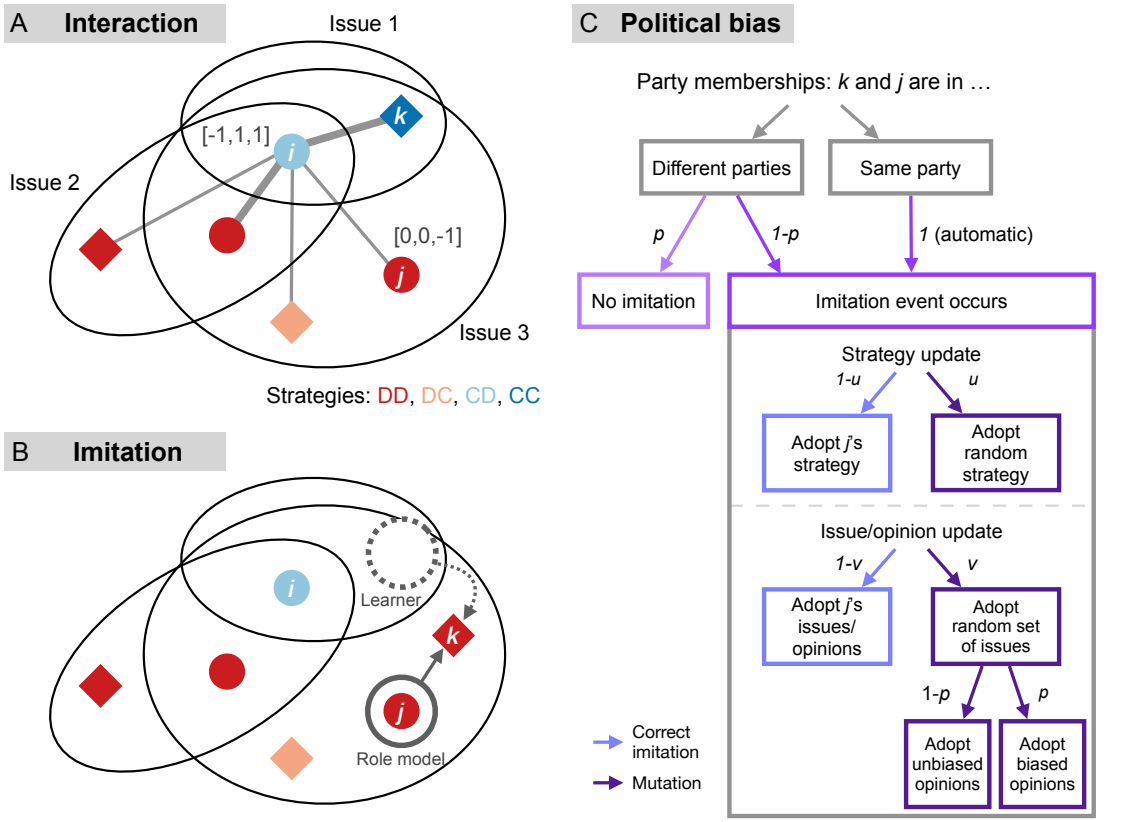

Figure 2.1: Schematic illustration of the model. (A) $N = 6$ individuals (nodes) are distributed over $M = 3$ groups (black ovals), each representing a political issue. Colors represent behavioral strategies as indicated; shapes represent party affiliations (circle $= L$, diamond $= R$). Individual i cares about all three issues and has opinion vector $\mathbf{h}_i = [-1, 1, 1]$, where -1 and $+1$ correspond to liberal and conservative positions, respectively. Individual j cares only about issue 3 and takes a liberal position; hence, $\mathbf{h}_j = [0, 0, -1]$. In each round, every pair plays one-shot donation games as many times as they have issues in common. Edges represent the interactions of focal individual i; widths are proportional to the number of interactions. Opinions determine interaction patterns: whether i cooperates with j in a given group depends on whether they agree on that issue. (B) Once all games in a round are played, a learner k is chosen uniformly at random to imitate a role model j, chosen proportional to fitness. Strategies and issues/opinions can both be imitated but not party affiliation, which is fixed. (C) Whether learner k imitates role model j depends on partisan bias p and on the pair's party affiliations: an imitation event occurs with probability $1 - p$ if k and j belong to different parties and with probability 1 otherwise. In an imitation event, k adopts j's strategy with probability $1 - u$ or a random strategy with probability u; independently, k abandons her issues and adopts j's issues and opinions with probability $1 - v$; with probability v, k picks a random set of issues and adopts opinions on those issues that are biased toward k's party with probability p.

many times as they have shared interests (Fig.2.1A). This dynamic reflects, for instance, social media interactions, where an individual will only respond to someone else if they are talking about the same issue, and will do so regardless of whether they have the same opinion on that issue. How they choose to respond will, however, be determined by their opinions. In our model, an individual can employ one of four strategies depending on her own opinion and that of the donor: unconditional defector (DD), unconditional cooperator (CC), homophilous cooperator (CD; cooperates with those who share the same opinion but defects against those who have the opposite opinion), or heterophilous cooperator (DC; defects against those who share the same opinion but cooperates with those who have the opposite opinion).

2.4.3 Fitness

After all pairwise games for a given round have been played, the fitness f_i of individual i is computed as $f_i = 1 + \beta \cdot \pi_i$, where π_i denotes the total payoff accumulated by individual i and β denotes the intensity of selection, a quantity employed in evolutionary game theory to capture the impact of the dynamics under study on relative fitness. Most often, and in our case, the assumption is that selection is weak (i.e., $\beta \ll 1$), to reflect the fact that most peer interactions represent only a tiny fraction of an individual's overall fitness. This limit also facilitates analytical insights.

2.4.4 Imitation dynamics

The population updates dynamically according to a frequency-dependent Moran process (Nowak et al., 2004; Taylor et al., 2004; Traulsen et al., 2006), a standard approach in models of cultural evolution (Fig.2.1B). This framework describes a social learning process in which individuals preferentially copy the traits of successful others. In our model, both the strategy and the issues and associated opinions are subject to this updating process. However, we assume that individual party affiliations are fixed over

time because empirical evidence suggests that Americans rarely change their party affiliations (Green et al., 2004)—although future work can relax this assumption to explore the dynamics of party affiliations, possibly on longer timescales. This imitation process plays out at the individual level (i.e., individuals imitate peers). However, it mirrors the influence of political leaders and campaigns on public discourse (Wang et al., 2020), as exemplified by the empirically documented follow-the-leader phenomenon (Lenz, 2013); i.e., voters tend to first pick a political leader they deem successful and then adopt their policies, rather than choosing a leader whose policies match the voters' own preferences.

Once fitness is computed for all individuals, a learner k is chosen uniformly at random from the population. The learner then selects a role model j randomly with probability proportional to fitness (Fig. 2.1B). Importantly, the learner and the role model do not have to share any issues in common prior to the imitation event; i.e., the imitation network is the complete graph and there is a breaking in symmetry (Ohtsuki et al., 2007) between the interaction network (which is local) and the imitation network (which is global). Whether the learner proceeds to imitate the role model or not depends on their party affiliations (Fig. 2.1C), so that an imitation event is initiated with probability 1 if k and j belong to the same party, but only with probability $1 - p$ otherwise. When $p = 1$ the imitation graph completely segregates into two modules according to party affiliations. The exogenous parameter $0 \leq p \leq 1$—which, for simplicity, we assume to be the same for both parties—thus captures partisan bias: a larger p means that individuals are less willing to imitate across party lines, consistent with cognitive dissonance theory and partisan mediated reasoning (Bakker et al., 2020); if $p = 1$, individuals only imitate those in their own party.

An imitation event also allows for the possibility of errors (e.g., incorrectly assessing someone's strategy or opinions) and for non-social learning or exploration (e.g., learning

about new issues from sources other than peers) (see Fig.2.1C). Let $0 \leq u \leq 1$ and $0 \leq v \leq 1$ be the strategy mutation rate and the issue and opinion exploration rate, respectively. Learner k adopts either role model j's strategy with probability $1 - u$ or a random strategy with probability u. Similarly, with probability $1 - v$, k adopts j's opinion vector $\mathbf{h}_j$; with probability v, however, k explores a new and random set of issues and opinions, $\mathbf{h}_k$. The lower the exploration rate, the more reliant individuals are on their peers as sources of information. When an individual explores a completely new and random set of issues, party affiliation can still play a role in determining what opinions that individual will take on the newly adopted issues. With probability $1 - p$, learner k adopts a random set of opinions. With probability p, however, learner k adopts a biased set of opinions aligned with her party membership.

2.5 Individual- and collective-level metrics

We define the following metrics to characterize the three phenomena of interest: cooperation, opinion alignment, and interest alignment. See *Materials and methods* for full mathematical definitions.

2.5.1 Cooperation

To quantify the amount of interindividual cooperation in the population, we define the effective cooperation to be the population-level mean fraction of cooperative interactions averaged over the stationary distribution of the dynamical process. To characterize individual behaviors in more detail, we also measure the steady-state strategy distribution, i.e., the frequency (or relative abundance) of each of the four possible behavioral strategies averaged over the stationary distribution.

2.5.2 Opinion alignment

We use as a measure of factionalization the ability of a party to act as a monolith on issues of interest, i.e., the extent to which within-party opinions are aligned. Though this metric has been primarily used to describe party unity, some have argued that it can be used to characterize the degree of societal polarization (Baldassarri and Gelman, 2008).

To quantify opinion alignment, we define the average opinion distance in a given subpopulation as the average city block distance—also known as Manhattan distance or ℓ_1 norm—between pairs of opinion vectors. The opinion distance between individuals i and j thus represents the total magnitude of their opinion differences across all issues and is computed as $\sum_{k=1}^{M} |h_{ik} - h_{jk}|$. We define average opinion distance for three subpopulations: among members of the same party (within-party), among members of different parties (between-party), and among all individuals (population-level). A lower average opinion distance in a subpopulation indicates greater opinion alignment within that subpopulation.

2.5.3 Interest alignment

Interest alignment refers to the degree to which individuals share overlapping interests and therefore interact with one another. To quantify it, we define the average interest distance within a given subpopulation as the average pairwise Hamming distance between issue interest vectors. The interest distance between i and j thus measures the number of issues they do not have in common (i.e., issues that either i or j cares about but not both) and is computed as $\sum_{k=1}^{M} \big|(|h_{ik}| - |h_{jk}|)\big|$. We define average interest distance within parties, between parties, and within the whole population. A lower average interest distance within a subpopulation indicates greater interest alignment within that subpopulation.

To illustrate how opinion distance and interest distance work in tandem, consider a population in which each individual cares about three out of five available issues (i.e., $M = 5$, $K = 3$). Suppose individuals i, j, and k have opinion vectors $[0, 0, 1, 1, 1]$, $[1, 1, 1, 0, 0]$, and $[1, 1, -1, 0, 0]$, respectively. Even though both pairs ij and ik have the largest possible divergence in issues for the given M and K (pairwise interest distance 4, since they only share one issue of interest), the opinion distance for pair ik (given by $1 + 1 + 2 + 1 + 1 = 6$) is greater than that for pair ij (given by $1 + 1 + 0 + 1 + 1 = 4$) because i and j have the same opinion on the one issue that they do have in common. Thus, the two quantities together capture not only the overlap in issues but also the divergence in opinion on those overlapping issues.

2.6 Results and discussion

We conducted computational simulations (see *Materials and methods*) and, where possible, analytical calculations (see *Supplementary analyses*).

2.6.1 Regardless of political bias, increasing the number of available issues (M) but decreasing the number of issues that each individual cares about (K) promotes inter-individual cooperation and reduces polarization

At the individual level, pairwise cooperation tended to increase when there were more available issues (higher M; Fig. 2.2A) or when individuals cared about fewer issues (lower K; Fig. 2.2B). The latter had a stronger effect on cooperation, particularly when individuals were not exploratory ($v = 0.001$), but the effect of the former became clear as v increased to an intermediate level (see effect sizes in Tables B.2 and B.3). Analytical calculations provide insight into the relative effects of M and K (*Supplementary analyses*,

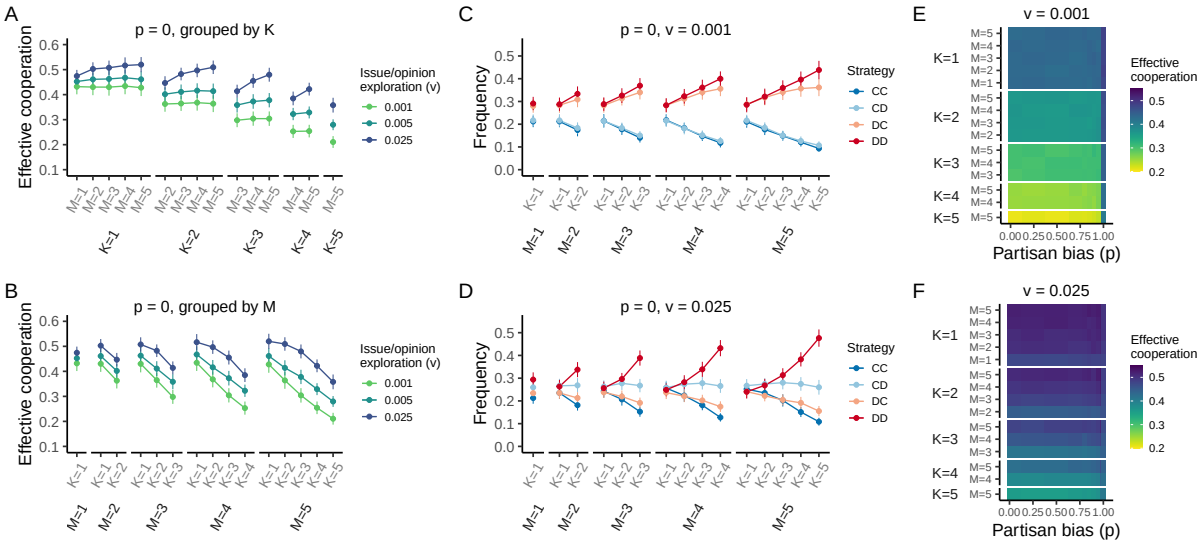

Figure 2.2: Cooperation increases with increasing number of available issues (M) and decreasing number of issues individuals care about (K). For each parameter setting, we ran an ensemble of 150 simulations with population size $N = 40$, each lasting 2×10^7 generations, the first 10% of which were disregarded to account for potential initialization effects. (A and B) Effective cooperation as a function of M and K in the absence of partisan bias ($p = 0$), grouped by K (A) or by M (B). Within a simulation, effective cooperation was measured as the fraction of cooperative actions among all interactions in a generation, averaged across generations. Each circle represents the mean effective cooperation ($\pm$SD) averaged across the ensemble. Colors indicate issue/opinion exploration rates (v). (C and D) Steady-state strategy distributions as a function of M and K in the absence of partisan bias ($p = 0$). Each circle represents the average frequency ($\pm$SD) of the corresponding strategy (indicated by its color) across generations, averaged across the ensemble; error bars indicate s.e. within the ensemble. Parameter v is as indicated. (E and F) Effective cooperation as a function partisan bias across combinations of M and K. Color indicates degree of effective cooperation, from low (yellow) to high (purple). Parameter v is as indicated. See Table B.1 for other parameter values and Tables B.2 and B.3 for the effect sizes corresponding to A–D.

Eqs. (B.21) and (B.22)). Whereas K affects the frequencies of both unconditional (CC) and conditional (CD, DC) cooperators, M only affects the former and only positively (*Supplementary analyses*, Eq. (B.22)). Consequently, effective cooperation always increases with the number of available issues, consistent with the simulations. However, M impacts the frequency of CC via a term proportional to $1/M$, and therefore the positive effect of increasing M is vanishingly small. In contrast, K impacts all frequencies at least

linearly, and therefore the effects of varying K are much stronger than those of varying M.

Consistent with previous work (Tarnita et al., 2009a), these findings capture the essence of why structured populations promote cooperation: the greater the possibility for assortment with like-minded individuals, the higher the chance for cooperation to thrive (Antal et al., 2009a; Cavaliere et al., 2012; Nowak et al., 2010; Tarnita et al., 2009a,b). Having more available issues but few of those issues claimed by any one individual increases the possibility for cooperators to find refugia from free-riders (i.e., unclaimed issues that cooperators can make their own and thrive). This increased assortment leads to a lower frequency of unconditional defection (DD) relative to unconditional cooperation (CC) (Fig. 2.2C and D).

At the collective level, within-party average opinion distance increased (and the potential for a party to act as a monolith decreased) with increasing M and decreasing K (see Fig. 2.4A). When there are more issues to explore, individuals have the possibility to adopt a wider variety of opinions and therefore are less confined to a small cluster of opinions. This reduces the chances of high within-party opinion alignment and the potential for polarization.

2.6.2 A moderate rate of issue/opinion exploration optimally promotes cooperation while also reducing polarization relative to a rigid population

At the individual level, relative to our low-exploration baseline ($v = 0.001$), effective cooperation tended to be higher as the exploration rate increased up to a moderate rate of issue/opinion exploration ($v = 0.025$), after which it began to decrease (Fig. 2.2A and B, Fig. 2.3A). These results are consistent with previous work showing that an intermediate level of stochasticity in the imitation of the population structure optimally

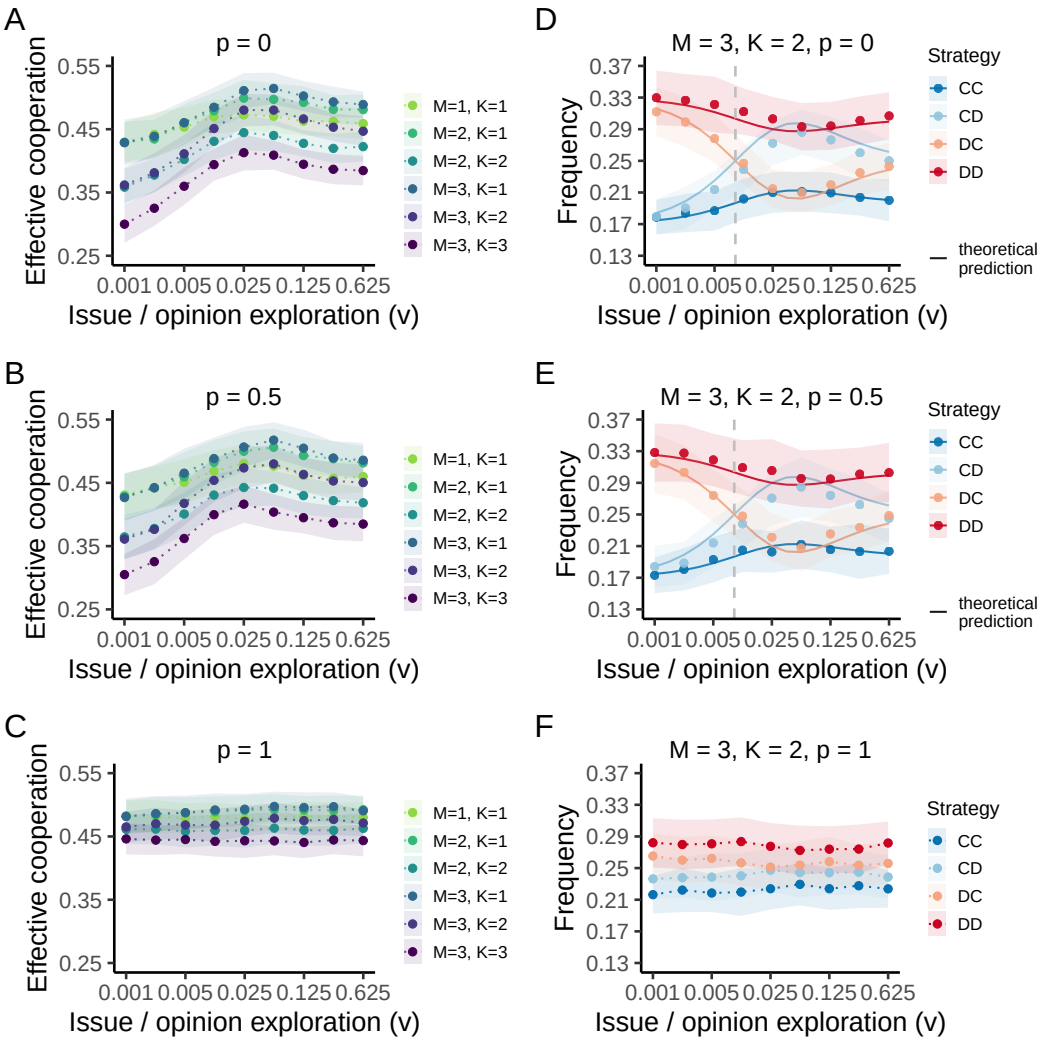

Figure 2.3: Moderate rates of issue/opinion exploration promote cooperation. For each parameter setting, we ran an ensemble of 150 simulations with population size $N = 40$, each lasting 2×10^7 generations, the first 10% of which were disregarded to account for potential initialization effects. (A–C) Mean effective cooperation ($\pm$SD) across the ensemble as a function of issue/opinion exploration rate v log scale. Colors indicate combinations of M and K. (D–F) Steady-state strategy distributions for $M = 3$, $K = 2$ as a function of issue/opinion exploration rate v log scale. Each circle represents the average frequency ($\pm$SD) of the corresponding strategy (indicated by color), averaged across the ensemble. Solid curves in D and E show the corresponding theoretical predictions in the limit of small $\mu = Nu$ (*Supplementary analyses*, Eqs. (B.19) and (B.20)) and dashed gray lines show the critical exploration rate v^* computed from Eq. (2.1), both showing excellent agreement with the simulation results. Partisan bias p is as indicated in each panel. See Table B.1 for other parameter values and Fig. B.3 for an expanded figure with $p = 0.25, 0.75$ and $M = 1, K = 1$.

promotes cooperation (Nowak et al., 2010; Tarnita et al., 2009a). That such an intermediate optimum also arises in our system is to be expected, since too little exploration limits the cooperators' ability to take advantage of 'empty' issues while too much exploration scrambles the population structure and renders it virtually well-mixed.

To understand how changes in individual behavior drive the rising effective cooperation, we investigated the effect of issue/opinion exploration rate v on the steady-state strategy distribution (Fig. 2.2C and D, Fig. 2.3D). Notably, while heterophilous cooperators (DC) were more frequent than homophilous cooperators (CD) at low exploration rates (Fig. 2.2C), this ordering was eventually reversed at intermediate exploration rates (Fig. 2.2D, Fig. 2.3D). Analytical calculations confirm that these simulation results hold for any benefit-to-cost ratio (b/c), as long as $b > c > 0$ (*Supplementary analyses*; fitted to simulation data in Fig. 2.3D). Selection favors CD (and simultaneously disfavors DC) when the effective population-level exploration rate ($\nu = Nv$) satisfies

$$\nu > \nu^* = \frac{-2(b/c) + 3 + \sqrt{4(b/c)^2 - 3}}{2(b/c - 1)}, \tag{2.1}$$

where ν^* is the critical threshold, which is independent of M and K. Importantly, although selection always favors CD when Eq. (2.1) holds, the frequency of CD has a maximum as a function of ν (Fig. B.1), which likely contributes to the existence of an optimum exploration rate for the effective cooperation (Fig. 2.3).

Intuitively, the order reversal in the frequencies of CD and DC occurs because, as the exploration rate increases, so does the possibility of assortment with 'like' individuals. This favors those who cooperate with others who are the same (share the same opinion) and penalizes those who cooperate with others who are different: both CD and DC are more likely to encounter their own type (same strategy and, importantly, same opinions),

but two *CD*'s with the same opinions will mutually cooperate and gain benefits, whereas two *DC*'s with the same opinions will mutually defect and forgo benefits.

At the collective level, exploration introduces new issues and opinions into a subpopulation, thus continually increasing opinion diversity. This, in turn, helps shuffle the opinion clusters, thereby mitigating polarization. Unlike at the individual level, where eventually too much scrambling of opinions and issues diminishes the possibility of assortment and reduces cooperation, polarization at the collective level will continue to decrease with increasing shuffling of sets and opinions. We therefore did not expect an intermediate optimum level of exploration, past which polarization would begin to increase again. Accordingly, we found that the average opinion distance increased with the issue/opinion exploration rate v, regardless of the subpopulation (Fig. 2.4, Fig. B.2A).

2.6.3 Strong partisan bias promotes within-party opinion and interest alignment at the cost of global alignment

At the collective level, partisan bias tended to promote stronger opinion alignment by party regardless of M and K (Fig. 2.4A, Fig. B.2A), though these trends were more striking with increasing exploration rate (Fig. 2.4A). Extreme partisan bias ($p = 1$) corresponded to maximum alignment among members of the same party (minimum within-party average opinion distance) and minimum alignment among those of different parties (maximum between-party average opinion distance). Within-party alignment decreased nonlinearly and between-party alignment increased nonlinearly as partisan bias declined; at reasonable values of issue/opinion exploration ($0.001 \leq v \leq 0.125$), the most pronounced change occurred between $p = 1$ and $p = 0.75$. This pattern suggests that, while extreme partisan bias ($p = 1$) leads to strong assortment in the opinion space and therefore polarization, a fairly small amount of cross-party imitation can mitigate this adverse effect.

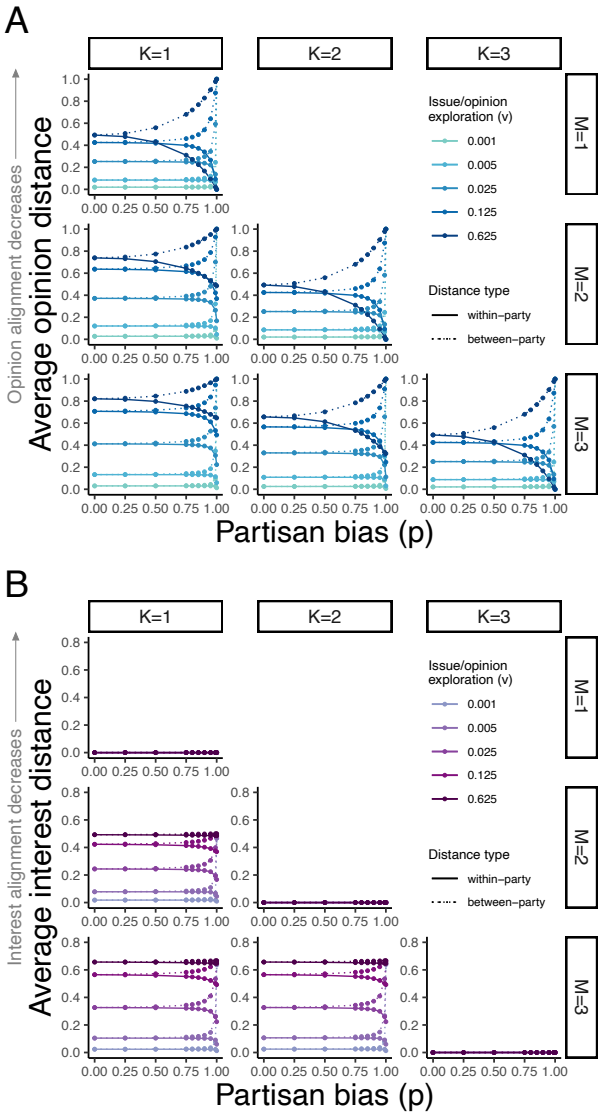

Figure 2.4: Opinion and interest alignment as a function of partisan bias. For each parameter setting, we ran an ensemble of 150 simulations with population size $N = 40$, each lasting 2×10^7 generations, the first 10% of which were disregarded to account for potential initialization effects. Each circle within a panel represents the mean value ($\pm$SD) of the corresponding metric averaged across generations and across the ensemble. Dotted and dashed lines indicate values within and between parties, respectively. Values of M and K are as indicated; the values of p between 0.75 and 1.00 are p = 0.8, 0.85, 0.9, 0.95, 0.99. (A) Opinion alignment as measured by normalized average opinion distance. Opinion alignment decreases with increasing average opinion distance. (B) Interest alignment as measured by normalized average interest distance. Interest alignment decreases with increasing average interest distance. See *Materials and methods* for definitions and the normalization procedure; Fig. B.5 for population-level values; and Table B.1 for other parameter values.

As expected from its definition (see *Individual- and collective-level metrics*), the average interest distance was zero when individuals cared about all available issues ($K = M$), irrespective of subpopulation, partisan bias, or opinion mutation rate (Fig. 2.4B). When $K < M$, partisan bias increased interest alignment within, but not between, parties: average interest distance rose within parties but declined between parties. As in the opinion case, the within-party and between-party curves quickly converged as p decreased. This outcome showed that, while the population strongly tended to fragment into party-based clusters when partisan bias was extreme ($p = 1$), a fairly small likelihood of cross-party imitation made the population more cohesive. Unlike in the opinion case, however, partisan bias had no effect on interest distance at high exploration rates (Fig. 2.4B).

2.6.4 Extreme partisan bias promotes pairwise cooperation while maximizing polarization, but only if individuals are not sufficiently exploratory

At the individual level, there was a marked difference in steady-state behavior between the cases $p < 1$ and $p = 1$ (extreme bias): while $p < 1$ behaved the same as $p = 0$ (compare Fig. 2.3A vs. B and D vs. E) and this was confirmed by our analytical calculations (see *Supplementary analyses*, Eqs. (B.19) and (B.20)), $p = 1$ had a markedly different dynamics (Fig. 2.3C and F). This is because p qualitatively modifies the imitation network: when $p = 0$, the imitation network is the complete graph (anyone can imitate anyone else in the population); as p increases, the imitation structure becomes modular (according to party label) with an increasingly weak connection between the two modules. Ultimately, when partisan bias is extreme ($p = 1$), the two modules become disconnected, as individuals can only imitate those of their own party. Importantly, when $p = 1$, even if the exploration rate v is nontrivial, individuals'

opinions on new issues are perfectly aligned with their party (Fig. 2.1C). This gives rise to a discontinuity in the system behavior: for $p = 1$, even though individuals continue to interact according to issue membership, the world becomes segregated according to party labels when it comes to learning and exploration (i.e., at $p = 1$, there is zero probability to be influenced by an individual outside of one's party or to adopt an opinion misaligned with one's party).

Consequently, this party-based segregation and alignment further boosted spatial assortment by both strategy and opinion, maximizing effective cooperation (Fig. 2.2E). Unlike when $p < 1$, cooperation was not primarily boosted by a strong positive effect on CD, but rather by a positive effect on CC and a negative effect on DD (Fig. 2.3F vs. Fig. 2.3D, Fig. B.4C and D vs. Fig. 2.2C and D): when assortment by opinion is very high, the second letter of each strategy matters less because individuals will mostly encounter others of the same opinion. However, this trend largely disappeared around the optimum issue/opinion exploration rates (Fig. 2.2F and compare Fig. 2.3A–C), where the moderate exploration sufficiently boosts cooperation at low p as to match the positive effect of extreme bias. Moreover, when $p = 1$, the effect of extreme bias overshadowed the effect of exploration: effective cooperation changed minimally with increasing v (Fig. 2.2C; Fig. B.4A and B; and see also Fig. B.4C and D for corresponding steady-state strategy distributions).

These results—together with the fact that, at the collective level, extreme partisan bias maximized both opinion and interest alignment among members of the same party and minimized alignment among those of different parties—suggest a potential tension between the individual and collective levels when the population is away from the optimum issue/opinion exploration rate. This tension disappears when the issue/opinion exploration rate is around the optimum: then, extreme bias still increases polarization but without increasing effective cooperation.

2.7 Conclusion

Our results demonstrate that partisan bias interacts in unexpected ways with the diversity of issues that people care about. If partisan bias is not too high, increasing issue diversity both increases interindividual cooperation and prevents a monolithic majority. Interestingly, decreasing the number of issues that any given individual cares about has an even stronger positive effect than increasing the total number of available issues, suggesting that the extent to which individuals engage with available issues can dramatically impact cooperation and cohesion even when the scope of issues considered in the society stays the same. Thus, when partisanship is not too high, our results support Madison's argument that a diverse set of issues can prevent a monolithic majority, but they further suggest that, counter to some contemporary democratic theories (Sunstein, 2001), the splintering of attention driven by information abundance could, in fact, further improve outcomes.

However, increasingly high partisan bias induces party-based assortment of issues and opinions, thereby reducing issue diversity and making the collective worse off. When bias is extreme ($p = 1$), individuals become completely closed off to influence from ideologically divergent peers, and the emergent tribalism boosts interindividual cooperation at the cost of a weakened, polarized collective. This suggests that, in a highly polarized state, there will be an emergent tension between the individual and the collective levels, with little incentive for individuals to reduce the collective polarization. This emergent tension could hinder bottom-up efforts to reduce polarization endogenously until, eventually, the cost of living in a polarized, dysfunctional society outweighs the high individual benefits of tribalism (Finkel et al., 2020; Fu et al., 2012). But our results offer a silver lining: not only do the boost to cooperation and associated appeal of tribalism occur only when partisanship is extreme, they are also substantial

only in a society whose members are primarily learning from peers and are limited in their independent exploration.

Although, a priori, issues in our model are completely independent of each other (i.e., uncorrelated), high partisanship leads to emergent alignment of issues according to party labels and, thus, to emergent correlations among them (e.g., if i and j are both left leaning and i cares about issue X, there is a very high likelihood that j does too). The associated dimensionality reduction is, as hypothesized, a driver of the observed factioning. However, it is not the sole driver. Even when individuals care about all available issues and therefore cannot sort themselves across issues by party affiliation, we still observe between-party divergence in opinions when partisanship is high (i.e., if i holds a right-leaning position on issue X, then i is likely to also be right leaning on issue Y). This latter scenario seems to capture the current state of US politics: Democrats and Republicans care about the same set of hot-button issues, such as gun control and immigration, but they hold opposing views (McCarty et al., 2016). To understand how the waxing and waning of a society's interest in politics affect both individual and collective-level dynamics, future work could allow individuals to dynamically change the number of issues they care about and/or their party memberships. Such extensions would further our understanding of how independents or the politically indifferent might impact polarization (Jones et al., 2022).

Given well-known differences in openness to experience between the right and the left (Malka et al., 2014) and documented patterns of asymmetric polarization in the United States, wherein the right is more polarized than the left (Leonard et al., 2021; Mann and Ornstein, 2016; McCarty, 2019; Pierson and Schickler, 2020), future work needs to explore individual-level or party-level differences in partisan bias, openness to experience, and other attributes that might affect issue exploration and social learning. A simple extension would have p be party dependent (i.e., individuals of different parties

could perceive different levels of partisan bias), with independents experiencing an altogether different level. To study the endogenous evolution of partisanship, individuals could exhibit different levels of partisanship independent of their party labels and partisanship could be subject to learning and imitation, just as the issues and opinions are. If individual fitnesses then depend both on pairwise interaction payoffs and on the collective-level factioning, this approach could allow the study of endogenous waxing and waning of partisanship and polarization.

While our model focuses on two major parties because third parties have minimal influence in the United States–they typically get $<5\%$ of the popular vote in a presidential election (Goff and Lee, 2019)–it can be extended to include three or more parties. This would allow one to explore dynamics of coalition formation, including the possibility of logrolling ("I will cooperate with you on issue 1 if you cooperate with me on issue 2"), which would constitute a key step toward understanding polarization in multiparty parliamentary systems. Here, an important question would be whether and how the introduction of a third party could destabilize the system. To answer this question, one could consider parties with dynamic memberships, that is, where individuals can migrate from a party to another via social learning or based on shifts in party platforms, among other factors.

Despite the simplicity of our model, the results comport with recent evidence of polarization, factionalization, and party bias. Using data from 1972 to 2004, Baldassarri and Gelman (Baldassarri and Gelman, 2008) do not find increases in within-group issue alignment. However, since then, factionalism has markedly increased (Kozlowski and Murphy, 2021), as has partisan bias (Donovan et al., 2020; Iyengar et al., 2019) and subsequent polarization (McCarty, 2019). At the same time, the first decades of the 21st century were also accompanied by exponential increases in information production and consumption, driven by digital technology (Dhamdhere and Dovrolis, 2011). Our study

uncovers how these trends may not be in opposition and prompts us to reevaluate the effectiveness of Madison's suggested cure for the mischiefs of faction. Issue diversity, in the absence of strong partisan bias, promotes individual and collective welfare. Issue diversity, in the presence of extreme partisan bias and a rigid society, promotes individual cooperation while intensifying polarization.

2.8 Materials and methods

2.8.1 Full model description

Opinions and political affiliations. We consider N individuals distributed over M potentially overlapping groups, each representing a political issue. As described in *Model description*, let $\mathbf{h}_i = [h_{i1}, \cdots, h_{iM}] \in \{-1, 0, 1\}^M$ denote the M-element opinion vector of individual i, where h_{ik} represents i's opinion on issue k: liberal (-1), neutral (0), or conservative ($+1$). Individual i cares about issue k if she takes either a liberal or conservative position on it. Thus, the issue interest vector of i is given by $\overline{\mathbf{h}}_i = [|h_{i1}|, \cdots, |h_{iM}|] \in \{0, 1\}^M$, where $|h_{i1}| = 1$ if i cares about issue k and 0 otherwise.

Individuals also have political affiliations. Let $a_i \in \{1, \cdots, P\}^N$ denote the party affiliation of i. For simplicity, we focus on a two-party system ($P = 2$) with individuals distributed equally across the parties and, without loss of generality, assume that party 1 is liberal-leaning (L) and that party 2 is conservative-leaning (R).

Interactions and payoffs. Opinions determine the patterns of interaction. In each round, two individuals play one-shot pairwise donation games as many times as the number of shared interests. In a given game, the donor can choose to either cooperate (C)—incur a cost c to provide a benefit b to the recipient—or defect (D)—incur no cost and provide no benefit to the recipient.

Whether a donor cooperates or defects depends both on her behavioral strategy and on the agreement between her and the recipient. The strategy of individual i is given by $\mathbf{s}_i = [s_{ia}, s_{id}] \in \{0,1\}^2$, where 0 corresponds to defection and 1 to cooperation. When i interacts with j in group k, i plays strategy s_{ia} if i and j agree on issue k (e.g., both have opinion -1) and plays s_{id} if they disagree. Thus, an individual can be an unconditional defector ($DD = [0,0]$), a homophilous cooperator ($CD = [1,0]$), a heterophilous cooperator ($DC = [0,1]$), or an unconditional cooperator ($CC = [1,1]$). Note that i does not interact with j in group k if j does not care (i.e., is neutral) about issue k.

In sum, an individual i is characterized by three variables: (i) party affiliation ($a_i \in \{1\ (=L), 2\ (=R)\}$ under our simplifying assumptions), (ii) opinions $\mathbf{h}_i$, and (iii) behavioral strategy $\mathbf{s}_i$. These, together with benefit b and cost c, determine the total payoff π_i of i in a given round:

$$\pi_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{k=1}^{M} |h_{ik} h_{jk}| \left[\delta_{ij}^k \left[-cs_{ia} + bs_{ja} \right] + (1 - \delta_{ij}^k) \left[-cs_{id} + bs_{jd} \right] \right], \tag{2.2}$$

with $\delta_{ij}^k = \mathbb{1}_{h_{ik}=h_{jk}} = 1$ if i and j agree on issue k and 0 otherwise; $\left(-cs_{i*} + bs_{j*}\right)$ is i's payoff when i and j agree ($* = a$) or disagree ($* = d$).

Fitness and its nonnegativity. After all the games for a given round are played, we compute the fitness f_i of i as $f_i = 1 + \beta \cdot \pi_i$, where β denotes the intensity of selection. To guarantee nonnegativity of f_i, we consider the scenario that results in minimum possible fitness and derive the parameter conditions under which $f_i \geq 0$. When i interacts with every other individual in every issue group and loses c in every interaction, i's fitness is: $f_i = 1 + \beta\pi_i \geq 1 - \beta(N-1)Mc \geq 0$. Thus, β and c must satisfy $1 \geq \beta(N-1)Mc$, that is, $\beta \leq \beta^* = 1/(N-1)Mc$. We chose simulation parameters ($c = 0.2$, $N = 40$, $1 \leq M \leq 5$, and $\beta = 0.001$; Table B.1) satisfying this condition.

2.8.2 Simulation details

We implemented our model as stochastic agent-based simulations in Julia (Bezanson et al., 2017). For the simulations, we assumed that every individual cares about exactly K issues. Under these assumptions, the population was initialized as follows: without loss of generality, individuals 1 through $N/2$ were assigned to party L and $N/2+1$ through N were assigned to party R. To initialize an individual's opinions, we first selected K out of the M issues at random. We assigned her an opinion corresponding to her party affiliation (-1 for party L, $+1$ for party R) for each of these K issues and a neutral opinion 0 for the remaining $M-K$ issues. This process was repeated independently for all N individuals. Finally, each individual was also assigned a strategy (DD, DC, CD, CC) at random.

2.8.3 Measuring opinion alignment

Polarization was characterized using average opinion distance. The *opinion distance* between individuals i and j is defined as the city block distance between their opinion vectors $\mathbf{h}_i$ and $\mathbf{h}_j$: $d_{\text{opinion}}(\mathbf{h}_i, \mathbf{h}_j) = \sum_{k=1}^{M} |h_{ik} - h_{jk}|$. Then, population-level, within-party, and between-party average opinion distances are defined as

$$d_{\text{opinion}}^{\text{population}} = \frac{1}{\binom{N}{2}} \sum_{i<j}^{N} d_{\text{opinion}}(\mathbf{h}_i, \mathbf{h}_j) \,, \tag{2.3}$$

$$d_{\text{opinion}}^{\text{within}} = \frac{1}{2} \sum_{a \in \{L,R\}} \frac{1}{\binom{N/2}{2}} \sum_{i<j, a_i = a_j = a}^{N} d_{\text{opinion}}(\mathbf{h}_i, \mathbf{h}_j) \,, \tag{2.4}$$

$$d_{\text{opinion}}^{\text{between}} = \frac{1}{(N/2)^2} \sum_{i<j, a_i \neq a_j}^{N} d_{\text{opinion}}(\mathbf{h}_i, \mathbf{h}_j) \,, \tag{2.5}$$

respectively. Equation (2.3) computes the average opinion distance between every pair of individuals i and j in the population. In Eq. (2.4), the bracketed term sums the opinion distance between every pair i and j within party a ($a_i = a_j = a$). In Eq. (2.5), the

bracketed term sums the opinion distance between every pair i and j whose party affiliations differ ($a_i \neq a_j$).

Normalization. The range of possible pairwise opinion distances depends on K: given K, the maximum possible opinion distance is $d^{\max}_{\text{opinion}}(K) = 2K$. To allow for meaningful comparisons of opinion alignment across values of K, we divided each raw average opinion distance by $d^{\max}_{\text{opinion}}(K)$.

2.8.4 Measuring interest alignment

Interest alignment was characterized using average interest distance . The interest distance between i and j is defined as the Hamming distance between their issue interest vectors $\bar{\mathbf{h}}_i$ and $\bar{\mathbf{h}}_j$: $d_{\text{interest}}(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j) = \sum_{k=1}^{M} \big||h_{ik}| - |h_{jk}|\big|$. Then, population-level, within-party, and between-party average interest distances are defined as

$$d^{\text{population}}_{\text{interest}} = \frac{1}{\binom{N}{2}} \sum_{i<j}^{N} d_{\text{interest}}(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j) \ , \tag{2.6}$$

$$d^{\text{within}}_{\text{interest}} = \frac{1}{2} \sum_{a \in \{L,R\}} \frac{1}{\binom{N/2}{2}} \sum_{i<j, a_i = a_j = a}^{N} d_{\text{interest}}(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j) \ , \tag{2.7}$$

$$d^{\text{between}}_{\text{interest}} = \frac{1}{(N/2)^2} \sum_{i<j, a_i \neq a_j}^{N} d_{\text{interest}}(\bar{\mathbf{h}}_i, \bar{\mathbf{h}}_j) \ , \tag{2.8}$$

respectively. Equation (2.6) computes the average interest distance between every pair of individuals i and j in the population. In Eq. (2.7), the bracketed term sums the interest distance between every pair i and j within party a ($a_i = a_j = a$). In Eq. (2.8), the bracketed term sums the interest distance between every pair i and j whose party affiliations differ ($a_i \neq a_j$).

Normalization. In contrast to opinion distance, the range of possible pairwise interest distances depends on both M and K: the maximum possible interest distance is $d^{\max}_{\text{interest}}(M, K) = 0$ if $K = M$ and $2\big|\lfloor M/2 \rfloor - \lfloor |M/2 - K| \rfloor\big|$ if $K < M$, where $\lfloor \cdot \rfloor$ is the

floor function. To allow for meaningful comparisons of interest alignment across values of K, we divided each raw average interest distance by $d^{\max}_{\text{interest}}(M, K)$ when $K < M$.

Data availability

All simulation data and code for simulations, figures, and analytical calculations are available at Github (https://github.com/marikawakatsu/CooperationPolarization2).

Chapter 3

Stereotypes, moral reputations, and indirect reciprocity in group-structured populations

3.1 Notes

This chapter is adapted from:

> Mari Kawakatsu*, Sebastián Michel-Mata*, Taylor A. Kessinger, Corina E. Tarnita**, Joshua B. Plotkin**. Stereotypes, moral reputations, and indirect reciprocity in group-structured populations. Manuscript in preparation.

Author contributions. S. Michel-Mata and I contributed equally to this work as co-first authors, and C. E. Tarnita and J. B. Plotkin as co-senior authors. All authors designed this study. S. Michel-Mata and I conducted computational simulations. T. A. Kessinger, J. B. Plotkin, and I performed mathematical analyses. All authors interpreted the results. I drafted the manuscript, and all authors provided comments.

Prior presentation. I have given a talk on this work at the following seminar:

- Theoretical Ecology Lab Tea, Princeton University (online; May 2021). Joint presentation with Sebastián Michel-Mata.

Acknowledgments. We thank Joe Sartini for helpful discussions on earlier versions of this work. M. Kawakatsu gratefully acknowledges support from the Army Research Office Grant W911NF-18-1-0325.

3.2 Abstract

Moral reputations facilitate cooperation: altruistic behavior may improve individuals' standings in their community, making them more likely to receive help in future interactions. However, moral reputations based on individual actions may be costly to observe, assess, and remember. Instead of relying solely on reputations based on past actions, people may also use stereotypes—generalized reputations of individuals based on their group affiliations. But it remains unclear whether stereotypical views boost or undermine cooperation, as stereotypes reduce cognitive burden while also diminishing the precision of perceived reputations. Here we investigate the effect of stereotype use on indirect reciprocity. We develop a theoretical model of group-structured populations in which individuals are assigned both individual reputations based on their actions and stereotyped reputations based on their group memberships. Among individuals with a uniform propensity to use stereotypes, group-level sharing of reputation information generates emergent in-group favoritism, even when behavior is unconditional on group identity. Stereotype use can spread via social imitation and remain robust against replacement when access to individual reputations is costly, reputation assessments and strategy execution are error-prone, and individuals hold private views of others. Finally, we show that whether stereotyping benefits or harms society depends on the extent to which individual and stereotyped reputations are public knowledge. Under some scenarios, increased stereotyping can simultaneously promote altruism and in-group

bias, resulting in an asymmetric improvement in within- and between-group cooperation.

3.3 Introduction

Reputations are critical to cooperation in human societies (Alexander, 1987; Nowak and Sigmund, 2005; Trivers, 1971). When individuals behave altruistically, their perceived reputations improve, predisposing others to help them in the future. This feedback loop, termed indirect reciprocity, is a powerful motivator and mechanism for cooperation (Boyd and Richerson, 1989; Leimar and Hammerstein, 2001; Nowak and Sigmund, 1998a, 2005). Indeed, people are more likely to offer help when being observed (Bereczkei et al., 2007).

Theoretical studies of indirect reciprocity typically assume that individuals decide whether to cooperate with others based on their moral reputations. Classical models assume reputations are public knowledge (Nowak and Sigmund, 1998b; Ohtsuki and Iwasa, 2004, 2006), facilitated by informal processes such as rapid gossip (Balliet et al., 2020; Nowak and Sigmund, 2005; Sommerfeld et al., 2007) or formal structures such as e-commerce websites that collect customer reviews. Under public knowledge, individuals never disagree about the reputations of one another, and this high level of agreement helps facilitate cooperation. However, in realistic settings, people may hold private—and often differing—opinions about the moral standings of others. Recent work has shown that, under private information, disagreements in how individuals see each other can arise and retard cooperation (Hilbe et al., 2018; Okada et al., 2017, 2018; Uchida, 2010). But studies have also identified mechanisms that can rescue cooperation in such scenarios, including empathetic moral evaluation (Radzvilavicius et al., 2019), generous moral evaluation (Schmid et al., 2021), or the evolution of adherence to public monitoring systems that broadcast public reputations (Radzvilavicius et al., 2021).

However, the underlying assumption in these models—that cooperative behavior is conditional on the receivers' individual reputations—is not always realistic. Although people are adept at processing social information, cognitive constraints limit their ability to attend to and recall information (Fiske and Taylor, 1991). These constraints may be particularly costly under private assessments: mental burden can arise from costs of evaluation (i.e., individuals must form their own judgments of others by either observing their interactions or talking to those who have previously done so) or from costs of memory (i.e., individuals must remember and recall each reputation). Public monitoring systems help ease this burden, but they may introduce economic and administrative costs: in modern societies, establishing and maintaining public institutions often requires taxation and legislation (Radzvilavicius et al., 2021). Altogether, these costs may make individual reputations less accessible.

In reality, instead of relying solely on individual reputations, people may also use heuristics based on social association. For instance, individuals may use biological or social "tags" (i.e., the hypothetical "green beard" (Hamilton, 1964)) to identify and preferentially help those who are like them. Theoretical studies on tag-based cooperation find that altruism can evolve if individuals cooperate only with sufficiently similar others (Riolo et al., 2001). However, in well-mixed populations, tag-based cooperation can fail unless individuals of similar tags always help each other (Roberts and Sherratt, 2002). Without this assumption, successful tag-based cooperation requires additional mechanisms, such as spatial structure (Hammond and Axelrod, 2006; Jansen and van Baalen, 2006), a large number of tags (Jansen and van Baalen, 2006; Traulsen and Nowak, 2007), or knowledge of others' strategies (Masuda and Ohtsuki, 2007)—each of which likely increases rather than decreases the cognitive burden of choosing whom to help.

Stereotypes are another type of reputation based on social affiliation. Social psychology defines stereotypes in various ways (Ashmore and Del Boca, 1981; Hilton

and von Hippel, 1996); here, we adopt the widely accepted view that stereotypes are sets of beliefs about the characteristics, attributes, or behaviors of members of certain social groups (Ashmore and Del Boca, 1981; Hilton and von Hippel, 1996; Judd and Park, 1993). Stereotypes reduce cognitive burden by simplifying how people process information about others: they provide mental shortcuts that are readily learnable and highly structured, whereby group memberships indicate a few associated individual attributes (Macrae and Bodenhausen, 2000; Martin et al., 2014). However, stereotyped reputations may be less accurate relative to individual standings. Although often based on empirical reality, stereotypes are generalizations and may even entail exaggerations (Bordalo et al., 2016; Judd and Park, 1993). For example, the stereotype that members of a particular academic department are collegial may be accurate in general but may not apply to all members.

Given this trade-off between cognitive cost and accuracy, how do individual and stereotyped reputations differentially affect indirect reciprocity? And under what conditions do populations evolve to use stereotyped reputations? Despite the prevalence of stereotypes in human societies, their role in the evolution of cooperation has not been thoroughly studied.

Here we investigate how stereotype use and its evolution impact cooperative behavior in group-structured populations. To do so, we extend a game-theoretic framework of indirect reciprocity in two key ways. First, we consider a group-wise monitoring system, in which members within a group agree on their views of others but might disagree across groups. Although natural and realistic to consider in human populations, such group-level sharing of information remains under-explored in the literature (but see Kessinger and Plotkin (2022) for a theoretical framework for intermediate levels of information availability between strictly private and fully public).

Second, we consider monitoring systems for stereotyped reputations that are analogous to individual reputations. We define a stereotyped reputation as a reputation assigned to a sub-group of the population. In the simplest model of stereotypes, which we study here, to form the stereotyped reputation of a given group, a third-party observer observes a random individual in that group and assesses her moral standing based on her action toward another random individual in the population. Although elementary, this approach captures basic features of stereotypes: stereotypes are generalized beliefs about members of a group based on some empirically observed behavior, and, contrary to the closely related concept of prejudice, stereotypes may be positive or negative (Judd and Park, 1993). However, in some contexts, stereotype assessments may be conditioned on group memberships; for example, an observer might judge a focal individual based on her behavior toward the observer's own group or toward the focal individual's own peers. These alternative models of forming stereotypes remain a topic for future research.

To establish baseline dynamics under stereotyping, we first investigate the effects of stereotype use on the level of sustained cooperation among individuals with a uniform propensity to use stereotyped reputations. Then, we study the evolution of stereotype use using the framework of adaptive dynamics (Geritz et al., 1998). We identify conditions under which stereotype use can evolve and remain evolutionarily stable, and determine the evolutionary consequences of stereotyping for cooperation. This chapter summarizes our results to date; we outline future directions in *Discussion*.

3.4 Model description

We consider a population of N individuals, each belonging to one of K groups. Let ν_k be the fraction of the total population in group k, satisfying $\sum_{k=1}^{K} \nu_k = 1$. For simplicity, our analyses will focus on the scenario with two groups of equal size ($K = 2$, $\nu_1 = \nu_2 = 0.5$).

However, the model implementation described below can accommodate any K and ν_k; future research could analyze the effects of heterogeneous group structures.

3.4.1 Games and behavioral strategies

Individuals play a sequence of pairwise, one-shot donation games, also known as a simplified prisoner's dilemma. In a game, the *donor* must choose whether to cooperate with the *recipient*. If the donor cooperates, she pays a cost c, and the recipient receives a benefit $b > c > 0$; if the donor defects, she incurs no cost, and the recipient receives no benefit. For simplicity, we fix $c = 1$ throughout our analyses.

Whether the donor i cooperates or not depends on her current strategy s_i. We consider three strategies commonly explored in game-theoretic models of reputations (Santos et al., 2018; Sasaki et al., 2017): Always Cooperate (ALLC), which cooperates with any recipient; Always Defect (ALLD), which defects against any recipient; and Discriminate (pDISC), which cooperates when the donor considers the recipient as "good" but defects when the donor considers the recipient as "bad". The parameter p in pDISC modulates the type of information the donor uses to judge whether the recipient is good or bad: with probability $1 - p$, a donor with the pDISC strategy uses the recipient's *individual* reputation, as in commonly used models of indirect reciprocity (Nowak and Sigmund, 2005; Ohtsuki and Iwasa, 2004; Santos et al., 2018; Sasaki et al., 2017); whereas with probability p, the donor instead uses the recipient's *stereotyped* reputation, i.e., the donor's view of the entire group to which the recipient belongs (see *Monitoring systems for reputations* for how reputations are updated and shared).

A key difference between individual and stereotyped reputations is that using the former may carry a higher cognitive cost (Macrae and Bodenhausen, 2000). We operationalize this difference in the simplest possible manner: we assume that a pDISC donor pays an *access cost* $\eta \geq 0$ per interaction when using individual reputations. As we

show, we find rich model dynamics even under this simple assumption; however, because cognitive cost many depend nonlinearly on the number of interactions, future research should consider a variety of cost functions.

Every individual interacts with every other individual in each round, once as a donor and once as a recipient. After all games in a given round are complete, reputations are updated according to a monitoring system (see *Monitoring systems for reputations*) and a social norm (see *Social norms*).

3.4.2 Monitoring systems for reputations

A monitoring system specifies how widely individuals' views of others are shared within the population. Prior studies on indirect reciprocity assume that reputations are either public knowledge, i.e., individuals are assessed publicly and agree about the moral standings of others (Nowak and Sigmund, 1998b; Ohtsuki and Iwasa, 2004, 2006), or private knowledge, i.e., individuals make private judgments about others and may disagree about who is good or bad (Hilbe et al., 2018; Okada et al., 2017, 2018; Uchida, 2010). Here we introduce a *group-wise* monitoring system, in which members within a group agree on their views of others but might disagree across groups. In other words, under group-wise assessment, knowledge is shared less widely than under public assessment, but shared more widely than under private assessment.

Our model also introduces monitoring systems for stereotyped reputations. We define a stereotype as a reputation assigned to a group based on the behavior of a subset of its members. As the simplest implementation of this definition, our model assumes that, to assign a stereotyped reputation to a given group, an observer observes a randomly chosen donor in that group and assesses whether she is good or bad based on her action toward a randomly chosen recipient in the population. This stereotypical assessment of the group is then shared with a subset of the population as defined by the

stereotype system (see below and Table 3.1). Importantly, we assume that reputations can operate at different scales of information sharing: for instance, members of an academic department might disagree on their views of individual colleagues (private reputations) but collectively subscribe to the stereotype that their department is good and a another department is bad (group-wise stereotypes).

Altogether, we consider three levels of individual and stereotyped reputations: *public*, *group-wise*, and *private*, as defined below.

1. *Public*: There is a single public view of each individual or each group.

   Public individual reputations: One randomly chosen observer assesses each donor based on their action toward a random recipient and the recipient's *individual reputation*. The observer then broadcasts the reputations of each individual to the entire population.

   Public stereotyped reputations: One randomly chosen observer assesses a randomly chosen donor in each group based on her action toward a random recipient and the *stereotyped reputation* of the recipient's group, and assigns the resulting reputation as a stereotype to the donor's group. The observer then broadcasts the stereotype of each group to the entire population.

2. *Group-wise*: Each group has a commonly-held view of each individual or group.

   Group-wise individual reputations: One randomly chosen observer from each group assesses every donor based on her action toward a random recipient and the recipient's individual reputation. Each observer then broadcasts the reputations of individuals to her group.

   Group-wise stereotyped reputations: One randomly chosen observer from each group assesses a randomly chosen donor in each group based on her action toward a random recipient and the stereotyped reputation of the recipient's group, and

	Type	No. of observers	No. of assessments	Individual / stereotyped reputations with $K = 2$, $\nu_1 = \nu_2 = 0.5$
Reputations	Public	1	N	$\begin{pmatrix} r_{1\star} & r_{1\star} & \cdots & r_{1\star} \\ r_{2\star} & r_{2\star} & \cdots & r_{2\star} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N\star} & r_{N\star} & \cdots & r_{N\star} \end{pmatrix}$
	Group-wise	K	$N \times K$	$\begin{pmatrix} r_{1A} & \cdots & r_{1A} & r_{1B} & \cdots & r_{1B} \\ r_{2A} & \cdots & r_{2A} & r_{2B} & \cdots & r_{2B} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ r_{NA} & \cdots & r_{NA} & r_{NB} & \cdots & r_{NB} \end{pmatrix}$
	Private	N	$N \times N$	$\begin{pmatrix} r_{11} & \cdots & r_{1N} \\ \vdots & \ddots & \vdots \\ r_{N1} & \cdots & r_{NN} \end{pmatrix}$
Stereotypes	Public	1	K	$\begin{pmatrix} s_{A\star} & s_{A\star} & \cdots & s_{A\star} \\ s_{B\star} & s_{B\star} & \cdots & s_{B\star} \end{pmatrix}$
	Group-wise	K	$K \times K$	$\begin{pmatrix} s_{AA} & \cdots & s_{AA} & s_{AB} & \cdots & s_{AB} \\ s_{BA} & \cdots & s_{BA} & s_{BB} & \cdots & s_{BB} \end{pmatrix}$
	Private	N	$K \times N$	$\begin{pmatrix} s_{A1} & \cdots & s_{AN} \\ s_{B1} & \cdots & s_{BN} \end{pmatrix}$

Table 3.1: Summary of monitoring systems for individual and stereotyped reputations. The rightmost column show corresponding matrices of individual reputations (N-by-N) and stereotyped reputations (K-by-N) in a population of N individuals in two groups of equal size ($K = 2$, $\nu_1 = \nu_2 = 0.5$). Top three rows: r_{ij} is the individual reputation of individual i in the eyes of the whole population ($j = \star$), a group ($j \in \{A, B\}$), or an individual ($j \in \{1, \ldots, N\}$). Bottom three rows: s_{kj} is the stereotyped reputation of group $k \in \{A, B\}$ in the eyes of the whole population ($j = \star$), a group ($j \in \{A, B\}$), or an individual ($j \in \{1, \ldots, N\}$). Without loss of generality, we assume that individuals $1, \ldots, N/2$ belong to group A and $N/2 + 1, \ldots, N$ to group B. Highlighted cells show how widely individual (light blue) and stereotyped (light pink) reputations are shared within each system.

assigns the resulting reputation as a stereotype to the donor's group. Each observer then broadcasts the stereotyped reputations to her group.

3. *Private*: Every individual has a private view of every other individual or group. No information is broadcast.

   Private individual reputations: Every individual assesses every donor based on her action toward a random recipient and the recipient's individual reputation.

   Private stereotyped reputations: Every individual assesses a randomly chosen donor in each group based on her action toward a random recipient and the stereotyped reputation of the recipient's group, and assign the donor's reputation as the stereotype of her group.

Table 3.1 summarizes these three levels of monitoring systems. The structure of individual and stereotyped reputation matrices under the assumption of two equally sized groups are described in the rightmost column of Table 3.1.

3.4.3 Social norms

An observer assesses a donor according to a prescribed social norm, i.e., a set of rules that determine how the donor's moral standing (good or bad) depends on her behavior. We consider four second-order norms, which depend only on the donor's action and the recipient's reputation, that are commonly explored in the literature (Radzvilavicius et al., 2019; Santos et al., 2018; Sasaki et al., 2017). While more complex norms are possible, they often produce less cooperation and mean payoff than the four simple simple norms we consider (Santos et al., 2018). Each norm can be expressed as a binary matrix, whose row indicates the donor's action and whose column indicates the recipient's reputation

(individual or stereotyped):

$$\text{Stern Judging (SJ):} \quad \begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} G & B \end{array} \\ \begin{pmatrix} G & B \\ B & G \end{pmatrix} \end{array} \qquad \text{Simple Standing (SS):} \quad \begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} G & B \end{array} \\ \begin{pmatrix} G & G \\ B & G \end{pmatrix} \end{array}$$

$$\text{Scoring (SC):} \quad \begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} G & B \end{array} \\ \begin{pmatrix} G & G \\ B & B \end{pmatrix} \end{array} \qquad \text{Shunning (SH):} \quad \begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} G & B \end{array} \\ \begin{pmatrix} G & B \\ B & B \end{pmatrix} \end{array}$$

For example, under Stern Judging, an observer will (1) endorse (with a good reputation) a donor who either cooperates with a recipient with a good reputation (according to the observer) or defects against a recipient with a bad reputation but (2) condemn (with a bad reputation) a donor who cooperates with a bad recipient or defects against a good recipient.

Following Sasaki et al. (2017) and Radzvilavicius et al. (2019), we allow for errors in both strategy execution and reputation assessment. With probability $0 \leq u_e \leq 1$, a donor makes an *execution error*, erroneously defecting while intending to cooperate. With probability $0 \leq u_a \leq 1$, an observer makes an *assessment error*, erroneously assigning a good reputation instead of a bad reputation, and vice versa.

3.5 Results

To investigate how the use of stereotypes affects cooperation, we use replicator dynamics (Taylor and Jonker, 1978) to describe how the frequencies of strategies change over time in an infinite population. Under these dynamics, strategies spread in the population at rates proportional to their relative payoffs (see *Materials and methods*). Following Sasaki et al. (2017), Radzvilavicius et al. (2019), Radzvilavicius et al. (2021),

and Kessinger and Plotkin (2022), we assume that the timescale of reputations is faster than that of strategy dynamics; in other words, reputations (either individualized or stereotyped) are assumed to equilibrate before individuals consider updating their strategies (see *Materials and methods*).

3.5.1 Cooperation and in-group favoritism in monomorphic populations

To establish baseline dynamics under stereotyping, we first study monomorphic populations of discriminators with a fixed propensity to use stereotypes. This effectively turns off strategy evolution (all individuals use the same pDISC strategy), allowing us to isolate how equilibrium behavior and payoff depends on both the monitoring systems (public, group-wise, private) and the likelihood of stereotyping (p). In particular, the equilibrium reputations described below are independent of the payoff parameters b, c, and η.

Under the Scoring (SC) norm, the equilibrium level of cooperation is independent of whether the monitoring systems are public, group-wise, or private (Fig.C.1). But under all other norms, we find that the more public the individual or stereotyped reputation system, the higher the resulting level of cooperation (Fig.3.1 for Stern Judging; Fig.C.2 for Simple Standing; and Fig.C.3 for Shunning). In other words, for a given reputation system (e.g., public) and probability of using stereotypes (e.g., $p = 0.5$), average cooperation is lowest when stereotypes are private ($\sim$62%, Fig.3.1A), intermediate when group-wise ($\sim$72%, Fig.3.1B), and highest when public ($\sim$83%, Fig.3.1C). These results are consistent with previous work on reputations, which has found that public monitoring outperform private assessment, because the former facilitates greater agreement among individuals on their views of one another (Hilbe et al., 2018; Okada et al., 2017). Our results suggest that the same intuition extends to group-wise

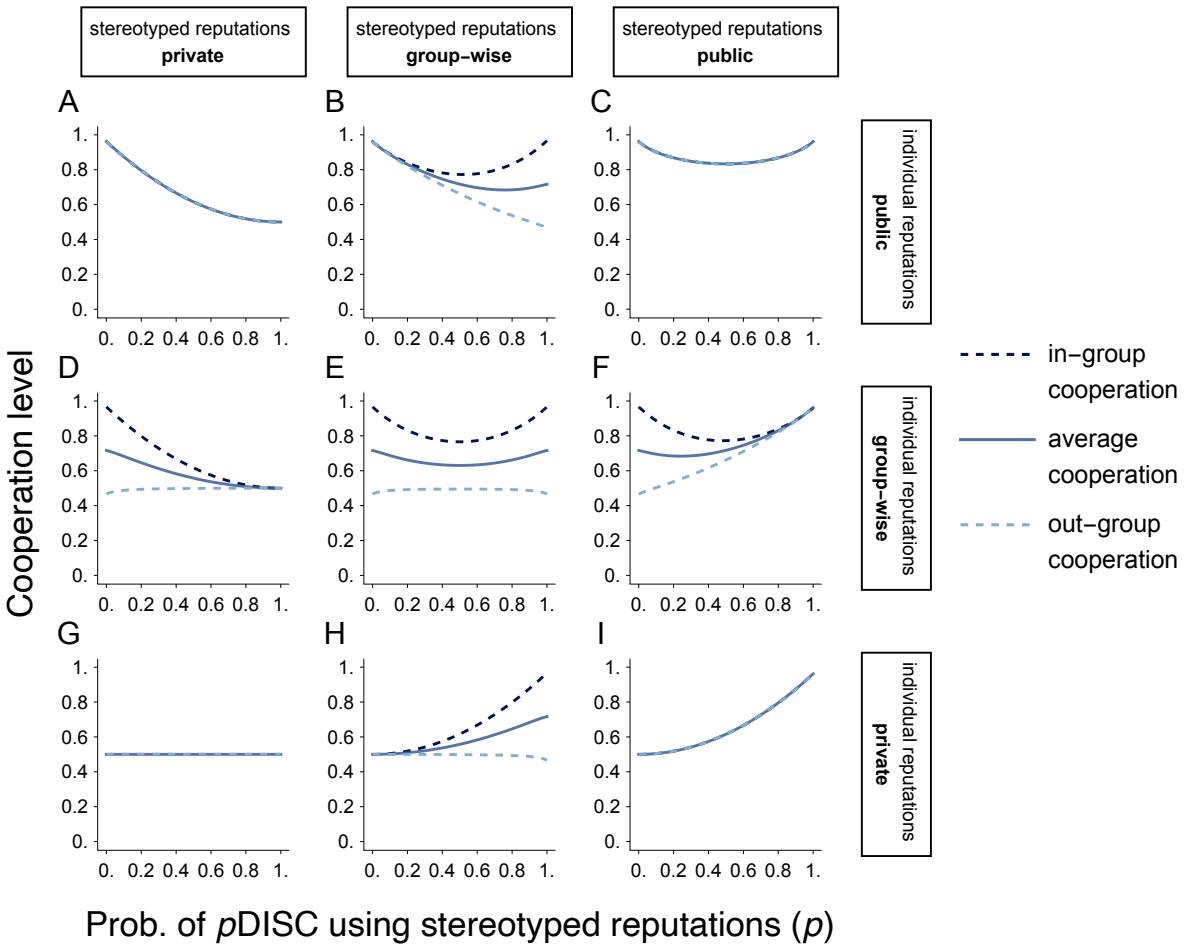

Figure 3.1: Group-wise monitoring promotes intermediate levels of cooperation and in-group favoritism in monomorphic populations. We analyzed equilibrium levels of cooperation under the Stern Judging norm among pDISC strategists with a uniform propensity p to use stereotypes. Individuals are in two groups of equal size ($K = 2$, $\nu_1 = \nu_2 = 0.5$). Each panel corresponds to a combination of monitoring systems for reputations (rows) and stereotypes (columns). Solid lines indicate average levels of cooperation in the population; dashed lines indicate average levels of cooperation within (navy) and between (light blue) groups. Error rates are $u_e = u_a = 0.02$. For analogous results for the Simple Standing, Scoring, and Shunning norms, see Figs. C.1–C.3.

monitoring: a system in which knowledge about others' standing is semi-public, i.e., shared within but not necessarily between groups, outperforms private monitoring but underperforms fully public monitoring. And the benefits of monitoring systems that

broadcast information, either to groups or to the entire populations, hold for both for individual and stereotypical reputations.

In fact, in monomorphic populations, individual and stereotypical reputations systems (public, group-wise, or private) have symmetric effects on cooperation (Fig.3.1), as indicated by the symmetry across the diagonal in Fig.3.1 (Fig.3.1A vs. I, B vs. F, D vs. H). For example, the level of cooperation at $p = 0.2$ under public individual reputations and private stereotypes (Fig.3.1A) matches the level of cooperation at $p = 0.8$ under private individual reputations and public stereotypes (Fig.3.1I). This suggests that, in the absence of competing strategies, neither individual nor stereotypical reputations are more effective at promoting a cooperative society per se; instead, cooperation increases with more frequent use of whichever type of reputational information is shared more widely across the population.

Moreover, the population maximizes the level of cooperation when individuals use only individual ($p = 0$) or only stereotyped ($p = 1$) reputations. If individual and stereotypical information are held at different scales, then cooperation is maximized at either $p = 0$ or 1 (Fig.3.1A, B, D, F, H, I). If reputations both group-wise or both public, cooperation is equally maximized at both ends (Fig.3.1C, E). This highlights that, in general, the population does best when everyone uses individual assessment or everyone stereotypes, at least when strategies are not evolving. The only exception is when both types of reputations are held privately, in which case the cooperation level is independent of p (Fig.3.1G).

Group-wise monitoring gives rise to a phenomenon not found under public or private monitoring: in-group favoritism, in which individuals cooperate preferentially with members of their own group (Fig.3.1B, D–F, H). This phenomenon is emergent: we assume no behavioral strategies whose cooperation is conditional on group memberships. Instead, group-wise monitoring ensures agreement among members of

the same group (group members share the same view of all individuals in the population) while creating possibility for disagreement among members of different groups (groups can differ in how they view each individual in the population). These different levels of agreement within and between groups, in turn, produce differential in- and out-group cooperation. The gap between in- and out-group cooperation is most pronounced when individuals are using only individual reputations ($p = 0$; Fig.3.1D–F) or only stereotypes ($p = 1$; Fig.3.1B, E, H). The value(s) of p that maximizes this divergence also maximizes cooperation in some (Fig.3.1D, E, H) but not all cases, highlighting that in-group bias can but does not always boost global cooperation.

In-group favoritism is particularly strong under Stern Judging, relative to the other three norms (Figs.C.1–C.3). This is likely because Stern Judging harshly punishes discrepancies in assessment: under Stern Judging, a donor garners a bad reputation if she cooperates with a recipient who is considered bad in the eyes of a third-party observer, whereas under Simple Standing, for example, the donor would retain a good reputation. Moreover, prior literature has found that the Stern Judging norm is also the most effective at stimulating cooperation under indirect reciprocity (Santos et al., 2018); indeed, Stern Judging is naturally favored over other norms in models of multi-level selection (Cooney et al., in prep). To study the most salient effects of the group-wise monitoring on cooperation, our analysis will hereafter focus exclusively on the Stern Judging norm, but our approach is general and covers all four norms (*Materials and methods*).

3.5.2 Evolution of stereotype use

We have shown that, under Stern Judging, each combination of monitoring systems has an optimal level (or levels) of stereotype use that maximizes collective cooperation in monomorphic populations. However, it remains unclear whether a stereotyping propensity that is best for the collective will actually evolve in a population under

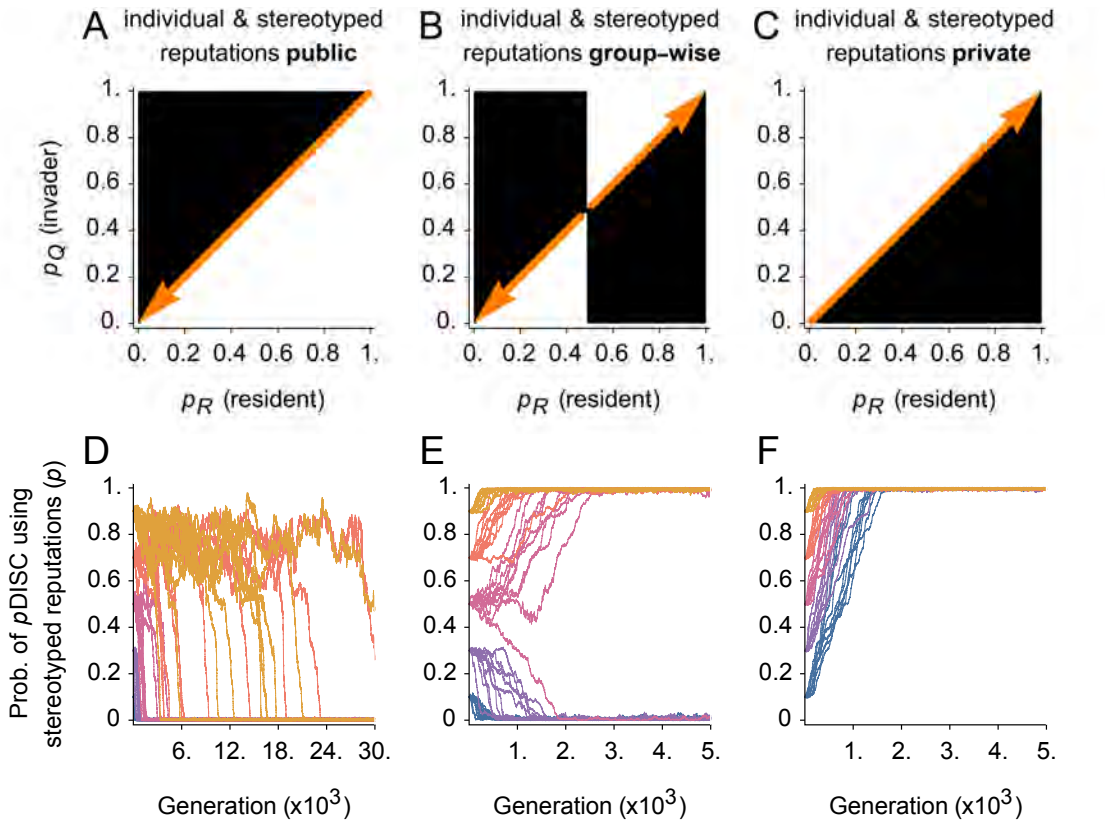

Figure 3.2: Evolution of stereotype use. We use adaptive dynamics to predict which invader types p_Q can invade which resident types p_R under Stern Judging. Results shown are with individual and stereotype reputations held publicly (A, D), group-wise (B, E), or privately (C, F); see Figs. C.4 and C.5 for expanded versions. Individuals are in two groups of equal size ($K = 2$, $\nu_1 = \nu_2 = 0.5$). (A–C) Pairwise invasibility plots indicate parameter regions in which p_Q can invade p_R (white), i.e., invader payoff Π_Q exceeds resident payoff Π_R in the limit of negligible invader frequency, or not (black) (*Pairwise invasibility*). Orange arrows indicate predicted directions for the evolution of p. (D–F) Stochastic simulations in finite populations of $N = 50$ with small, local mutations (*Stochastic simulations*) support the predictions of pairwise invasibility analysis. Lines indicate mean p in the population over time. Colors distinguish initial conditions (monomorphic populations with uniform p), with 10 simulation runs per initial condition. Payoff parameters are $b = 3$, $c = 1$, and $\eta = 0.3$; error rates are $u_a = u_e = 0.02$.

selection for increasing (individual-level) payoffs. Under what conditions do individuals evolve to adopt stereotyping, a cognitively inexpensive heuristic, and how what effect does the evolution of stereotype propensities have on cooperative in the resulting

society? Conversely, is a population that adopts the more cognitively costly approach of individual assessment resist invasion by those who use stereotypes?

To investigate whether individual-level selection favors stereotype use, we apply the framework of adaptive dynamics (Geritz et al., 1998). We restrict our analysis to discriminators (pDISC), and we let the the propensity for an individual to use stereotypes p evolve according to biased imitation of individuals with high payoff. We define Π_R and Π_Q as the per-round expected average payoff of the resident and invader types with probabilities $0 < p_R, p_Q < 1$ of using stereotypes, respectively. To determine the conditions under which an invader with p_Q can invade a resident population with p_R, we derive an analytical expression for the invasion fitness $\Pi_R - \Pi_Q$ in the limit of negligible invader frequencies, and we evaluate it numerically across a range of model parameters (Fig. 3.2; see also *Materials and methods*). To determine long-term population dynamics, we also identify singular points $0 < p^* < 1$ and we characterize their evolutionary stability (Fig. 3.3; *Materials and methods*). Since p is restricted to $p \in [0, 1]$, we also consider whether the extremal values ($p = 0$ and 1) are evolutionary attractors. Finally, to support the analysis of singular points and their stability in an infinite population, we perform stochastic simulations in finite populations of $N = 50$ individuals with small, local mutations in p (Fig. 3.2; *Materials and methods*).

Selection can favor stereotyping. But whether and to what extent stereotyping evolves depends on the monitoring system—individual, group-wise, or public. Figure 3.2 focuses on the cases where reputations are monitored at the same scale (variable scales are included in Figs. C.4 and C.5). In this example, stereotyping can evolve under group-wise or private monitoring but not under public monitoring (Fig. 3.2): under public monitoring, $p = 0$ (i.e., no stereotyping) is the only attractor (Fig. 3.2A, D); under group-wise and private monitoring, by contrast, maximum stereotype use ($p = 1$) is an attractor (Fig. 3.2B, C, E, F). In fact, under group-wise

monitoring, there is a single repulsive singular point (Fig.3.2B, E): if the (monomorphic) resident population starts out with p_R below this value, then the population will evolve toward complete reliance on individual reputations ($p = 0$); otherwise, the population will adopt a high level of stereotyping ($p = 1$). Thus, unlike for cooperation levels—in which group-wise assessments achieve intermediate performance between public and private assessments (Fig.3.1)—the evolution of stereotyping propensity exhibits bistability: a population may evolve to one or another stable propensity, depending on the initial conditions (Fig.3.2).

These results suggest that the evolutionary dynamics of stereotype use are potentially more complex than the dynamics of cooperation in monomorphic populations. To better understand these dynamics, we analyze how the evolutionary stability of stereotyping depends on key model parameters, namely the cost of accessing reputations and the rate of errors in reputation assessment and strategy execution.

3.5.3 Evolutionary stability of stereotype use

Across monitoring systems, evolution favors stereotyping when access to individual reputations are costly (Fig.3.3). In general, when η is small, $p = 0$ is the unique evolutionary attractor; when η is large, $p = 1$ is the unique evolutionary attractor. Thus, the population evolves toward no stereotyping when access to individual reputations is inexpensive but towards full stereotyping when they are expensive. The one exception is when both reputations are private, in which case $p = 1$ is the unique attractor for any $\eta > 0$; selection is neutral in p for $\eta = 0$ (Fig.3.3G; see also Fig.C.6G).

This trend makes sense in light of the cost-precision trade-off between individual and stereotyped reputations: individual reputations are more costly to use than stereotypes, but they are simultaneously more precise an indicator each individuals' standing because, in our model, the stereotype as a group results from the assessment of a

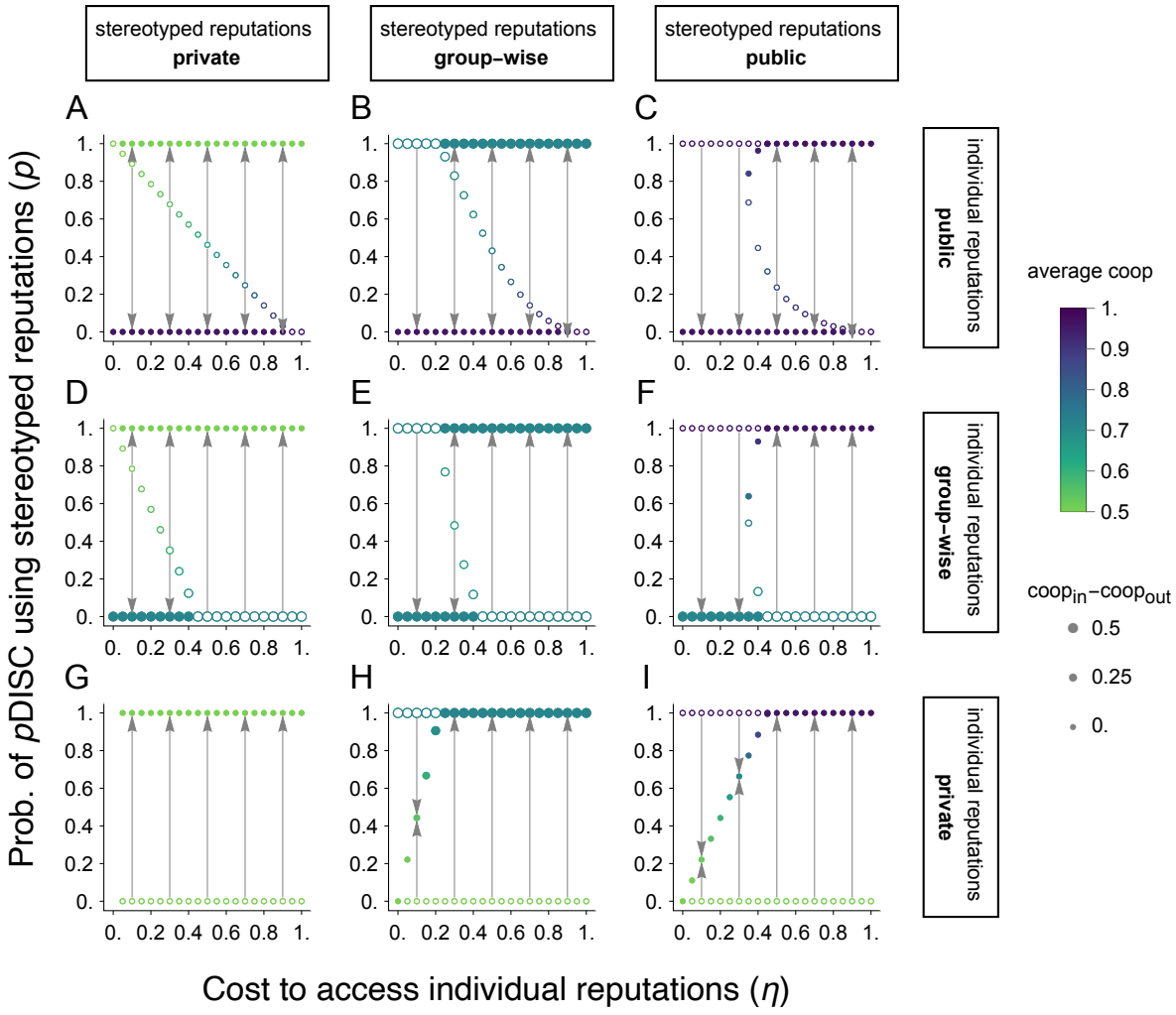

Figure 3.3: Evolutionarily stable levels of stereotyping and corresponding outcomes for cooperation. We analyzed the stability of the singular points (p^*) of the adaptive dynamics as well as the extremal values of p (0 and 1) as a function of reputation access cost η in two groups of equal size ($K = 2$, $\nu_1 = \nu_2 = 0.5$). Each panel corresponds to a combination of monitoring systems for individual reputations (rows) and stereotyped reputations (columns). Solid and empty circles are attractive and repulsive points, respectively. Gray arrows indicate predicted directions for the evolution of p. The color of each circle indicates the average level of cooperation in a monomorphic population of pDISC with p corresponding to that circle (Fig.3.1). Size indicates difference between in- and out-group cooperation levels (i.e., average in-group cooperation − average out-group cooperation); the greater the difference, the larger the circle, and the stronger the preference toward cooperating with one's own group. Parameters: $b = 3$, $c = 1$, $u_e = u_a = 0.02$.

randomly sampled individual in that group. When the cost to access individual reputations η is sufficiently high, though, it exceeds benefit provided by increased reputational precision, so that stereotyping becomes favorable.

While costly individual reputations promote stereotyping in general, the nature of the transition from no stereotyping to full stereotyping varies across monitoring systems. When individual reputations are held privately but stereotypes are not (i.e., they are held group-wise or publicly), the evolutionarily stable level of stereotype use increases gradually from $p = 0$ to 1 with increasing η (Fig. 3.3H, I). Intermediate values of η lead to unique intermediate equilibria ($0 < p^* < 1$): regardless of the initial conditions, individuals will converge to a strategy that uses both individual and stereotyped reputations with some probability. This strategy is sensible and intuitive: by using both reputations in some proportions, one can strike a balance between the benefits of reputational precision (higher for individual reputations) and costs of cognitive burden (lower for stereotyped reputations).

When individual reputations are held group-wise or publicly, by contrast, there are regions of bistability facilitated by backward bifurcations: at an intermediate value of η, an attractive singular point emerges at a high level of stereotyping (e.g., $p^* \approx 0.95$ at $\eta = 0.4$ in Fig. 3.3C). Further increasing η increases the range of initial conditions (i.e., p_R of the initial monomorphic resident population) from which the population will evolve towards an attractor at $p > 0$—until $p = 0$ loses stability, when the cost η is sufficiently large.

This hysteretic patterns suggest that the use of individual reputations can be resistant to small changes in access cost (Fig. 3.3A–F): under group-wise or public reputations, if one starts from a society whose individuals rely solely on reputations ($p = 0$), then a small increase in η may not immediately trigger a shift toward stereotyping. But this shift, when it does occur, will be sudden: under public reputations, for example, our

analysis predicts a jump from $p = 0$ to 1 as η crosses ~0.90 (Fig. 3.3C). Stereotyping behavior can also be 'sticky': in a population that relies only on stereotypes and never on reputations ($p = 1$), merely decreasing the cost of accessing reputations would not curtail use of stereotypes, at least not immediately.

So far, we have identified the cost of accessing individual reputations as a key driver of stereotyping. This raises two questions: How do the other parameters in the payoff function affect the evolution of stereotyping? And how does the evolution of stereotype use propensity affect cooperation? We address the former first and return to the latter in *Consequences for cooperation*.

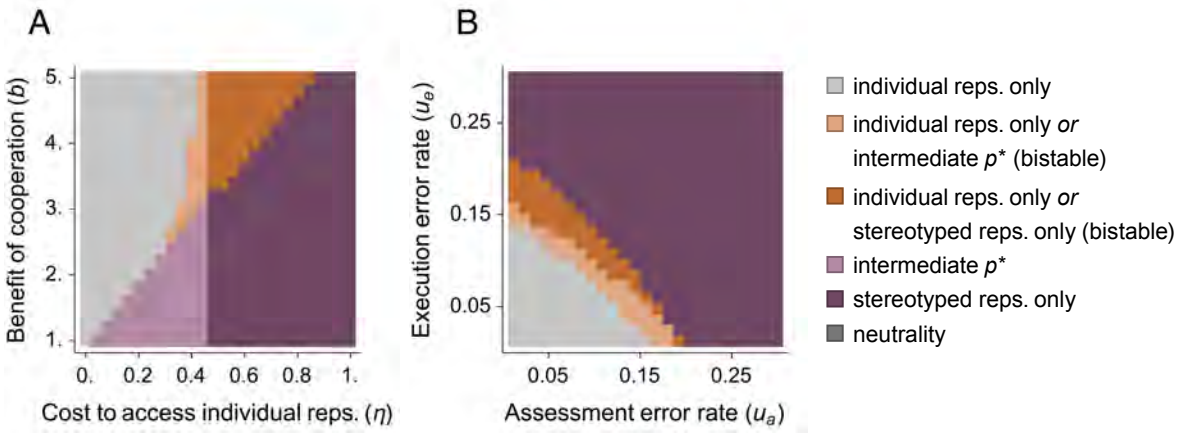

Figure 3.4: Errors and costly reputations promote stereotype use. We show the number and type of evolutionarily attractive values of p as a function of (A) benefit of cooperation (b) and cost of accessing reputations (η) and (B) rates of assessment (u_a) and execution (u_e) errors. We assume two groups of equal size ($K = 2, \nu_1 = \nu_2 = 0.5$) Results shown are with stereotyped reputations held publicly and individual reputations held group-wise; see Fig. C.6 and Fig. C.7 for expanded versions with all 9 combinations of monitoring systems. Light gray indicates parameter regions in which stereotype use does not evolve ($p = 0$ in the only stable outcome). Hues of purple indicate regions in which stereotype use is guaranteed to evolve ($p^* > 0$ is the only stable outcome). Hues of orange indicate regions of bistability ($p = 0$ and $p^* > 0$ are both stable outcomes), in which stereotype use may evolve depending on initial conditions. Parameters: $\eta = 0.3, b = 3, c = 1, u_e = u_a = 0.02$.

We find that, for a fixed cooperation cost ($c = 1$), decreasing the benefit b of cooperation promotes stereotyping behavior (Fig. 3.4A, Fig. C.6). A lower b effectively

increases the relative cost of using individual reputations, thus making stereotypes more beneficial. As a result, given a fixed η, decreasing b shifts the system from a regime that does not support stereotyping (light gray regions in Fig. 3.4A) through bistable regimes (light and dark orange regions in Fig. 3.4A, in which both non-stereotyping and stereotyping are possible outcomes depending on initial conditions) to regimes with a single attractive point (light and dark purple regions in Fig. 3.4A) in which stereotyping will evolve regardless of initial conditions.

Interestingly, the evolutionary outcome is independent of the benefit of cooperation, b, under private individual reputations (Fig. C.6, bottom row). Individuals gain b when donors view them as having good reputations and therefore cooperate with them (*Individual reputations of discriminators*). Under private monitoring, on average, invader and resident individuals have identical individual reputations. (This is because when two observers judge the same individual privately, their assessments are uncorrelated, regardless of whether the focal individual uses p_Q or p_R.) As a result, for any value of b residents and invaders receive equal amounts of cooperation. And so changing b has no impact on their relative fitness and, consequently, on the evolution of stereotype propensity.

In addition to the costs and benefits of cooperation, errors in reputation assessment and strategy execution also facilitate the evolution of stereotyping (Fig. 3.4B, Fig. C.7). Because each assessment introduces a possibility for an erroneous judgment, assessment errors are more harmful for individual reputations than for stereotyped reputations: a single observation is required to assign a stereotype to a group of N/K individuals, whereas N/K observations are required to assign N/K individual reputations. Stereotyping thus confers a roughly N/K-fold increase in the accuracy of evaluations. Execution errors also penalize payoffs under individual reputations more than they do under stereotypes: if a donor defects erroneously, she becomes much more likely to

garner a bad individual reputation—at least under Stern Judging—which, in turn, reduces the likelihood that others cooperate with her; however, the group to which she belongs could maintain a good stereotype if the donor sampled for stereotype assessment is viewed as 'good.' Thus, relying on stereotypes can help mitigate the vicious cycle of bad reputations and diminished cooperation.

Our evolutionary analysis thus far has focused on parameters that affect individual-level payoffs (η, b) or accuracy (u_a, u_e). We now return to the question of how the monitoring systems affect the evolutionary stability of stereotyping. To compare evolutionary outcomes across scenarios with and without bistability, we compute the *expected level of stereotyping (expected p)* (Fig. 3.5) as the weighted average of the evolutionarily attractive values of p, where the weight for each p is the range of initial values of $p_0 \in [0, 1]$ that will evolve toward that p—that is, the weight of each attractive point is given by the volume of its basin of attraction. This procedure is equivalent to sampling initial conditions—monomorphic pDISC populations—uniformly at random, letting p evolve, and computing the average long-term p in the population.

Plotting expected p against access cost η reveals that, in general, the more private the monitoring of individual reputations, the greater the expected propensity to use stereotypes (Fig. 3.5). Given a fixed stereotype monitoring system (private, group-wise, or public), expected p becomes non-zero at a lower η when individuals reputations are held privately (light olive lines in Fig. 3.5) than when they are shared (i.e., held group-wise or publicly; olive and dark olive lines, Fig. 3.5).

We can understand these results in terms of the trade-off between precision and disagreement under each monitoring system. As discussed previously, in our model, stereotyped reputations are less precise than individual reputations because the former is based on the behavior of a single randomly sampled donor. However, stereotypes provide a benefit that potentially counteracts the cost of diminished precision: given a

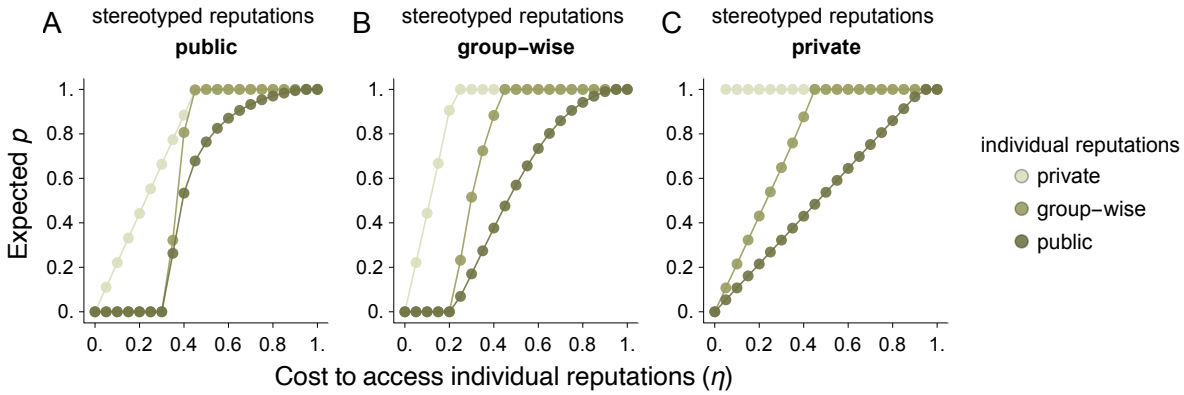

Figure 3.5: Private assessments of individual reputations promote the use of stereotypes. We analyzed the expected level of stereotyping (expected p) as a function of the cost to access individual reputations (η) under Stern Judging and in two groups of equal size ($K = 2$, $\nu_1 = \nu_2 = 0.5$). Expected p is computed as the weighted average of evolutionarily attractive values of p, where the weight for each value is the range of initial values of p that will evolve toward that value (see text for an expanded discussion). Stereotyped reputations are public (A), group-wise (B), or private (C). Colors indicate the monitoring system for individual reputations. Parameters: $b = 3$, $c = 1$, $u_a = u_e = 0.02$.

scale of monitoring, stereotypes constrain the space of possible disagreement relative to individual reputations simply because the former requires fewer assessments (Table 3.1). This reduction in disagreement is particularly beneficial when the underlying reputation monitoring system allows high levels of disagreement, i.e., when reputations are monitored privately. As a result, for a fixed access cost to individual reputations η and stereotype system, the expected use of stereotypes is highest under privately held reputations (Fig. 3.5A–C).

3.5.4 Consequences for cooperation

We have shown that selection can favor the use of stereotypes when individual reputations are relatively costly to access; when strategy execution and moral judgments are prone to errors; and when individuals hold private view of others. These findings mean that, in principle, stereotyping behavior could be reduced by eliminating these conditions. But we may not always want to curb the use of stereotypes. In fact, as we

saw in Fig. 3.1, stereotyping behavior can be either beneficial or harmful from the perspective of cooperation. When selection acts on stereotyping propensity, how cooperative is the resulting society? To address this question, we study the average level of cooperation in a monomorphic population of discriminators pDISC at each evolutionarily attractive value of p (Fig. 3.3).

We find that whether the transition from no stereotyping to full stereotyping improves cooperation depends on the underlying reputation and stereotype systems. In general, when stereotypes are more public than reputations (panels below the diagonal, Fig. 3.3F, H, I), the greater the access cost η, the higher the level of cooperation achieved at the corresponding evolutionarily attractive p. Conversely, when stereotypes are less public than reputations (panels above the diagonal, Fig. 3.3A, B, D), reducing η improves cooperation at the attractor. When stereotypes and reputations are monitored at the same scale, η has little effect on the level of cooperation at attractive p (with the exception of the bistable region in Fig. 3.3C, in which cooperation dips slightly at $p^* < 1$).

But group-wise monitoring systems introduce an important subtlety. Under private individual reputations and group-wise stereotypes, higher access cost (η) increases cooperation while simultaneously increasing the gap between in-group and out-group cooperation (Fig. 3.3H); the case with individual reputations shared group-wise and stereotypes held privately exhibits a similar effect but in reverse direction of η (Fig. 3.3D). This highlights an asymmetry that arises from the interplay between the evolutionary dynamics of reputation use and the emergence of in-group bias (Fig. 3.1): when cooperation increases via the increased use of group-wise reputations (individual or stereotyped), the boost in cooperation benefits in-group interactions more than it helps out-group interactions.

3.6 Discussion

Social norms and reputations catalyze altruistic behavior in human societies. Most theoretical studies on indirect reciprocity assume that individuals decide whether to cooperate with others based on their individual moral standings, i.e., reputations based on their prior actions (Hilbe et al., 2018; Nowak and Sigmund, 1998b, 2005; Ohtsuki and Iwasa, 2004, 2006; Okada et al., 2017, 2018; Uchida, 2010). However, individualized reputations may be cognitively costly to assess and access. This challenge is particularly relevant in modern societies, in which cooperative interactions often involve strangers whose reputational information may be difficult to obtain. As a result, individuals may resort to heuristic, generalized reputations based on social affiliations: stereotypes.

In this chapter, we have explored how stereotype use and its evolution affect indirect reciprocity. In our model, conditional cooperators (discriminators) probabilistically use either individual or stereotyped reputations of potential interaction partners to determine whether to cooperate. A discriminator is thus characterized by her propensity to use stereotypes.

Among discriminators with a uniform stereotype use propensity, we find that group-wise reputations promote cooperation more successfully than public reputations but less so than public reputations. Interestingly, group-wise monitoring systems also facilitate emergent in-group favoritism: individuals cooperate preferentially with members of their own group, despite playing a strategy conditional only on reputations and not on group memberships. This suggests that in-group bias—a phenomenon widely observed in both real-world and laboratory settings (Tajfel, 1982; Tajfel et al., 1971; Yamagishi and Mifune, 2008)—can arise as a function of differential access to information, even if individuals harbor no preferences toward in- or out-groups a priori.

Future work further needs to investigate structured reputation systems, particularly those with multiple groups or public institutions (Kessinger and Plotkin, 2022).

We have also analyzed whether stereotype use can spread via social imitation. We find a rich set of outcomes: populations can evolve to never, always, or sometimes use stereotypes, or arrive at two different levels of stereotyping depending on initial conditions. As expected from the cost-precision trade-off between individual and stereotyped reputations, players tend to adopt stereotyping behavior when individual standings are costly to access.

However, high access costs are not the sole driver. Stereotype use is also favored when strategy execution and moral assessments are error-prone. Stereotyping reduces the number of opportunities for assessment errors: fewer observations are required to assign a stereotype to a group than to assign individual reputations to the corresponding members. Also, while a single erroneous action can have immediate effects on the donor's status, it has comparatively little impact on the status of her group. For these reasons, indirect reciprocity based on stereotypes may be less sensitive to noisy actions and assessments than its classical counterpart based on individualized reputations (Okada et al., 2017, 2018; Uchida, 2010). For the reader distressed by the prevalence of quick, harsh judgments cast on misdeeds—i.e., cancel culture—this may be good news: in a world of imperfect individuals, generalized reputations can provide a buffer against mistaken actions and judgments. But stereotypes can be a double-edged sword: group-based reputations may obscure individual errors that merit punishment, allowing offenders to remain in good standing when they do not deserve to.

But the question remains: can the spread of stereotyping behavior via social imitation ever improve cooperation? Our analysis reveals that, on average, stereotyping can indeed lead to higher cooperation levels, but only if stereotype information is more publicly available than individual reputations. However, group-wise monitoring

introduces an important nuance: under private individual reputations and group-wise stereotypes, increased use of stereotypes promotes cooperation and in-group favoritism simultaneously. This results in an asymmetric improvement in cooperation levels, with individuals cooperating more on average but primarily with their in-group members. This finding bodes ill in a world where political tribes hold different, antagonistic stereotypes of one another: in the United States, for example, both Democrats and Republicans say that members of the other party are hypocritical, selfish, and closed-minded (Iyengar et al., 2019). Our results suggest that, under such polarization, increased reliance on stereotypes may only entrench in-group preference without improving society-wide cooperation.

Our model assumes arguably the most straightforward mechanism for assigning stereotypes. To form the stereotyped reputation of a group, an observer assesses a random member of that group and applies her reputation to the whole group. Because few studies have explored stereotypes in the context of cooperation, we have used this simple setup to establish their baseline effects. But this approach neglects the complex factors on which stereotypes may depend. For example, evidence shows that individuals preferentially recall information consistent with existing stereotypes (Hilton and von Hippel, 1996; Kashima, 2000). As a result, stereotypes may be slower to change than individual reputations: observing a single "bad" behavior by a member of a stereotypically "good" group may not alter the observer's stereotype of that group. Future research could explore these additional mechanisms for stereotype formation.

Our study also focuses on a simple population structure. While we have assumed that strategic interactions are well-mixed, group memberships can skew interaction patterns so that individuals interact more frequently with in-group members than with out-group members. A simple extension could consider how such interaction insularity affects the dynamics of stereotyping. Variation in group size may also play a key role in

forming stereotypes. Social cognition research shows that people tend to view minority groups more negatively than majority groups, even if they behave identically (Hilton and von Hippel, 1996). However, it remains unclear how this minority bias affects indirect reciprocity. While our analysis has focused on two equally sized groups, our full model allows for more complex structures, including arbitrarily many groups of arbitrary sizes. Future work could take advantage of this flexibility to investigate how group size variation influences the use of individual versus stereotyped reputations. Finally, contrary to our assumption that each individual remains a member of one social group, group memberships can be dynamic and overlapping—even within a lifetime, one can belong to different cultural, familial, or occupational groups at different times. Although we have tools for studying cooperation in temporal social networks (Cavaliere et al., 2012; Fu et al., 2012; Tarnita et al., 2009a), little is known about the co-evolution of population structure, individual reputations, and stereotypes. These topics remain important directions for future research.

3.7 Materials and methods

The notations used in our analyses are summarized in Table 3.2.

3.7.1 Replicator dynamics

We study the evolutionary dynamics of strategies using replicator dynamics in an infinitely large population (Taylor and Jonker, 1978). We consider three main types of strategies: Always Cooperate (ALLC), denoted X; Always Defect (ALLD), denoted Y; and Discriminate (pDISC), denoted Z. Discriminators condition their behavior on their views of others: pDISC cooperates with recipients who are "good" and defects against those who are "bad," where good or bad is determined based on the recipients'

Parameter	Definition
K	number of groups
q_C	probability that cooperating with a bad individual yields a good reputation (barring errors)
q_D	probability that defecting against a bad individual yields a good reputation (barring errors)
b	benefit of cooperation
c	cost of cooperation
η	cost of accessing individual reputations
p	probability that a pDISC uses stereotyped reputations rather than individual reputations (stereotype use propensity)
u_a	probability that a bad donor is accidentally assigned a good reputation (assessment error)
u_e	probability that a donor intending to cooperate accidentally defects (execution error)
ϵ	probability that an individual who intends to cooperate with a recipient with a good reputation is assigned a good reputation ($\epsilon = (1-u_e)(1-u_a) + u_e u_a$)
P_{XY}	probability that a donor intending to $Y \in$ {cooperate, defect} with a recipient whom the observer views as $X \in$ {good, bad} is viewed as good

Variable	Definition
g_X	probability that an ALLC has a good reputation
g_Y	probability that an ALLD has a good reputation
g_Z	probability that a pDISC has a good reputation
ν_I	fraction of the population in group I
f_i^I	frequency of strategy i in group I
Π_i^I	fitness of a strategy i individual in group I
$\bar{\Pi}^I$	average fitness of individuals in group I
$g_i^{I,J}$	fraction of strategy i individuals in group I who have good individual reputations in the eyes of J
$g^{I,J}$	fraction of I individuals who have good individual reputations in the eyes of J
$g_S^{I,J}$	fraction of I individuals who have good stereotyped reputations in the eyes of J

Table 3.2: Parameters and variables used in pairwise invasibility analysis.

stereotyped reputations with probability p and based on their individual reputations with probability $1-p$. We assume that p is fixed for a given pDISC.

Individuals are distributed among K non-overlapping groups, with each group I containing fraction ν_I of the population ($\sum_{I=1}^{K} \nu_I = 1$). However, we assume that interactions are well-mixed: every individual plays a game with every other individual. We also assume global imitation: individuals can choose to imitate anyone in the population, not just those within their groups.

Let f_i be the frequency of strategy $i \in \{X, Y, Z\}$ in the total population and f_i^I be the frequency of strategy i in group $I \in \{1, \ldots, K\}$, such that $f_i = \sum_I \nu_I f_i^I$. The replicator dynamics (see Kessinger and Plotkin (2022) for the derivation) follows

$$\dot{f}_i^I = f_i^I \sum_J \nu_J \left(\Pi_i^J - \bar{\Pi}^J \right) . \tag{3.1}$$

Here, Π_i^I is the average fitness of strategy i individuals in group I; $\bar{\Pi}^I$ is the average fitness of group I, given by $\bar{\Pi}^I = \sum_i f_i^I \Pi_i^I$.

The fitness of each individual is the average expected payoff earned from the donation game. In each round, each individual interacts once as a donor and once as a recipient with every other individual in the population. As a recipient, each individual earns a benefit b from every interaction with a donor who cooperates, either a cooperator (ALLC) or a discriminator (pDISC) who views the recipient (the focal individual) as good. As a donor, a cooperator (ALLC) pays a cost c to cooperate in every interaction; a defector (ALLD) never pays the cost; and a discriminator (pDISC) pays the cost only when interacting with a recipient with a good standing. A pDISC pays an additional η per interaction when using a recipient's individual reputation to determine her standing but not when using her stereotyped reputation. Finally, with probability $0 \leq u_e \leq 1$, a donor erroneously defects while intending to cooperate.

Altogether, the average fitness of each strategy in group I is given by

$$\begin{aligned}
\Pi_X^I &= (1-u_e)\left[b\sum_J \nu_J\left(f_X^J + f_Z^J[(1-p)g_X^{I,J} + pg_S^{I,J}]\right) - c\right] , \\
\Pi_Y^I &= (1-u_e)\left[b\sum_J \nu_J\left(f_X^J + f_Z^J[(1-p)g_Y^{I,J} + pg_S^{I,J}]\right)\right] , \\
\Pi_Z^I &= (1-u_e)\left[b\sum_J \nu_J\left(f_X^J + f_Z^J[(1-p)g_Z^{I,J} + pg_S^{I,J}]\right) - c[(1-p)g^{\bullet,I} + pg^{\star,I}]\right] - \eta(1-p) ,
\end{aligned} \tag{3.2}$$

where we adopt and extend the notation in Kessinger and Plotkin (2022):

- $g_i^{I,J}$, $i \in \{X, Y, Z\}$, is the fraction of group I members using strategy i who have good individual reputations in the eyes of group J;
- $g_S^{I,J}$ is the fraction of group I members who have good stereotyped reputations in the eyes of group J;
- $g^{\bullet,I} = \sum_J \nu_J \sum_i f_i^J g_i^{J,I}$ is the fraction of individuals in the whole population who have good individual reputations in the eyes of group I; and
- $g^{\star,I} = \sum_J \nu_J g_S^{J,I}$ is the fraction of individuals in the whole population who have good stereotyped reputations in the eyes of group I.

In Eq. (3.2), the quantity $(1-p)g_i^{I,J} + pg_S^{I,J}$ is the probability that a strategy i individual in group I receives cooperation from discriminators in group J. The quantity $(1-p)g^{\bullet,I} + pg^{\star,I}$ gives the fraction of individuals in the population a discriminator pDISC sees as good and therefore cooperates with.

3.7.2 Reputation dynamics

We assume that reputations equilibrate more quickly than strategies; in other words, the timescale of reputations is faster than that of strategy dynamics (Radzvilavicius et al., 2021, 2019; Sasaki et al., 2017). After all games in a round are complete, each observer—specified for each monitoring system as described in Table 3.1—observes an

independent, random interaction of each donor (in the case of individual reputations) or a random interaction of a randomly selected donor in each group (in the case of stereotyped reputations). In the former, the observer evaluates each donor according to the social norm and the individual reputation of the recipient; in the latter, the observer applies the social norm to the stereotype of the recipient's group instead.

Social norms and probability of being assigned a good reputation. The four second-order norms considered in our model share two entries in the norm matrix: cooperating with a recipient who is good is considered good, as is defecting against a recipient who is in bad is considered bad. But the norms differ in (i) whether cooperating with a bad recipient is considered good and (ii) whether defecting against a bad recipient is considered good. Suppose that cooperating with a bad recipient yields a good standing with probability q_C and defecting against a bad individual yields a a good standing with probability q_D. Then the four norms can be parameterized as in Table 3.3.

Norm	q_C	q_D
Stern Judging	0	1
Simple Standing	1	1
Scoring	1	0
Shunning	0	0

Table 3.3: A parameterization of the four norms. Here q_C is the probability that cooperating with a bad recipient yields a good standing and q_D is the probability that defecting against a bad individual yields a a good standing.

Computing the equilibrium reputations involves keeping track of observations with different combinations of (a) observer view (does the observer view the recipient as good or bad?) and (b) donor intent (did the donor view the recipient as good (or bad) and therefore intend to cooperate (or defect)?). To facilitate this, we define the following

quantities:

$$\begin{aligned} P_{GC} &= (1-u_e)(1-u_a) + u_e u_a = \varepsilon \ , \\ P_{GD} &= u_a \ , \\ P_{BC} &= q_C(\varepsilon - u_a) + q_D(1-\varepsilon-u_a) + u_a \ , \\ P_{BD} &= q_D(1-2u_a) + u_a \ , \end{aligned} \tag{3.3}$$

where P_{XY} is the probability that a donor who intends to $Y \in$\{cooperate (C), defect (D)\} with a recipient viewed as $X \in$\{good, bad\} by the observer is assigned a good reputation (individual or stereotyped). For example, consider P_{GC}: a donor who intends to cooperate with a recipient who has a good individual reputation in the eyes of the observer can maintain a good individual reputation when the donor either (i) successfully cooperates (with probability $1-u_e$) and is correctly assigned a good individual reputation (with probability $1-u_a$), or (ii) erroneously defects (with probability u_e) and is erroneously assigned a good individual reputation (with probability u_a).

Individual reputations of cooperators and defectors. A cooperator (ALLC) gains a good individual reputation by either

- interacting with someone with a good individual reputation (probability $g^{\bullet,I}$), intending to cooperate, and successfully being assigned a good individual reputation (probability P_{GC}), or
- interacting with someone with a bad individual reputation (probability $1-g^{\bullet,I}$), intending to cooperate, and erroneously being assigned a good individual reputation (probability P_{BC}).

Thus, the average individual reputation for cooperators is given by

$$g_X^{I,I} = g_X^{J,I} = g^{\bullet,I} P_{GC} + (1 - g^{\bullet,I}) P_{BC} \,. \tag{3.4}$$

Similarly, a defector (ALLD) gains a good individual reputation by either

- interacting with someone with a good individual reputation (probability $g^{\bullet,I}$), intending to defect, and erroneously being assigned a good individual reputation (probability P_{GD}), or
- interacting with someone with a bad individual reputation (probability $1 - g^{\bullet,I}$), intending to defect, and successfully being assigned a good individual reputation (probability P_{BD}).

Thus, the average individual reputation for defectors is given by

$$g_Y^{I,I} = g_Y^{J,I} = g^{\bullet,I} P_{GD} + (1 - g^{\bullet,I}) P_{BD} \,. \tag{3.5}$$

Individual reputations of discriminators. Throughout the following, we assume, without loss of generality, that the observer is in group I, the donor in group J, and the recipient in group L.

A discriminator (pDISC) in group I with probability p of using stereotypes can gain a good individual reputation by

(i) using individual reputations (probability $1 - p$)

  (a) that are shared between the donor and the observer (i.e., individual reputations are public, or they are group-wise and donor and observer belong to the same group), and

* interacting with someone with a good individual reputation (probability $g^{\bullet,I}$), intending to cooperate, and successfully being assigned a good individual reputation (probability P_{GC}), or

* interacting with someone with a bad individual reputation (probability $1 - g^{\bullet,I}$), intending to defect, and successfully being assigned a good individual reputation (probability P_{BC}).

(b) that are not shared between the donor and the observer (i.e., individual reputations are private, or they are group-wise and donor and observer belong to different groups), and

* interacting with someone the donor (the focal discriminator) views as good (probability $g_i^{L,J}$) and the observer views as good (probability $g_i^{L,I}$), intending to cooperate, and being assigned a good individual reputation (probability P_{GC}), or

* interacting with someone the donor views as bad (probability $1 - g_i^{L,J}$) and the observer views as good (probability $g_i^{L,I}$), intending to defect, and being assigned a good individual reputation (probability P_{GD}).

* interacting with someone the donor views as good (probability $g_i^{L,J}$) and the observer views as bad (probability $1 - g_i^{L,I}$), intending to cooperate, and being assigned a good individual reputation (probability P_{BC}).

* interacting with someone the donor views as bad (probability $1 - g_i^{L,J}$) and the observer views as bad (probability $1 - g_i^{L,I}$), intending to defect, and being assigned a good individual reputation (probability P_{BD}).

(ii) using stereotyped reputations (probability p), and

– interacting with someone the donor (the focal discriminator) views as good (probability $g_S^{L,J}$) and the observer views as good (probability $g_i^{L,I}$), intending

to cooperate, and being assigned a good individual reputation (probability P_{GC}).

- interacting with someone the donor views as bad (probability $1 - g_S^{L,J}$) and the observer views as good (probability $g_i^{L,I}$), intended to defect, and being assigned a good individual reputation (probability P_{GD}).
- interacting with someone the donor views as good (probability $g_S^{L,J}$) and the observer views as bad (probability $1 - g_i^{L,I}$), intending to cooperate, and being assigned a good individual reputation (probability P_{BC}).
- interacting with someone the donor views as bad (probability $1 - g_S^{L,J}$) and the observer views as bad (probability $1 - g_i^{L,I}$), intending to defect, and being assigned a good individual reputation (probability P_{BD}).

Altogether, a pDISC in group I will have a good individual reputation

(i)(a) with probability $g_{public}^{J,I} = g^{\bullet,I} P_{GC} + (1 - g^{\bullet,I}) P_{BD}$, when using individual reputations that are shared,

(i)(b) with probability $g_{private}^{J,I} = g_{\alpha,1}^{J,I} P_{GC} + g_{\beta,1}^{J,I} P_{GD} + g_{\gamma,1}^{J,I} P_{BC} + g_{\delta,1}^{J,I} P_{BD}$, when using individual reputations that are not shared, and

(ii) with probability $g_{independent}^{J,I} = g_{\alpha,2}^{J,I} P_{GC} + g_{\beta,2}^{J,I} P_{GD} + g_{\gamma,2}^{J,I} P_{BC} + g_{\delta,2}^{J,I} P_{BD}$, when using stereotyped reputations,

where we define the following disagreement terms for convenience:

$$
\begin{aligned}
g^{J,I}_{\alpha,1} &= \sum_L \nu_L \sum_i f_i^L g_i^{L,I} g_i^{L,J} \\
g^{J,I}_{\beta,1} &= \sum_L \nu_L \sum_i f_i^L g_i^{L,I} (1 - g_i^{L,J}) = g^{\bullet,I} - g^{J,I}_{\alpha,1} \\
g^{J,I}_{\gamma,1} &= \sum_L \nu_L \sum_i f_i^L (1 - g_i^{L,I}) g_i^{L,J} = g^{\bullet,J} - g^{J,I}_{\alpha,1} \\
g^{J,I}_{\delta,1} &= \sum_L \nu_L \sum_i f_i^L (1 - g_i^{L,I})(1 - g_i^{L,J}) = 1 - g^{\bullet,I} - g^{\bullet,J} + g^{J,I}_{\alpha,1}
\end{aligned}
\tag{3.6}
$$

$$
\begin{aligned}
g^{J,I}_{\alpha,2} &= \sum_L \nu_L g_S^{L,J} \sum_i f_i^L g_i^{L,I} = \sum_L \nu_L g_S^{L,J} g^{L,I} \\
g^{J,I}_{\beta,2} &= \sum_L \nu_L (1 - g_S^{L,J}) \sum_i f_i^L g_i^{L,I} = \sum_L \nu_L (1 - g_S^{L,J}) g^{L,I} = g^{\bullet,I} - g^{J,I}_{\alpha,2} \\
g^{J,I}_{\gamma,2} &= \sum_L \nu_L g_S^{L,J} \sum_i f_i^L (1 - g_i^{L,I}) = \sum_L \nu_L g_S^{L,J} (1 - g^{L,I}) = g^{\star,J} - g^{J,I}_{\alpha,2} \\
g^{J,I}_{\delta,2} &= \sum_L \nu_L (1 - g_S^{L,J}) \sum_i f_i^L (1 - g_i^{L,I}) = \sum_L \nu_L (1 - g_S^{L,J})(1 - g^{L,I}) = 1 - g^{\bullet,I} - g^{\star,J} + g^{J,I}_{\alpha,2}
\end{aligned}
\tag{3.7}
$$

Putting these together, we obtain equations for the average individual reputation of each discriminator subtype:

$$
\begin{aligned}
g_Q^{J,I} &= (1 - p_Q) \left[(1 - A_{IJ}) \left(g^{J,I}_{private} \right) + A_{IJ} \left(g^{J,I}_{public} \right) \right] + p_Q \left[g^{J,I}_{independent} \right] , \\
g_R^{J,I} &= (1 - p_R) \left[(1 - A_{IJ}) \left(g^{J,I}_{private} \right) + A_{IJ} \left(g^{J,I}_{public} \right) \right] + p_R \left[g^{J,I}_{independent} \right] ,
\end{aligned}
\tag{3.8}
$$

where

$$
A_{IJ} = \begin{cases} 0 & \text{private individual reputations ,} \\ \delta_{IJ} & \text{group-wise individual reputations ,} \\ 1 & \text{public individual reputations .} \end{cases}
$$

Finally, we can write the average individual reputation of individuals in group J in the eyes of I as

$$\begin{aligned} g^{J,I} &= f_Q^J g_Q^{J,I} + \left(1 - f_Q^J\right) g_R^{J,I} \\ &= \left[f_Q^J \left(1 - p_Q\right) + \left(1 - f_Q^J\right) \left(1 - p_R\right)\right] \left[(1 - A_{IJ})(g_{private}^{J,I}) + A_{IJ}(g_{public}^{J,I})\right] \\ &\qquad + \left[f_Q^J p_Q + \left(1 - f_Q^J\right) p_R\right] \left[g_{independent}^{J,I}\right] \end{aligned} \tag{3.9}$$

where $f_Q^J \left(1 - p_Q\right) + \left(1 - f_Q^J\right) \left(1 - p_R\right)$ is the probability that a J individual uses an individual reputation in a given interaction.

Stereotyped reputations. By similar logic, a pDISC in group I will have a good stereotyped reputation

(iii) with probability $g_{S,independent}^{J,I} = g_{\alpha,3}^{J,I} P_{GC} + g_{\beta,3}^{J,I} P_{GD} + g_{\gamma,3}^{J,I} P_{BC} + g_{\delta,3}^{J,I} P_{BD}$, when using individual reputations,

(iv)(a) with probability $g_{S,public}^{J,I} = g^{\star,I} P_{GC} + (1 - g^{\star,I}) P_{BD}$, when using shared stereotyped reputations, and

(iv)(b) with probability $g_{S,private}^{J,I} = g_{\alpha,4}^{J,I} P_{GC} + g_{\beta,4}^{J,I} P_{GD} + g_{\gamma,4}^{J,I} P_{BC} + g_{\delta,4}^{J,I} P_{BD}$, when using private stereotyped reputations,

where we define the following disagreement terms for convenience:

$$\begin{aligned} g_{\alpha,3}^{J,I} &= \sum_L \nu_L g_S^{L,I} \sum_i f_i^L g_i^{L,J} = \sum_L \nu_L g_S^{L,I} g^{L,J} \\ g_{\beta,3}^{J,I} &= \sum_L \nu_L g_S^{L,I} \sum_i f_i^L (1 - g_i^{L,J}) = \sum_L \nu_L g_S^{L,I} (1 - g^{L,J}) = g^{\star,I} - g_{\alpha,3}^{J,I} \\ g_{\gamma,3}^{J,I} &= \sum_L \nu_L (1 - g_S^{L,I}) \sum_i f_i^L g_i^{L,J} = \sum_L \nu_L (1 - g_S^{L,I}) g^{L,J} = g^{\bullet,J} - g_{\alpha,3}^{J,I} \\ g_{\delta,3}^{J,I} &= \sum_L \nu_L (1 - g_S^{L,I}) \sum_i f_i^L (1 - g_i^{L,J}) = \sum_L \nu_L (1 - g_S^{L,I})(1 - g^{L,J}) = 1 - g^{\star,I} - g^{\bullet,J} + g_{\alpha,3}^{J,I}, \end{aligned} \tag{3.10}$$

$$
\begin{aligned}
g^{J,I}_{\alpha,4} &= \sum_L \nu_L g^{L,I}_S g^{L,J}_S \\
g^{J,I}_{\beta,4} &= \sum_L \nu_L g^{L,I}_S (1 - g^{L,J}_S) = g^{\star,I} - g^{J,I}_{\alpha,4} \\
g^{J,I}_{\gamma,4} &= \sum_L \nu_L (1 - g^{L,I}_S) g^{L,J}_S = g^{\star,J} - g^{J,I}_{\alpha,4} \\
g^{J,I}_{\delta,4} &= \sum_L \nu_L (1 - g^{L,I}_S)(1 - g^{L,J}_S) = 1 - g^{\star,I} - g^{\star,J} - g^{J,I}_{\alpha,4}.
\end{aligned}
\tag{3.11}
$$

Using these terms, we can write an equation for the average stereotyped reputation of group J in the eyes of I:

$$
\begin{aligned}
g^{J,I}_S &= f^J_Q g^{J,I}_{S,Q} + \left(1 - f^J_Q\right) g^{J,I}_{S,R} \\
&= \left[f^J_Q \left(1 - p_Q\right) + \left(1 - f^J_Q\right)\left(1 - p_R\right)\right] \left[g^{J,I}_{S,independent}\right] \\
&\quad + \left[f^J_Q p_Q + \left(1 - f^J_Q\right) p_R\right] \left[\left(1 - B_{IJ}\right)\left(g^{J,I}_{S,private}\right) + B_{IJ}\left(g^{J,I}_{S,public}\right)\right].
\end{aligned}
\tag{3.12}
$$

with

$$
B_{IJ} = \begin{cases} 0 & \text{private stereotyped reputations ,} \\ \delta_{IJ} & \text{group-wise stereotyped reputations ,} \\ 1 & \text{public stereotypes .} \end{cases}
$$

3.7.3 Pairwise invasibility

We now turn to the evolution of stereotype use. Our goal is to determine the level(s) of stereotyping that are evolutionarily attractive. To do so, we use the framework of adaptive dynamics (Geritz et al., 1998) and perform pairwise invasibility analysis in p—that is, we investigate which invaders p_QDISC (with stereotyping probability $0 < p_Q < 1$, denoted Z_Q) can invade a given resident population p_RDISC (with stereotyping propensity $0 < p_R < 1$, denoted Z_R).

Let f_Q^I and ${f_R}^I$ be the frequency of p_QDISC and p_RDISC individuals in group I, respectively. The replicator dynamics for $\dot{f}_Q^I$ is given by Eq. (3.1) with $i \in \{Z_Q, Z_R\}$. Then, the frequency of Z_Q individuals in the full population follows

$$\dot{f}_Q = \sum_I \nu_I \dot{f}_Q^I = \sum_I \nu_I f_Q^I \sum_J \nu_J \left(\Pi_Q^J - \bar{\Pi}^J\right) = f_Q \sum_J \nu_J \left(1 - f_Q^J\right) \left(\Pi_Q^J - \Pi_{Z_R}^J\right) , \tag{3.13}$$

where the last equality follows from the fact that $\bar{\Pi}^J = f_Q^J {\Pi_Q}^J + (1 - f_Q^J){\Pi_R}^J$.

To determine when p_QDISC can invade p_RDISC, we first compute the partial derivative of $\dot{f}_Q$ with respect to f_Q, evaluated at $f_Q = 0$. Noting that $\partial f_Q^I / \partial f_Q = (\partial f_Q / \partial f_Q^I)^{-1} = \nu_I^{-1}$ and that $f_Q = 0$ means $f_Q^I = 0$ for all I, we have

$$\begin{aligned}
\left.\frac{\partial \dot{f}_Q}{\partial f_Q}\right|_{f_Q=0} &= \left.\sum_J \nu_J \left(1 - f_Q^J - \nu_J^{-1} f_Q\right) \left({\Pi_Q}^J - {\Pi_R}^J\right)\right|_{f_Q=0} + f_Q \sum_J \nu_J \left(1 - f_Q^J\right) \left.\frac{\partial({\Pi_Q}^J - {\Pi_R}^J)}{\partial f_Q}\right|_{f_Q=0} \\
&= \left.\sum_J \nu_J \left({\Pi_Q}^J - {\Pi_R}^J\right)\right|_{f_Q=0}
\end{aligned}$$

Thus, p_QDISC will invade resident p_RDISC if and only if

$$\left.\frac{\partial \dot{f}_Q}{\partial f_Q}\right|_{f_Q=0} = \left.\sum_J \nu_J \left({\Pi_Q}^J - {\Pi_R}^J\right)\right|_{f_Q=0} > 0 . \tag{3.14}$$

As in Eq. (3.2), the average fitness of p_QDISC and p_RDISC in group I are, respectively,

$$\begin{aligned}
{\Pi_Q}^I = (1 - u_e) \Bigg[b \sum_J \nu_J \left\{ \left(f_Q^J (1 - p_Q) + (1 - f_Q^J)(1 - p_R)\right) g_Q^{I,J} + \left(f_Q^J p_Q + (1 - f_Q^J) p_R\right) g_S^{I,J} \right\} \\
- c \left((1 - p_Q) g^{\bullet,I} + p_Q g^{\star,I}\right) \Bigg] - \eta (1 - p_Q) , \\
{\Pi_R}^I = (1 - u_e) \Bigg[b \sum_J \nu_J \left\{ \left(f_Q^J (1 - p_Q) + (1 - f_Q^J)(1 - p_R)\right) g_R^{I,J} + \left(f_Q^J p_Q + (1 - f_Q^J) p_R\right) g_S^{I,J} \right\} \\
- c \left((1 - p_R) g^{\bullet,I} + p_R g^{\star,I}\right) \Bigg] - \eta (1 - p_R) .
\end{aligned}$$

Evaluating these at $f_Q = 0$ and substituting them into Eq. (3.14), we can express the condition for invasibility as

$$(1-u_e)\sum_I \nu_I \left[b \sum_J \nu_J (1-p_R)\left(g_Q^{I,J} - g_R^{I,J}\right) - c(p_Q - p_R)\left(-g^{\bullet,I} + g^{\star,I}\right) \right] \Bigg|_{f_Q=0} + \eta(p_Q - p_R) > 0 \,. \tag{3.15}$$

Equivalently, the critical benefit-to-cost ratio for invasibility is given by

$$\left(\frac{b}{c}\right)^* = \left. \frac{(p_Q - p_R)\left(\sum_I \nu_I \left(g^{\star,I} - g^{\bullet,I}\right) - \frac{\eta}{c(1-u_e)}\right)}{(1-p_R)\sum_I \sum_J \nu_I \nu_J \left(g_Q^{I,J} - g_R^{I,J}\right)} \right|_{f_Q=0} . \tag{3.16}$$

The direction of inequality (e.g., whether invasion is possible with b/c above or below this quantity) depends on the sign of the denominator. The denominator corresponds to the 'benefit' of switching from Z_R to Z_Q resulting from the increased probability that others cooperate with the focal individual (i.e., viewed as having a good individual reputation). The numerator corresponds to the 'cost' of this switch, given by (change in probability of using individual reputations) $\times$ (difference in probability that a donor sees a recipient as good when using stereotyped vs individual reputations—i.e., effective reduction in cost when using the former instead of the latter).

This threshold provides some intuition for when Z_Q can invade Z_R:

- If the 'benefit' of switching from Z_R to Z_Q is positive, then Z_Q can invade Z_R iff $(b/c) < (b/c)^*$, which requires $p_Q > p_R$. Invasibility is enhanced ($(b/c)^*$ is larger) when
  (a) invader Z_Q uses individual reputations much more frequently relative to Z_R (larger $p_Q - p_R$);
  (b) individual reputations are cheaper to access (smaller η); or
  (c) resident Z_R tends to use stereotypes more often (larger p_R).

- If the 'benefit' of switching from Z_R to Z_Q is negative, then Z_Q can invade Z_R iff $(b/c) > (b/c)^*$.

Note that the stereotype term $g_S^{I,J}$ does not appear explicitly in Eq. (3.15) or Eq. (3.16). This makes sense because switching strategies has no immediate impact on one's stereotype in the eyes of others—stereotypes are tied to group memberships, not strategies.

Finding equilibrium reputations at $f_Q = 0$. When evaluating the reputations at $f_Q = 0$, we can make several simplifications. First, the fraction of group J individuals with a good individual reputation in the eyes of I is

$$g^{J,I}\big|_{f_Q=0} = \sum_i f_i^J g_i^{J,I}\big|_{f_Q=0} = g_R^{J,I}\,, \tag{3.17}$$

which implies $g^{\bullet,I}\big|_{f_Q=0} = \sum_J \nu_J g^{J,I}\big|_{f_Q=0} = \sum_J \nu_J g_R^{J,I}$. Also, in Eq. (3.6), we have

$$g_{\alpha,1}^{J,I} = \sum_L \nu_L \sum_i f_i^L g_i^{L,I} g_i^{L,J} = \sum_L \nu_L g_R^{L,I} g_R^{L,J}\,. \tag{3.18}$$

This means that $g_R^{J,I}$ appears nowhere in the equations for individual (Eq. (3.9)) and stereotyped (Eq. (3.12)) reputations (which makes sense in the limit of negligibly few invaders Z_R). Hence, we only need to solve a system of eight equations for $(g_R^{J,I} =) g^{J,I}$ (Eq. (3.9)) and $g_S^{J,I}$ (Eq. (3.12)) with $I, J \in \{1, 2\}$ to find the equilibrium reputations. We then substitute these values into Eq. (3.15) to numerically identify parameter conditions under which Z_Q can invade Z_R.

3.7.4 Stochastic simulations

We also perform stochastic simulations in finite populations of $N = 50$ discriminators (pDISC). We assume that, initially, all individuals are characterized by a single stereotype

use propensity p, but allow for subsequent variation in p arising from the stochastic evolutionary dynamics. Both individual and stereotyped reputations are initialized randomly, i.e., each is either good or bad with equal probability. All individuals in a given simulation follow the same prescribed social norm (*Social norms*) and adhere to the prescribed monitoring systems for reputations (*Monitoring systems for reputations*).

In each generation, individuals undergo multiple rounds of games and reputation updates. A round consists of two steps. First, every individual interacts with everyone in the population (including herself), once as a donor and once as a recipient; whether the donor cooperates with the recipient depends on the donor's p, the recipient's individual or stereotyped reputation, and the execution error rate u_e (see *Games and behavioral strategies*). Second, all reputations are updated according to the monitoring systems, taking into account possible assessment errors occurring at rate u_a; for simplicity, we assume all updates within a round occur synchronously. We then repeat these steps over 2,500 rounds; that is, within each generation, every individual plays 2,500 games with $N = 50$ individuals in the population, for a total of 125,000 pairwise games. This ensures that reputations equilibrate sufficiently before strategy updating, approximating the time-scale separation assumed in the numerical treatment.

Strategy updating follows a pairwise comparison process. After all rounds in a generation are complete, we compute payoff π_i for each individual, with a fixed benefit b and cost c of cooperation as well as a fixed access cost η of using reputations (*Games and behavioral strategies*). Here we use per-generation average payoff (i.e., cumulative payoff across 50 games in a generation, averaged over generations), a scaled version of the per-game average payoff used in the numerical treatment (*Replicator dynamics*). Then, 5 random pairs of individuals are chosen from the population. Within each pair i and j, j adopts i's strategy with probability $1/\left(1+\exp\{-w(\pi_i-\pi_j)\}\right)$; parameter w

denotes the intensity of selection (Traulsen et al., 2007), which captures the impact of the game payoffs on relative success.

The population is also subject to recurring local mutations in p. In each generation, the stereotype use propensity p of a randomly selected individual changes by some Δp with probability $u_s = 10/N = 0.2$. Since p is a continuous parameter, the deviation Δp is sampled from a normal distribution with mean 0 and standard deviation 0.05.

Chapter 4

Emergence of hierarchy in networked endorsement dynamics

4.1 Notes

This chapter is adapted from:

> Mari Kawakatsu*, Philip S. Chodrow*, Nicole Eikmeier*, Daniel B. Larremore. Emergence of hierarchy in networked endorsement dynamics. *Proceedings of the National Academy of Sciences*, 118(16):e2015188118 (2021). doi:10.1073/pnas.2015188118

Author contributions. P. S. Chodrow, N. Eikmeier, and I contributed equally to this study as co-first authors. All authors designed the study and developed the model framework. P. S. Chodrow, N. Eikmeier, and I performed mathematical analysis. P. S. Chodrow and I conducted numerical and computational simulations, and P. S. Chodrow analyzed data. All authors drafted the paper and provided comments.

Prior presentations. I have given talks on this work at the following conferences and seminars:

- Tenth International Conference on Complex Systems (online; July 2020).
- International School and Conference on Network Science (NetSci 2020) (online; September 2020).
- Graduate Student Seminar, Program in Applied and Computational Mathematics, Princeton University (online; October 2020).
- Women in Networks Science Seminar, University of Washington (online; November 2020).
- Canadian Mathematical Society Winter Meeting (online; December 2021).

Acknowledgments. We thank Dakota S. Murray, Kate Wootton, V. P. S. Ritwika, and Rodrigo Migueles Ramírez for extremely helpful discussions during the early phase of this work. We are also grateful to the organizers of the Complex Networks Winter Workshop in Quebec City, QC, at which this work was conceived. M. Kawakatsu was supported by Army Research Office Grant W911NF-18-1-0325. P. S. Chodrow was supported by NSF Award 1122374. D. B. Larremore was supported by NSF Award SMA 1633791 and Air Force Office of Scientific Research Award FA9550-19-1-0329.

4.2 Abstract

Many social and biological systems are characterized by enduring hierarchies, including those organized around prestige in academia, dominance in animal groups, and desirability in online dating. Despite their ubiquity, the general mechanisms that explain the creation and endurance of such hierarchies are not well understood. We introduce a generative model for the dynamics of hierarchies using time-varying networks, in which new links are formed based on the preferences of nodes in the current network and old links are forgotten over time. The model produces a range of hierarchical structures, ranging from egalitarianism to bistable hierarchies, and we derive critical points that

separate these regimes in the limit of long system memory. Importantly, our model supports statistical inference, allowing for a principled comparison of generative mechanisms using data. We apply the model to study hierarchical structures in empirical data on hiring patterns among mathematicians, dominance relations among parakeets, and friendships among members of a fraternity, observing several persistent patterns as well as interpretable differences in the generative mechanisms favored by each. Our work contributes to the growing literature on statistically grounded models of time-varying networks.

4.3 Introduction

Hierarchies—stable sets of dominance relationships among individuals (Fushing et al., 2011; Hobson and DeDeo, 2015; Hobson et al., 2021)—structure many human and animal societies. Among animals, hierarchical rank may determine access to resources such as food, grooming, and reproduction (Holekamp and Strauss, 2016). Among humans, rank shapes the epistemic capital and employment prospects of researchers (Clauset et al., 2015; Morgan et al., 2018), susceptibility of adolescents to bullying (Garandeau et al., 2014), messaging patterns in online dating (Bruch and Newman, 2018), and influence in group decision-making (Cheng and Tracy, 2014).

A central question concerns how enduring hierarchies shape and are shaped by interactions between individuals. Empirical studies have indicated the presence of a winner effect: an individual who participates in a favorable interaction, such as winning a fight or receiving an endorsement, increases their likelihood of being favored in future interactions (Chase et al., 1994; Hogeweg and Hesper, 1983). Both theoretical work (Ben-Naim and Redner, 2005; Bonabeau et al., 1995, 1996a; Hemelrijk, 1999; Hickey and Davidsen, 2019; Miyaguchi et al., 2020; Pósfai and D'Souza, 2018; Sánchez-Tójar et al., 2018; Vehrencamp, 1983) and controlled experiments in humans (Salganik et al., 2006)

suggest that winner effects are sufficient (though not necessary) to form stable hierarchies. Mechanistic explanations of winner effects vary. A common approach postulates that each individual possesses an intrinsic strength, which may depend on factors such as size, skill, or aggression levels. For instance, physiological mechanisms, such as changes in hormone levels following confrontational interactions (Mehta and Prasad, 2015), can alter an individual's strength, causing the strong to get stronger.

However, intrinsic strengths are not necessary to produce winner effects. If a politician endorses a rival candidate, the latter does not become intrinsically more fit for office; instead, the endorsee builds support for their candidacy that may lead to future endorsements. The fame of the endorser is key: the better-known the endorser, the more valuable the endorsement. We refer to such prestige by proxy as transitive prestige. Since transitive prestige enables hierarchical rank to flow through interactions between individuals, networks provide a natural lens through which to study its role. Recent empirical studies have emphasized the networked nature of hierarchy in biological and social groups (Ball and Newman, 2013; Hobson and DeDeo, 2015; Hobson et al., 2021; Pinter-Wollman et al., 2014; Shizuka and McDonald, 2015). Several theoretical studies (Bardoscia et al., 2013; König and Tessone, 2011; König et al., 2014; Krause et al., 2013) have also investigated reinforcing hierarchy using time-varying network models called adaptive networks (Porter, 2020; Sayama et al., 2013). In this class of models, edges, representing interactions, evolve in response to node states and vice versa. Edges tend to accrue to important or highly central nodes, leading to self-reinforcing hierarchical network structures. Despite their recent uses, adaptive networks are often difficult to analyze analytically or compare to empirical data.

We present a flexible adaptive network model of social hierarchy that addresses these challenges. Winner effects in our model are driven entirely by social reinforcement, rather than intrinsic strengths. We allow arbitrary matrix functions to determine rank or

prestige of nodes in the network and introduce parameters governing the behavior of individuals in response to rank. A key feature of our model is that it is amenable both to mathematical analysis and to statistical inference. We analytically characterize a critical transition separating egalitarian and hierarchical model states for several choices of ranking function. We also explore hierarchical patterns in four biological and social datasets, using our model to perform principled selection between competing ranking methods in each dataset, and highlight persistent macroscopic patterns. We conclude with a discussion of potential model extensions and connections to recent work on centrality in temporal networks.

4.4 Modeling emergent hierarchy

In our adaptive network model, new directed edges are formed based on existing, node-based hierarchy, after which they decay over time. We conceptualize a directed edge $i \to j$ as an endorsement, in which i affirms that j is fit, prestigious, or otherwise of high quality. For example, endorsements could capture contests won by j over i, retweets of j by i, or comparisons in which a third party ranks j above i. We collect endorsements in a weighted directed network on n nodes summarized by its adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, where entry a_{ij} is the weighted number of interactions $i \to j$. The matrix $\mathbf{A}$ evolves in discrete time via the iteration

$$\mathbf{A}(t+1) = \lambda \mathbf{A}(t) + (1-\lambda)\mathbf{\Delta}(t) \; . \tag{4.1}$$

Here, the update matrix $\mathbf{\Delta}(t)$ contains new endorsements at time t. The memory parameter $\lambda \in [0, 1]$ represents the rate with which memories of old endorsements decay; the smaller the value of λ, the more quickly previous endorsements are forgotten.

The new endorsements in $\mathbf{\Delta}(t)$ depend on previous endorsements through a ranking of the n nodes, which we call the score vector (or simply score) $\mathbf{s} \in \mathbb{R}^n$. The score vector is the output of a score function $\sigma : \mathbf{A} \mapsto \mathbf{s} \in \mathbb{R}^n$, which may be any rule that assigns a real number to each node.

We consider three score functions chosen for analytical tractability and relevance in applications. Let $\mathbf{D}^{\text{in}}$ and $\mathbf{D}^{\text{out}}$ be diagonal matrices whose entries are the weighted in- and out-degrees of the network, i.e., $\mathbf{D}^{\text{in}}_{ii} = \sum_j \mathbf{A}_{ij}$ and $\mathbf{D}^{\text{out}}_{ii} = \sum_j \mathbf{A}_{ji}$. First, the Root-Degree score is the square root of the in-degree—the weighted number of endorsements—of each node i, defined as $s_i = \sqrt{\mathbf{D}^{\text{in}}_{ii}}$. The Root-Degree score function does not model transitive prestige, since only the number of endorsements is considered, not the prestige of the agents from which they come. Second, PageRank (Brin and Page, 1998) is a recursive notion of rank in which high-rank nodes are those whose endorsers are numerous, and themselves high rank. The foundational algorithm used by Google in ranking webpages, PageRank computes a value for each node interpretable as the proportion of time that a random surfer following the network of endorsements would spend on that node. We define PageRank score $\mathbf{s}$ as the PageRank vector of $\mathbf{A}^T$, which is the unique solution to the system

$$\left[\alpha_p \mathbf{A}^T (\mathbf{D}^{\text{out}})^{-1} + (1 - \alpha_p) n^{-1} \mathbf{e}\,\mathbf{e}^T\right] \mathbf{s} = \mathbf{s} \tag{4.2}$$

up to scalar multiplication. Here, $\alpha_p \in [0, 1]$ is the so-called teleportation parameter, for which we use the customary value $\alpha_p = 0.85$. We normalize the PageRank vector so that $\mathbf{e}^T \mathbf{s} = n$, where $\mathbf{e}$ is the vector of ones. Finally, SpringRank (De Bacco et al., 2018) is another recursive definition of rank in which endorsers are ranked one unit below endorsees, and disagreements are resolved using an analogy to a physical system of springs: the ranking of nodes minimizes the total energy of the system. Mathematically,

the SpringRank score $\mathbf{s}$ is the unique solution to the linear system (De Bacco et al., 2018)

$$\left[\mathbf{D}^{\text{in}} + \mathbf{D}^{\text{out}} - (\mathbf{A} + \mathbf{A}^T) + \alpha_s \mathbf{I}\right] \mathbf{s} = \left[\mathbf{D}^{\text{in}} - \mathbf{D}^{\text{out}}\right] \mathbf{e} \,, \tag{4.3}$$

with the identity matrix $\mathbf{I}$ and a regularization parameter $\alpha_s > 0$ which ensures the uniqueness of $\mathbf{s}$. Unlike the Root-Degree score, both PageRank and SpringRank scores model transitive prestige: the impact of an endorsement depends on the prestige of the endorser. These three score functions can all be interpreted as rankings or centrality measures, although this property is not required of score functions in our model.

Given score vector $\mathbf{s}$, new endorsements $\mathbf{\Delta}$ are chosen using a random utility model, a standard framework in discrete choice theory which has recently been applied in models of growing networks (Overgoor et al., 2019). At time step t, node i is selected uniformly at random. We suppose that endorsing j has utility $u_{ij}(\mathbf{s})$ for i, which depends on the current scores. In this work, we focus on utilities of the functional form

$$u_{ij}(\mathbf{s}) = \beta_1 s_j + \beta_2 (s_i - s_j)^2 \,, \tag{4.4}$$

where we generally assume that $\beta_1 > 0$ and $\beta_2 < 0$. The parameter β_1 captures a preference for prestige; a positive value of β_1 indicates a tendency to endorse others with high scores. The parameter β_2 captures a preference for proximity; a negative value of β_2 indicates a tendency to endorse others with scores relatively similar to their own. Many other choices of utility functions are possible; we prove a stability theorem for a large class of these functions in Appendix D.1.

In the random utility model, node i observes all possible utilities subject to noise. Traditionally, this noise is chosen to be Gumbel-distributed, in which case the probability

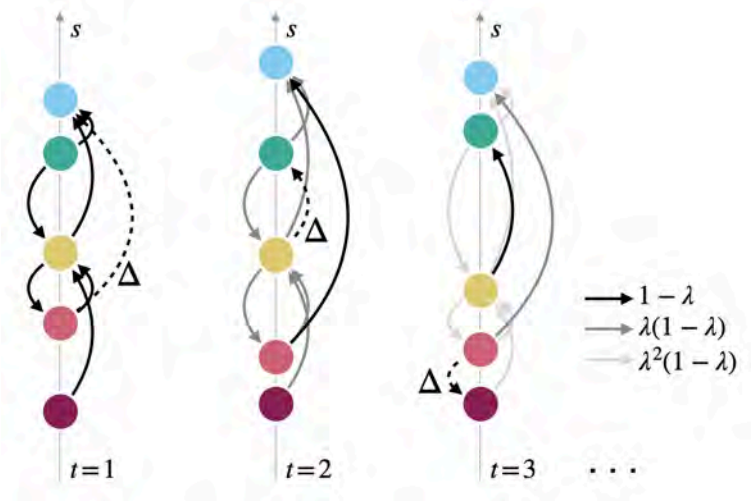

Figure 4.1: Schematic illustration of our model dynamics. Nodes are initialized at time $t = 1$ with a set of pre-existing endorsements logged in $\mathbf{A}$ (solid arrows) and the score $\mathbf{s} = \sigma(\mathbf{A})$ is computed (vertical axis). Then, a new edge logged in $\mathbf{\Delta}$ is added (dashed line). In the next time step $t = 2$, old interactions decay by a factor of λ (gray arrows). The new endorsement and decay of previous endorsements lead to an updated score function, which then informs the next time step.

that endorsing j yields the greatest utility is given by the multinomial logit (Train, 2009)

$$p_{ij}(\mathbf{s}) = \frac{e^{u_{ij}(\mathbf{s})}}{\sum_{j=1}^{n} e^{u_{ij}(\mathbf{s})}} . \tag{4.5}$$

We collect $m \in \mathbb{N}$ endorsements in an update matrix $\mathbf{\Delta}$, where the entry $\mathbf{\Delta}_{ij}$ gives the number of times that i endorses j in the time step. More complex random utility models can lead to more realistic structures in networks with a growing number of nodes (Gupta and Porter, 2020); we do not pursue these complications here because our model does not focus on network growth, and because these complications obstruct analytical insight.

Equation (4.5) can also be derived from an alternative model in which node i makes a randomized choice among n nodes to endorse. In this model, the option to endorse j is assigned a deterministically-observed weight proportional to $e^{u_{ij}(\mathbf{s})}$. In this case, β_1 and

β_2 signify inverse temperatures that tune the degree of randomness in this choice, with lower values corresponding to greater randomness. Although this alternative model—in which node i makes a noisy choice between deterministically-observed utilities—and the random utility model—in which node i makes a deterministic choice between noisily-observed utilities—are mathematically equivalent, the two formulations can lead to different interpretations of system behavior. In the case of institutional faculty hiring discussed below (see *Hierarchies in data*), the random utility model assumes that a hiring committee makes imperfect observations of the utilities of the institutions from which they could hire, and then deterministically chooses the highest of these imperfectly-observed qualities. In contrast, the alternative framework assumes that the committee makes a perfect observation of the utilities, but then chooses among them with some degree of randomness, which may reflect dissension on the hiring committee, search-specific priorities, or other factors.

Equations (4.1) and (4.5) capture key features of our model. First, the dynamics in Eq. (4.1) imply that past interactions decay geometrically at rate λ. This global, gradual decay contrasts with another rank-based relinking model in which single edges fully disappear within each time step (König and Tessone, 2011). Second, Eq. (4.5) implies that the likelihood of a node being endorsed at a given time step depends only on the distribution of previous endorsements and not on intrinsic strength or desirability. Those who receive more endorsements and therefore obtain higher scores are more likely to be endorsed in the future—a mechanistic instantiation of winner effects via social reinforcement.

Figure 4.1 schematically illustrates model dynamics with $m = 1$ endorsement per time step. At time $t = 1$, the model is initialized with a small number of endorsements logged in $\mathbf{A}$. The score function takes $\mathbf{A}$ as an input and outputs the score vector $\mathbf{s}$, which, in turn, determines a new interaction according to Eq. (4.5). Logged in $\mathbf{\Delta}$, this new

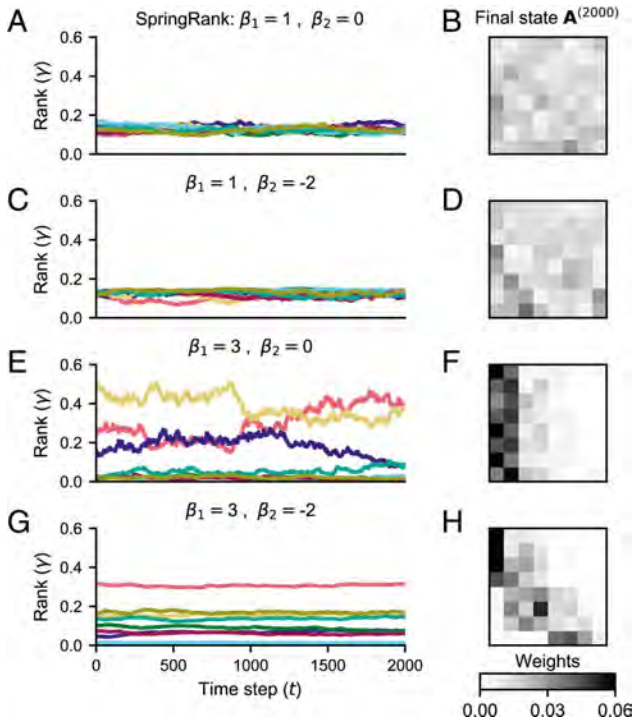

Figure 4.2: Representative dynamics of the proposed model. Each column shows a population of $n = 8$ nodes simulated for 2,000 time steps using the SpringRank score function with $m = 1$ update per time step, varying the preference parameters β_1 and β_2. A, C, E, and G show the simulated rank vector γ over time; different colors track the ranks of different nodes. B, D, F, and H show the adjacency matrix $\mathbf{A}$ at time step $t = 2000$ for the corresponding parameter combinations. See Fig. D.1 for additional examples with SpringRank; Fig. D.2 for examples PageRank; and Fig. D.3 for examples with Root-Degree. See Fig. D.4 for the dependence of rank variance on β_1 and β_2 jointly. Parameters: $\lambda = 0.995$, $\alpha_s = 10^{-8}$.

interaction is weighted by $1 - \lambda$ and added to the previous endorsements, which are discounted by λ. This process repeats over time with new endorsements gradually replacing old ones in the system's memory, sequentially updating the score vector $\mathbf{s}$. Figure 4.1 also depicts in stylized fashion the operation of both a winner effect ($\beta_1 > 0$), in which endorsements tend to flow in the direction of increasing score, and a proximity effect ($\beta_2 < 0$), in which endorsements tend to flow between nodes of similar scores. The net effect is that most endorsements are "short hops" up the hierarchy. As we will discuss, this is a common pattern in empirical data.

Despite its simplicity, the model displays a wide range of behaviors. To observe them, we define a rank vector $\boldsymbol{\gamma}$, whose jth entry $\gamma_j = n^{-1} \sum_i p_{ij}$ gives the likelihood that a new endorsement flows to j. We say that the system state is egalitarian when all ranks γ_j are equal and hierarchical otherwise. Figure 4.2 illustrates representative behaviors when the SpringRank score is used. When β_1 is relatively small, winner effects are overtaken by noise, and the system settles into an approximately egalitarian state (Fig. 4.2A and B). When β_1 is relatively large, persistent hierarchies emerge (Fig. 4.2C–F). Moreover, the distribution and stability of ranks depend on the strength of proximity effects, modeled by the quadratic term in the utilities. For $\beta_2 = 0$ (no proximity preference), a single node garners more than half of endorsements in a hierarchy with significant fluctuations (Fig. 4.2C and D). Adding a proximity preference leads to a marginally more equitable hierarchy with ranks that are nearly constant in time (Fig. 4.2E and F).

4.5 The long-memory limit

The behavior observed in Fig. 4.2 suggests the presence of qualitatively distinct regimes depending on prestige preference β_1. For small β_1 (Fig. 4.2A), the winner effect is weak, and approximate egalitarianism prevails. For larger β_1, a stronger winner effect enforces a stable hierarchy. We characterize the boundary between these regimes analytically in the long-memory limit $\lambda \to 1$ by defining a function $\mathbf{f}$, which is analogous to a deterministic time-derivative for the dynamics of our discrete-time stochastic process. Let

$$\mathbf{f}(\mathbf{s}, \mathbf{A}) = \lim_{\lambda \to 1} \frac{\mathbb{E}\left[\sigma\left(\lambda \mathbf{A} + (1-\lambda)\,\boldsymbol{\Delta}\right)\right] - \mathbf{s}}{1-\lambda} \tag{4.6}$$

where the expectation is taken with respect to $\boldsymbol{\Delta}$. If $\mathbf{f}(\mathbf{s}, \mathbf{A}) = 0$ for all $\mathbf{A}$, the score vector $\mathbf{s}$ is a fixed point of the model dynamics in expectation. Our choices of Root-Degree, PageRank, and SpringRank score functions admit closed-form expressions for $\mathbf{f}$,

allowing us to analytically derive the conditions for the stability of egalitarianism in the limit of long memory.

Theorem 4.1. *For each of the Root-Degree, PageRank, and SpringRank score functions,* $\mathbf{f}$ *has a unique egalitarian root. This root is linearly stable if and only if* $\beta_1 < \beta_1^c$*, where*

$$\beta_1^c = \begin{cases} 2\sqrt{\dfrac{n}{m}} & \textit{Root-Degree}, \\ 1/\alpha_p & \textit{PageRank}, \\ 2 + \alpha_s \dfrac{n}{m} & \textit{SpringRank}. \end{cases}$$

In Appendix D.1, we prove Theorem 4.1, as well as a generalization to arbitrary smooth utility functions. In each case, the proof of uniqueness exploits the algebraic structure of the score function, and the critical value β_1^c is obtained via the linearization of $\mathbf{f}$ about the egalitarian state. Interestingly, only β_1 plays a role in the stability of the egalitarian root. While proximity preference β_2 does not determine where the hierarchical regime begins, it does influence the structure of and the transient dynamics toward nonegalitarian equilibria (Fig. 4.2E and G).

Figure 4.3 illustrates the destabilization of egalitarianism predicted by Theorem 4.1 in the case of $n = 8$ nodes. Although not required by Theorem 4.1, we fix $\beta_2 = 0$ for simplicity. Curves show fixed points of the model dynamics in the long-memory limit. We show only fixed points in which nodes separate into two groups, each of which have identical rank. For $\beta_1 < \beta_1^c$, the egalitarian regime is stable and the long-run state deviates from egalitarianism only slightly. For $\beta_1 > \beta_1^c$, in contrast, the long-run state switches to an inegalitarian, stable fixed point.

In the Root-Degree and PageRank models, there is a single stable inegalitarian equilibrium with one node absorbing nearly all endorsements (Fig. 4.3A and B). Interestingly, there is a bistable regime in which both egalitarian and inegalitarian states

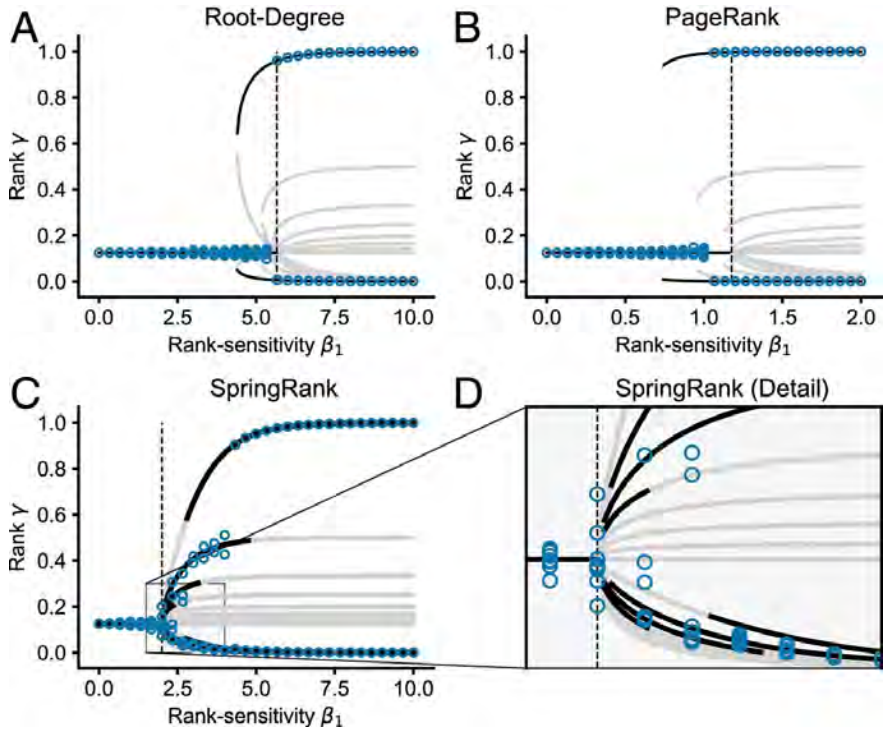

Figure 4.3: Bifurcations in models with Root-Degree (A), PageRank (B), and SpringRank (C and D) score functions with $\beta_2 = 0$ and $m = 1$ update per time step. Points give the value of the rank vector γ averaged over the final 500 time-steps of a 5×10^4-step simulation with $n = 8$ nodes, memory parameter $\lambda = 0.9995$, and varying β_1 specified by the horizontal axis. Solid curves show stationary points of the long-memory dynamics obtained by numerically solving the equation $\mathbf{f}(\mathbf{s}, \mathbf{A}) = \mathbf{0}$, subject to the restriction that nodes separate into two groups with identical ranks in each. Black curves are linearly stable, while gray curves are unstable. Stability was determined by studying the spectrum of the Jacobian matrix of $\mathbf{f}$. Vertical lines give the critical value β_1^c at which the egalitarian solution becomes linearly unstable according to Theorem 4.1. Parameters: $\alpha_p = 0.85$, $\alpha_s = 10^{-8}$.

are attracting. Whether the system converges to one or the other depends on initial conditions. The SpringRank model displays qualitatively distinct behavior (Fig. 4.3C and D). Past β_1^c, we observe staggered multistable regimes. As β_1 increases, equilibria with multiple elite (i.e., highly ranked) nodes become sequentially unstable until eventually only a single elite node remains. The long-term behavior of the system again depends on initial conditions, but now there are many more possible stable states. This behavior would seem to make the SpringRank score function especially appropriate for modeling

empirical systems with multiple distinct hierarchical regimes and sensitivity to initial conditions, an intuition which we confirm empirically in the following section.

4.6 Hierarchies in data

In addition to being amenable to analytical treatment, our model has a tractable likelihood function, described in Appendix D.1. This allows us to study hierarchical structures in empirical data using principled statistical inference. The likelihood function not only supports maximum-likelihood parameter estimates of λ, β_1, and β_2, but also enables direct comparisons of different score functions in a statistically rigorous framework: score functions with higher likelihoods provide more predictive low-dimensional summaries of observed interactions. This, in turn, allows us to explore the relative value of competing mechanistic explanations of observed data.

Several mathematical features of the model facilitate the exploration of real data. First, the predictive distribution Eq. (4.5) is in the linear exponential family, making the estimation of β a convex optimization with a unique solution. Second, the estimation problem in $\hat{\lambda}$ is, in general, nonconvex, but can be tractably solved via first-order optimization methods with multiple starting points. Finally, while model likelihoods evaluated on training data may, in principle, be inflated due to overfitting, our model uses only three parameters to fit hundreds or thousands of observations, suggesting that overfitting is not a major concern.

We conducted a comparative study of model behavior on four datasets: an academic exchange network in math, two networks of parakeet interactions, and a network of friendships among members of a fraternity. The Math PhD Exchange dataset is extracted from The Mathematics Genealogy Project (North Dakota State University Department of Mathematics, 2003; Taylor et al., 2017a,b). Nodes are universities. An interaction $i \to j$ at time t occurs when a mathematician who received their degree from university j at time

t supervises one or more PhD theses at university i. This event is a proxy for university i hiring a graduate from university j at a time near t. We view this as an endorsement by j that graduates of i are of high quality (Clauset et al., 2015). We restricted our analysis to the activity of the 70 institutions that placed the most graduates between 1960 and 2000. Doing so helped to avoid singularities produced by institutions with no placements early in the time period and to minimize temporal boundary effects associated with the beginning and end of data collection.

The two Parakeet datasets (Hobson and DeDeo, 2015, 2016) record aggression events in two distinct groups of birds studied over four observation quarters (weeks). An interaction $i \to j$ at time t occurs when parakeet i loses a fight to parakeet j in period t. Since there are just four observation periods, estimates of the memory parameter λ should be approached with caution.

Lastly, the Newcomb Fraternity dataset was collected by the authors of refs. (Newcomb, 1961; Nordlie, 1958) and accessed via the KONECT network database (Batagelj and Mrvar, 2007; Kunegis, 2013). The dataset documents friendships among members of a fraternity at the University of Michigan. Each week during a fall semester, excluding a week for fall break, each of 17 cohabiting brothers ranked every other brother according to friendship preference, with ranks 1 and 16 referring to that brother's most and least preferred peers, respectively. An endorsement $i \to j$ is logged when brother i ranks j among his top $k = 5$ peers (small changes to k did not significantly alter the results). While friendship is often viewed as a symmetric relationship, expressed friendship preferences may be asymmetric (Carley and Krackhardt, 1996).

We studied these data using the Root-Degree, PageRank, and SpringRank score functions. Table 4.1 summarizes our results, including parameter estimates, SEs (obtained by inverting the numerically calculated Fisher information matrix), and optimized log-likelihoods for each combination of score and dataset. Several features

		Root-Degree	PageRank	SpringRank
Math PhD	$\hat{\lambda}$	0.87 (0.01)	0.96 (0.01)	0.91 (0.01)
Exchange	$\hat{\beta}_1$	1.28 (0.02)	0.74 (0.01)	2.99 (0.04)
(N = 6,019)	$\hat{\beta}_2$	-0.18 (0.01)	-0.07 (0.00)	-1.12 (0.04)
	$\mathcal{L}$	**-14,379**	-15,001	-14,927
Parakeets (G1)	$\hat{\lambda}$	0.97 (0.08)	0.59 (0.08)	0.67 (0.14)
(N = 838)	$\hat{\beta}_1$	0.84 (0.05)	1.82 (0.08)	3.03 (0.16)
	$\hat{\beta}_2$	-0.12 (0.01)	-0.50 (0.03)	-1.74 (0.12)
	$\mathcal{L}$	-1,106	-1,053	**-964**
Parakeets (G2)	$\hat{\lambda}$	0.42 (0.07)	0.13 (0.03)	0.40 (0.06)
(N = 961)	$\hat{\beta}_1$	0.62 (0.03)	0.82 (0.04)	2.86 (0.14)
	$\hat{\beta}_2$	-0.06 (0.01)	-0.12 (0.01)	-1.46 (0.12)
	$\mathcal{L}$	-975	-1029	**-924**
Newcomb	$\hat{\lambda}$	0.56 (0.13)	0.81 (0.19)	0.71 (0.14)
Fraternity	$\hat{\beta}_1$	0.95 (0.05)	1.21 (0.07)	2.33 (0.14)
(N = 1,428)	$\hat{\beta}_2$	-0.08 (0.03)	-0.25 (0.05)	-0.86 (0.16)
	$\mathcal{L}$	-1,850	-1,865	**-1,841**

Table 4.1: Parameter estimates and likelihood scores using each of three score functions for the four data sets described in the main text. Parenthetical values are standard errors for each parameter estimate. For each data set, the largest log-likelihood $\mathcal{L}$ is indicated in **bold**. All parameter estimates are statistically distinct from zero at 95% confidence. N gives the total number of interactions in the data. See Fig. D.5 for simulated trajectories with the inferred parameters.

stand out. In all four datasets and across all three score functions, we find $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$. This suggests a persistent pattern in time-dependent hierarchies: while endorsements do flow upward ($\hat{\beta}_1 > 0$), nodes are more likely to endorse those close to them in rank ($\hat{\beta}_2 < 0$). Endorsements tend to flow a few rungs up the laddernot directly to the top. The reasons for this pattern likely vary across datasets. In the Math PhD Exchange, this may indicate that low-ranked schools struggle to recruit graduates of high-ranked ones due to a limited supply of elite candidates. In parakeet populations, proximal aggression may facilitate inference of dominance hierarchies through transitive inference (Hobson and DeDeo, 2015). In Newcomb's Fraternity, we postulate that

implicit social norms may encourage friendships between those of similar standing. Similar results have been reported in static social-network data among adolescents (Ball and Newman, 2013). Thus, while we do not attribute this pattern in the parameter estimates to a universal mechanism, we suggest its persistence as an interesting observation worthy of future study.

Because different score functions capture distinct qualitative features of the data, quantitative comparisons yield insights into the generating mechanisms at work. In general, parameters from models using differing score functions should not be directly compared, since these parameters are sensitive to the scale of the score vector. However, we can compare models on the basis of their likelihoods. In the Math PhD Exchange, the Root-Degree model was strongly favored over either SpringRank or PageRank. In the context of this dataset, the Root-Degree score is a measure of faculty production: a school that places more candidates has a higher score, regardless of the prestige of the institutions at which the candidates land. The strong fit from the Root-Degree score is consistent with previous findings that raw faculty production plays a major role in structuring the hierarchy of academic hiring within computer science, business, and history (Clauset et al., 2015). But as Clauset et al. (2015) note, transitive prestige also plays an important role. It would be of significant interest to extend our study to include multiple score functions, enabling an inferential analysis of the relative roles of production and transitive prestige.

In contrast, the SpringRank score was favored by large margins in both Parakeet datasets and by a smaller margin in the Newcomb Fraternity dataset, suggesting that transitive prestige plays a more prominent role. Among parakeets, it may matter not only how many confrontations one wins, but also against whom, with victories over high-ranking birds counting more toward one's own prestige. This finding is consistent with those of Hobson and DeDeo (2015), which found, using different methodology, that

parakeet behavior suggests the ability to draw sophisticated, transitive inferences about location in the hierarchy. Similarly, in Newcomb's Fraternity, friendships with highly ranked brothers may confer greater prestige than those with lower-ranked ones.

In addition to the likelihoods, we can also compare the memory estimate $\hat{\lambda}$ across models and datasets. Since the model assumes that the impact of past endorsements decays at rate λ, the quantity $t_{1/2} = -\log(2)/\log(\hat{\lambda})$ represents the half-life of system information according to the inferred dynamics, in units of observation periods. In the Math PhD data, the favored Root-Degree score gave a half-life of $t_{1/2} \approx 5$ years. In the Parakeets data, the half-life estimated under SpringRank is $t_{1/2} \approx 1.7$ weeks for the first group and $t_{1/2} \approx 0.8$ weeks for the second. The small number of observation periods implies that these estimates should be approached with caution. Finally, in the Newcomb Fraternity data, the SpringRank half-life was $t_{1/2} \approx 2$ weeks. This suggests that the friendships in this dataset evolved on timescales much shorter than the full semester. This likely reflects the fact that the brothers did not know each other prior to data collection, requiring them to form their social relationships from scratch. An important caveat in interpreting these estimated half-lives is that the indirect influence of an interaction may extend far beyond its direct influence. In the Math PhD data, for instance, while the half-life indicates that only a quarter of hiring events will be directly "remembered" in the system after a decade, those events will have influenced 10 cycles of hiring, which may further reinforce the patterns established by the earlier events.

As described in Theorem D.1, in the long-memory limit, our model has distinct egalitarian and hierarchical regimes, separated by a critical value β_1^c. The model's estimate of β_1 allows us to roughly locate empirical systems within these regimes. There are two necessary points of caution. First, when the estimate $\hat{\lambda}$ is far from the idealized long-memory limit, hierarchical and egalitarian regimes may not be sharply distinguished. Second, in the Math PhD and Parakeet data, the number of updates m

		Root-Degree	PageRank	SpringRank
Math PhD	β_1^c	**1.36**	1.18	2.00
Exchange	$\hat{\beta}_1$	**1.28**$_*$ (0.02)	0.74$_*$ (0.01)	2.99* (0.04)
Parakeets (G1)	β_1^c	0.55	1.18	**2.00**
	$\hat{\beta}_1$	0.84* (0.05)	1.82* (0.08)	**3.03*** (0.16)
Parakeets (G2)	β_1^c	0.49	1.18	**2.00**
	$\hat{\beta}_1$	0.62* (0.03)	0.82$_*$ (0.04)	**2.86*** (0.14)
Newcomb	β_1^c	0.89	1.18	**2.00**
Fraternity	$\hat{\beta}_1$	0.95 (0.05)	1.21 (0.07)	**2.33*** (0.14)

Table 4.2: Estimates of β_1 (identical to those in Table 4.1) compared to the mean critical value β_1^c for each system. β_1^c is calculated as in Theorem 4.1, using as m the mean number of interactions per time-step in the observed data. As in Table 4.1, the parameters corresponding to the highest log-likelihood are shown in **bold**. Estimates shown with an upper asterisk (*) exceed the approximate critical value by two standard errors, while estimates shown with a lower asterisk ($_*$) are smaller than the approximate critical value by two standard errors. See Fig. D.5 for simulated trajectories using the inferred parameters.

varies between time steps. Here, a reasonable approximation is to use the average number of updates $\bar{m}$ per time step. Using this average and Theorem D.1, we computed an approximate long-memory critical value β_1^c for each empirical system.

Comparing the data-derived preference estimates $\hat{\beta}_1$ to the approximate critical values β_1^c reveals that all four empirical systems are in or near the hierarchical regime (Table 4.2). The Root-Degree estimates of β_1 tend to be very close to the approximate critical point. For the Math PhD data, in which Root-Degree is the preferred model, the estimate of β_1 is slightly, but statistically significantly, below the critical value. In each of the other three datasets, the estimate is slightly above the critical value, and significantly so in the two Parakeet groups. Given the presence of a bistable regime in the Root-Degree model (Fig. 4.3A), the estimate of β_1 for the Math PhD data is consistent with persistent hierarchy, despite the fact that the estimate falls slightly below the critical threshold. Indeed, simulations with the inferred parameters produce persistent hierarchical structure similar to that observed in the data (Fig. D.5). The PageRank

estimates behave similarly to Root-Degree, although the finding in Parakeets (G2) is reversed. The presence of a bistable regime in the PageRank model (Fig. 4.3B) indicates that these findings are consistent with persistent hierarchy in any of these datasets (see Fig. D.5 for simulated dynamics). Finally, in the SpringRank model, which obtains the highest likelihood for both Parakeet datasets and the Newcomb Fraternity dataset, the estimated values of β_1 significantly exceed the estimated critical values and tend to lie in or near the range $[2, 3]$. In summary, all three models suggest that the system corresponding to each dataset is in or near the regime of self-reinforcing hierarchy.

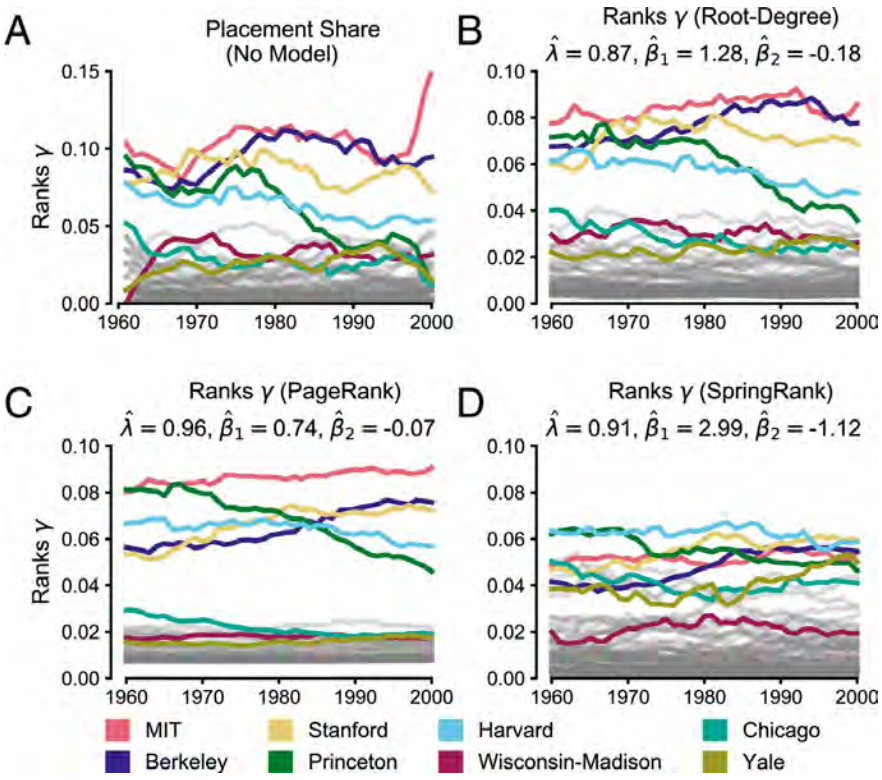

Figure 4.4: Visualization of evolving ranking functions in the Math PhD Exchange. (A) Fraction of all placements (number of graduates hired) from each school, shown as a moving average with bin-width 8 years for visualization purposes. (B) Inferred rank vector γ as a function of time using the Root-Degree score function. (C and D) As in B, with PageRank and SpringRank score functions, respectively. Parameters for B–D are shown in the first section of Table 4.1.

Our model also assigns interpretable, time-dependent ranks to empirical data (Fig. 4.4). For the Math PhD Exchange network, for example, the raw placement share

(Fig. 4.4A) and Root-Degree model (Fig. 4.4B) show strong qualitative agreement, with institutions that place the most candidates occupying higher ranks. Due to the relatively large estimates $\hat{\lambda}$, both the Root-Degree and PageRank models (Fig. 4.2B and C) produce smoother rank trajectories than the purely descriptive placement share with 8-year rolling average. In contrast, the SpringRank score generates qualitatively different trajectories that are less sensitive to raw volume (Fig. 4.4D). For instance, SpringRank places Harvard at the top over most of the time period, while the other scores prefer MIT. This difference reflects SpringRank's sensitivity to where Harvard's graduates were placed, a consideration that Root-Degree entirely ignores. Similarly, SpringRank places Chicago and Yale noticeably higher than Wisconsin-Madison, despite all three having similar numbers of placements.

4.7 Discussion

We have proposed a simple and flexible model of persistent hierarchy as an emergent feature of networked endorsements with feedback. When the preference for high status exceeds a critical value, egalitarian states destabilize, and hierarchies emerge. The location of this transition depends on the structure of the score function and of the node's preferences. Our findings emphasize that winner effects do not require internal, rank-enhancing feedback mechanisms. Social reinforcement through prestige preference is sufficient to generate social hierarchies.

Crucially, our model has a tractable likelihood function, supporting principled statistical inference of parametersfor both preferences and memory strengthfrom empirical data. In the four datasets analyzed, we found that links are typically formed in alignment with the hierarchy ($\hat{\beta}_1 > 0$), but that they are preferentially created to other nodes with similar ranks ($\hat{\beta}_2 < 0$). The likelihood also opens the door to model selection to determine relevant score functions. We found that networked ranking methods that

capture transferable prestige are preferred over nonnetworked methods in some, but not all, systems. Due to its flexibility, our framework can be applied to additional datasets, score functions, and/or preference models to test the generality of these empirical observations.

There are limitations to our approach. First, we specified a fixed parametric form for the utilities with Eq. (4.4) and Gumbel-distributed noise with Eq. (4.5). Other choices may be more justified in particular applications, ideally informed by domain-specific considerations. Importantly, our inferential framework allows for quantitative evaluation and comparison of these choices. Taking advantage of this, future work could systematically explore the most appropriate functional forms in systems from diverse scientific domains. Second, our model assumes that all nodes use identical preference parameters β_1, β_2 and score vector s when computing utilities. The latter is an especially strong assumption, since it requires each node to have global knowledge of the endorsement network, or at least of the score vector. This is unlikely to be true in real systems and should be regarded as a modeling device. Future work, along the lines of Hobson and DeDeo (2015), could explore the interplay between the cognitive capabilities of individuals represented by nodes and the information available to them in the formation of social hierarchies.

Our model points to several other avenues for further work. A crucial step would be to extend extant network-based models (König and Tessone, 2011; König et al., 2014; Krause et al., 2013) so that their parameters could be statistically learned from data. This would enable comparative validation of different modeling frameworks. Studies of the relationship between measures of time-dependent centralities (Liao et al., 2017; Taylor et al., 2017b, 2019) and dynamic models of hierarchy would also be valuable. In particular, the theory of time-dependent centralities faces an important methodological issue: different reasonable ranking methods can yield directionally different orderings of

nodes when applied to the same dataset (Mariani and Lü, 2020). Their performance on external validation tasks, such as the prediction of central nodes in spreading processes (Lü et al., 2016), may also vary significantly. Because the theories of centrality and generative networks have evolved largely separately, evaluating the suitability of a centrality metric for a given dynamic system can be difficult. Our inferential approach offers a candidate validation task to overcome this challenge: good centrality metrics are those that most effectively predict the future evolution of the system. This approach enables us to not only compare different score and utility functions in a principled manner, but also explore their relative importance in observed networks. For instance, one could study the relative influence of degree-based and SpringRank scores by incorporating both into our model and then analyzing their distinct coefficients. Further work in this direction could reveal how different forms of centrality combine to govern the evolution of interaction networks. We anticipate that a fruitful dialogue between centrality theory and generative models of time-varying networks will deepen our understanding of the feedback mechanism between local interactions and hierarchical structures.

Data availability

A repository containing all data used in our analyses, model implementation, and figure-generation scripts is available at GitHub (https://github.com/PhilChodrow/prestige_reinforcement). Raw data are available at https://sites.google.com/site/danetaylorresearch/data (Math PhD Exchange), https://datadryad.org/stash/dataset/doi:10.5061/dryad.p56q7 (Parakeets G1 and G2), and http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm#newfrat (Newcomb Fraternity).

Appendix A

Supplementary materials for Chapter 1

A.1 Supplementary analyses

A.1.1 Analytical treatment of the model

In addition to simulating the fixed threshold model (*Theoretical model*), we derive analytical predictions for its long-term behavior.

Recall that the model considers a colony of N individuals performing M tasks; we assume that there are two tasks ($M = 2$; see *Theoretical model*). Individuals can be of one of two types, X or Y. To analytically study how individual behavior depends on the ratio of the types, we define f and $1 - f$ to be the fractions of the colony consisting of individuals of type X and Y, respectively.

The model assumes that individual i's internal threshold θ_{ij} is drawn from a normal distribution with mean μ_j and normalized standard deviation σ_j. For our analytical analysis, we assume that $\sigma_j = 0$ for all tasks. In other words, the type- and task-specific thresholds are assumed to be given by the constant parameters, μ_j^X and μ_j^Y. Under this assumption, the probabilities $P_{ij,t}^X$ and $P_{ij,t}^Y$ that inactive individuals i of types X and Y

begin to perform task j at time t are, respectively,

$$P_{j,t}^{X}\left(s_{j,t}\right)=\frac{s_{j,t}^{\eta}}{s_{j,t}^{\eta}+\left(\mu_{j}^{X}\right)^{\eta}}, \qquad P_{j,t}^{Y}\left(s_{j,t}\right)=\frac{s_{j,t}^{\eta}}{s_{j,t}^{\eta}+\left(\mu_{j}^{Y}\right)^{\eta}}. \tag{A.1}$$

Because we assume that there are two tasks ($M = 2$), the numbers of X and Y individuals performing task j at time $t + 1$ are governed by the following equations, in which j' denotes the other task:

$$\begin{aligned}
n_{j,t+1}^{X}-n_{j,t}^{X} &= \frac{1}{2}\left[P_{j,t}^{X}(s_{j,t})+\left(1-P_{j',t}^{X}\left(s_{j',t}\right)\right)P_{j,t}^{X}\left(s_{j,t}\right)\right]\left(fN-\left(n_{j,t}^{X}+n_{j',t}^{X}\right)\right)-\tau n_{j,t}^{X} \\
n_{j,t+1}^{Y}-n_{j,t}^{Y} &= \frac{1}{2}\left[P_{j,t}^{Y}(s_{j,t})+(1-P_{j',t}^{Y}(s_{j',t}))P_{j,t}^{Y}(s_{j,t})\right]\left((1-f)N-\left(n_{j,t}^{Y}+n_{j',t}^{Y}\right)\right)-\tau n_{j,t}^{Y},
\end{aligned}$$

where $n_{j,t}^{X}$ and $n_{j,t}^{Y}$ are the numbers of X and Y individuals performing task j at time t, respectively, and τ is the probability of quitting a task. The sums in the larger parentheses represent the pool of individuals who could possibly initiate task j—that is, the total number of inactive individuals. The sums in square brackets then capture the possible ways in which these inactive individuals can initiate task j: they can either encounter the stimulus for task j immediately and begin performing that task, or they can first encounter the stimulus for the other task j', not perform that task, subsequently encounter the stimulus for task j, and begin performing task j. Lastly, recall that the dynamics of stimulus $s_{j,t}$ associated with task j is governed by Eq. (A.2):

$$s_{j,t+1}-s_{j,t}=\delta_{j}-\frac{\alpha_{j}^{X}n_{j,t}^{X}+\alpha_{j}^{Y}n_{j,t}^{Y}}{N}, \tag{A.2}$$

where δ_j is the task-specific demand rate, and α_j^X and α_j^Y are the task-specific performance efficiencies of X and Y individuals, respectively.

In the subsequent sections, we compute the long-term behavior of the system of six difference equations in Eq. (A.2) (for $n_1^X, n_1^Y, n_2^X, n_2^Y$) and Eq. (A.2) (for s_1, s_2) and compare the results to simulations.

Theoretical maximum activity level. In the model, individuals have a latency period of one time step between when they quit a task and when they recommence working. This means that, on average, only a fraction of the colony can be working at any given time.

To find this maximum activity level, let $Z_t = (n_{1,t}^X + n_{1,t}^Y + n_{2,t}^X + n_{2,t}^Y)/N$ be the fraction of active individuals in a colony at time t. Note that $0 \leq Z_t \leq 1$. At time $t+1$, on average, a fraction τZ_t of the colony becomes inactive. Therefore, we have $(\text{fraction active}) + (\text{fraction inactive}) = X_{t+1} + \tau Z_t \leq 1$. At steady state, the system satisfies $X_{t+1} = X_t = Z^*$. Thus, the theoretical maximum activity level is

$$Z^* \leq \frac{1}{1+\tau} \,. \tag{A.3}$$

For example, when $\tau = 0.2$ (Fig. 1.3), at most 83.33% of colony members can be active at steady state. A similar condition has been noted by Gautrais et al. (2002).

Pure colonies. Without loss of generality, we consider pure colonies that consist of type X individuals only: let $f = 1$ and $n_{j,t}^Y = 0$ for all t. By setting Eq. (A.2) to zero, we obtain the fraction of X individuals performing task j at steady state, given by

$$\frac{n_j^X}{N} = \frac{\delta_j}{\alpha_j^X} \,. \tag{A.4}$$

Notably, the steady-state values of n_j^X are independent of the mean threshold (μ_j^X) or the quit probability (τ^X). This agrees with our simulation results in which differences in μ

(Fig. 1.1C) or τ (Fig. A.3) alone did not change the mean task performance levels in pure colonies.

According to (A.3), this steady state is biologically possible only if

$$(Z^* =) \frac{n_1^X}{N} + \frac{n_2^X}{N} = \frac{\delta_1}{\alpha_1^X} + \frac{\delta_2}{\alpha_2^X} \leq \frac{1}{1+\tau} .$$

If this condition is not met, then the stimuli are expected to continue growing (i.e., system would not reach a steady state).

Now, suppose that the demand rate and task performance efficiency are the same for both tasks ($\delta_1 = \delta_2 = \delta$, $\alpha_1^X = \alpha_2^X = \alpha^X$). Equation (A.4) implies that the fractions of X individuals performing tasks 1 and 2 at steady state would be

$$\frac{n_1^X}{N} = \frac{n_2^X}{N} = \frac{\delta}{\alpha^X} .$$

Similarly, in pure colonies of type Y, if $\delta_1 = \delta_2 = \delta$ and $\alpha_1^Y = \alpha_2^Y = \alpha^Y$, then $n_1^Y/N = n_2^Y/N = \delta/\alpha^Y$ at steady state. Thus, in order for pure colonies of type X and type Y to have different average task performance levels (i.e., $\delta/\alpha^X \neq \delta/\alpha^Y$) under the assumptions above—i.e., that the tasks are equally demanding and that a given type of individual *is* equally efficient at both tasks—the two types must differ in task performance efficiency ($\alpha^X \neq \alpha^Y$) (see main text).

Mixed colonies with 1:1 mixes. We now consider mixed colonies consisting of X and Y individuals in equal proportions ($f = 0.5$). We assume that the mean thresholds and the quit probabilities are identical for both tasks and ant types ($\mu_1^X = \mu_2^X = \mu_1^Y = \mu_2^Y$ and $\tau^X = \tau^Y$)[1]. Setting Eq. (A.2) equal to zero, we find that the steady-state numbers of

[1]The parameters μ and τ do not explicitly appear in Eq. (A.5) when we assume that the mean thresholds are identical for all individuals and both tasks. However, based on Eq. (A.2), we expect the general form of steady state fractions of active individuals to be explicit functions of μ_j^X and μ_j^Y as well as τ^X and τ^Y. While the steady states can be computed numerically for the case when these parameters differ between types or tasks, the analytical expressions are too complicated to write down.

individuals performing task j are given by

$$n_j^X = n_j^Y = N\left(\frac{\delta_j}{\alpha_j^X + \alpha_j^Y}\right).$$

This quantity can also be expressed as a fraction of each type of individuals:

$$\frac{n_j^X}{(N/2)} = \frac{n_j^Y}{(N/2)} = \frac{2\delta_j}{\alpha_j^X + \alpha_j^Y}\,. \tag{A.5}$$

Applying condition (A.3), this steady-state is only biologically relevant when

$$(Z^* =) \sum_{j=1}^{2} \frac{n_j^X}{N} + \frac{n_j^Y}{N} = \sum_{j=1}^{2} \frac{2\delta_j}{\alpha_j^X + \alpha_j^Y} \leq \frac{1}{1+\tau}\,. \tag{A.6}$$

Again, if this condition is not met, then we would expect the stimuli to continue growing over time (i.e., the colony is unable to keep up with the demand) and for the individuals to be working at maximum capacity.

Mixed colonies with non-1:1 mixes. We now generalize to the case in which a fraction f of individuals ($0 \leq f \leq 1$) in a mixed colony are of type X. In the simplified case where $\mu_1^X = \mu_2^X = \mu_1^Y = \mu_2^Y$ and $\tau^X = \tau^Y$, the steady-state fractions of individuals performing task j are

$$n_j^X = \frac{fn\delta_j}{f\alpha_j^X + (1-f)\alpha_j^Y}, \quad n_j^Y = \frac{(1-f)n\delta_j}{f\alpha_j^X + (1-f)\alpha_j^Y}\,.$$

Since there are fn individuals of type X and $(1-f)n$ individuals of type Y, these quantities can be expressed as fractions of individuals of type X and Y individuals performing task j:

$$\frac{n_j^X}{fN} = \frac{n_j^Y}{(1-f)N} = \frac{\delta_j}{f\alpha_j^X + (1-f)\alpha_j^Y}\left(= \frac{n_j^X + n_j^Y}{N}\right). \tag{A.7}$$

The last equality highlights the fact that, at steady state, the fraction of individuals of each type performing task j is identical to the fraction of the whole colony performing that task, i.e., both X and Y perform task j at equal rates. As expected, the expressions Eq.(A.7) reduce to Eq.(A.5) when $f = 0.5$ (1:1 mixes) and to Eq.(A.4) when $f = 1$ (pure colonies with X individuals only). Again, we expect to see this equilibrium only when condition (A.3) is satisfied. Moreover, from Eq.(A.7), we expect the steady-state task j performance frequency to depend non-linearly on the fraction f of X individuals.

Mixed colonies with symmetric mean thresholds. So far we have assumed that the mean task thresholds μ_j^X and μ_j^Y are identical for both ant types and tasks ($\mu_1^X = \mu_2^X = \mu_1^Y = \mu_2^Y$). While Eq.(A.2) can be solved numerically when we introduce between-type differences in μ, the steady-state expressions become too difficult to write down. In the following special case, however, we can express the steady-state values exactly. Assume that

1. colonies consist of type X and Y individuals in equal proportions ($f = 0.5$);
2. task efficiency is the same for both ant types and tasks ($\alpha_1^X = \alpha_2^X = \alpha_1^Y = \alpha_2^Y = \alpha$);
3. demand rate is the same for both tasks ($\delta_1 = \delta_2 = \delta$); and
4. mean task thresholds are symmetric, such that one type has a low threshold for one task and a high threshold for the other while this ordering is reversed in the other type: $\mu_1^X = \mu_2^Y = a$ and $\mu_2^X = \mu_1^Y = b$.

Importantly, the symmetry between the two tasks and between the two types imply that the stimulus levels for the tasks would be identical at stead state ($s_1 = s_2 = s^*$). Moreover, at steady state, the number of X individuals performing task 1 would be identical to the number of Y individuals performing task 2 ($n_1^X = n_2^Y$); similarly, we would expect that $n_1^Y = n_2^X$. Substituting these conditions into Eq.(A.2) and setting it

equal to zero, we find that, at steady state,

$$n_1^X + n_1^Y = n_2^X + n_2^Y = n_1^X + n_2^X = n_1^Y + n_2^Y = N\left(\frac{\delta}{\alpha}\right).$$

By substituting this into Eq. (A.2) and following the symmetry argument above, we derive an expression for the steady-state stimulus level s^*:

$$s^*(= s_1 = s_2) = \left[\frac{1}{2}\left(-(a^\eta + b^\eta) \pm \sqrt{(a^\eta + b^\eta)^2 + (a^\eta b^\eta)\cdot\frac{8\delta\tau}{\alpha - 2\delta(1+\tau)}}\right)\right]^{\frac{1}{\eta}}.$$

The corresponding steady-state fractions of X and Y individuals performing tasks 1 and 2 are, respectively,

$$\begin{aligned}\frac{n_1^X}{(N/2)} = \frac{n_2^Y}{(N/2)} &= \frac{1}{\tau}\left(\frac{(s^*)^\eta}{(s^*)^\eta + a^\eta}\right)\left[2 - \frac{(s^*)^\eta}{(s^*)^\eta + b^\eta}\right]\left(\frac{1}{2} - \frac{\delta}{\alpha}\right), \\ \frac{n_2^X}{(N/2)} = \frac{n_1^Y}{(N/2)} &= \frac{1}{\tau}\left(\frac{(s^*)^\eta}{(s^*)^\eta + b^\eta}\right)\left[2 - \frac{(s^*)^\eta}{(s^*)^\eta + a^\eta}\right]\left(\frac{1}{2} - \frac{\delta}{\alpha}\right).\end{aligned} \tag{A.8}$$

When $a = b$ (i.e., when all μ's are identical), these expressions reduce to the steady states predicted in Eq. (A.5).

Downward vs. upward convergence. Both our experiments (Fig. 1.2A) and theoretical analyses (Fig. 1.3A–B) demonstrated patterns of asymmetric behavioral convergence between the types, in which individuals of different types were behaviorally more similar to each other when mixed. Here we combine our analytical predictions for pure and mixed colonies to investigate conditions under which such convergence patterns arise. Consider two pure colonies consisting of X and Y individuals, respectively, and a third, mixed colony consisting of a 1:1-ratio of X and Y individuals. Let us assume that each colony reaches a steady state (i.e., each colony satisfies condition (A.3)). We show analytically that, under these conditions, if the ant

types only differ in task efficiency (α_j^X, α_j^Y), then the system can exhibit a downward convergence but not an upward convergence.

We can directly apply the steady-state fractions of active individuals in Eqs. (A.4) and (A.5) because the mean threshold (μ) and the quit probability (τ) are assumed to be identical across types. The behavioral convergence is *downward* if

$$\frac{1}{2}\left(\frac{\delta_j}{\alpha_j^X} + \frac{\delta_j}{\alpha_j^Y}\right) > \frac{2\delta_j}{\alpha_j^X + \alpha_j^Y} \tag{A.9}$$

and *upward* if the inequality is reversed (see also Fig. 1.3a–b).

By manipulating the inequality (A.9), we see that the left-hand side is always at least as large as the right-hand side:

$$\frac{1}{2}\left(\frac{\delta_j}{\alpha_j^X} + \frac{\delta_j}{\alpha_j^Y}\right) - \frac{2\delta_j}{\alpha_j^X + \alpha_j^Y} = \frac{\delta_j}{2}\left(\frac{(\alpha_j^X - \alpha_j^Y)^2}{\alpha_j^X \alpha_j^Y (\alpha_j^X + \alpha_j^Y)}\right) \geq 0\,.$$

The equality holds if and only if $\alpha_j^X = \alpha_j^Y$, in which case the types are indistinguishable with respect to task j. If $\alpha_j^X \neq \alpha_j^Y$, then only downward convergence is possible under our assumptions (in particular, we assume that condition Eq. (A.3) is satisfied). Note that the threshold between upward and downward convergence, Eq. (A.9) is agnostic to between-task differences in task efficiency or task demand; in other words, it holds even when $\alpha_1^X \neq \alpha_2^X$, $\alpha_1^Y \neq \alpha_2^Y$, and $\delta_1 \neq \delta_2$.

Contextualizing the analytical calculations. To put our analytical results into context, consider scenarios (i) with two high-demand tasks (e.g., $\delta_1 = \delta_2 = 1.3$, as in Fig. 1.3a), (ii) with two low-demand tasks (e.g., $\delta_1 = \delta_2 = 0.6$, as in Fig. 1.3b), and (iii) with one high-demand task and one low-demand task (e.g., $\delta_1 = 1.3$, $\delta_2 = 0.6$). We suppose that all type-specific parameters are identical across scenarios and that mean

thresholds and quit probabilities are identical for both tasks and ant types (i.e., $\mu_1^X = \mu_2^X = \mu_1^Y = \mu_2^Y$, $\tau^X = \tau^Y$, as in *Mixed colonies with 1:1 mixes*).

Equation (A.5) and condition (A.6) (or, equivalently, condition (A.3)) suggest that, in the absence of differences in mean threshold, the total task performance frequency of all ants in a colony (i.e., $\sum_{j=1}^{2} 2\delta_j/(\alpha_j^X + \alpha_j^Y)$) would be highest in (i), lowest in (ii), and intermediate in (iii); and that a colony is most likely to keep up with the demand in (iii), less likely in (ii), and least likely in (i). In this sense, we predict that the outcome in (iii) will be quantitatively intermediate between (i) and (ii). However, this would not alter the possible qualitative outcomes of mixing: if condition (A.6) is satisfied under scenario (iii), we would expect mixing to produce a downward convergence (see *Downward vs. upward convergence*); if not, mixing could lead to either downward or upward convergence depending on demand values. See also *An expanded model of DOL* in the main text.

A.1.2 Theoretical predictions for mean task performance in non-1:1 mixes

We further explored expected patterns of task allocation in colonies with different ratios of ant types. For the parameter combinations in Fig. 1.3—which collectively captured all experimentally observed patterns—we investigated how the mean task performance of colonies changed as we varied the ratio of the two ant types.

Simulations predicted a striking range of patterns. For the parameter combination that produced no effect in the mixed colonies with equal proportions of the two ant types ('1:1 mixes'), the model produced an approximately linear relationship between mean task performance and the ratio of ant types (Fig. A.7A). In all other cases, the mean task performance followed nonlinear functions the ratio of the types, but their shapes differed among the cases. In the cases corresponding to behavioral convergence in the

1:1 mixes, the relationship followed a convex decreasing function, so long as there were enough individuals of the more efficient type such that the colony could keep up with the demand (Fig. A.7B; *Analytical treatment of the model*); otherwise the colony performed the tasks at a fixed maximum capacity that depended only on the average task duration (Fig. A.7C). In the case corresponding to behavioral divergence, the relationship followed a concave decreasing function (Fig. A.7D). Hence, despite one type being more efficient than the other in all cases considered, replacing an individual of the former type with one of the latter type would lead to qualitatively different outcomes depending on the differences in mean threshold.

Regardless of the case studied, the ratio of the types did not alter the qualitative effect of mixing on individual behavior (behavioral convergence, divergence, or no effect); for example, the case that led to behavioral divergence in 1:1 mixes predicted behavioral divergence for all non-1:1 mixes tested (Fig. A.7D).

A.2 Supplementary methods

A.2.1 Experimental design

Four experiments were performed to investigate the effect of genetic composition (2 experiments differing in the brood genotype used), age composition (1 experiment), and morphological composition (1 experiment). Each experiment comprised three treatments (2 with pure colonies, 1 with mixed colonies; Table A.1). All colonies within one experiment were monitored in parallel, but the different experiments were performed separately.

Experimental colonies were composed of workers of controlled age, genotype, and morphology (Table A.1), as well as larvae of controlled genotype and age. Colonies were housed in airtight Petri dishes 5 cm in diameter (corresponding to about 25 ant body

lengths) with a plaster of Paris floor, in which the workers formed a nest by freely choosing a location where they piled their larvae. To control individual genotype, clonally related workers were sourced from the same stock colony. We used two commonly used genotypes, A and B (Kronauer et al., 2012; Oxley et al., 2014; Teseo et al., 2014; Ulrich et al., 2018). To control individual age, workers were sourced from a single age cohort from the same stock colony. Owing to the synchronized reproduction of *O. biroi*, all age-matched workers collected this way had eclosed within a day of each other (Ravary and Jaisson, 2002). Young ants were 1 cycle old (approximately 1 month old), and old ants were 3 cycles old (approximately 3 months old). The estimated life span of workers of this species under laboratory conditions is approximately 1 year. To control individual morphology, age-matched regular workers and intercastes from the same stock colony were screened based on body size (small or large) and the absence or presence of vestigial eyes, respectively. From the time they were collected (1 to 3 days after eclosion) until the start of experiments, workers of a given type were kept as a group. All workers were tagged with color marks on the thorax and gaster using oil paint markers. Experimental colonies contained 16 (genetic composition and age composition experiments) or 8 (morphological composition experiment) workers and a matching number of age-matched larvae (4 to 5 days old). This 1:1 larvae-to-workers ratio corresponds to the estimated ratio found in a typical laboratory stock colony. We used 8 (genetic composition and age composition experiments) or 16 (morphological composition experiment) replicate colonies for each group composition, for a total of 120 colonies.

Colony number and size varied across experiments due to constraints on the number of available slots in the tracking system at the time each experiment was performed. However, all colony sizes employed here were previously shown to have high fitness and stable DOL (Ulrich et al., 2018), and all experiments were analyzed separately so that variation in colony size should not impact the results.

The experiments took place in a climate room at 25C and 75% relative humidity under constant light (*O. biroi* is blind, and its behavior is not affected by light). Every three days, we cleaned and watered the plaster and added one prey item (live pupae of fire ant minor workers) per live *O. biroi* larva at a random location within the Petri dish.

A.2.2 Behavioral data acquisition and analysis

Image acquisition and analysis were performed as in Ulrich and colleagues (Ulrich et al., 2018). We used an automated scan sampling approach, in which a picture of each colony was acquired every approximately 400 seconds throughout the experiment by a custom setup comprising 28 webcams (B910 or C910; Logitech, Lausanne, Switzerland) and controlled LED lighting. Each webcam acquired images (5 megapixels, RGB) of 4 colonies, and the position of colonies within the setup was randomized. Custom software (available at https://doi.org/10.5281/zenodo.1211644) was used to detect individual ants in images. For all behavioral analyses, ants were excluded from the dataset if they were detected in less than 30% of the frames acquired within the considered time frame (brood care phase or day); for ants that died during the brood care phase, the considered time frame was the portion of the brood care phase preceding death.

O. biroi colonies switch between reproductive phases (of approximately 18 days), in which all workers stay in the nest and lay eggs, and brood care phases (of approximately 16 days), in which workers nurse the larvae in the nest but also leave the nest to scout, forage, or dispose of waste. For each colony, behavioral analyses were restricted to the brood care phase, which started at the beginning of the experiment and ended when all larvae had either reached the nonfeeding prepupal stage or died.

For each colony or subcolony, mean behavior was computed as the average of individual r.m.s.d. values, and behavioral variation was computed as the standard

deviation of individual r.m.s.d. values. Both metrics were then compared across treatments.

To quantify specialization, we use a metric appropriate for use on continuous behavioral data (here, r.m.s.d.). For each colony, specialization was defined as the Spearman correlation coefficient between individual r.m.s.d. ranks on consecutive days of the brood care phase, averaged over days. Mean rank-correlation coefficients were then compared across treatments.

A.2.3 Statistical analyses

Statistical analyses were performed in R (R Core Team, 2019) separately for each of the four experiments. As the experiments were performed at different times using different cohorts of ants, we cannot rule out batch effects and therefore avoid any statistical analyses comparing treatments across experiments.

Effects of individual traits on behavior. To evaluate whether type-specific behavior depended on colony composition, we tested for a statistical interaction between the effects of individual attributes (genotypes A vs. B, Young vs. Old, or Regular worker vs. Intercaste) and of colony composition (pure vs. mixed) on individual behavior (individual r.m.s.d.) using linear mixed effects (LMEs, function *lmer* of package *lme4* (Bates et al., 2015)) models with colony as a random factor. If a significant interaction between colony composition and individual attributes was detected, we then used a second LME model with a four-level independent fixed variable combining colony composition and individual attributes (X_p, Y_p, X_m and Y_m, where X_k and Y_k are the mean behavior of ant types X and Y, respectively, in pure ($k = p$) or mixed colonies ($k = m$)), followed by a Tukey post hoc test with Bonferroni-Holm correction (function *glht* of package *multcomp* (Hothorn et al., 2008)) for the following planned pairwise comparisons: X_p vs. X_m, Y_p vs. Y_m, X_p vs. Y_p, and X_m vs. Y_m. The two models are

functionally equivalent but were used to test different hypotheses regarding interaction between terms (first model) and pairwise differences between groups (second model). When needed, the response variable was transformed (r.m.s.d.^{2} in the genotype experiment with brood of genotype A and the age experiment, r.m.s.d$^{3/5}$ in the genotype experiment with brood of genotype B; no transformation for the morphology experiment) to satisfy model assumptions. We evaluated the significance of terms by comparing pairs of nested models using χ^2 log-likelihood ratio tests following deletion of the term of interest (the interaction in the first model and the four-level variable combining colony composition and individual attributes in the second model) using the function *drop1* in R.

Effects of genetic, demographic, and morphological mixing on DOL. The effects of the treatment (a three-level variable: pure X, pure Y, and mixed XY on colony-level DOL (behavioral variation and specialization) were investigated using generalized linear models (GLMs). The significance of treatment was evaluated as above. Pairwise comparisons between treatments were evaluated using Tukey post hoc tests with Bonferroni-Holm correction. Behavioral variation was square root-transformed in the genotype experiment with B larvae to satisfy model assumptions.

Effects of genetic, demographic, and morphological mixing on individual behavior. To assess how type-specific behavior was affected by mixing, and more specifically, whether the difference in behavior between types of ants was affected by mixing, we compared the difference in mean behavior (type-specific mean r.m.s.d. in each colony) between types across pure colonies to the difference in mean behavior between the same types within mixed colonies (i.e., $Y_p - X_p$ vs. $Y_m - X_m$, where $Y_p > X_p$ and $Y_m > X_m$; see below for definitions of behavioral patterns), using unpaired t tests, after verifying assumptions of normality. In mixed colonies, the difference in mean behavior was calculated between types of ants within a colony (e.g., old and young

workers from the same colony); in pure colonies, the difference in mean behavior was calculated between arbitrary pairs of pure colonies (e.g., old workers from the pure colony #1 and young ants from pure colony #1, where 1 is a replicate number assigned randomly at the beginning of the experiment). We further tested whether the amplitude of the effect differed across types by comparing the magnitude of change in type-specific behavior between pure and mixed colonies across the 2 ant types (i.e., $|X_m - X_p| \neq |Y_m - Y_p|$) with unpaired t tests, after verifying assumptions of normality.

A.3 Supplementary tables and figures

Table A.1: Parameter settings for model simulations.

Parameter	*Description*	*Baseline values*
T	Simulation length in time steps	10,000
N	Number of individuals	16
M	Number of tasks	2
$\delta_j = \delta$	Brood-specific rate of stimulus increase (i.e., demand rate); taken to be the same for all tasks	0.6
$\alpha_j^X = \alpha^X$, $\alpha_j^Y = \alpha^Y$	Type-specific performance efficiency of active individuals for task j; taken to be the same for all tasks	2
$\mu_j^X = \mu^X$, $\mu_j^Y = \mu^Y$	Mean of the type-specific threshold distribution for task j; taken to be the same for all tasks	10
$\sigma_j^X = \sigma^X$, $\sigma_j^Y = \sigma^Y$	Variance of the type-specific threshold distribution for task j as a fraction of the corresponding mean; taken to be the same for all tasks	0.1
η	Threshold stochasticity	7
$\tau_j^X = \tau^X$, $\tau_j^Y = \tau^Y$	Type-specific probability of quitting task j once active (inverse of average task performance duration); taken to be the same for all tasks	0.2

Table A.2: List of experimental treatments. Text in bold denotes the variable of interest for each experiment. All mixed colonies contained a 1:1 ratio of each ant type.

Experiment	Worker genotype	Brood genotype	Age (cycles)	Subcaste	Colony size N	n replicates / composition	# colonies
Genetic composition 1	**A, B, mixed**	A	1	Regular workers	16	8	24
Genetic composition 2	**A, B, mixed**	B	1	Regular workers	16	8	24
Age composition	B	B	**1, 3, mixed**	Regular workers	16	8	24
Morphological composition	B	B	1	**Regular workers, intercastes, mixed**	8	16	48

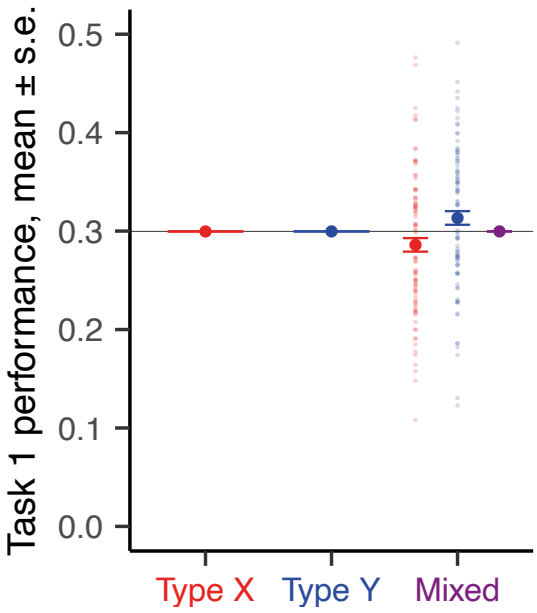

Figure A.1: Theoretical predictions with differences in threshold variance. Task performance frequency for a single task as a function of colony composition. Opaque circles represent individual replicate colonies ($N = 16$; $n = 100$ replicates per composition); solid circles represent average value ($\pm$SE) across replicates. Horizontal gray lines represent the average of the pure colonies (first two columns). Types X and Y differ in threshold variance: $\sigma^X = 0.1$, $\sigma^Y = 0.5$; all other parameters are identical (see Table A.1).

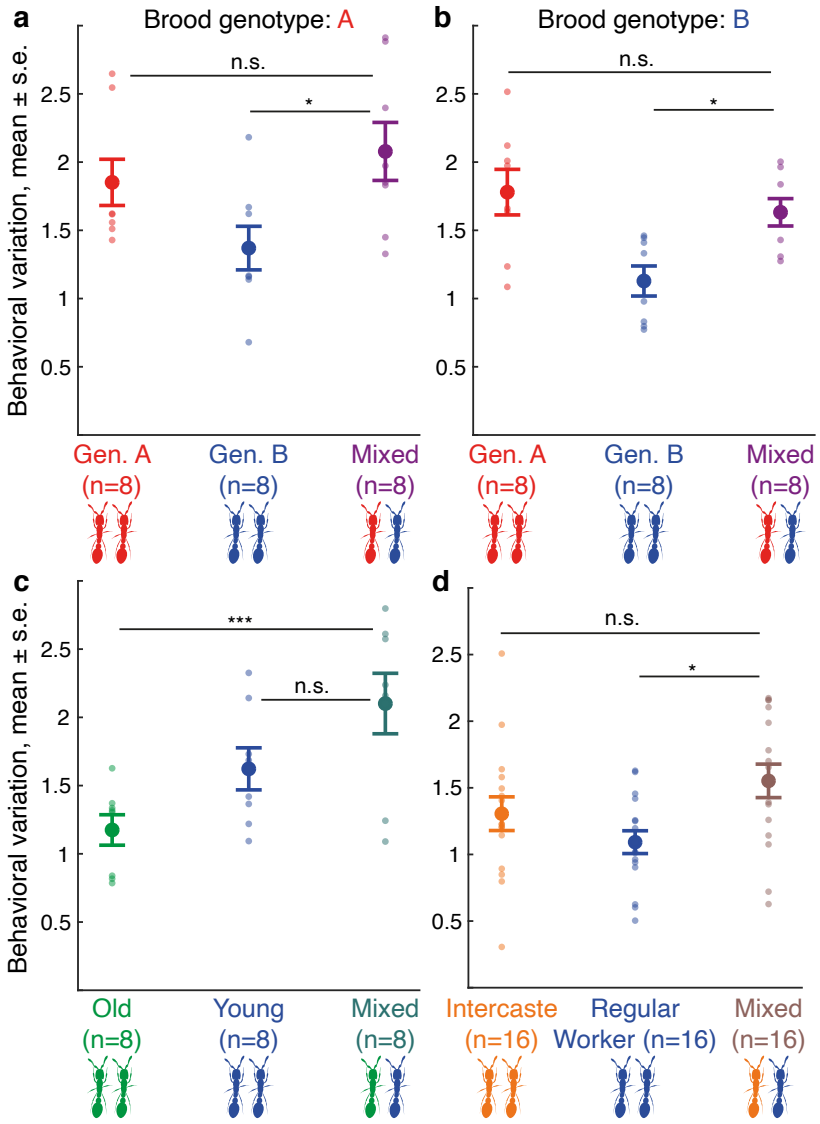

Figure A.2: Behavioral variation (standard deviation in r.m.s.d. across colony members) as a function of colony composition. Small opaque circles represent individual colonies; large solid circles represent the average values across n replicate colonies. Identical colors across panels indicate ants of the same genotype, age, and morphological types. (a) Behavioral variation as a function of colony genetic composition in colonies with A brood ($N = 16$; $\mathrm{B_{hom}}$ vs. Mixed: $z = -2.85$, $p = 0.013$; $\mathrm{A_{hom}}$ vs. Mixed: $z = 0.81, p = 0.421$). (b) Behavioral variation as a function of colony genetic composition in colonies with B brood ($N = 16$; $\mathrm{B_{hom}}$ vs. Mixed: $z = -2.76, p = 0.012$; $\mathrm{A_{hom}}$ vs. Mixed: $z = -0.81, p = 0.419$). (c) Behavioral variation as a function of colony age composition ($N = 16$; $\mathrm{Young_{hom}}$ vs. Mixed: $z = 2.01, p = 0.090$; $\mathrm{Old_{hom}}$ vs. Mixed: $z = 3.89, p = 3.07 \cdot 10^{-04}$). (d) Behavioral variation as a function of colony morphological composition ($N = 8$; $\mathrm{Regular\ Worker_{hom}}$ vs. Mixed: $z = -2.85$, $p = 0.013$, $\mathrm{Intercaste_{hom}}$ vs. Mixed: $z = 1.53, p = 0.254$). n.s.: non-significant, *:$p < 0.05$, **:$p < 0.01$, ***:$p < 0.001$.

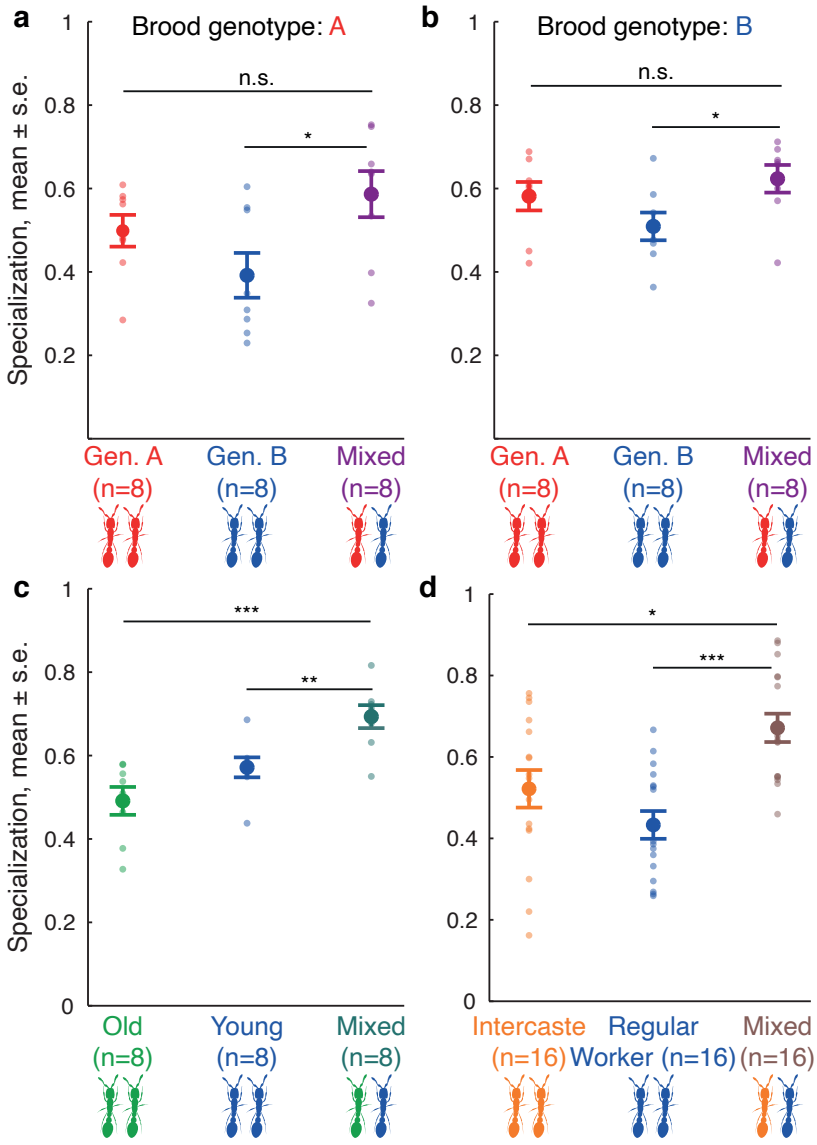

Figure A.3: Colony-level specialization (day-to-day rank correlation in r.m.s.d.) as a function of colony composition. Small opaque circles represent individual colonies; large solid circles represent the average values across n replicate colonies. Identical colors across panels indicate ants of the same genotype, age, and morphological types. (a) Specialization as a function of colony genetic composition in colonies with A brood ($N = 16$; GLM post hoc tests; B_p vs. mixed: $z = -2.78$, $p = 0.017$; A_p vs. mixed: $z = 1.25$, $p = 0.256$). (b) Specialization as a function of colony genetic composition in colonies with B brood ($N = 16$; B_p vs. mixed: $z = -2.41$, $p = 0.048$; A_p vs. mixed: $z = 0.88$, $p = 0.378$). (c) Specialization as a function of colony age composition ($N = 16$; $Young_p$ vs. mixed: $z = 3.01$, $p = 0.005$; Old_p vs. mixed: $z = 5.01$, $p = 1.63 \cdot 10^{-06}$). (d) Specialization as a function of colony morphological composition ($N = 8$; Regular $Worker_p$ vs. mixed: $z = -4.35$, $p = 4.07 \cdot 10^{-05}$, $Intercaste_p$ vs. mixed: $z = 2.73$, $p = 0.013$). n.s.: non-significant, *:$p < 0.05$, **:$p < 0.01$, ***:$p < 0.001$.

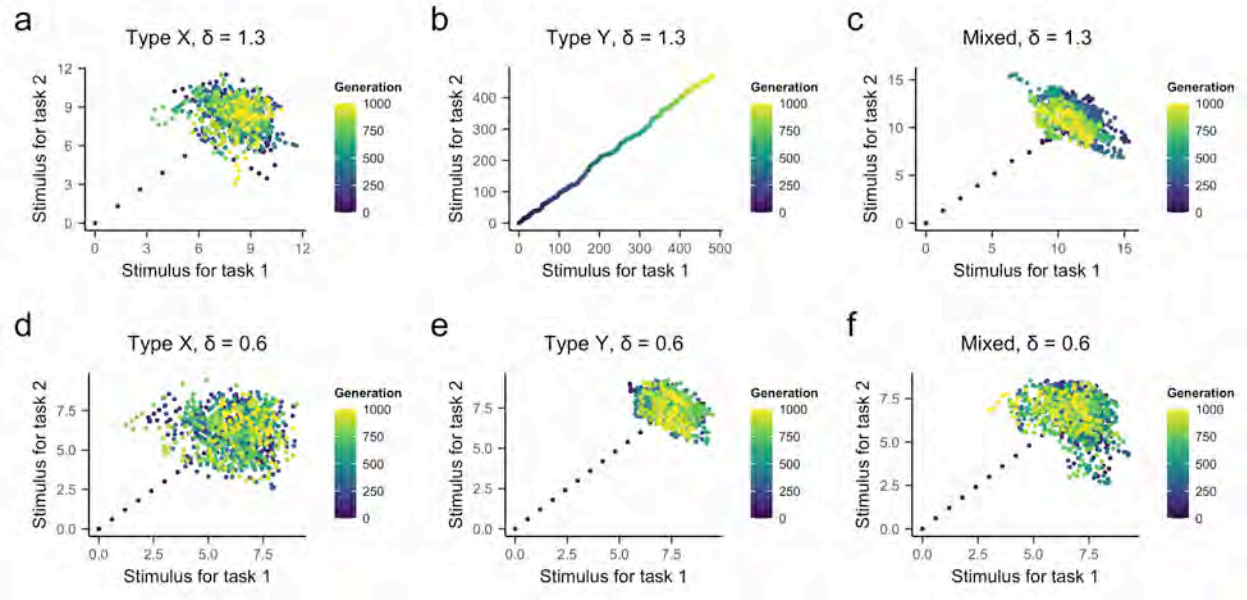

Figure A.4: Dynamics of stimulus levels in pure and mixed colonies. Each point shows the simulated stimulus level for the two tasks (task 1 on the horizontal axes, task 2 on the vertical axes) in the generation indicated by its color. Each of panels a, b, d, and e shows a pure colony of the type indicated; each of panels c and f shows a mixed colony of Types X and Y. Panels a–c ($\delta = 1.3$) correspond to Fig. 1.3A and d–f ($\delta = 0.6$) to Fig. 1.3b. (a–c) When the demand is higher ($\delta = 1.3$), the more efficient type (Type X) can keep up with the demand on its own (a) but the less efficient type (Type Y) cannot, as demonstrated by the continual growth of the stimuli (b); however, mixed colonies can keep up with the higher level of demand (c). (d–f) When the demand is lower ($\delta = 0.6$), the stimulus levels grow quickly at first but then stabilizes to an oscillatory pattern around a point, demonstrating that both pure and mixed colonies can keep up with the demand. Each simulation ran for 1000 time steps; all other parameters are identical to those in the corresponding Fig. 1.3 panels.

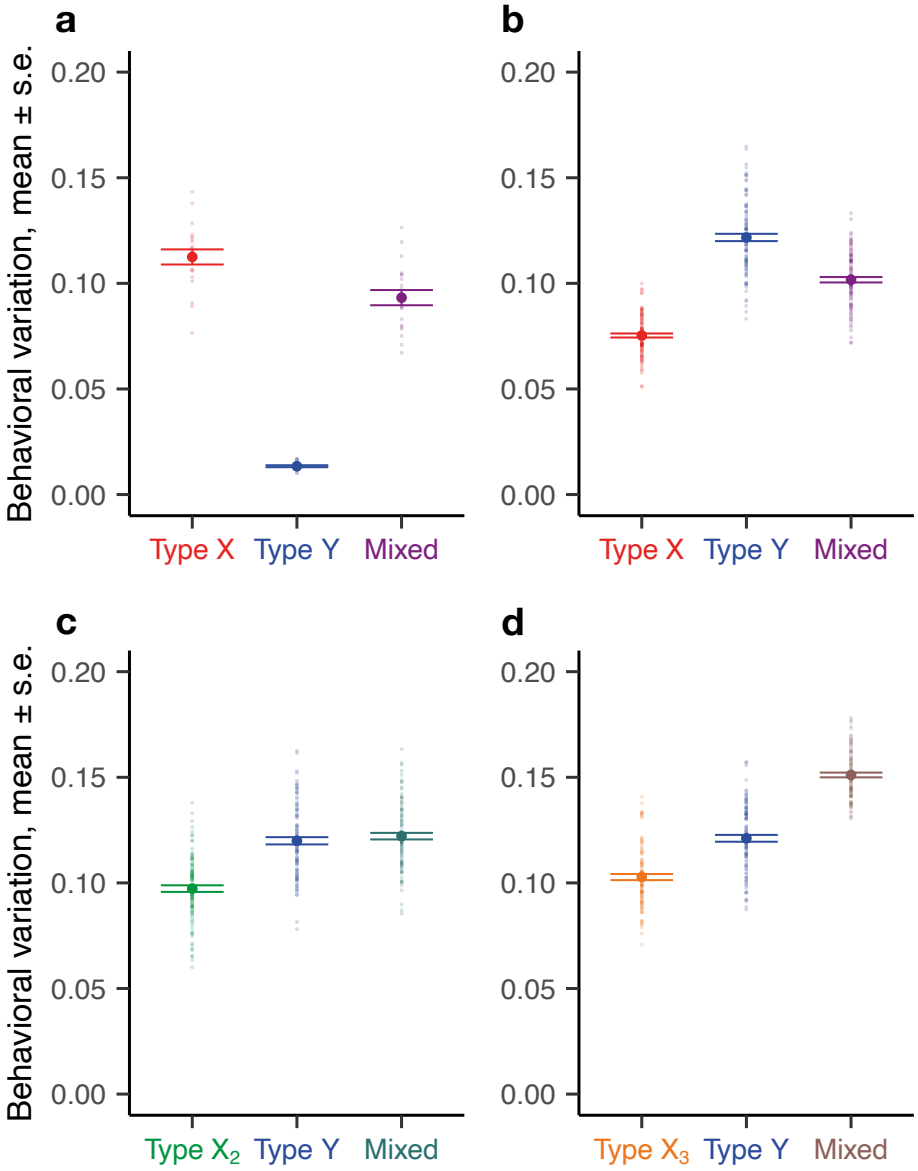

Figure A.5: Theoretical predictions of the expanded model on behavioral variation. Behavioral variation was quantified as the standard deviation of task performance frequency across individuals in a colony. Opaque circles represent individual replicate colonies ($N = 16$; $n = 100$ replicates per composition); solid circles represent the average value (mean ±SE) across replicates. Types X_1, X_2, X_3, Y and their corresponding parameters are identical to those in Fig. 1.3. See Table A.1 for other parameters.

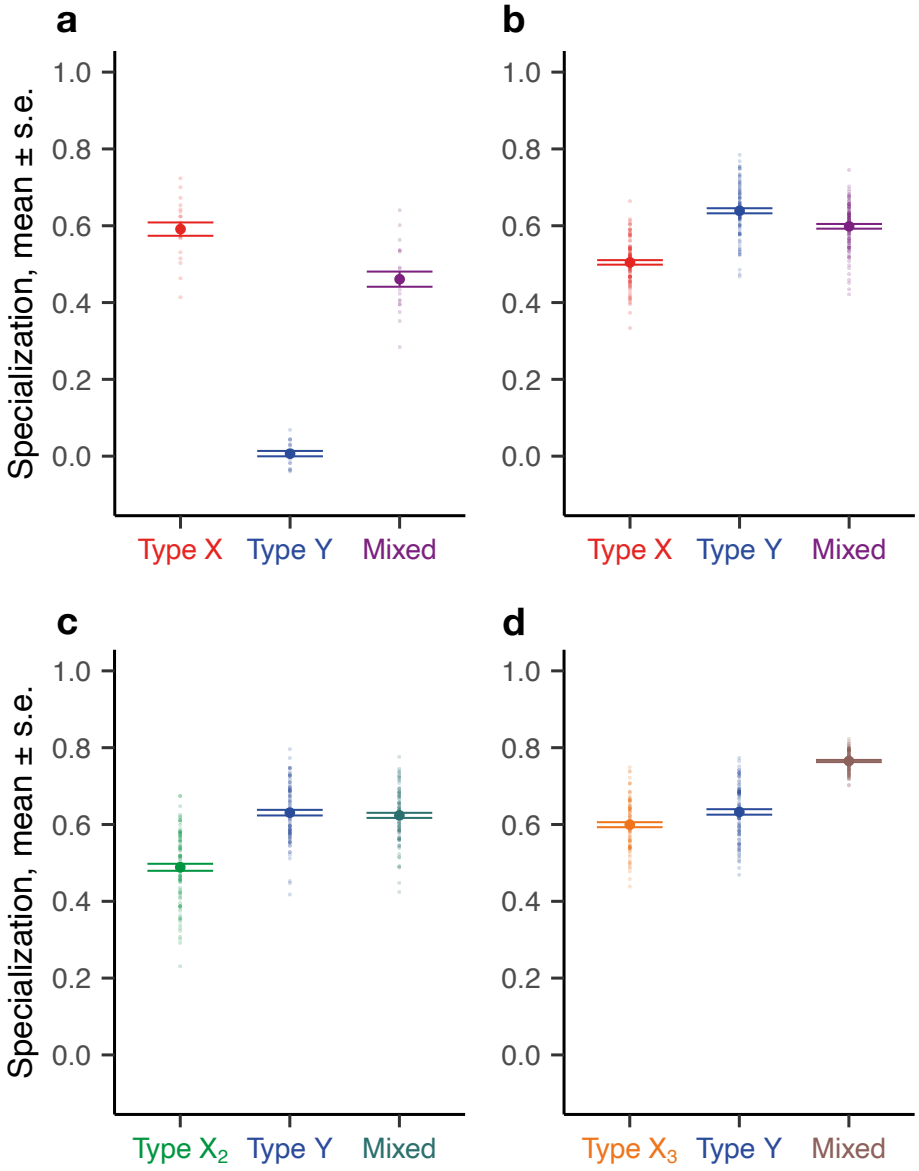

Figure A.6: Theoretical predictions of the expanded model on behavioral specialization. Colony-level specialization was quantified using Spearman rank correlation on consecutive windows of 200 time steps. Opaque circles represent individual replicate colonies ($N = 16$; $n = 100$ replicates per composition); solid circles represent the average value (mean ±SE) across replicates. Types X_1, X_2, X_3, Y and their corresponding parameters are identical to those in Fig. 1.3. See Table A.1 for other parameters.

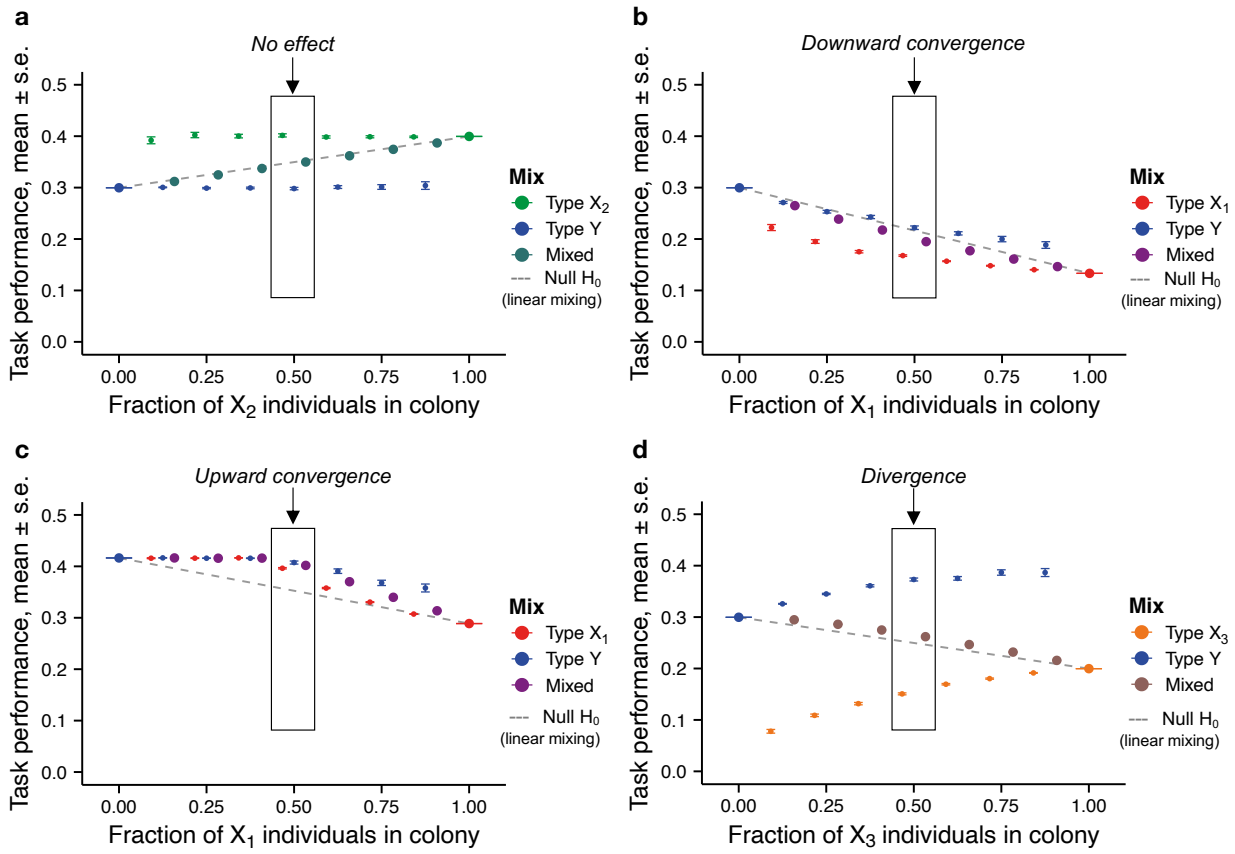

Figure A.7: Model predictions for non-1:1 mixes. Colonies with varying ratios of X and Y individuals were simulated under different conditions of threshold values, task-performance efficiency, and task demand ($n = 100$ replicates per colony composition). Each large circle represents the mean task performance (task 1) for that mix of X and Y individuals; the neighboring smaller circles represent the means of X and Y individuals, respectively, within that mix. Dashed lines indicate the null hypothesis of linear behavioral effects of mixing types. The boxes highlight the behavioral patterns characterizing the 1:1-mixes, and their labels indicate correspondence with Fig. 1.3 (a: Fig. 1.3e, b: Fig. 1.3b, c: Fig. 1.3a, d: Fig. 1.3d). Types X_1, X_2, X_3, Y and their corresponding parameters are identical to those in Fig. 1.3. See Table A.1 for other parameters.

Appendix B

Supplementary materials for Chapter 2

B.1 Supplementary analyses

B.1.1 Selection-mutation equilibrium in the limit of weak selection

Our model considers a population of N individuals distributed over M potentially overlapping groups, each representing a political issue of interest. Let $\mathbf{h}_i = [h_{i1}, h_{i2}, \cdots, h_{iM}] \in \{-1, 0, 1\}^M$ denote the M-element opinion vector of individual i, where h_{ik} represents i's opinion on issue k: liberal (-1), neutral (0), or conservative ($+1$). We say that individual i cares about issue k if she takes either a liberal or conservative position on it and that she does not care about issue k if she takes a neutral position. We assume that every individual cares about exactly K issues, i.e., $\sum_{k=1}^{M} |h_{ik}| = K$ for all $i \in \{1, \ldots, N\}$.

Individuals also have political affiliations. Let $a_i \in \{0, \cdots, P\}$ denote the party affiliation of individual i. For simplicity, we focus on a two-party system ($P = 2$), i.e., $a_i \in \{1, 2\}$, with individuals distributed equally across the parties.

Interactions and fitness. In each round, two individuals play one-shot pairwise donation games as many times as the number of issues they both care about. In a given game, the donor can choose to either cooperate (C), incurring a cost c to provide a benefit b to the recipient, or defect (D), incurring no cost and providing no benefit to the recipient. Individual i's strategy is given by $\mathbf{s}_i = [s_{ia}, s_{id}] \in \{0,1\}^2$, where 0 corresponds to D and 1 to C. When i interacts with j in group k, i plays strategy s_{ia} if i and j agree on issue k ($h_{ik} = h_{jk}$) and s_{id} if they disagree ($h_{ik} \neq h_{jk}$). We refer to s_{ia} and s_{id} as *agreement strategy* and *disagreement strategy*, respectively.

The fitness of individual i in a given round is given by $f_i = 1 + \beta\pi_i$, where π_i is the total payoff accrued by individual i in that round:

$$\begin{aligned}\pi_i &= \sum_{\substack{j=1 \\ j\neq i}}^{N}\sum_{k=1}^{M} |h_{ik}h_{jk}| \left[\delta_{ij}^k \left[-cs_{ia} + bs_{ja}\right] + (1-\delta_{ij}^k)\left[-cs_{id} + bs_{jd}\right]\right] \\ &= -Ks_{ia}(b-c) + \sum_{j=1}^{N}\sum_{k=1}^{M} |h_{ik}h_{jk}| \left[\delta_{ij}^k \left[-cs_{ia} + bs_{ja}\right] + (1-\delta_{ij}^k)\left[-cs_{id} + bs_{jd}\right]\right],\end{aligned} \tag{B.1}$$

where $|h_{ik}h_{jk}|$ filters the shared issues ($|h_{ik}h_{jk}| = 1$ if $h_{ik}, h_{jk} = 1$ or -1, $|h_{ik}h_{jk}| = 0$ otherwise). The function δ_{ij}^k indicates whether i and j agree on a given issue: $\delta_{ij}^k = \mathbb{1}\{h_{ik} = h_{jk}\} = \frac{1}{2}(1 + h_{ik}h_{jk}) = 1$ if $h_{ik} = h_{jk}$ (that is, if i and j both care about issue k and have the same opinion) and 0 if $h_{ik} = -h_{jk}$ (that is, if i and j both care about issue k but have opposing opinions). Note that, when $M = K$, $|h_{ik}h_{jk}| = 1$ for every $k \in \{1, \ldots, M\}$.

Frequencies of strategies. Let $x_{\mathbf{s}}$ be the frequency of strategy $\mathbf{s} = [s_{ia}, s_{id}] \in \{CC, CD, DC, DD\}$, satisfying $\sum_{\mathbf{s}} x_{\mathbf{s}} = 1$. The frequency of each strategy

can be expressed in terms of s_{ia}, s_{id} as

$$x_{CC} = \frac{1}{N}\sum_{i=1}^{N} s_{ia}s_{id} \ , \qquad x_{CD} = \frac{1}{N}\sum_{i=1}^{N} s_{ia}(1-s_{id}) \ ,$$
$$x_{DC} = \frac{1}{N}\sum_{i=1}^{N} (1-s_{ia})s_{id} \ , \quad x_{DD} = \frac{1}{N}\sum_{i=1}^{N} (1-s_{ia})(1-s_{id}) \ .$$

Evolutionary updating for $p < 1$. The population evolves according to the Moran process. In each generation, a learner is selected uniformly at random, and the learner randomly chooses a role model with probability proportional to her relative fitness. The average probability with which the learner considers imitating the role model depends on their party affiliations and the partisan bias p. Let this probability be $Q(p)$.

When the learner attempts to imitate, she either adopts the strategy of the role model with probability $1-u$ (*selection*) or a random strategy with probability u (*mutation*).

- The average change in $x_{\mathbf{s}}$ due to *mutation* is given by $\Delta x_{\mathbf{s}}^{\text{mut}} = \frac{1}{N}\left(\frac{1}{4}(1-x_{\mathbf{s}}) - \frac{3}{4}x_{\mathbf{s}}\right) = \frac{1}{N}\left(\frac{1}{4} - x_{\mathbf{s}}\right)$, where the number 4 in the denominator corresponds to the number of possible strategies.

- The average change in $x_{\mathbf{s}}$ due to *selection* is given by $\Delta x_{\mathbf{s}}^{\text{sel}} = \frac{1}{N}\sum_i \mathbb{1}_{\mathbf{s}_i=\mathbf{s}}(\omega_i - 1)$, where ω_i is the expected number of copies of individual i after one generation. When $p < 1$, this quantity is given by

$$\omega_i = 1 - \frac{1}{N} + \frac{f_i}{\sum_j f_j} \ .$$

  In the limit of weak selection (β small), we can linearize $\Delta x_{\mathbf{s}}^{\text{sel}}$ to the leading order in β:

$$\Delta x_{\mathbf{s}}^{\text{sel}} = \frac{1}{N}\cdot\frac{\beta}{N}\sum_i \mathbb{1}_{\mathbf{s}_i=\mathbf{s}}\left(\pi_i - \frac{1}{N}\sum_j \pi_j\right) + \mathcal{O}(\beta^2) \ .$$

Thus, at the mutation-selection equilibrium, we have

$$Q(p)\left(u\left\langle \Delta x_{\mathbf{s}}^{\text{mut}}\right\rangle +(1-u)\left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle \right)=0 \quad \Longrightarrow \quad \frac{u}{N}\left\langle \frac{1}{4}-x_{\mathbf{s}}\right\rangle +(1-u)\left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle =0\,, \tag{B.2}$$

where the average $\langle\cdot\rangle$ is taken over the stationary population state. In particular, in the weak selection limit, we can approximate $\left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle$—the expected change in frequency of strategy $\mathbf{s}$ due to selection—by taking the average over the neutral stationary state (i.e., with $\beta=0$) (Antal et al., 2009a,b; Tarnita et al., 2009a):

$$\left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle =\beta\left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle _{0}+\mathcal{O}(\beta^{2})\,,$$
$$\text{with} \quad \left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle _{0}=\left\langle \frac{1}{N}\sum_{i}\mathbb{1}_{\mathbf{s}_{i}=\mathbf{s}}\frac{d\omega_{i}}{d\beta}\right\rangle _{0}=\left\langle \frac{1}{N^{2}}\sum_{i}\mathbb{1}_{\mathbf{s}_{i}=\mathbf{s}}\left(\pi_{i}-\frac{1}{N}\sum_{j}\pi_{j}\right)\right\rangle _{0}\,,$$

where the subscript 0 refers to neutral drift ($\beta=0$). Thus, to compute the stationary distribution of strategies in the weak selection limit (β small), we only need to consider selection under neutrality; we refer the reader to (Antal et al., 2009a,b; Tarnita et al., 2009a) for a full justification of this approach.

Solving (B.2) for $\langle x_{\mathbf{s}}\rangle$, we obtain the stationary frequency of strategy $\mathbf{s}$ in the weak selection limit as

$$\langle x_{\mathbf{s}}\rangle=\frac{1}{4}+\left(\frac{1-u}{u}\right)\cdot N\beta\left\langle \Delta x_{\mathbf{s}}^{\text{sel}}\right\rangle _{0}\,. \tag{B.3}$$

Thus a strategy $\mathbf{s}$ is favored by selection if $\langle x_{\mathbf{s}} \rangle > 1/4$, i.e., if $\left\langle \Delta x_{\mathbf{s}}^{\text{sel}} \right\rangle_0 > 0$. Hence, our objective is to compute the following quantities:

$$\begin{aligned}
\left\langle \Delta x_{CC}^{\text{sel}} \right\rangle_0 &= \left\langle \frac{1}{N^2} \sum_{i=1}^{N} s_{ia} s_{id} \left(\pi_i - \frac{1}{N} \sum_j \pi_j \right) \right\rangle_0 , \\
\left\langle \Delta x_{CD}^{\text{sel}} \right\rangle_0 &= \left\langle \frac{1}{N^2} \sum_{i=1}^{N} s_{ia} (1 - s_{id}) \left(\pi_i - \frac{1}{N} \sum_j \pi_j \right) \right\rangle_0 , \\
\left\langle \Delta x_{DC}^{\text{sel}} \right\rangle_0 &= \left\langle \frac{1}{N^2} \sum_{i=1}^{N} (1 - s_{ia}) s_{id} \left(\pi_i - \frac{1}{N} \sum_j \pi_j \right) \right\rangle_0 , \\
\left\langle \Delta x_{DD}^{\text{sel}} \right\rangle_0 &= \left\langle \frac{1}{N^2} \sum_{i=1}^{N} (1 - s_{ia})(1 - s_{id}) \left(\pi_i - \frac{1}{N} \sum_j \pi_j \right) \right\rangle_0 .
\end{aligned} \tag{B.4}$$

B.1.2 Computing the expected change in strategy frequency due to selection

To compute expected change in strategy frequencies due to selection, we begin by substituting Eq. (2.2) into Eq. (B.4) (see Mathematica scripts in the accompanying Github repository at https://github.com/marikawakatsu/CooperationPolarization2 for detailed calculations). We obtain:

$$\left\langle \Delta x_{CC}^{\text{sel}} \right\rangle_0 = - \left\langle \Delta x_{DD}^{\text{sel}} \right\rangle_0 = \frac{1}{4} M (b(g' - h') + c(h' - z')) , \tag{B.5}$$

$$\left\langle \Delta x_{CD}^{\text{sel}} \right\rangle_0 = - \left\langle \Delta x_{DC}^{\text{sel}} \right\rangle_0 = \frac{1}{4} M (b(g - h) + c(h - z)) , \tag{B.6}$$

where, for convenience, we have defined the following correlations (see, e.g., Tarnita et al., 2009a):

$$
\begin{aligned}
y &= \mathbb{P}\left(s_{ia} = s_{ja} \mid i \neq j\right) \ , \\
z &= \left\langle h_{ik} h_{jk} \mid i \neq j \right\rangle_0 \ , \\
g &= \left\langle h_{ik} h_{jk} \mathbb{1}_{s_{ia}=s_{ja}} \mid i \neq j \right\rangle_0 \ , \\
h &= \left\langle h_{ik} h_{jk} \mathbb{1}_{s_{ia}=s_{\ell a}} \mid i \neq j \neq \ell \right\rangle_0 \ , \\
z' &= \left\langle |h_{ik} h_{jk}| \mid i \neq j \right\rangle_0 \ , \\
g' &= \left\langle |h_{ik} h_{jk}| \mathbb{1}_{s_{ia}=s_{ja}} \mid i \neq j \right\rangle_0 \ , \\
h' &= \left\langle |h_{ik} h_{jk}| \mathbb{1}_{s_{ia}=s_{\ell a}} \mid i \neq j \neq \ell \right\rangle_0 \ .
\end{aligned}
\tag{B.7}
$$

When $M = K$, every individual cares about every issue ($|h_{ik}| = 1$ for all i, k). This means $z' = 1$ and $g' = h' = y$, which reduce Eq. (B.5) to

$$
\left\langle \Delta x^{\text{sel}}_{CC} \right\rangle_0 = -\left\langle \Delta x^{\text{sel}}_{DD} \right\rangle_0 = -\frac{1}{4} Mc(1-y) < 1 \ , \tag{B.8}
$$

meaning that, when $M = K$, selection never favors CC and always favors DD.

Time to most recent common ancestor (MRCA). To compute (B.7), we use coalescent theory as described in (Antal et al., 2009a,b; Fu et al., 2012; Tarnita et al., 2009a). The key idea of this approach is that, given two or more individuals, we can always find their most recent common ancestor (MRCA) by tracing back their lineages.

We first compute the probability that two randomly chosen individuals i and j have their MRCA at time $\Delta_{ij} = t$. Following the derivation in (Antal et al., 2009b), the probability that i and j share an ancestor in the step immediately prior (i.e., "parent") is

$$
\mathbb{P}\left(\Delta_{ij} = 1\right) = \left(1 - \frac{1}{N}\right) \cdot 2 \cdot \frac{1}{N} \cdot \frac{1}{N-1} \cdot \gamma = \frac{2\gamma}{N^2} \ ,
$$

where $\gamma = \gamma(N, p)$ is the probability that an imitation event occurs between i and j. Specifically, an imitation event occurs with probability 1 if i and j are in the same party or with probability $1 - p$ if they are in different parties. Thus we have

$$\gamma(N, p) = 1 \cdot \frac{N/2 - 1}{N - 1} + (1 - p) \cdot \frac{N/2}{N - 1} = 1 - \frac{Np}{2(N - 1)} .$$

Note that $1 \geq \gamma(N, p) \geq 1 - \frac{Np}{2(N-1)}$, with the lower bound approaching 0.5 as $N \to \infty$. Then the probability that i and j have their MRCA at time $\Delta_{ij} = t$ is

$$\mathbb{P}(\Delta = t) = \left(1 - \frac{2\gamma}{N^2}\right)^{t-1} \frac{2\gamma}{N^2} .$$

Let $\tau_2 = t/\left(N^2/2\gamma\right)$ be the rescaled time and $\rho_2(\tau_2)$ be its probability density function. By change of variables, the distribution of coalescence times τ_2 in the continuous time limit $(N \to \infty)$ is given by

$$\rho_2(\tau_2) = e^{-\tau_2} .$$

Similarly, the density function $\rho_3(\tau_3, \tau_2)$ for the coalescence time among three randomly chosen individuals (Antal et al., 2009b; Tarnita et al., 2009a) is,

$$\rho_3(\tau_3, \tau_2) = 3e^{-3\tau_3} e^{-\tau_2} .$$

Here, two of the three individuals coalesce first in time $\tau_3 = t_3/(N^2/2\gamma)$, and then this lineage coalesces with the remaining individual after additional time $\tau_2 = t/(N^2/2\gamma)$.

In the rescaled time, i.e., $\tau = t/(N^2/2\gamma)$, the opinion mutation and issue/opinion rates rescale to $\mu = Nu$ and $\nu = Nv$, respectively.

Probability that two individuals have the same agreement strategy. We begin by computing y, the probability that two randomly selected individuals i and j have the same agreement strategy (s_{*a}).

Starting from the MRCA, each lineage mutates with rate $\mu/2$. If neither lineage has mutated after time τ_2 since the MRCA (which occurs with probability $e^{-\mu\tau_2}$), i and j have the same agreement strategy (i.e., $s_{ia} = s_{ja}$). If at least one has mutated (which occurs with probability $1 - e^{-\mu\tau_2}$), i and j have the same agreement strategy with probability $1/2$. Thus, the probability that two randomly chosen individuals have the same agreement strategy after time τ_2 since their MRCA is

$$y(\tau_2) = \mathbb{P}\left(s_{ia} = s_{ja} \mid i \neq j, \Delta_{ij} = \tau_2\right) = e^{-\mu\tau_2} + \frac{1}{2}\left(1 - e^{-\mu\tau_2}\right) = \frac{1}{2}\left(1 + e^{-\mu\tau_2}\right) \ . \quad \text{(B.9)}$$

Using the distribution of coalescence times τ_2 computed above, we obtain

$$y = \int_0^\infty \rho_2(\tau_2) y(\tau_2) \ \mathrm{d}\tau_2 = \int_0^\infty e^{-\tau_2} \cdot \frac{1}{2}\left(1 + e^{-\mu\tau_2}\right) \ \mathrm{d}\tau_2 = \frac{\mu + 2}{2(\mu + 1)}$$

in the continuous time limit ($N \to \infty$).

Average opinion agreement between two individuals. The quantity $z = \left\langle h_{ik} h_{jk} \mid i \neq j \right\rangle_0$ represents the average opinion agreement between two randomly selected individuals i and j on a randomly selected issue k. To compute z, we first compute $z(\tau_2)$, the average opinion agreement between i and j at time τ_2 from their MRCA.

At the MRCA, the probability that i and j both care about and agree on a randomly selected issue k (i.e., $h_{ik} h_{jk} = 1$) is K/M; this is because each individual cares about exactly K of the M issues. From there, each lineage mutates with rate $\nu/2$. If neither lineage has mutated after time τ_2 since the MRCA (with probability $e^{-\nu\tau_2}$), i and j agree on a given issue k (i.e., $h_{ik} = h_{jk}$, which gives $h_{ik} h_{jk} = 1$). If at least one has mutated, i and j still care about issue k with probability K/M, but can have either the same opinion ($h_{ik} = h_{jk}$) or opposite opinions ($h_{ik} = -h_{jk}$) with equal probability. Hence, we can write

$z(\tau_2)$ as

$$\begin{aligned} z(\tau_2) &= \left\langle h_i h_j \mid i \neq j, \Delta_{ij} = \tau_2 \right\rangle_0 \\ &= \frac{K}{M}\left(e^{-\nu\tau_2} + \left(1 - e^{-\nu\tau_2}\right)\left(\frac{K}{2M} - \frac{K}{2M}\right)\right) = \frac{K}{M}e^{-\nu\tau_2} \,. \end{aligned} \tag{B.10}$$

Finally, integrating over all possible coalescence times,

$$z = \int_0^\infty \rho_2(\tau_2) z(\tau_2) \, \mathrm{d}\tau_2 = \int_0^\infty e^{-\tau_2} \cdot e^{-\nu\tau_2} \, \mathrm{d}\tau_2 = \frac{K}{M}\left(\frac{1}{\nu+1}\right) \,.$$

Average opinion agreement between two individuals with the same agreement strategy. Next, $g = \left\langle h_{ik} h_{jk} \mathbb{1}_{s_{ia}=s_{ja}} | i \neq j \right\rangle$ can be interpreted as the average opinion agreement between two randomly selected individuals on a randomly selected issue given that they have a non-zero contribution to the average only if they have the same agreement strategy.

To compute g, we first fix the time to MRCA: $g(\tau_2) = \left\langle h_{ik} h_{jk} \mathbb{1}_{s_{ia}=s_{ja}} \mid i \neq j, \Delta_{ij} = \tau_2 \right\rangle_0$. Since opinion mutations and strategy mutations are independent from $\Delta_{ij} = \tau_2$ onward, we can write $g(\tau_2) = \left\langle h_{ik} h_{jk} \mid i \neq j, \Delta_{ij} = \tau_2 \right\rangle_0 \mathbb{P}\left(s_{ia} = s_{ja} \mid i \neq j, \Delta_{ij} = \tau_2\right) = z(\tau_2) y(\tau_2)$. Thus, in the continuous limit, we substitute in the expressions in Eq. (B.9) and Eq. (B.10) to obtain

$$\begin{aligned} g &= \int_0^\infty \rho_2(\tau_2) z(\tau_2) y(\tau_2) \, \mathrm{d}\tau_2 \\ &= \int_0^\infty e^{-\tau_2} \cdot \frac{1}{2}\left(1 + e^{-\mu\tau_2}\right) \cdot \frac{K}{M} e^{-\nu\tau_2} \, \mathrm{d}\tau_2 = \frac{K}{2M}\left(\frac{1}{\mu+\nu+1} + \frac{1}{\nu+1}\right) \,. \end{aligned}$$

Average opinion agreement between two out of three individuals, provided that a different pair shares the agreement strategy. The quantity $h = \left\langle h_{ik} h_{jk} \mathbb{1}_{s_{ia}=s_{\ell a}} \mid i \neq j \neq \ell \right\rangle$, can be interpreted as follows: given three randomly selected individuals i, j, and ℓ, h is the average agreement between i and j on a randomly

selected issue k, provided that they have a non-zero contribution to the average only if i and ℓ have the same agreement strategy. Mathematically, this is identical to the scenario worked out in (Tarnita et al., 2009a), which considers the three different orders in which i, j, and ℓ can coalesce. In this case, we can write

$$h(\tau_3, \tau_2) = \frac{1}{3}\left(y\left(\tau_3\right) z\left(\tau_3+\tau_2\right) + y\left(\tau_3+\tau_2\right) z\left(\tau_3\right) + y\left(\tau_3+\tau_2\right) z(\tau_3+\tau_2)\right) \ .$$

Integrating this expression, we obtain

$$\begin{aligned} h &= \int_0^\infty \int_0^\infty h\left(\tau_3, \tau_2\right) \, \mathrm{d}\tau_3 \, \mathrm{d}\tau_2 \\ &= \frac{K}{2M}\left[\frac{\mu+3}{(\mu+2)(\nu+1)} - \frac{\mu(\mu+3)}{2(\mu+1)(\mu+2)(\mu+\nu+3)} + \frac{1}{2(\mu+\nu+1)}\right] \ . \end{aligned}$$

Probability that two individuals care about a given issue. The quantity $z' = \left\langle |h_{ik} h_{jk}| \mid i \neq j \right\rangle_0$ represents the probability that two randomly selected individuals i and j both care about a randomly selected issue k.

At the MRCA, the probability that i and j both care about a given issue k (i.e., $|h_{ik}| = |h_{jk}| = 1$) is K/M. From there, each lineage mutates with rate $\nu/2$. If neither lineage has mutated after time τ_2 since the MRCA, then $|h_{ik}| = |h_{jk}| = 1$. If at least one has mutated, i and j still care about issue k with probability K/M. Hence, we can write $z'(\tau_2)$ as

$$z'(\tau_2) = \frac{K}{M}\left(e^{-\nu\tau_2} + \left(1 - e^{-\nu\tau_2}\right)\frac{K}{M}\right) \ . \tag{B.11}$$

Finally, integrating over all possible coalescence times,

$$z' = \int_0^\infty \rho_2(\tau_2) z'(\tau_2) \, \mathrm{d}\tau_2 = \frac{K}{M}\left[\frac{K}{M} + \frac{M-K}{M} \cdot \frac{1}{\nu+1}\right] \ .$$

Probability that two individuals care about a given issue and have the same agreement strategy. The quantity $g' = \left\langle |h_{ik}h_{jk}| \mathbb{1}_{s_{ia}=s_{ja}} | i \neq j \right\rangle$ can be interpreted as the probability that two randomly selected individuals both care about a randomly selected issue and have the same agreement strategy. We compute g' by first fixing the time to MRCA:

$$g'(\tau_2) = \left\langle |h_{ik}h_{jk}| \mid i \neq j, \Delta_{ij} = \tau_2 \right\rangle_0 \mathbb{P}\left(s_{ia} = s_{ja} \mid i \neq j, \Delta_{ij} = \tau_2\right) = z'(\tau_2)y(\tau_2).$$

Substituting Eq. (B.9) and Eq. (B.10) and integrating over coalescence times τ_2, we obtain

$$g' = \int_0^\infty \rho_2(\tau_2) z'(\tau_2) y(\tau_2)\, \mathrm{d}\tau_2 = \frac{K}{2M}\left[\frac{K}{M}\cdot\frac{2+\mu}{1+\mu} + \frac{M-K}{M}\cdot\left(\frac{1}{\mu+\nu+1} + \frac{1}{\nu+1}\right)\right].$$

Probability that two out of three individuals care about a given issue, provided that a different pair shares the agreement strategy. The last quantity $h' = \left\langle |h_{ik}h_{jk}| \mathbb{1}_{s_{ia}=s_{\ell a}} \mid i \neq j \neq \ell \right\rangle$ can be interpreted as follows: given three randomly selected individuals i, j, and ℓ, h' is the probability that i and j both care about a randomly selected issue k, provided that i and ℓ share the agreement strategy. Similarly to h, we can write $h'(\tau_3, \tau_2) = \frac{1}{3}\left(y\left(\tau_3\right) z'\left(\tau_3+\tau_2\right) + y\left(\tau_3+\tau_2\right) z'\left(\tau_3\right) + y\left(\tau_3+\tau_2\right) z'(\tau_3+\tau_2)\right)$. Integrating this expression, we obtain

$$\begin{aligned} h' &= \int_0^\infty \int_0^\infty h'\left(\tau_3, \tau_2\right) \mathrm{d}\tau_3\, \mathrm{d}\tau_2 \\ &= \frac{K}{2M}\left[\frac{K}{M}\cdot\frac{2+\mu}{1+\mu} + \frac{M-K}{M}\cdot\left(\frac{\mu+3}{2(\mu+2)(\nu+1)} - \frac{\mu(\mu+3)}{2(\mu+1)(\mu+2)(\mu+\nu+3)} + \frac{1}{2(\mu+\nu+1)}\right)\right]. \end{aligned}$$

Summary: computing the correlations. For $p < 1$, we have obtained

$$y = \frac{\mu+2}{2(\mu+1)}\,, \tag{B.12}$$

$$z = \frac{K}{M}\left(\frac{1}{\nu+1}\right)\,, \tag{B.13}$$

$$g = \frac{K}{2M}\left(\frac{1}{\mu+\nu+1}+\frac{1}{\nu+1}\right)\,, \tag{B.14}$$

$$h = \frac{K}{2M}\left(\frac{\mu+3}{(\mu+2)(\nu+1)}-\frac{\mu(\mu+3)}{2(\mu+1)(\mu+2)(\mu+\nu+3)}+\frac{1}{2(\mu+\nu+1)}\right)\,, \tag{B.15}$$

$$z' = \frac{K}{M}\left[\frac{K}{M}+\frac{M-K}{M}\cdot\frac{1}{\nu+1}\right] = \frac{K^2}{M^2}+\frac{M-K}{M}\cdot z\,, \tag{B.16}$$

$$g' = \frac{K}{2M}\left[\frac{K}{M}\cdot\frac{2+\mu}{1+\mu}+\frac{M-K}{M}\left(\frac{1}{\mu+\nu+1}+\frac{1}{\nu+1}\right)\right] = \frac{K^2}{M^2}\cdot y+\frac{M-K}{M}\cdot g\,, \tag{B.17}$$

$$\begin{aligned} h' &= \frac{K}{2M}\left[\frac{K}{M}\cdot\frac{2+\mu}{1+\mu}+\frac{M-K}{M}\left(\frac{\mu+3}{2(\mu+2)(\nu+1)}-\frac{\mu(\mu+3)}{2(\mu+1)(\mu+2)(\mu+\nu+3)}+\frac{1}{2(\mu+\nu+1)}\right)\right] \\ &= \frac{K^2}{M^2}\cdot y+\frac{M-K}{M}\cdot h\,, \end{aligned} \tag{B.18}$$

where $\mu = Nu$ and $\nu = Nv$ are the rescaled strategy mutation and issue/opinion exploration rates, respectively. We note that the quantity y is independent of K and M. Figures B.6 and B.7 compare numerical evaluations of these quantities with simulated quantities (for $M = K = 1$ and $M = 3, K = 2$, as in Fig. B.3), showing excellent agreement.

B.1.3 Computing the stationary strategy frequencies

Substituting Eqs. (B.12)–(B.18) into Eqs. (B.5) and (B.6), we obtain the change in frequency of each strategy due to selection under neutrality:

$$\begin{aligned} \left\langle \Delta x_{CC}^{\text{sel}} \right\rangle_0 &= -\left\langle \Delta x_{DD}^{\text{sel}} \right\rangle_0 \\ &= -\frac{K\mu\left(-b\nu(M-K)(\mu+\nu+2)+cK\nu(\mu+\nu+2)^2+cM\left(\mu^2+2\mu(\nu+2)+\nu(\nu+3)+3\right)\right)}{8(\mu+1)(\nu+1)M(\mu+\nu+1)(\mu+\nu+3)}\,, \\ \left\langle \Delta x_{CD}^{\text{sel}} \right\rangle_0 &= -\left\langle \Delta x_{DC}^{\text{sel}} \right\rangle_0 \\ &= -\frac{K\mu\left(-b\nu(\mu+\nu+2)+c\left(\mu^2+2\mu(\nu+2)+\nu(\nu+3)+3\right)\right)}{8(\mu+1)(\nu+1)(\mu+\nu+1)(\mu+\nu+3)}\,. \end{aligned}$$

In the limit of small μ ($\mu \ll 1$), these simplify to

$$\begin{aligned}\left\langle \Delta x_{CC}^{\text{sel}} \right\rangle_0 &= -\left\langle \Delta x_{DD}^{\text{sel}} \right\rangle_0 \\ &= -K \cdot \frac{\mu\left(-(1-K/M)b\nu(\nu+2) + (K/M)c\nu(\nu+2)^2 + c(\nu(\nu+3)+3)\right)}{8(\nu+1)^2(\nu+3)} , \end{aligned} \tag{B.19}$$

$$\begin{aligned}\left\langle \Delta x_{CD}^{\text{sel}} \right\rangle_0 &= -\left\langle \Delta x_{DC}^{\text{sel}} \right\rangle_0 \\ &= -K \cdot \frac{\mu(-b\nu(\nu+2) + c(\nu(\nu+3)+3))}{8(\nu+1)^2(\nu+3)} . \end{aligned} \tag{B.20}$$

Substituting these expressions into Eq. (B.3) completes the computation of the stationary strategy frequencies $\langle x_{\mathbf{s}} \rangle$, which are plotted against simulation data in Fig. 2.3D–E and Fig. B.3F–I.

Moreover, we can deduce the following:

1. We can rewrite Eq. (B.19) and Eq. (B.20) as

$$\left\langle \Delta x_{CC}^{\text{sel}} \right\rangle_0 = -\left\langle \Delta x_{DD}^{\text{sel}} \right\rangle_0 = \mu K \cdot R(\nu, b, c) - \mu \frac{K^2}{M} \cdot S(\nu, b, c) , \tag{B.21}$$

$$\left\langle \Delta x_{CD}^{\text{sel}} \right\rangle_0 = -\left\langle \Delta x_{DC}^{\text{sel}} \right\rangle_0 = \mu K \cdot R(\nu, b, c) , \tag{B.22}$$

where

$$\begin{aligned} R(\nu, b, c) &= \frac{b\nu(\nu+2) - c(\nu(\nu+3)+3)}{8(\nu+1)^2(\nu+3)} , \\ S(\nu, b, c) &= \frac{\nu(\nu+2)(b + c(\nu+2))}{8(\nu+1)^2(\nu+3)} . \end{aligned} \tag{B.23}$$

These expressions also provide insight into the relative effects of M and K on effective cooperation (see Fig. 2.2). Whereas K affects the frequencies of both unconditional (CC) and conditional (CD, DC) cooperators, M only affects the former. In fact, since $\nu, b, c > 0$ and therefore $S > 0$, Eq. (B.21) reveals that increasing M always increases the frequency of CC (and decreases that of DD).

However, the positive effect of increasing M is vanishingly small, as it affects CC (and DD) via the term proportional to $1/M$. In contrast, K impacts all frequencies at least linearly, which helps explain why the effects of varying K are much stronger than those of varying M in Fig. 2.2.

2. When $M = K$, Eq. (B.19) reduces to $\langle \Delta x^{\text{sel}}_{CC} \rangle_0 = -\langle \Delta x^{\text{sel}}_{DD} \rangle_0 = -cM\mu/8 < 0$, meaning that (1) selection never favors CC and always favors DD (see also Eq. (B.8)) and that (2) the frequencies of CC and DD are independent of issue/opinion exploration rate ν (v).

   When $M > K$, selection favors CC (and disfavors DD) when $b/c > (z' - h')/(g' - h')$ (see Eq. (B.5)). Substituting in Eqs. (B.12) and (B.16)–(B.18), we can express this condition as

$$\left(\frac{b}{c}\right) > \left(\frac{b}{c}\right)^* = \frac{(K/M)\nu(\nu+2)^2 + (\nu(\nu+3)+3)}{(1-K/M)\nu(\nu+2)}\,, \tag{B.24}$$

   where the RHS is (1) a U-shaped function of ν ($\nu > 0$) and (2) an increasing function of K/M ($0 < K/M \leq 1$). Hence, CC is favored for intermediate values of v if M is sufficiently large relative to K (i.e., K/M is sufficiently small) such that $(b/c) > (b/c)^*$.

3. CD is favored by selection (and DC is disfavored) when $b/c > (z-h)/(g-h)$ (Eq. (B.6)), where the fraction on the RHS is decreasing in ν. Substituting Eqs. (B.12)–(B.15) into Eq. (B.6) and solving for ν, we can express this condition in the limit of small μ as

$$\nu > \nu^*_{CD} = N v^*_{CD} = \frac{-2(b/c) + 3 + \sqrt{4(b/c)^2 - 3}}{2(b/c - 1)}\,. \tag{B.25}$$

Note that, unlike Eq. (B.24), this condition depends only on b and c; in other words, for fixed b and c, the value of v beyond which CD is favored by selection is independent of M and K. For $b = 1, c = 0.2$, the critical threshold v^*_{CD} is 0.00890268.

B.2 Supplementary tables and figures

Table B.1: Parameter settings for model simulations. Baseline values are those used in the simulations unless otherwise indicated.

Parameter	*Description*	*Values (Baseline)*
N	Number of individuals	40
M	Number of available issues	1–5 (1)
K	Number of issues each individual cares about ($K \leq M$)	1–5 (1)
P	Number of parties	2
β	Strength of selection	0.001
p	Partisan bias in mutation	0–1 (0)
u	Strategy mutation rate	0.001
v	Issue and opinion exploration rate	0.001–0.625 (0.001)
ε	Bias attenuation in opinion mutation	1

Table B.2: Effect of changing M on effective cooperation. Values shown correspond to the simulated data shown in Fig. 2.2A, for each value of K and v, we compute the relative change (%) in effective cooperation between extreme values of M as $(\text{cooperation}_{M_{\max}} - \text{cooperation}_{M_{\min}})/\text{cooperation}_{M_{\min}} \times 100$.

v	K	$M_{\min}$	$\text{cooperation}_{M_{\min}}$	$M_{\max}$	$\text{cooperation}_{M_{\max}}$	% change
	1	1	0.432	5	0.428	-1.0
	2	2	0.363	5	0.364	0.3
0.001	3	3	0.298	5	0.304	2.1
	4	4	0.253	5	0.254	0.5
	5	5	0.211	5	0.211	0
	1	1	0.452	5	0.461	2.0
	2	2	0.402	5	0.414	3.0
0.005	3	3	0.358	5	0.378	5.4
	4	4	0.322	5	0.329	2.0
	5	5	0.280	5	0.280	0
	1	1	0.474	5	0.520	9.6
	2	2	0.447	5	0.509	14.0
0.025	3	3	0.414	5	0.480	15.9
	4	4	0.385	5	0.422	9.7
	5	5	0.358	5	0.358	0

Table B.3: Effect of changing K on effective cooperation. Values shown correspond to the simulated data shown in Fig. 2.2B. For each value of M and v, we compute the relative change (%) in effective cooperation between extreme values of M as $(\text{cooperation}_{K_{\max}} - \text{cooperation}_{K_{\min}})/\text{cooperation}_{K_{\min}} \times 100$.

v	M	$K_{\min}$	$\text{cooperation}_{K_{\min}}$	$K_{\max}$	$\text{cooperation}_{K_{\max}}$	% change
	1	1	0.432	1	0.432	0
	2	1	0.430	2	0.363	-15.7
0.001	3	1	0.430	3	0.298	-30.8
	4	1	0.434	4	0.253	-41.7
	5	1	0.428	5	0.211	-50.8
	1	1	0.452	1	0.452	0
	2	1	0.461	2	0.402	-12.9
0.005	3	1	0.463	3	0.358	-22.6
	4	1	0.468	4	0.322	-31.0
	5	1	0.461	5	0.280	-39.4
	1	1	0.474	1	0.474	0
	2	1	0.503	2	0.447	-11.1
0.025	3	1	0.507	3	0.414	-18.5
	4	1	0.516	4	0.385	-25.5
	5	1	0.520	5	0.358	-31.2

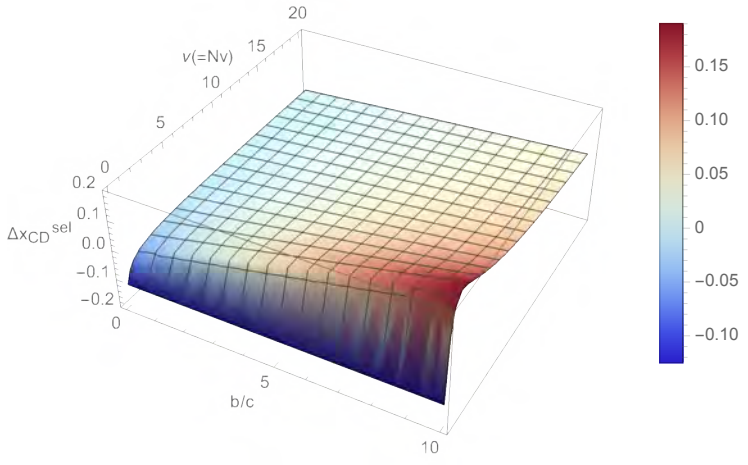

Figure B.1: Change in x_{CD} due to selection as a function of the benefit-to-cost ratio and issue/opinion exploration ($M = K = 1$). We plot Eq. (B.20) as a function of b/c and $\nu = Nv$, fixing $c = 1$ and $\mu = 1$. Colors correspond to the values of $\langle \Delta x_{CD}^{sel} \rangle_0$.

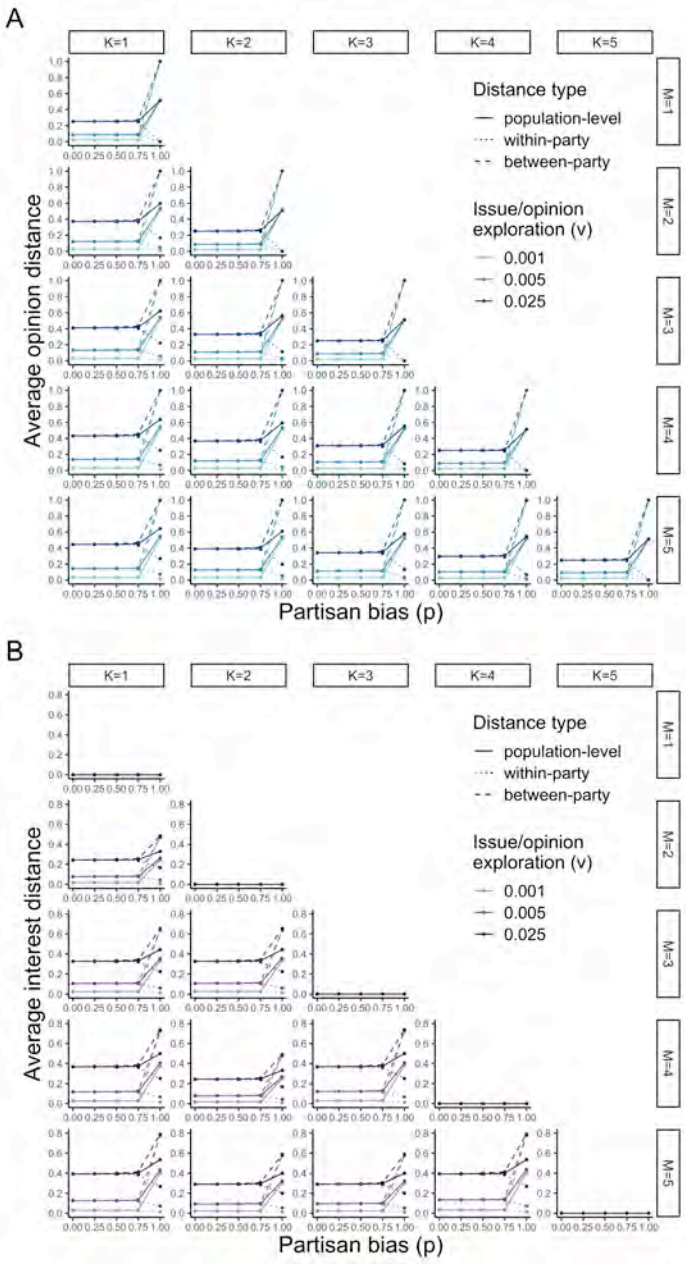

Figure B.2: Opinion and interest alignment as a function of partisan bias. An extended version of Fig. 2.4 with identical metrics. For each parameter setting, we ran an ensemble of 150 simulations with population size $N = 40$, each lasting 2×10^7 generations. First 10% of each simulation were disregarded to account for potential initialization effects. Each circle within a panel represents the mean value ($\pm$SD) of the corresponding metric averaged across generations and across the ensemble. Solid lines indicate values computed for the full population; dotted and dashed lines indicate values within and between parties, respectively. Values of M and K are as indicated. Values of M, K, and v are as indicated. See Table B.1 for other parameter values.

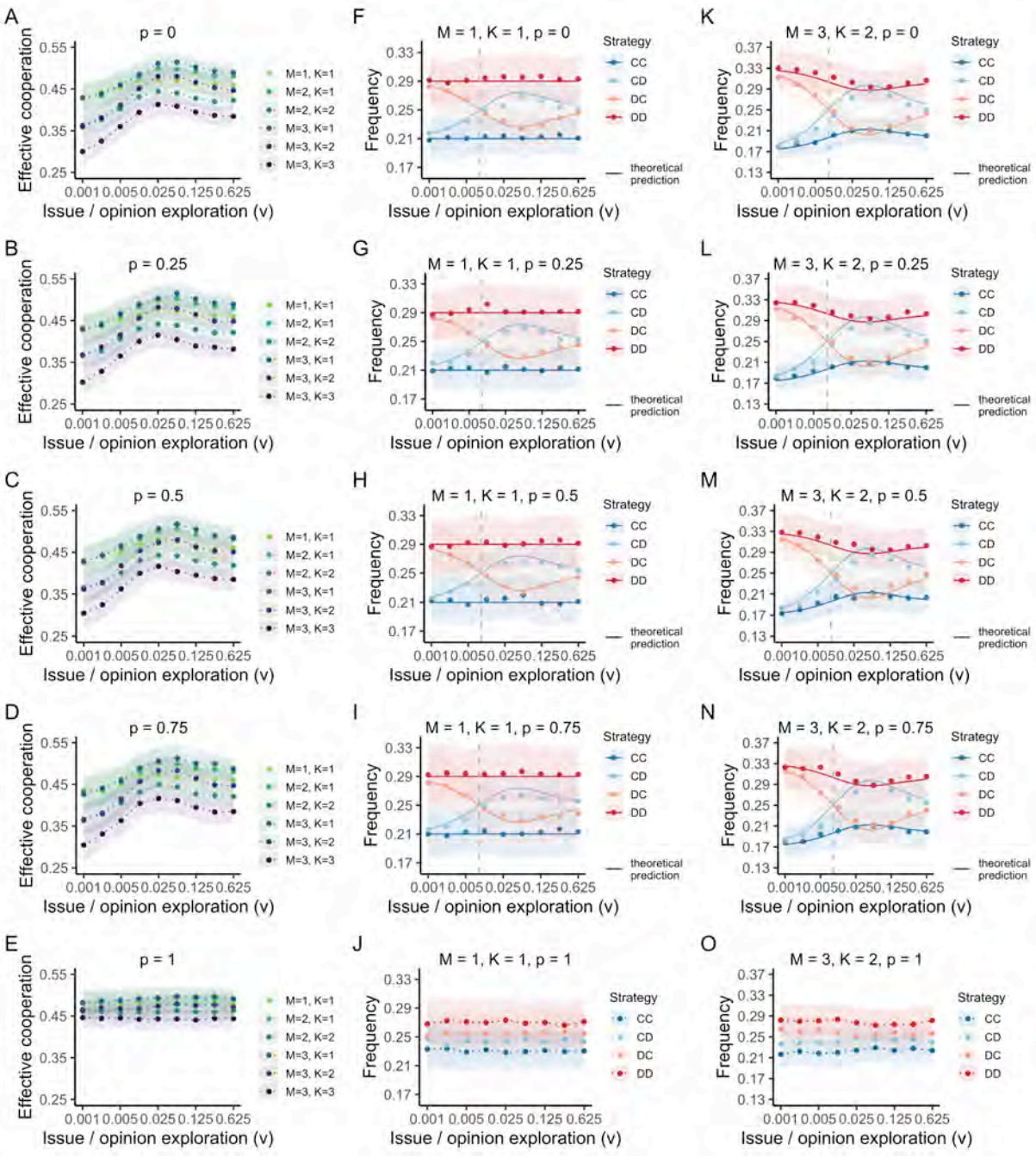

Figure B.3: Interplay between issue/opinion exploration and partisan bias. An extended version of Fig. 2.3. A, C, E and K, M, P are identical to Fig. 2.3A–C and Fig. 2.3D–F, respectively, with $M = 3, K = 2$. The middle column (F–J) show cases with $M = 1, K = 1$. For each parameter setting, we ran an ensemble of 150 simulations with population size $N = 40$, each lasting 2×10^7 generations. We disregarded the first 10% of the generations in each simulation to account for potential initialization effects. All other parameters and the quantities plotted are as in Fig. 2.3.

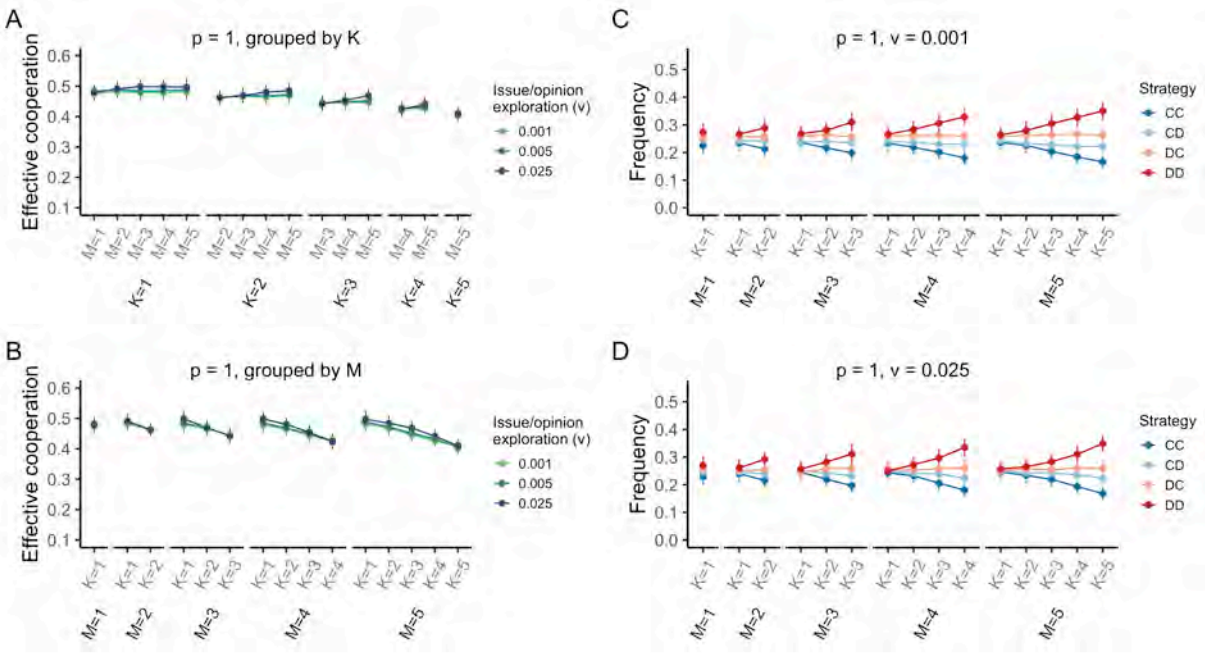

Figure B.4: Evolutionary dynamics of cooperation. An extended version of Fig. 2.2, with A–D mirroring Fig. 2.2A–D, respectively, but with $p = 1$. For each parameter setting, we ran an ensemble of 150 simulations with population size $N = 40$, each lasting 2×10^7 generations. We disregarded the first 10% of the generations in each simulation to account for potential initialization effects. All other parameters and the quantities plotted are as in Fig. 2.2.

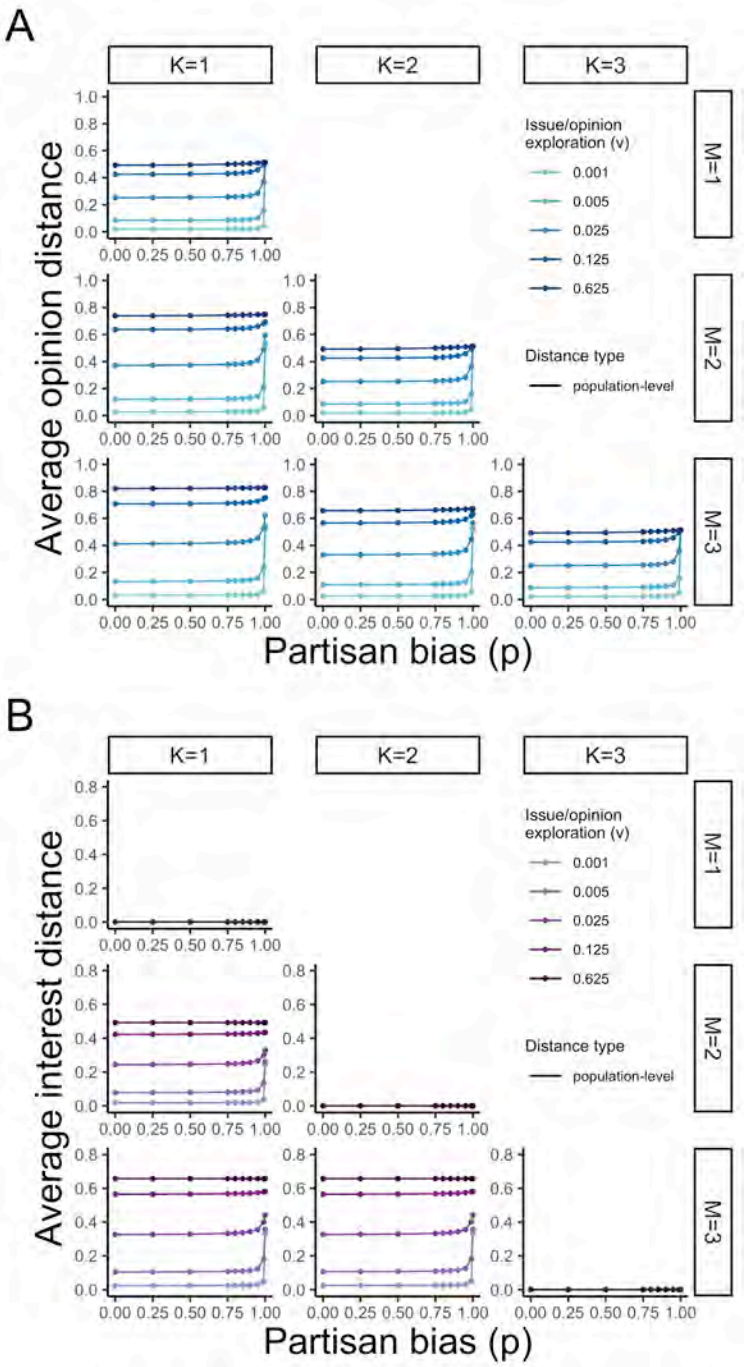

Figure B.5: Opinion and interest alignment as a function of partisan bias. Average opinion distance (A) and average interest distance (B) measured in the full population. Values shown are computed from the same set of simulations as in Fig. 2.4. Each circle within a panel represents the mean value (±SD) of the corresponding metric averaged across generations and across the ensemble. See *Materials and methods* for definitions and the normalization procedure and Table B.1 for parameter values.

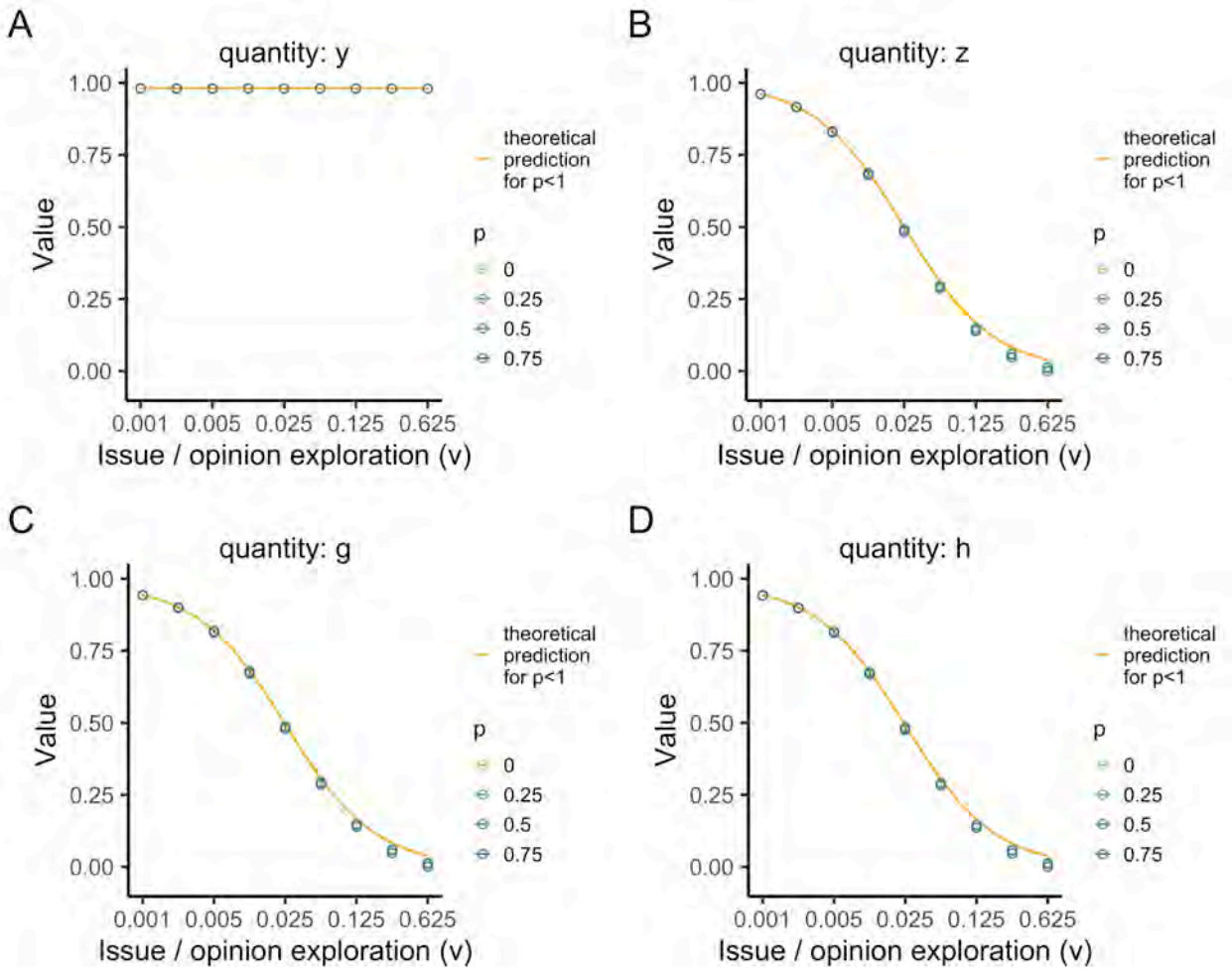

Figure B.6: Comparison between simulated and analytically derived values of y, z, g, and h at neutrality ($M = K = 1$). For each parameter setting, we ran an ensemble of 200 simulations under neutral selection ($\beta = 0$) with population size $N = 40$, each lasting 2×10^7 generations. Each circle represents the mean value ($\pm$SD) the quantity indicated in the corresponding title, averaged across the ensemble. Colors indicate values of partisan bias p; see Table B.1 for all other parameters. Each solid line represents the theoretical prediction for the corresponding quantity (see *Computing the expected change in strategy frequency due to selection*).

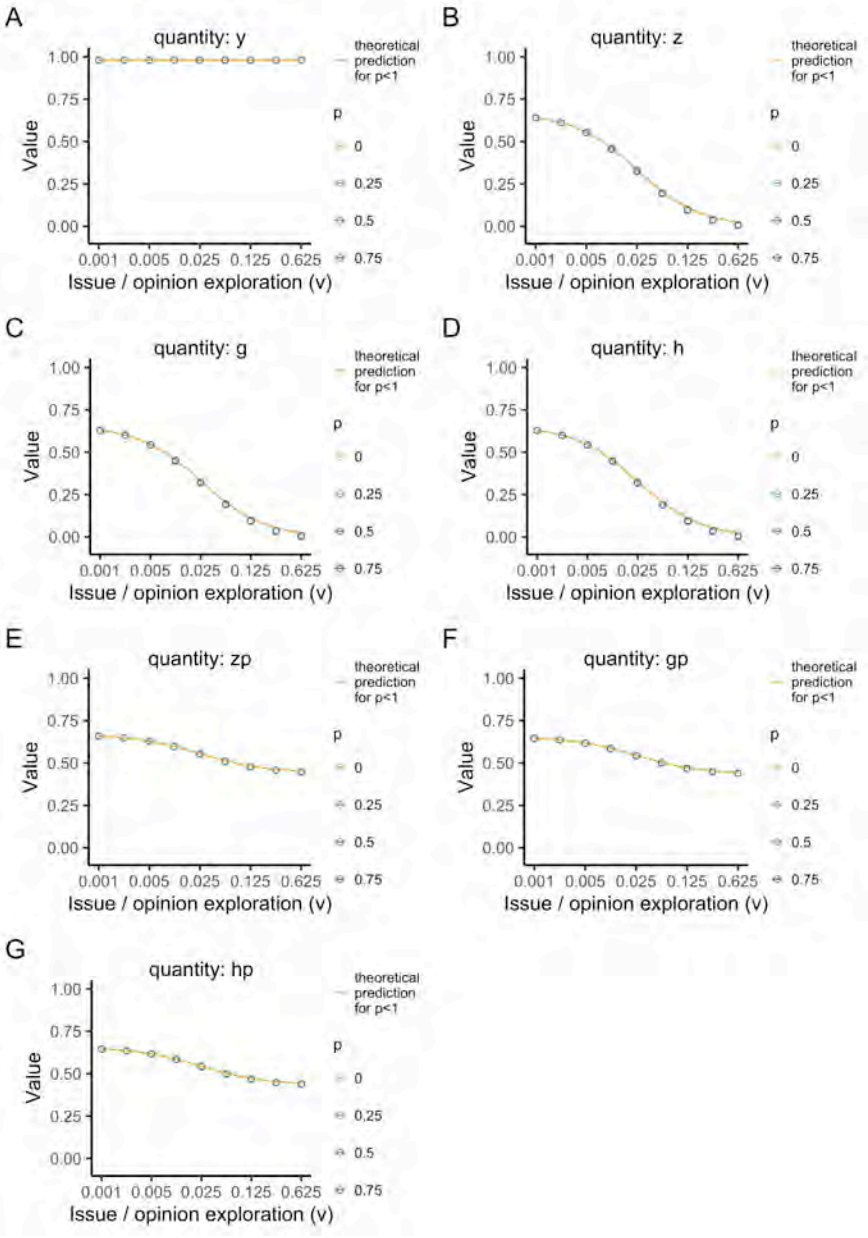

Figure B.7: Comparison between simulated and analytically derived values of y, z, g, h, z', g', and h' at neutrality ($M = 3, K = 2$). For each parameter setting, we ran an ensemble of 75 simulations under neutral selection ($\beta = 0$) with population size $N = 40$, each lasting 2×10^7 generations. Each circle represents the mean value ($\pm$SD) the quantity indicated in the corresponding title, averaged across the ensemble. Colors indicate values of partisan bias p; see Table B.1 for all other parameters. Each solid line represents the theoretical prediction for the corresponding quantity (see *Computing the expected change in strategy frequency due to selection*).

Appendix C

Supplementary materials for Chapter 3

C.1 Supplementary figures

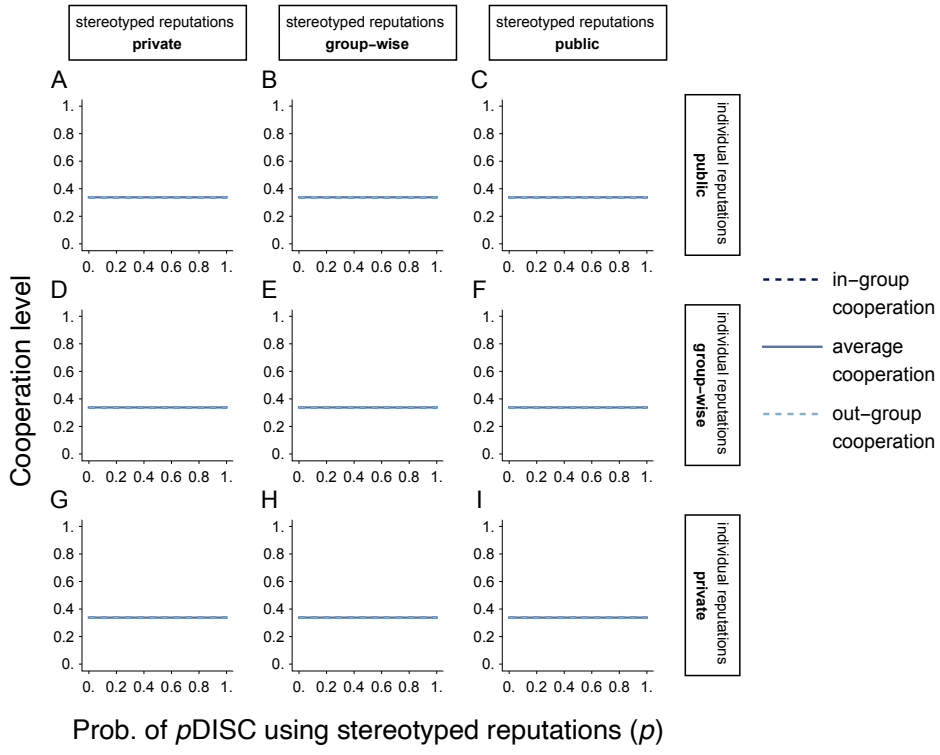

Figure C.1: Cooperation in monomorphic populations of pDISC under Scoring. Same as in Fig. 3.1 but under the Scoring norm.

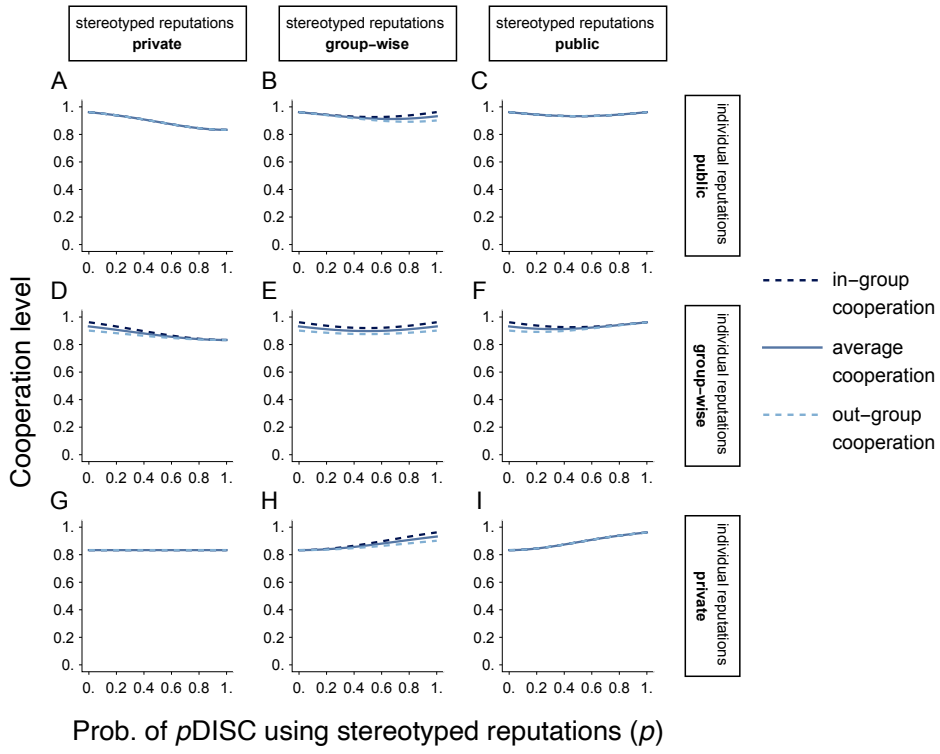

Figure C.2: Cooperation in monomorphic populations of pDISC under Simple Standing. Same as in Fig.3.1 but under the Simple Standing norm.

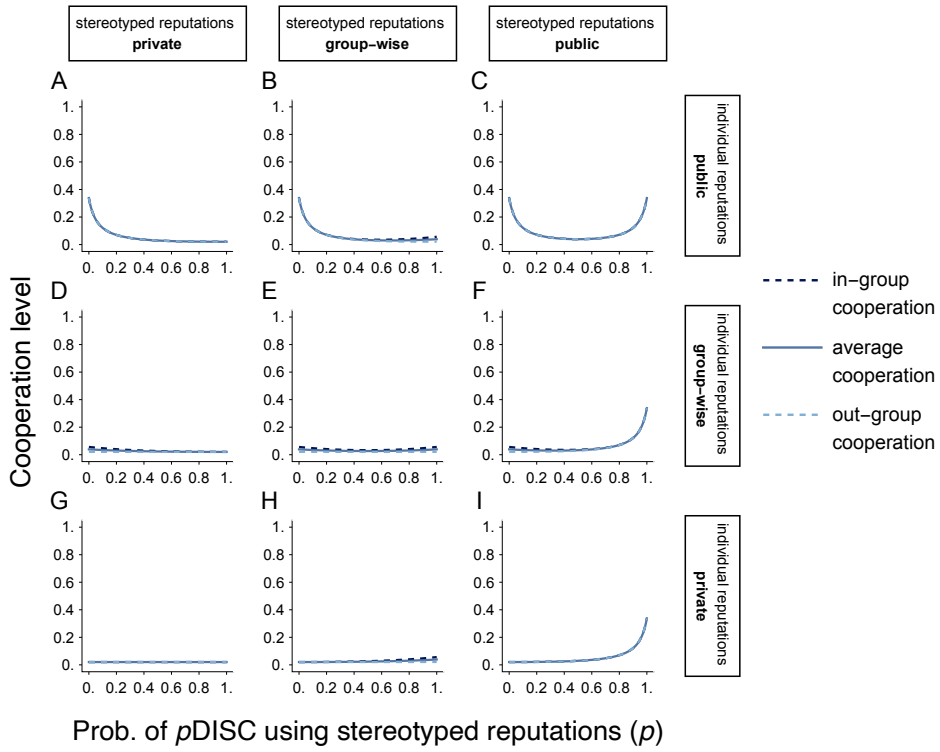

Figure C.3: Cooperation in monomorphic populations of pDISC under Shunning. Same as in Fig.3.1 but under the Shunning norm.

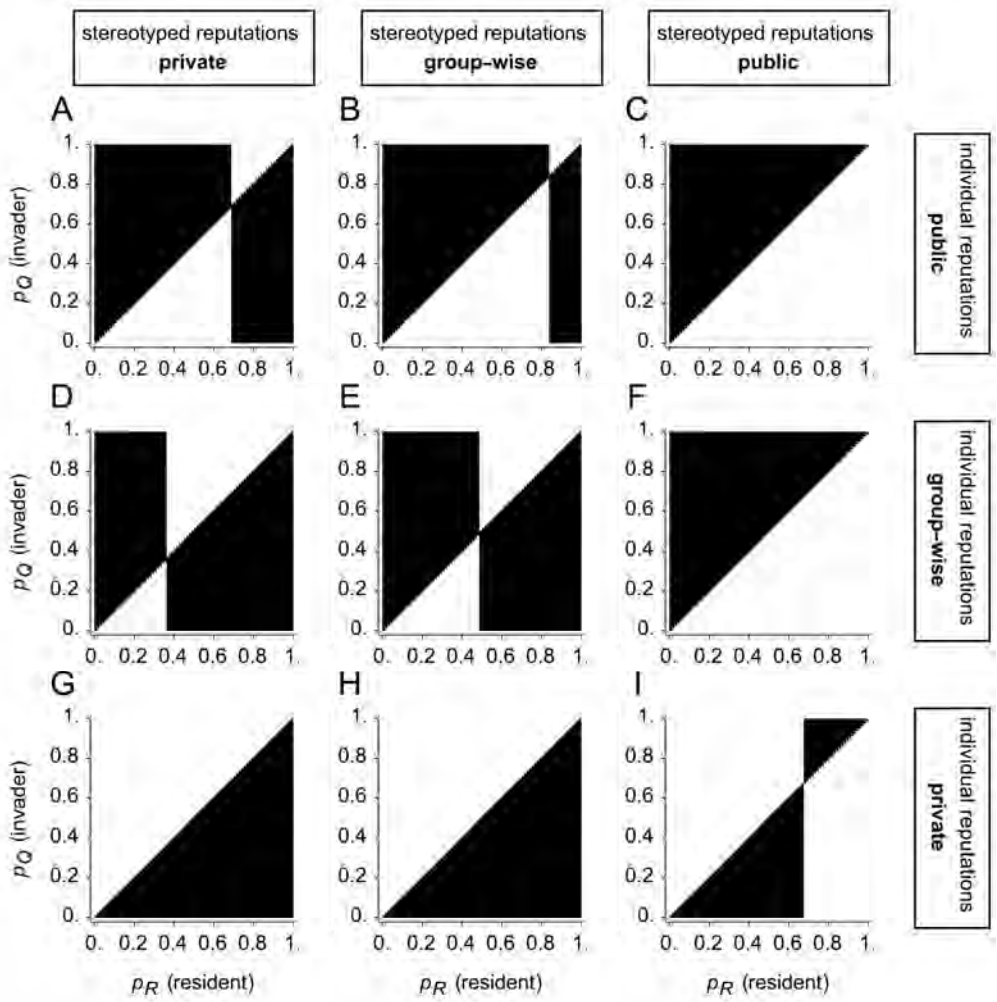

Figure C.4: Pairwise invasibility of pDISC strategies under Stern Judging. An expanded version of Fig. 3.2; C, E, and G correspond to Fig. 3.2A, B, C, respectively. Each panel corresponds to a combination of monitoring systems for individual reputations (rows) and stereotyped reputations (columns). Parameters are as in Fig. 3.2.

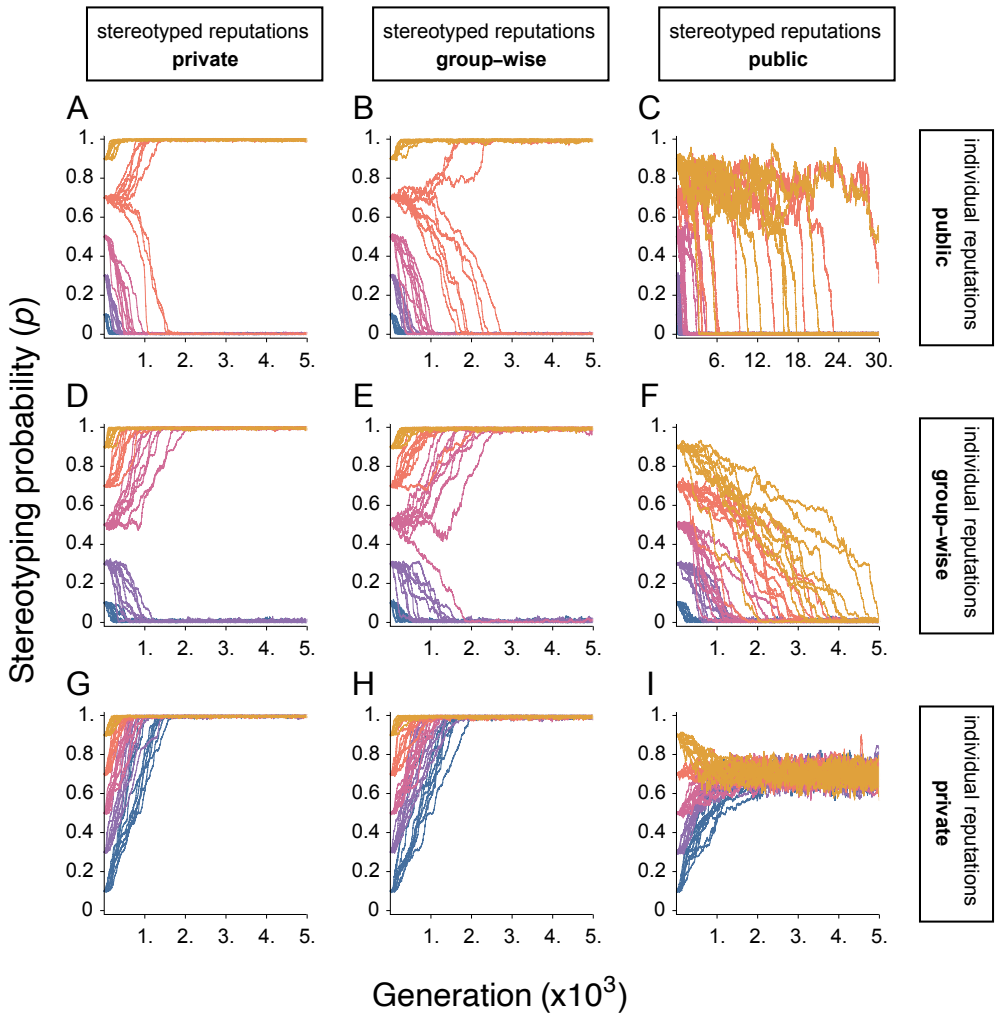

Figure C.5: Stochastic evolutionary dynamics of pDISC strategies under Stern Judging. An expanded version of Fig.3.2; C, E, and G correspond to Fig.3.2D, E, and F, respectively. Each panel corresponds to a combination of monitoring systems for individual reputations (rows) and stereotyped stereotypes (columns). Parameters are as in Fig.3.2.

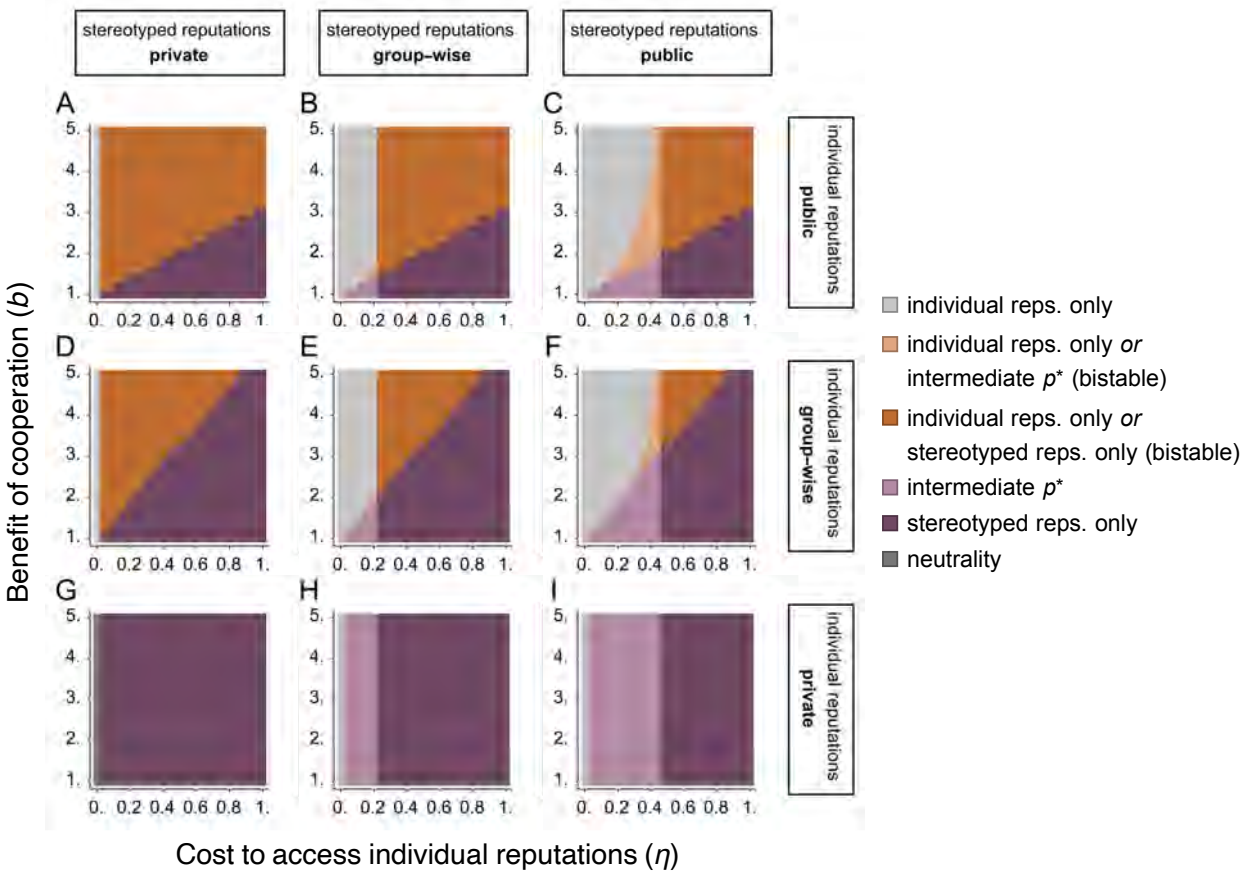

Figure C.6: Costly reputations promote the use of stereotypes. An expanded version of Fig. 3.4; F corresponds to Fig. 3.4A. Each panel corresponds to a combination of monitoring systems for reputations (rows) and stereotypes (columns). Parameters are as in Fig. 3.4.

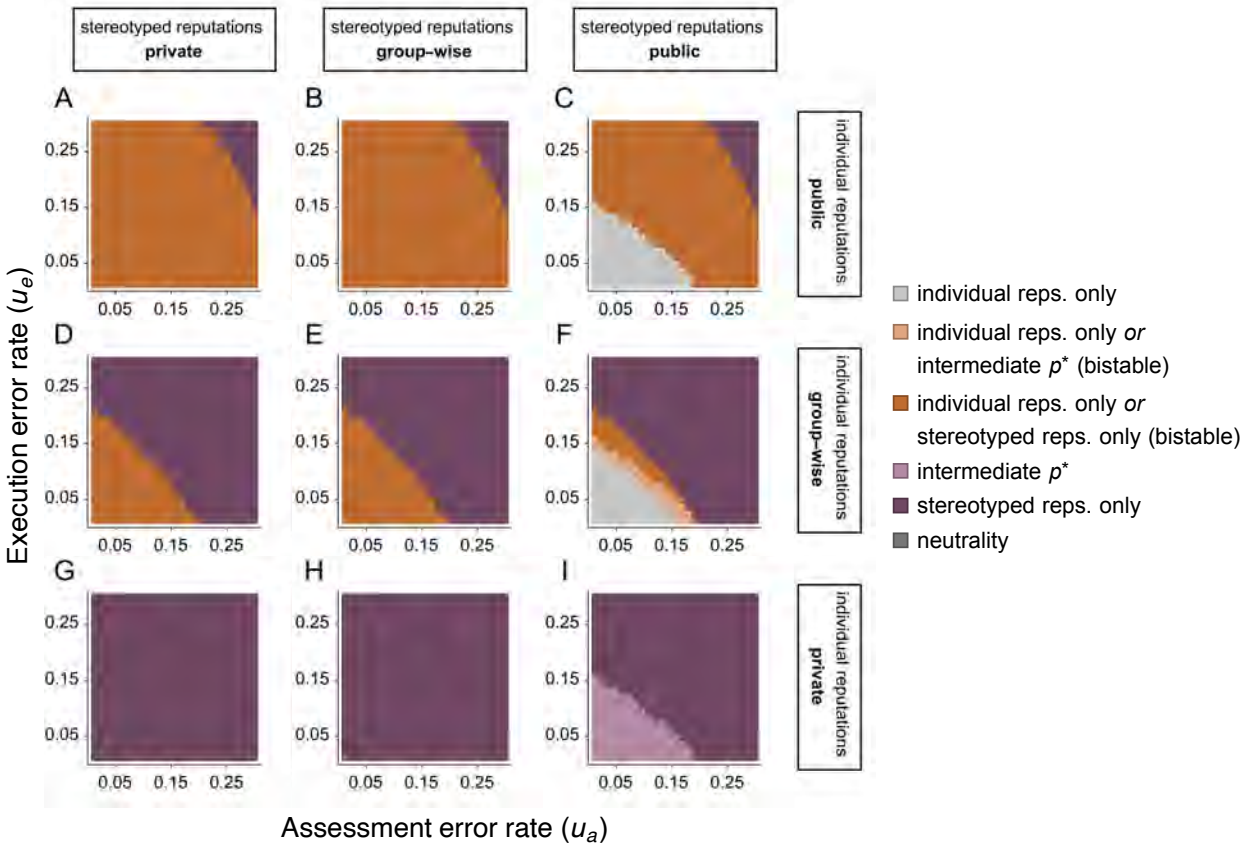

Figure C.7: Errors in strategy execution and reputation assessment promote the use of stereotypes. An expanded version of Fig.3.4; F corresponds to Fig.3.4B. Each panel corresponds to a combination of monitoring systems for individual reputations (rows) and stereotyped reputations (columns). Parameters are as in Fig.3.4.

Appendix D

Supplementary materials for Chapter 4

D.1 Supplementary analyses

In this section, we prove a set of linear stability results that generalize Theorem 4.1 in the main text. Our generalizations account for (a) nonlinear features and (b) multiple updates per round.

Throughout this section, we consider a utility function of the form

$$u_{ij}(\mathbf{s}) = \sum_{\ell=1}^{k} \beta_\ell \phi_{ij}^\ell(\mathbf{s}) \ , \tag{D.1}$$

where each $\phi^\ell : \mathbb{R}^n \mapsto \mathbb{R}^{n \times n}$ is a smooth *feature map*; $\beta_\ell \in \mathbb{R}$ is a *preference parameter* indicating relative importance of the ℓth feature; and $\phi_{ij}^\ell(\mathbf{s})$ is the ijth entry of $\phi^\ell(\mathbf{s})$. We collect the parameters β in a vector $\boldsymbol{\beta} \in \mathbb{R}^k$. The utility function in Eq. (4.4) from the main text is a special case with linear feature map $\phi_{ij}^1(\mathbf{s}) = s_j$, and quadratic feature map, $\phi_{ij}^2(\mathbf{s}) = (s_i - s_j)^2$. We also define the *rate matrix* $\mathbf{G} = [n^{-1} p_{ij}]$, whose $(i,j)th$ entry gives the probability that, in a given time step, node i chosen uniformly at random endorses node j (see Eq. (4.5) in the main text for the definition of p_{ij}).

Since we aim to characterize the linear stability of egalitarian fixed points, we will consider the Jacobian of the rank vector γ evaluated at egalitarian fixed points. We will therefore evaluate the Jacobian at $\mathbf{s}_0 = \theta\mathbf{e}$, where $\theta \in \mathbb{R}$. By definition, $\gamma = n^{-1}\mathbf{G}^T\mathbf{e} = n^{-1}\sum_i \gamma_i$, where γ_i is the ith column of $\mathbf{G}$. Differentiating and applying the chain rule, we have

$$\frac{\partial\gamma(\mathbf{s}_0)}{\partial\mathbf{s}} = \sum_i \left(\mathbf{\Gamma}_i - \gamma_i\gamma_i^T\right)\sum_{\ell=1}^{k}\beta_\ell\frac{\partial\phi_i^\ell}{\partial\mathbf{s}} \ ,$$

where $\mathbf{\Gamma}_i = \text{diag}\gamma_i$ and $\phi_{i\cdot}^\ell(\mathbf{s}_0)$ is the ith row of the ℓth feature map evaluated at $\mathbf{s}_0$. At $\mathbf{s}_0 = \theta\mathbf{e}$, $\mathbf{G} = n^{-1}\mathbf{E}$. It follows that $\gamma_i = n^{-1}\mathbf{e}$ and $\mathbf{\Gamma}_i = n^{-1}\mathbf{I}$. We thus have

$$\frac{\partial\gamma(\mathbf{s}_0)}{\partial\mathbf{s}} = n^{-1}(\mathbf{I} - n^{-1}\mathbf{E})\sum_{i=1}^{n}\sum_{\ell=1}^{k}\beta_\ell\frac{\partial\phi_{i\cdot}^\ell(\mathbf{s}_0)}{\partial\mathbf{s}} \triangleq \mathbf{M}(\mathbf{s}_0;\boldsymbol{\beta}) \ . \tag{D.2}$$

We will express our primary results in terms of this matrix.

When writing proofs involving dynamics, we will typically repress the time-argument of quantities like $\mathbf{s}$ and $\mathbf{A}$. When time step t is implied, we will use the somewhat informal notation $\delta\mathbf{s} = s(t+1) - \mathbf{s}(t)$ and $\delta\mathbf{A} = \mathbf{A}(t+1) - \mathbf{A}(t)$ to denote the increments of these and other quantities in the current time step.

D.1.1 Degree scores

Theorem D.1 (Stable Egalitarianism with Degree Scores)**.** *When $\sigma(\mathbf{A}) = \mathbf{s} = \mathbf{A}^T\mathbf{e}$, the vector $\mathbf{s}_0 = d\mathbf{e}$ is a root of $\mathbf{f}$, where $d = \frac{m}{n}$, and is the only egalitarian root. Furthermore, $\mathbf{s}_0$ is linearly stable in the long-memory limit if and only if $\mathbf{M}(\mathbf{s}_0;\boldsymbol{\beta})$ has eigenvalues strictly smaller than $\frac{1}{m}$.*

Proof. We first derive the functional form of $\mathbf{f}$. We can write

$$\begin{aligned}\mathbb{E}[\mathbf{s}(t+1)|\mathbf{A}(t)] &= \mathbb{E}[\mathbf{A}(t+1)|\mathbf{A}(t)]^T\mathbf{e} \\ &= \lambda\mathbf{A}(t)\mathbf{e} + (1-\lambda)\mathbb{E}[\boldsymbol{\Delta}(t)]^T\mathbf{e} \\ &= \lambda\mathbf{A}(t)\mathbf{e} + (1-\lambda)mn^{-1}\mathbf{G}(t)^T\mathbf{e}\,.\end{aligned}$$

Inserting this expression into Eq. (4.6), and recognizing $n^{-1}\mathbf{G}(t)\mathbf{e} = \boldsymbol{\gamma}(t)$, we have

$$\mathbf{f}(\mathbf{s}) = mn^{-1}\mathbb{E}[\mathbf{G}]\mathbf{e} - \mathbf{A}(t)\mathbf{e} = m\boldsymbol{\gamma} - \mathbf{s}\,.$$

We can now check that $\mathbf{s}_0$ is indeed the unique egalitarian root of $\mathbf{f}$. Suppose that $\mathbf{s} = s\mathbf{e}$ for some scalar s. Then,

$$\mathbf{f}(\mathbf{s}) = m\boldsymbol{\gamma}(\mathbf{s}) - \mathbf{s} = (mn^{-1} - s)\mathbf{e}\,,$$

which is only equal to zero when $s = \frac{m}{n}$, as needed.

Now computing derivatives, we have

$$\frac{\partial\mathbf{f}(\mathbf{s})}{\partial\mathbf{s}} = m\mathbf{M}(\mathbf{s};\boldsymbol{\beta}) - \mathbf{I}\,.$$

This matrix has strictly negative eigenvalues provided that the eigenvalues of $\mathbf{M}(\mathbf{s_0};\boldsymbol{\beta})$ are strictly smaller than $\frac{1}{m}$, completing the proof. □

Corrolary 1. *Using the Root-Degree score function,* $\mathbf{s}_0 = \frac{m}{n}\mathbf{e}$ *is a linearly stable fixed point of* $\mathbf{f}$ *if and only if* $\beta < 2\sqrt{\frac{n}{m}}$.

Proof. It is convenient to treat the operation of taking the square root as part of the feature map, rather than part of the score function. We therefore suppose that s_j is the

in-degree of node j and that $\phi_j(\mathbf{s}) = \sqrt{s_j}$. Computing from (D.2), we obtain

$$\mathbf{M}(\mathbf{s}_0; \boldsymbol{\beta}) = \frac{1}{2}\frac{n^{-1}}{\sqrt{d}}\beta(\mathbf{I} - n^{-1}\mathbf{E}) \; .$$

This matrix again has a zero eigenvalue associated with the direction $\mathbf{e}$. For any direction $\mathbf{v} \perp \mathbf{e}$, there is an eigenvalue $\frac{1}{2}\frac{n^{-1}}{\sqrt{d}}\beta$. From Theorem D.1, $\mathbf{s}_0$ will be linearly stable provided that

$$\frac{1}{m} > \frac{1}{2}\frac{n^{-1}}{\sqrt{d}}\beta \; .$$

or

$$\beta < 2\sqrt{d}\frac{n}{m} = 2\sqrt{\frac{n}{m}} \; ,$$

as required. □

D.1.2 PageRank scores

The PageRank score (Brin and Page, 1998; Page et al., 1999) is the solution $\mathbf{s}$ of the linear system

$$\left[\alpha \mathbf{A}^T (\mathbf{D}^o)^{-1} + (1-\alpha) n^{-1} \mathbf{E}\right] \mathbf{s} = \mathbf{s} \; , \tag{D.3}$$

where $\mathbf{D}^o = \mathrm{diag}(\mathbf{A}\mathbf{e})$. The Perron-Frobenius Theorem (Horn and Johnson, 2012) ensures that $\mathbf{s}$ is strictly positive entrywise. We assume $\mathbf{s}$ to be normalized so that $\mathbf{s}^T\mathbf{e} = n$, which is contrary to the usual normalization $\mathbf{s}^T\mathbf{e} = 1$. This choice amounts to a rescaling of the parameters $\boldsymbol{\beta}$, and does not otherwise impact the analysis.

In the case of PageRank, it is difficult to derive a result for general features and we therefore work directly with the PageRank model with linear features.

Theorem D.2. *The vector $\mathbf{s}_0 = \mathbf{e}$ is the unique egalitarian root of $\mathbf{f}$ under PageRank scores. In the PageRank-Linear model, the egalitarian root is linearly stable if and only if $\beta < \frac{1}{\alpha}$.*

Proof. Uniqueness is a direct consequence of normalization: if $\mathbf{s} = \theta\mathbf{e}$ and $\mathbf{e}^T\mathbf{s} = n$, then we must have $\theta = 1$.

We next obtain a necessary condition describing roots of $\mathbf{f}$. We start with a useful simplification. At any fixed point of $\mathbf{f}$, we must have $\mathbf{D}^o = m\mathbf{I}$. This is because, at any such fixed point, we must have $\mathbf{A} = m\mathbf{G}$, and $n\mathbf{G}$ is row-stochastic. For the purposes of analysis in the long-memory limit, we can therefore consider $\mathbf{s}$ to be defined by the simplified equation

$$\left[\alpha m^{-1} n \mathbf{A}^T + (1-\alpha) n^{-1} \mathbf{E}\right] \mathbf{s} = \mathbf{s} \,. \tag{D.4}$$

In the next time step, we will have

$$\left[\alpha m^{-1} n (\mathbf{A}^T + \delta\mathbf{A}^T) + (1-\alpha) n^{-1} \mathbf{E}\right] (\mathbf{s} + \delta\mathbf{s}) = \mathbf{s} + \delta\mathbf{s} \,.$$

Expanding and canceling yields

$$\left[\alpha m^{-1} n \mathbf{A}^T + (1-\alpha) n^{-1} \mathbf{E}\right] \delta\mathbf{s} + \alpha m^{-1} n (\delta\mathbf{A}^T)\mathbf{s} + o(1-\lambda) = \delta\mathbf{s} \,.$$

The term $o(1-\lambda)$ includes terms involving the product $(\delta\mathbf{A}^T)(\delta\mathbf{s})$, and relies on the fact that $\delta\mathbf{s}$ is a smooth function of $\mathbf{A}$. Rearranging and dropping the asymptotic term, we obtain, in the long memory limit,

$$\left[\mathbf{I} - \alpha m^{-1} n \mathbf{A}^T - (1-\alpha) n^{-1} \mathbf{E}\right] \delta\mathbf{s} = \alpha m^{-1} n (\delta\mathbf{A}^T)\mathbf{s} \,. \tag{D.5}$$

This expression gives an implicit representation of $\mathbf{f}$ via the relation $\mathbf{f}(\mathbf{s}, \mathbf{A}) = \lim_{\lambda \to 1} \frac{\mathbb{E}[\delta\mathbf{s}]}{1-\lambda}$. We can therefore enforce $\mathbf{f}(\mathbf{s}, \mathbf{A}) = \mathbf{0}$ by setting $\mathbb{E}[\delta\mathbf{s}] = \mathbf{0}$,

obtaining the necessary condition $\mathbb{E}[\delta \mathbf{A}^T]\mathbf{s} = \mathbf{0}$ for roots of $\mathbf{f}$. Expanding this condition yields,

$$\mathbf{0} = \mathbb{E}[\delta \mathbf{A}^T]\mathbf{s} = (1-\lambda)(\mathbf{G}^T - \mathbf{A}^T)\mathbf{s} \, .$$

Inserting (D.4) and rearranging yields the nonlinear system

$$\left[\mathbf{G}^T + \alpha^{-1}(1-\alpha)n^{-2}\mathbf{E}\right]\mathbf{s} = \alpha^{-1}n^{-1}\mathbf{s} \, . \tag{D.6}$$

The largest eigenvalue of the matrix on the lefthand side is $\alpha^{-1}n^{-1}$. This allows us to numerically solve (D.6) iteratively, by alternating between solving for $\mathbf{s}$ via a standard eigenvalue solver and updating $\mathbf{G}$ with the new value of $\mathbf{s}$. This is the method implemented in the accompanying software and used to generate equilibria in Fig. 4.3.

In order to derive the linear stability criterion, we divide both sides of (D.5) by $1-\lambda$ and differentiate with respect to $\mathbf{s}$, obtaining

$$\left[\mathbf{I} - \alpha m^{-1} n \mathbf{A}^T - (1-\alpha)n^{-1}\mathbf{E}\right]\mathbf{J}(\mathbf{s}) = \alpha m^{-1} n \frac{\partial}{\partial \mathbf{s}}\left[\mathbf{G}^T\mathbf{s} - \mathbf{A}^T\mathbf{s}\right] \, .$$

After inserting (D.4) and simplifying, we have

$$\begin{aligned}
&\left[\mathbf{I} - \alpha m^{-1} n \mathbf{A}^T - (1-\alpha)n^{-1}\mathbf{E}\right]\mathbf{J}(\mathbf{s}) \\
&= \alpha m^{-1} n \frac{\partial}{\partial \mathbf{s}}\left[\mathbf{G}^T\mathbf{s} - \alpha^{-1}mn^{-1}\mathbf{s} + \alpha^{-1}(1-\alpha)mn^{-2}\mathbf{E}\mathbf{s}\right] \\
&= \alpha m^{-1} n \frac{\partial}{\partial \mathbf{s}}\left[\mathbf{G}^T\mathbf{s} - \alpha^{-1}mn^{-1}\mathbf{s}\right] \, .
\end{aligned}$$

The second line follows from the normalization of $\mathbf{s}$, which implies that $\mathbf{E}\mathbf{s} = n\mathbf{e}$, a constant vector which does not depend on $\mathbf{s}$. Differentiating the righthand side then

yields

$$\left[\mathbf{I} - \alpha m^{-1} n \mathbf{A}^T - (1-\alpha) n^{-1} \mathbf{E}\right] \mathbf{J}(\mathbf{s}) = \alpha m^{-1} n \left[\mathbf{G}^T + (\mathbf{e}^T \mathbf{s}) m n^{-1} \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}\right] - \mathbf{I}$$
$$= \alpha m^{-1} n \left[\mathbf{G}^T + m \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}\right] - \mathbf{I} \,.$$

Evaluated at the egalitarian solution $\mathbf{s}_0 = \mathbf{e}$, this becomes

$$\left[\mathbf{I} - \alpha m^{-1} n \mathbf{A}^T - (1-\alpha) n^{-1} \mathbf{E}\right] \mathbf{J}(\mathbf{s}_0) = \alpha m^{-1} n^{-1} \mathbf{E} + \alpha \mathbf{M}(\mathbf{s}_0; \boldsymbol{\beta}) - \mathbf{I} \,.$$

To complete the argument, we note that, at the egalitarian solution of our model dynamics, $\mathbf{A} = n^{-2}\mathbf{E}$. Inserting and simplifying, we have

$$\left[\mathbf{I} - \alpha m^{-1} n^{-1} \mathbf{E}\right] \mathbf{J}(\mathbf{s}_0) = \alpha n^{-1} m^{-1} \mathbf{E} + \alpha n \mathbf{M}(\mathbf{s}_0; \boldsymbol{\beta}) - \mathbf{I} \,.$$

Provided that $\alpha < 1$, the premultiplying matrix on the lefthand side is invertible, and $\left[\mathbf{I} - \alpha m^{-1} n^{-1} \mathbf{E}\right]^{-1} = \mathbf{I} + \alpha (m-\alpha)^{-1} n^{-1} \mathbf{E}$. This matrix has a single eigenvalue $1 + \alpha(m-\alpha)^{-1}$ with eigenvector $\mathbf{e}$, and additional eigenvalues equal to unity in orthogonal directions. We then have

$$\mathbf{J}(\mathbf{s}_0) = \alpha m^{-1} (1 + \alpha (m-\alpha)^{-1}) \mathbf{E} + \alpha n \left[\mathbf{I} + \alpha (m-\alpha)^{-1} n^{-1} \mathbf{E}\right] \mathbf{M}(\mathbf{s}_0; \boldsymbol{\beta}) - \mathbf{I} \,.$$

In the PageRank-Linear model, $\mathbf{M}(\mathbf{s}_0; \boldsymbol{\beta}) = \beta n^{-1} (\mathbf{I} - n^{-1}\mathbf{E})$, and we therefore have

$$\mathbf{J}(\mathbf{s}_0) = \alpha m^{-1} (1 + \alpha (m-\alpha)^{-1}) \mathbf{E} + \alpha\beta \left[\mathbf{I} + \alpha (m-\alpha)^{-1} n^{-1} \mathbf{E}\right] (\mathbf{I} - n^{-1}\mathbf{E}) - \mathbf{I} \,.$$

We can now read off the eigenvalues of $\mathbf{J}(\mathbf{s}_0)$ analytically. The eigenvector $\mathbf{e}$ has eigenvalue -1, while any vector orthogonal to $\mathbf{e}$ has eigenvalue $\alpha\beta - 1$. This latter eigenvalue is strictly negative if and only if $\beta < \frac{1}{\alpha}$, as was to be shown. □

D.1.3 SpringRank scores

We return to the general formalism of score functions and features introduced at the beginning of this section.

A SpringRank vector $\mathbf{s}$ for a matrix $\mathbf{A}$ with regularization $\alpha \in \mathbb{R}$ is a solution to the linear system

$$\left[\mathbf{D}^i + \mathbf{D}^o - (\mathbf{A} + \mathbf{A}^T) + \alpha \mathbf{I}\right] \mathbf{s} = \mathbf{d}^i - \mathbf{d}^o. \tag{D.7}$$

where, $\mathbf{d}^i = \mathbf{e}^T\mathbf{A}$, $\mathbf{d}^o = \mathbf{A}^T\mathbf{e}$, $\mathbf{D}^i = \mathrm{diag}(\mathbf{d}^i)$, and $\mathbf{D}^o = \mathrm{diag}(\mathbf{d}^o)$. When $\alpha > 0$, (4.3) is invertible and $\mathbf{s}$ is therefore unique. Thus, throughout this section we will assume that $\alpha > 0$, and correspondingly refer to $\mathbf{s}$ as "the" SpringRank vector of $\mathbf{A}$. It is convenient to define $\mathbf{L}_\alpha = \mathbf{D}^i + \mathbf{D}^o - (\mathbf{A} + \mathbf{A}^T) + \alpha\mathbf{I}$ and $\mathbf{\Lambda} = \mathbf{D}^i - \mathbf{D}^o$, in which case the SpringRank relation reads $\mathbf{L}_\alpha \mathbf{s} = \mathbf{\Lambda}\mathbf{e}$.

Theorem D.3 (Stable Egalitarianism with SpringRank Scores)**.** *When σ is the SpringRank map, the vector $\mathbf{s}_0 = \mathbf{0}$ is a fixed point of $\mathbf{f}$, and is the only egalitarian fixed point of the dynamics. This fixed point is linearly stable in the long-memory limit if and only if the matrix*

$$\mathbf{M}(\mathbf{0}; \boldsymbol{\beta}) - 2n^{-1}(\mathbf{I} - n^{-1}\mathbf{E})$$

has eigenvalues strictly smaller than $\frac{\alpha n}{m}$.

We will break the proof into a series of three lemmas. The first lemma calculates the analytical form of $\mathbf{f}$. The second shows that $\mathbf{s}_0 = \mathbf{0}$ is the unique egalitarian fixed point of the long-memory limiting dynamics $\mathbf{f}$. The third gives the criterion for linear stability.

Lemma 1. *The deterministic approximant $\mathbf{f}$ for the SpringRank vector is given by*

$$\mathbf{f}(\mathbf{s}, \mathbf{A}) = \mathbf{s} + \mathbf{L}_\alpha^{-1}\left(-\alpha\mathbf{s} - m\left(n^{-1}\mathbf{L}_\mathbf{G}\mathbf{s} - (n^{-1}\mathbf{e} - \boldsymbol{\gamma})\right)\right), \tag{D.8}$$

where $\mathbf{L_G} = \mathbf{\Gamma} + n^{-1}\mathbf{I} - (\mathbf{G} + \mathbf{G}^T)$.

Proof. Let us fix an implicit time step t. Here and below, we use the notational template $\delta M = M(t+1) - M(t)$ to refer to increments in various quantities under the dynamics (4.1). For example, $\delta\mathbf{A} = \mathbf{A}(t+1) - \mathbf{A}(t)$ refers to the increment in $\mathbf{A}$ under the dynamics. We compute directly

$$\begin{aligned}
\delta\mathbf{A} &= (\lambda - 1)(\mathbf{A} - \mathbf{\Delta}) \\
\delta\mathbf{D}^o &= (\lambda - 1)(\mathbf{D}^o - \mathrm{diag}(\mathbf{\Delta e})) \\
\delta\mathbf{D}^i &= (\lambda - 1)(\mathbf{D}^i - \mathrm{diag}(\mathbf{\Delta}^T\mathbf{e}))\ .
\end{aligned}$$

We can also explicitly write out formulae for the increments in $\mathbf{L}_\alpha$ and $\mathbf{\Lambda}$:

$$\begin{aligned}
\delta\mathbf{\Lambda} &= \delta\mathbf{D}^i - \delta\mathbf{D}^o \\
&= (\lambda - 1)\left[\mathbf{D}^i - \mathbf{D}^o + \mathrm{diag}((\mathbf{\Delta} - \mathbf{\Delta}^T)\mathbf{e})\right] \\
&= (\lambda - 1)\left[\mathbf{\Lambda} + \mathrm{diag}((\mathbf{\Delta} - \mathbf{\Delta}^T)\mathbf{e})\right]\ , \qquad \text{(D.9)} \\
\delta\mathbf{L}_\alpha &= \delta\mathbf{D}^i + \delta\mathbf{D}^o - (\delta\mathbf{A} + \delta\mathbf{A}^T) \\
&= (\lambda - 1)\left[\mathbf{D}^i + \mathbf{D}^o - \mathrm{diag}(\mathbf{\Delta}^T\mathbf{e} + \mathbf{\Delta e}) - (\mathbf{A} + \mathbf{A}^T) + \mathbf{\Delta} + \mathbf{\Delta}^T\right] \\
&= (\lambda - 1)\left[\mathbf{L} - \mathrm{diag}(\mathbf{\Delta}^T\mathbf{e} + \mathbf{\Delta e}) + \mathbf{\Delta} + \mathbf{\Delta}^T\right] \\
&\triangleq (\lambda - 1)\left[\mathbf{L} - \mathbf{L}_{\mathbf{\Delta}}\right]\ , \qquad \text{(D.10)}
\end{aligned}$$

where we have given a name to the Laplacian $\mathbf{L}_{\mathbf{\Delta}} = \mathrm{diag}(\mathbf{\Delta}^T\mathbf{e} + \mathbf{\Delta e}) - \mathbf{\Delta}^T - \mathbf{\Delta}$ of $\mathbf{\Delta}$. Note that $\delta\mathbf{L}_\alpha$ does not depend on α, and we therefore simply write $\delta\mathbf{L} = \delta\mathbf{L}_\alpha$.

We can now formulate a simple condition for equilibrium in expectation. We have

$$(\mathbf{L}_\alpha + \delta\mathbf{L})(\mathbf{s} + \delta\mathbf{s}) = (\mathbf{\Lambda} + \delta\mathbf{\Lambda})\mathbf{e}\ .$$

Subtracting the SpringRank relation $\mathbf{L}_\alpha \mathbf{s} = \mathbf{\Lambda e}$ from each side of this expression, we obtain

$$(\mathbf{L}_\alpha + \delta\mathbf{L})\delta\mathbf{s} = (\delta\mathbf{\Lambda})\mathbf{e} - (\delta\mathbf{L})\mathbf{s} \; .$$

Since $\delta\mathbf{L} = O(1-\lambda)$, the lefthand matrix is invertible in for small λ provided that $\alpha > 0$. We therefore obtain

$$\begin{aligned}\delta\mathbf{s} &= \left(\mathbf{L}_\alpha^{-1} + O(1-\lambda)\right)\left((\delta\mathbf{\Lambda})\mathbf{e} - (\delta\mathbf{L})\mathbf{s}\right) \\ &= \mathbf{L}_\alpha^{-1}((\delta\mathbf{\Lambda})\mathbf{e} - (\delta\mathbf{L})\mathbf{s}) + O((1-\lambda)^2) \; .\end{aligned}$$

The term $O((1-\lambda)^2)$ arises from the product of $O(1-\lambda)$ and the copy of $(\lambda - 1)$ within $\delta\mathbf{\Lambda}$ and $\delta\mathbf{L}$. Taking expectations,

$$\mathbb{E}[\delta\mathbf{s}] = \mathbf{L}_\alpha^{-1}(\mathbb{E}[\delta\mathbf{\Lambda}]\mathbf{e} - \mathbb{E}[\delta\mathbf{L}]\mathbf{s}) + O((1-\lambda)^2) \; .$$

We next insert the expressions (D.9) and (D.10) and use the fact that $\mathbb{E}[\mathbf{\Delta}] = m\mathbf{G}$. This gives

$$\mathbb{E}[\delta\mathbf{s}] = (1-\lambda)\mathbf{L}_\alpha^{-1}\left(\left[\mathbf{L} - m\mathbf{L}_\mathbf{G}\right]\mathbf{s} - \left[\mathbf{\Lambda} + m\cdot\mathrm{diag}((\mathbf{G}-\mathbf{G}^T)\mathbf{e})\right]\mathbf{e}\right) + O((1-\lambda)^2) \; .$$

We can simplify this expression by recalling that $(\mathbf{L} + \alpha\mathbf{I})\mathbf{s} = \mathbf{\Lambda e}$ by definition, as well as the identities $\mathbf{Ge} = n^{-1}\mathbf{e}$ and $\mathbf{G}^T\mathbf{e} = \boldsymbol{\gamma}$. Inserting these identities and simplifying yields

$$= (1-\lambda)\mathbf{L}_\alpha^{-1}\left(-\alpha\mathbf{s} - m\left(\mathbf{L}_\mathbf{G}\mathbf{s} + (n^{-1}\mathbf{e} - \boldsymbol{\gamma})\right)\right) + O((1-\lambda)^2) \; .$$

We now construct $\mathbf{f}$, obtaining Since $\mathbb{E}[\delta\mathbf{s}] = \mathbb{E}[\sigma(\lambda\mathbf{A} + (1-\lambda)\mathbf{\Delta})]$, we can write

$$\begin{aligned}\mathbf{f}(\mathbf{s}, \mathbf{A}) &= \mathbf{s} + \lim_{\lambda \to 1} \frac{\mathbb{E}[\delta\mathbf{s}]}{1-\lambda} \\ &= \mathbf{s} - \mathbf{L}_\alpha^{-1} \left[\alpha\mathbf{s} + m\left(\mathbf{L}_\mathbf{G}\mathbf{s} + (n^{-1}\mathbf{e} - \boldsymbol{\gamma})\right)\right] ,\end{aligned}$$

as was to be shown. □

Lemma 2. *When σ is the SpringRank map, the vector $\mathbf{s}_0 = \mathbf{0}$ is a root of $\mathbf{f}$, and is the only egalitarian fixed point.*

Proof. To show that $\mathbf{s}_0 = \mathbf{0}$ is a fixed point of $\mathbf{f}$, it suffices to insert this solution into (D.8) and simplify, noting that, when $\mathbf{s} = \mathbf{0}$, $\boldsymbol{\gamma} = n^{-1}\mathbf{e}$. To show that it is the unique egalitarian root realizable as a SpringRank score, suppose that $s\mathbf{e}$ were a SpringRank score for some $s \neq 0$. Inserting this into (4.3) and using the fact that $\mathbf{e}$ is a zero eigenvector of the unregularized Laplacian, we would have

$$\alpha s\mathbf{e} = \mathbf{d}^i - \mathbf{d}^o .$$

The total in-degree must equal the total out-degree. Pre-multiplying by $\mathbf{e}$ therefore zeros out the righthand, leaving:

$$\alpha s\mathbf{e}^T\mathbf{e} = \alpha s n = 0 ,$$

which is a contradiction unless $s = 0$. □

Lemma 3. *The egalitarian root $\mathbf{s} = \mathbf{0}$ is a linearly stable root of the SpringRank dynamics in the long-memory limit if and only if the matrix*

$$\mathbf{M}(\mathbf{0}; \boldsymbol{\beta}) - 2n^{-1}(\mathbf{I} - n^{-1}\mathbf{E})$$

has eigenvalues strictly smaller than $\frac{\alpha}{m}$.

Proof. We need to compute $\mathbf{J}(\mathbf{s}_0)$, the Jacobian matrix of $\mathbf{f}$ at $\mathbf{s}_0 = \mathbf{0}$. The fixed point will be stable provided that $\mathbf{J}(\mathbf{s}_0)$ has strictly negative eigenvalues. To compute this Jacobian, we compute derivatives in (D.8). Doing so and applying the product rule, we have

$$\frac{\partial f(\mathbf{s})}{\partial \mathbf{s}} = \mathbf{I} - \mathbf{L}_\alpha^{-1}\left(\alpha\mathbf{I} + m\left(n^{-1}\frac{\partial(\mathbf{L}_\mathbf{G}\mathbf{s})}{\partial \mathbf{s}} - \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}\right)\right) .$$

We calculate $\frac{\partial \mathbf{L}_\mathbf{G}}{\partial \mathbf{s}}$ in Equation (D.11), now obtaining

$$\frac{\partial f(\mathbf{s})}{\partial \mathbf{s}} = \mathbf{I} - \mathbf{L}_\alpha^{-1}\left(\alpha\mathbf{I} + m\left(n^{-1}\left[\mathbf{L}_\mathbf{G} + \boldsymbol{\Sigma}\frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}} - \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}(\mathbf{S}^T + (\mathbf{e}^T\mathbf{s})\mathbf{I})\right] - \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}\right)\right) .$$

Evaluating this expression at $\mathbf{s} = \mathbf{0}$, we have

$$\mathbf{J}(\mathbf{0}) = -\mathbf{L}_\alpha^{-1}\left(\alpha\mathbf{I} + m\left(n^{-1}\mathbf{L}_\mathbf{G} - \frac{\partial \boldsymbol{\gamma}(\mathbf{0})}{\partial \mathbf{s}}\right)\right) ,$$

where $\mathbf{L}_\mathbf{G}$ must also be evaluated at $\mathbf{s} = \mathbf{0}$. We have $\mathbf{G}(\mathbf{0}) = n^{-1}\mathbf{E}$, which implies $\mathbf{L}_\mathbf{G} = 2(\mathbf{I} - n^{-1}\mathbf{E})$. We insert this expression and the formula for $\frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}$ given in (D.2), obtaining

$$\mathbf{J}(\mathbf{0}) = -\mathbf{L}_\alpha^{-1}\left[\alpha\mathbf{I} + mn^{-1}(\mathbf{I} - n^{-1}\mathbf{E})\left(2\mathbf{I} - \sum_{i=1}^{n}\sum_{\ell=1}^{k}\beta_\ell\frac{\partial \phi_i^\ell(\mathbf{s}_0)}{\partial \mathbf{s}}\right)\right] .$$

Since $\mathbf{L}_\alpha$ is symmetric and positive-definite, $\mathbf{L}_\alpha^{-1}$ is as well. The stability of the egalitarian fixed point is therefore determined by the eigenvalues of the matrix inside the brackets. Multiplying by nm^{-1}, we find that a necessary and sufficient condition is that the matrix

$$(\mathbf{I} - n^{-1}\mathbf{E})\left(2\mathbf{I} - \sum_{i=1}^{n}\sum_{\ell=1}^{k}\beta_\ell\frac{\partial \phi_i^\ell(\mathbf{s}_0)}{\partial \mathbf{s}}\right) = \mathbf{M}(\mathbf{0};\boldsymbol{\beta}) - 2n^{-1}(\mathbf{I} - n^{-1}\mathbf{E})$$

have eigenvalues no larger than $\frac{\alpha}{m}$, completing the proof. □

Corrolary 2. *In the SpringRank-Linear model,* $\mathbf{s}_0 = \mathbf{0}$ *is a linearly stable fixed point of* $\mathbf{f}$ *if and only if* $\beta < 2 + \frac{\alpha n}{m}$.

Proof. It suffices to specialize Theorem D.3 to the case of linear features. In particular, we have $\mathbf{M}(\mathbf{0}; \beta) = \beta n^{-1}(\mathbf{I} - n^{-1}\mathbf{E})$. We therefore require that the matrix

$$\beta n^{-1}(\mathbf{I} - n^{-1}\mathbf{E}) - 2n^{-1}(\mathbf{I} - n^{-1}\mathbf{E}) = n^{-1}(\beta - 2)(\mathbf{I} - n^{-1}\mathbf{E})$$

have eigenvalues smaller than $\frac{\alpha}{m}$. We can compute the eigenvalues of this matrix analytically – there is a zero eigenvalue corresponding to the vector $\mathbf{e}$. Then, any vector $\mathbf{v} \perp \mathbf{e}$ is also an eigenvector with eigenvalue $n^{-1}(\beta - 2)$. We therefore require $n^{-1}(\beta - 2) < \frac{\alpha}{m}$, or $\beta < 2 + \frac{\alpha n}{m}$, completing the argument. □

Lemma 4. *We have*

$$\frac{\partial \mathbf{L}_\mathbf{G}\mathbf{s}}{\partial \mathbf{s}} = \mathbf{L}_\mathbf{G} + \mathbf{\Sigma}\frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}} - \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}(\mathbf{S}^T + (\mathbf{e}^T\mathbf{s})\mathbf{I}) \ . \tag{D.11}$$

Proof. We first compute the derivatives $\frac{\partial(\mathbf{G}\mathbf{s})}{\partial \mathbf{s}}$ and $\frac{\partial(\mathbf{G}^T\mathbf{s})}{\partial \mathbf{s}}$. The ith component of $\mathbf{G}\mathbf{s}$ is $v_i = \sum_j \gamma_j s_j$. The product rule for scalar functions of vectors gives the ith row of the derivative:

$$\frac{\partial \mathbf{G}\mathbf{s}_i}{\partial \mathbf{s}} = \sum_j \gamma_j \mathbf{e}_j + \sum_j s_j \frac{\partial \gamma_j}{\partial \mathbf{s}} = \boldsymbol{\gamma} + \sum_j s_j \frac{\partial \gamma_j}{\partial \mathbf{s}} \ .$$

Written in matrix notation, the first term is $\mathbf{G}$. To write the second term in matrix form, note that we need to multiply $\frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}$ by the matrix each of whose columns is a copy of $\mathbf{s}$. This matrix is $\mathbf{S}^T$. We therefore obtain

$$\frac{\partial(\mathbf{G}\mathbf{s})}{\partial \mathbf{s}} = \mathbf{G} + \frac{\partial \boldsymbol{\gamma}}{\partial \mathbf{s}}\mathbf{S}^T \quad .$$

To compute the second derivative, note that $\mathbf{G}^T\mathbf{s} = \boldsymbol{\gamma}(\mathbf{e}^T\mathbf{s})$, with ith component $\gamma_i\mathbf{e}^T\mathbf{s}$. Using the product rule for scalar functions of vectors, we have

$$\frac{\partial}{\partial \mathbf{s}}\gamma_i\mathbf{e}^T\mathbf{s} = \gamma_i\mathbf{e} + (\mathbf{e}^T\mathbf{s})\frac{\partial\gamma_i}{\partial\mathbf{s}}.$$

The first term will become the matrix whose ith row is γ_i, i.e. $\mathbf{G}^T$. This yields

$$\frac{\partial(\mathbf{G}^T\mathbf{s})}{\partial\mathbf{s}} = \mathbf{G}^T + (\mathbf{e}^T\mathbf{s})\frac{\partial\boldsymbol{\gamma}}{\partial\mathbf{s}}\ .$$

Combining these expressions yields our formula for $\frac{\partial\mathbf{L_G s}}{\partial\mathbf{s}}$:

$$\begin{aligned}\frac{\partial\mathbf{L_G s}}{\partial\mathbf{s}} &= \frac{\partial}{\partial\mathbf{s}}\left[\boldsymbol{\Gamma}\mathbf{s} + \mathbf{s} - \mathbf{G}\mathbf{s} - \mathbf{G}^T\mathbf{s}\right] \\ &= \boldsymbol{\Gamma} + \boldsymbol{\Sigma}\frac{\partial\boldsymbol{\gamma}}{\partial\mathbf{s}} + \mathbf{I} - \left(\mathbf{G} + \frac{\partial\boldsymbol{\gamma}}{\partial\mathbf{s}}\mathbf{S}^T + \mathbf{G}^T + (\mathbf{e}^T\mathbf{s})\frac{\partial\boldsymbol{\gamma}}{\partial\mathbf{s}}\right) \\ &= \mathbf{L_G} + \boldsymbol{\Sigma}\frac{\partial\boldsymbol{\gamma}}{\partial\mathbf{s}} - \frac{\partial\boldsymbol{\gamma}}{\partial\mathbf{s}}(\mathbf{S}^T + (\mathbf{e}^T\mathbf{s})\mathbf{I})\ ,\end{aligned}$$

as was to be shown. □

D.2 Supplementary figures

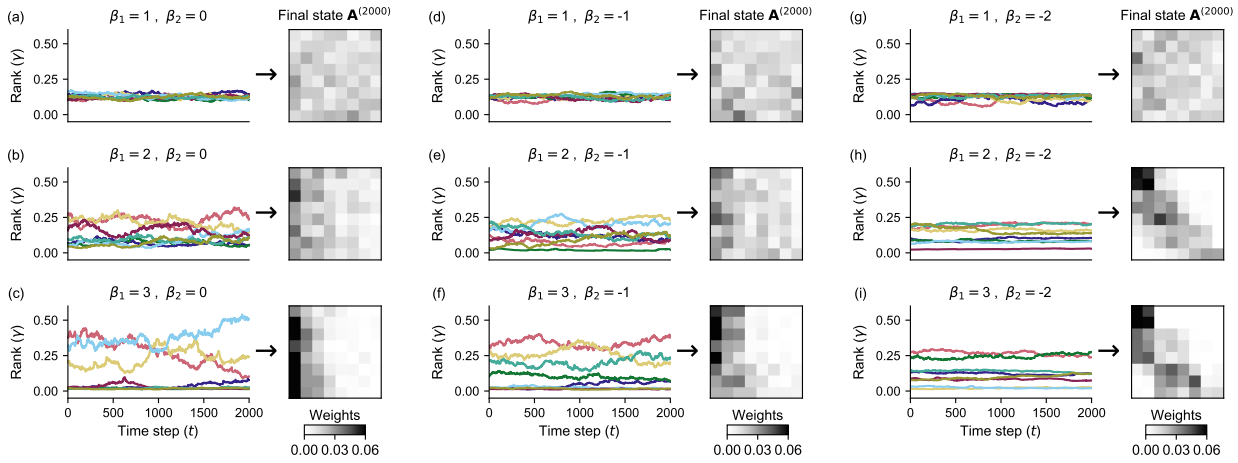

Figure D.1: Example dynamics of the model with SpringRank. Populations of $n = 8$ agents were simulated for 2000 time steps using the SpringRank score with linear and quadratic features, varying the preference parameters β_1 and β_2 as indicated in the panels. The memory parameter was fixed at $\lambda = 0.995$. In each panel, the plot on the left shows the simulated rank vector γ over time; different colors track the ranks of different agents. The heatmap on the right shows the adjacency matrix $\mathbf{A}$ at time step $t = 2000$ for the corresponding parameter values.

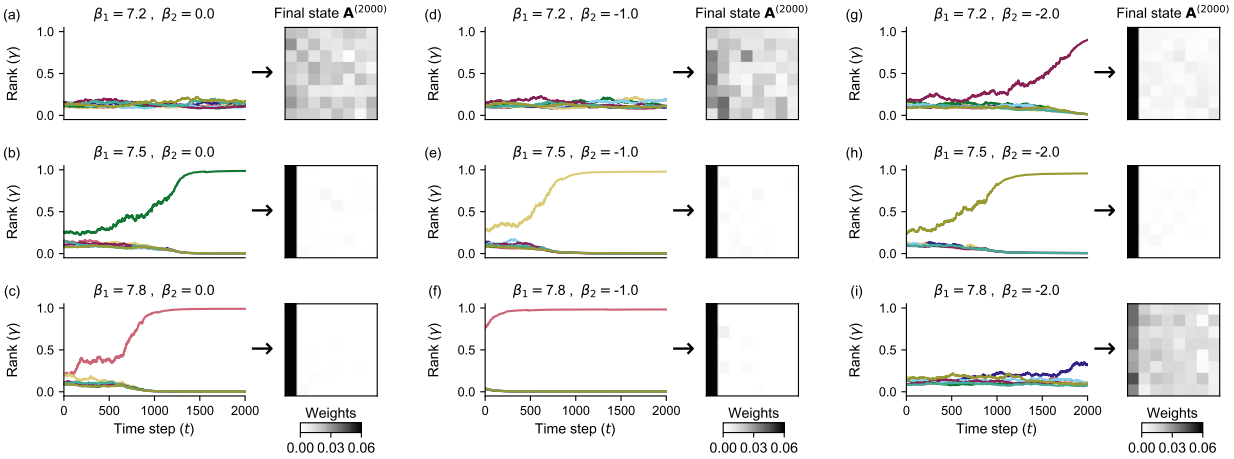

Figure D.2: As in Fig. D.1, using the PageRank score function.

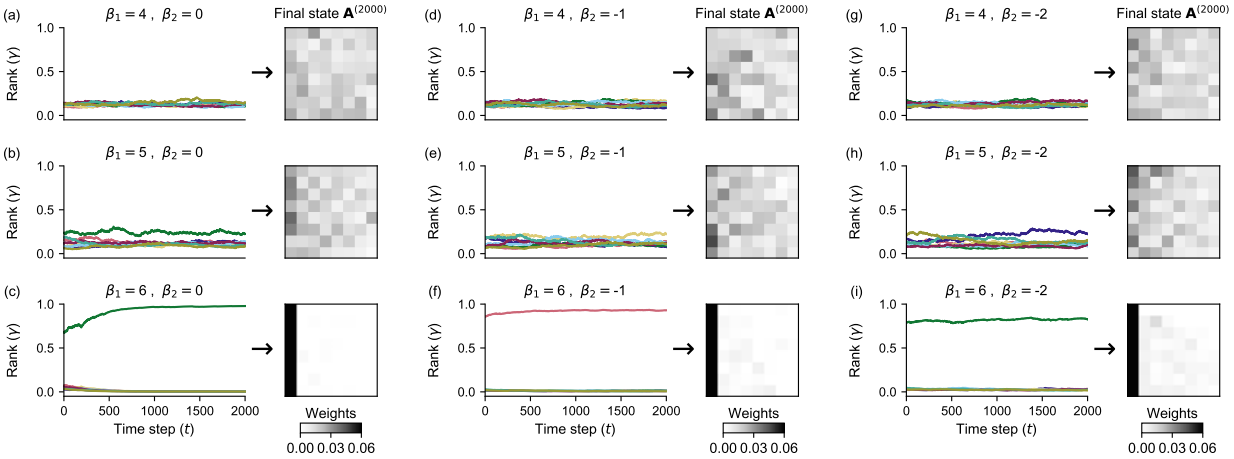

Figure D.3: As in Fig. D.1, using the Root-Degree score function.

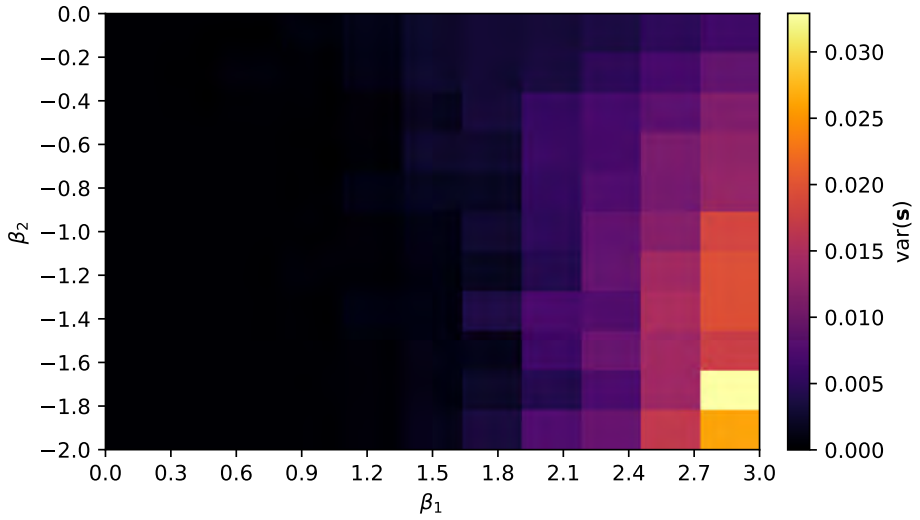

Figure D.4: Variance in the rank vector s over the final 500 iterations of a series of simulations with $n = 8$ and $\lambda = 0.995$ (as in Fig.4.2). The parameters β_1 and β_2 are allowed to vary. Higher variances correspond to more strongly hierarchical states.

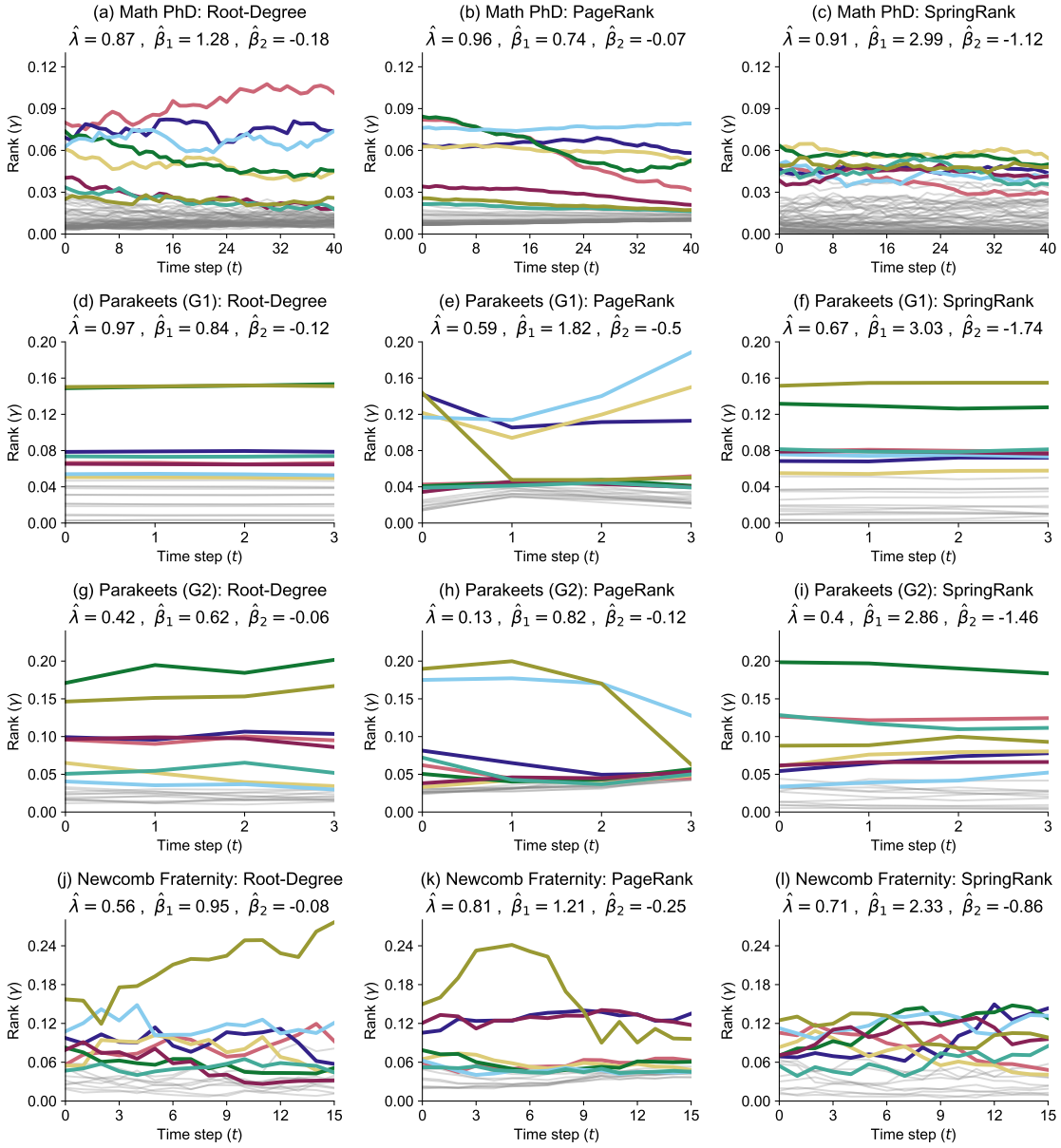

Figure D.5: Simulated dynamics of the model using inferred parameters $\hat{\lambda}$, $\hat{\beta}_1$, $\hat{\beta}_2$ in Table 4.1. The value of m for each row of panels corresponds to the average number of updates per time step in the corresponding data set, indicated in the panel title ($m = 150$ for Math PhD, $m = 279$ for Parakeets (G1), $m = 320$ for Parakeets (G2), and $m = 85$ for Newcomb Fraternity). Furthermore, the simulations in each row were initialized using the network at the relevant initial time step in the corresponding data set: the network of endorsements aggregated up to year 1960 for the Math PhD data set, and the network at time step 0 in each of the Parakeet and Newcomb Fraternity data sets. The traces in color correspond to nodes that rank among the top 8 on average over time; those in light gray track all other nodes. Other parameters: $\alpha_p = 0.85$, $\alpha_s = 10^{-8}$.

References

1. Alem, S., Perry, C. J., Zhu, X., Loukola, O. J., Ingraham, T., Søvik, E., and Chittka, L. (2016). Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect. *PLOS Biology*, 14(10):e1002564.

2. Alexander, R. (1987). *The Biology of Moral Systems*. Evolutionary Foundations of Human Behavior Series. Aldine de Gruyter, New York, NY.

3. Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., and Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191:104254.

4. Antal, T., Ohtsuki, H., Wakeley, J., Taylor, P. D., and Nowak, M. A. (2009a). Evolution of cooperation by phenotypic similarity. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21):8597–8600.

5. Antal, T., Traulsen, A., Ohtsuki, H., Tarnita, C. E., and Nowak, M. A. (2009b). Mutation-selection equilibrium in games with multiple strategies. *Journal of Theoretical Biology*, 258(4):614–622.

6. Aplin, L. M., Farine, D. R., Mann, R. P., and Sheldon, B. C. (2014). Individual-level personality influences social foraging and collective behaviour in wild birds. *Proceedings of the Royal Society B: Biological Sciences*, 281(1789):20141016–20141016.

7. Ashmore, R. D. and Del Boca, F. K. (1981). Cognitive processes in stereotyping and intergroup behavior. In Hamilton, D. L., editor, *Conceptual Approaches to Stereotypes and Stereotyping*, pages 1–35. Erlbaum, Hillsdale, NJ.

8. Bakker, B. N., Lelkes, Y., and Malka, A. (2020). Understanding partisan cue receptivity: Tests of predictions from the bounded rationality and expressive utility perspectives. *The Journal of Politics*, 82(3):1061–1077.

9. Baldassarri, D. and Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology*, 114(2):408–446.

10. Ball, B. and Newman, M. E. J. (2013). Friendship networks and social status. *Network Science*, 1(1):16–30.

11. Balliet, D., Wu, J., and Van Lange, P. A. M. (2020). Indirect reciprocity, gossip, and reputation-based cooperation. In Kruglanski, A. W., Higgins, E. T., and Van Lange, P. A. M., editors, *Social Psychology: Handbook of Basic Principles*, pages 265–287. The Guilford Press, New York.

12. Barber, M. and Lelkes, Y. (2015). Causes and consequences of polarization. In Mansbridge, J. and Martin, C., editors, *Political Negotiation: A Handbook*. Brookings Institution Press.

13. Bardoscia, M., De Luca, G., Livan, G., Marsili, M., and Tessone, C. J. (2013). The social climbing game. *Journal of Statistical Physics*, 151(3):440–457.

14. Batagelj, V. and Mrvar, A. (2007). Newcomb fraternity. http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm#newfrat. Accessed 18 Jul 2020.

15. Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

16. Bell, A. S., De. Roode, J. C., Sim, D., and Read, A. F. (2006). Within-host competition in genetically diverse malaria infections: Parasite virulence and competitive success. *Evolution*, 60(7):1358–1371.

17. Ben-Naim, E. and Redner, S. (2005). Dynamics of social diversity. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):L11002–L11002.

18. Berdahl, A., Torney, C. J., Ioannou, C. C., Faria, J. J., and Couzin, I. D. (2013). Emergent sensing of complex environments by mobile animal groups. *Science*, 339(6119):574–576.

19. Bereczkei, T., Birkas, B., and Kerekes, Z. (2007). Public charity offer as a proximate factor of evolved reputation-building strategy: An experimental analysis of a real-life situation. *Evolution and Human Behavior*, 28(4):277–284.

20. Bernadou, A., Schrader, L., Pable, J., Hoffacker, E., Meusemann, K., and Heinze, J. (2018). Stress and early experience underlie dominance status and division of labour in a clonal insect. *Proceedings of the Royal Society B: Biological Sciences*, 285(1885):20181468.

21. Beshers, S. N. and Fewell, J. H. (2001). Models of division of labor in social insects. *Annual Review of Entomology*, 46(1):413–440.

22. Bettenworth, V., Steinfeld, B., Duin, H., Petersen, K., Streit, W. R., Bischofs, I., and Becker, A. (2019). Phenotypic heterogeneity in bacterial quorum sensing systems. *Journal of Molecular Biology*, 431(23):4530–4546.

23. Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98.

24. Biedermann, P. H. W. and Taborsky, M. (2011). Larval helpers and age polyethism in ambrosia beetles. *Proceedings of the National Academy of Sciences*, 108(41):17064–17069.

25. Billig, M. and Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, 3(1):27–52.

26. Bizyaeva, A., Franci, A., and Leonard, N. E. (2022). Nonlinear opinion dynamics with tunable sensitivity. *IEEE Transactions on Automatic Control*. Advanced online publication. https://doi.org/10.1109/TAC.2022.3159527.

27. Blanchard, G. B., Orledge, G. M., Reynolds, S. E., and Franks, N. R. (2000). Division of labour and seasonality in the ant *Leptothorax albipennis*: Worker corpulence and its influence on behaviour. *Animal Behaviour*, 59(4):723–738.

28. Blatrix, R., Durand, J.-L., and Jaisson, P. (2000). Task allocation depends on matriline in the ponerine ant *Gnamptogenys striatula* Mayr. *Journal of Insect Behavior*, 13(4):553–562.

29. Bonabeau, E., Theraulaz, G., and Deneubourg, J.-L. (1995). Phase diagram of a model of self-organizing hierarchies. *Physica A: Statistical Mechanics and its Applications*, 217(3):373–392.

30. Bonabeau, E., Theraulaz, G., and Deneubourg, J.-L. (1996a). Mathematical model of self-organizing hierarchies in animal societies. *Bulletin of Mathematical Biology*, 58(4):661–717.

31. Bonabeau, E., Theraulaz, G., and Deneubourg, J.-L. (1996b). Quantitative study of the fixed threshold model for the regulation of division of labour in insect societies. *Proceedings of the Royal Society B: Biological Sciences*, 263(1376):1565–1569.

32. Bonabeau, E., Theraulaz, G., and Deneubourg, J.-L. (1998). Fixed response thresholds and the regulation of division of labor in insect societies. *Bulletin of Mathematical Biology*, 60(4):753–807.

33. Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

34. Boyd, R. and Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks*, 11(3):213–236.

35. Brahma, A., Mandal, S., and Gadagkar, R. (2018). Emergence of cooperation and division of labor in the primitively eusocial wasp *Ropalidia marginata*. *Proceedings of the National Academy of Sciences*, 115(4):756–761.

36. Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117.

37. Broly, P. and Deneubourg, J.-L. (2015). Behavioural contagion explains group cohesion in a social crustacean. *PLOS Computational Biology*, 11(6):1–18.

38. Bruch, E. E. and Newman, M. E. J. (2018). Aspirational pursuit of mates in online dating markets. *Science Advances*, 4:eaap9815.

39. Buttery, N. J., Thompson, C. R. L., and Wolf, J. B. (2010). Complex genotype interactions influence social fitness during the developmental phase of the social amoeba *Dictyostelium discoideum*. *Journal of Evolutionary Biology*, 23(8):1664–1671.

40. Carley, K. M. and Krackhardt, D. (1996). Cognitive inconsistencies and non-symmetric friendship. *Social Networks*, 18(1):1–27.

41. Carlin, R. E. and Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior*, 35(1):43–63.

42. Carlin, R. E. and Love, G. J. (2018). Political competition, partisanship and interpersonal trust in electoral democracies. *British Journal of Political Science*, 48(1):115–139.

43. Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646.

44. Cavaliere, M., Sedwards, S., Tarnita, C. E., Nowak, M. A., and Csikász-Nagy, A. (2012). Prosperity is associated with instability in dynamical networks. *Journal of Theoretical Biology*, 299:126–138.

45. Centola, D. (2018). *How Behavior Spreads: The Science of Complex Contagions*. Princeton University Press.

46. Chaffee, S. H. and Wilson, D. G. (1977). Media rich, media poor: Two studies of diversity in agenda-holding. *Journalism Quarterly*, 54(3):466–476.

47. Chandra, V., Gal, A., and Kronauer, D. J. C. (2021). Colony expansions underlie the evolution of army ant mass raiding. *Proceedings of the National Academy of Sciences*, 118(22):e2026534118.

48. Chase, I. D., Bartolomeo, C., and Dugatkin, L. A. (1994). Aggressive interactions and inter-contest interval: How long do winners keep winning?. *Animal Behaviour*, 48(2):393–400.

49. Cheng, J. T. and Tracy, J. L. (2014). Toward a unified science of hierarchy: Dominance and prestige are two fundamental pathways to human social rank. In Cheng, J. T., Tracy, J. L., and Anderson, C., editors, *The Psychology of Social Status*, pages 3–27. Springer, New York, NY.

50. Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379.

51. Chu, O. J. (2021). *Heterogeneity in human populations, from structure to personality—a modeling and data approach*. PhD thesis, Princeton University, NJ, United States.

52. Clauset, A., Arbesman, S., and Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005.

53. Cohen, D. J., James Nelson, W., and Maharbiz, M. M. (2014). Galvanotactic control of collective cell migration in epithelial monolayers. *Nature Materials*, 13(4):409–417.

54. Cooney, D. B., Kessinger, T. A., and Plotkin, J. B. (in prep.). Competition between social norms via multilevel selection.

55. Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., Conradt, L., Levin, S. A., and Leonard, N. E. (2011). Uninformed individuals promote democratic consensus in animal groups. *Science*, 334(6062):1578–1580.

56. Couzin, I. D., Krause, J., Franks, N. R., and Levin, S. A. (2005). Effective leadership and decision-making in animal groups on the move. *Nature*, 433(7025):513–516.

57. Couzin, I. D., Krause, J., James, R., Ruxton, G. D., and Franks, N. R. (2002). Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology*, 218(1):1–11.

58. Crall, J. D., Gravish, N., Mountcastle, A. M., Kocher, S. D., Oppenheimer, R. L., Pierce, N. E., and Combes, S. A. (2018). Spatial fidelity of workers predicts collective response to disturbance in a social insect. *Nature Communications*, 9(1):1–13.

59. De Bacco, C., Larremore, D. B., and Moore, C. (2018). A physical model for efficient ranking in networks. *Science Advances*, 4:8260.

60. DeDeo, S. and Hobson, E. A. (2021). From equality to hierarchy. *Proceedings of the National Academy of Sciences*, 118(21):e2106186118.

61. DellaPosta, D. (2020). Pluralistic collapse: The "oil spill" model of mass opinion polarization. *American Sociological Review*, 85(3):507–536.

62. Detrain, C. and Pasteels, J. M. (1991). Caste differences in behavioral thresholds as a basis for polyethism during food recruitment in the ant, *Pheidole pallidula* (Nyl.) (Hymenoptera: Myrmicinae). *Journal of Insect Behavior*, 4(2):157–176.

63. Dhamdhere, A. and Dovrolis, C. (2011). Twelve years in the evolution of the internet ecosystem. *IEEE/ACM Transactions on Networking*, 19(5):1420–1433.

64. Diggle, S. P., Griffin, A. S., Campbell, G. S., and West, S. A. (2007). Cooperation and conflict in quorum-sensing bacterial populations. *Nature*, 450(7168):411–414.

65. Dodds, P. S. and Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604.

66. Donovan, K., Kellstedt, P. M., Key, E. M., and Lebo, M. J. (2020). Motivated reasoning, public opinion, and presidential approval. *Political Behavior*, 42(4):1201–1221.

67. Dornhaus, A. (2008). Specialization does not predict individual efficiency in an ant. *PLOS Biology*, 6(11):2368–2375.

68. Duarte, A., Weissing, F. J., Pen, I., and Keller, L. (2011). An evolutionary perspective on self-organized division of labor in social insects. *Annual Review of Ecology, Evolution, and Systematics*, 42(1):91–110.

69. Edy, J. A. and Meirick, P. C. (2018). The fragmenting public agenda: Capacity, diversity, and volatility in responses to the "most important problem" question. *Public Opinion Quarterly*, 82(4):661–685.

70. Eyer, P.-A., Freyer, J., and Aron, S. (2012). Genetic polyethism in the polyandrous desert ant *Cataglyphis cursor*. *Behavioral Ecology*, 24(1):144–151.

71. Fewell, J. H. and Bertram, S. M. (1999). Division of labor in a dynamic environment: Response by honeybees (*Apis mellifera*) to graded changes in colony pollen stores. *Behavioral Ecology and Sociobiology*, 46(3):171–179.

72. Fewell, J. H. and Jr, R. E. P. (1999). The emergence of division of labour in forced associations of normally solitary ant queens. *Evolutionary Ecology Research*, 1(5):537–548.

73. Fewell, J. H. and Page, R. E. (1993). Genotypic variation in foraging responses to environmental stimuli by honey bees, *Apis mellifera*. *Experientia*, 49(12):1106–1112.

74. Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., and Druckman, J. N. (2020). Political sectarianism in America. *Science*, 370(6516):533–536.

75. Fiske, S. and Taylor, S. (1991). *Social Cognition*. McGraw-Hill, New York, NY, second edition.

76. Frumhoff, P. C. and Baker, J. (1988). A genetic component to division of labour within honey bee colonies. *Nature*, 333(6171):358–361.

77. Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., and Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific Reports*, 2:460.

78. Fushing, H., McAssey, M. P., Beisner, B., and McCowan, B. (2011). Ranking network of a captive rhesus macaque society: A sophisticated corporative kingdom. *PLOS ONE*, 6(3):e17817.

79. Garandeau, C. F., Lee, I. A., and Salmivalli, C. (2014). Inequality matters: Classroom status hierarchy and adolescents' bullying. *Journal of Youth and Adolescence*, 43(7):1123–1133.

80. Gautrais, J., Theraulaz, G., Deneubourg, J. L., and Anderson, C. (2002). Emergent polyethism as a consequence of increased colony size in insect societies. *Journal of Theoretical Biology*, 215(3):363–373.

81. Gell-Mann, M. (1994). Complex adaptive systems. In Cowan, G., Pines, D., and Meltzer, D., editors, *Complexity: Metaphors, Models, and Reality*, number 19 in Santa Fe Institute Studies in the Sciences of Complexity, pages 17–45. Addison-Wesley, Reading, MA.

82. Geritz, S., Kisdi, É., Meszéna, G., and Metz, J. (1998). Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology*, 12(1):35–57.

83. Goff, S. and Lee, D. J. (2019). Prospects for third party electoral success in a polarized era. *American Politics Research*, 47(6):1324–1344.

84. Gordon, D. (2010). *Ant Encounters: Interaction Networks and Colony Behavior*. Primers in Complex Systems. Princeton University Press.

85. Gordon, D. and Schwengel, M. (1999). *Ants at Work: How an Insect Society Is Organized*. Free Press.

86. Gordon, D. M. (1989). Dynamics of task switching in harvester ants. *Animal Behaviour*, 38(2):194–204.

87. Gordon, D. M. (1996). The organization of work in social insect colonies. *Nature*, 380(6570):121–124.

88. Gray, R., Franci, A., Srivastava, V., and Leonard, N. E. (2018). Multi-agent decision-making dynamics inspired by honeybees. *IEEE Transactions on Control of Network Systems*, 5(2):793–806.

89. Green, D., Palmquist, B., and Schickler, E. (2004). *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. Yale University Press.

90. Green, J. and Hobolt, S. B. (2008). Owning the issue agenda: Party strategies and vote choices in British elections. *Electoral Studies*, 27(3):460–476.

91. Greenwald, E. E., Baltiansky, L., and Feinerman, O. (2018). Individual crop loads provide local control for collective food intake in ant colonies. *eLife*, 7:e31730.

92. Guilbeault, D., Becker, J., and Centola, D. (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences*, 115(39):9714–9719.

93. Gupta, H. and Porter, M. A. (2020). Mixed logit models and network formation. *arXiv*, 2006.16516. Accessed 18 Jul 2020.

94. Hamilton, W. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1):17–52.

95. Hammond, R. A. and Axelrod, R. (2006). Evolution of contingent altruism when cooperation is expensive. *Theoretical Population Biology*, 69(3):333–338.

96. Hemelrijk, C. K. (1999). An individual–orientated model of the emergence of despotic and egalitarian societies. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 266(1417):361–369.

97. Hickey, J. and Davidsen, J. (2019). Self-organization and time-stability of social hierarchies. *PLOS ONE*, 14(1):e0211403.

98. Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., and Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, 115(48):12241–12246.

99. Hilton, J. L. and von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, 47(1):237–271.

100. Hinze, B. and Leuthold, R. H. (1999). Age related polyethism and activity rhythms in the nest of the termite *Macrotermes bellicosus* (Isoptera, Termitidae). *Insectes Sociaux*, 46(4):392–397.

101. Hobson, E. A. and DeDeo, S. (2015). Social feedback and the emergence of rank in animal society. *PLOS Computational Biology*, 11(9):e1004411.

102. Hobson, E. A. and DeDeo, S. (2016). Data from: Social feedback and the emergence of rank in animal society. https://doi.org/10.5061/dryad.p56q7. Accessed 18 Jul 2020.

103. Hobson, E. A., Mønster, D., and DeDeo, S. (2021). Aggression heuristics underlie animal dominance hierarchies and provide evidence of group-level social information. *Proceedings of the National Academy of Sciences*, 118(10):e2022912118.

104. Hofbauer, J., Sigmund, K., and Sigmund, P. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

105. Hogeweg, P. and Hesper, B. (1983). The ontogeny of the interaction structure in bumble bee colonies: A MIRROR model. *Behavioral Ecology and Sociobiology*, 12(4):271–283.

106. Holbrook, C. T., Kukuk, P. F., and Fewell, J. H. (2013). Increased group size promotes task specialization in a normally solitary halictine bee. *Behaviour*, 150(12):1449–1466.

107. Holekamp, K. E. and Strauss, E. D. (2016). Aggression and dominance: An interdisciplinary overview. *Current Opinion in Behavioral Sciences*, 12:44–51.

108. Holland, J. H. (2006). Studying complex adaptive systems. *Journal of Systems Science and Complexity*, 19(1):1–8.

109. Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

110. Horn, R. and Johnson, C. (2012). *Matrix Analysis*. Cambridge University Press.

111. Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363.

112. Huang, Z.-Y. and Robinson, G. E. (1996). Regulation of honey bee division of labor by colony age demography. *Behavioral Ecology and Sociobiology*, 39(3):147–158.

113. Huber, F. (1814). *Nouvelles Observations Sur Les Abeilles*. chez J.J. Paschoud, second edition.

114. Ito, F. and Higashi, S. (1991). A linear dominance hierarchy regulating reproduction and polyethism of the queenless ant *Pachycondyla sublaevis*. *Naturwissenschaften*, 78(2):80–82.

115. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1):129–146.

116. Iyengar, S. and Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3):690–707.

117. Jansen, V. A. A. and van Baalen, M. (2006). Altruism through beard chromodynamics. *Nature*, 440(7084):663–666.

118. Jeanson, R. (2019). Within-individual behavioural variability and division of labour in social insects. *The Journal of Experimental Biology*, 222(10):jeb190868.

119. Jeanson, R. and Fewell, J. H. (2008). Influence of the social context on division of labor in ant foundress associations. *Behavioral Ecology*, 19(3):567–574.

120. Jeanson, R., Fewell, J. H., Gorelick, R., and Bertram, S. M. (2007). Emergence of increased division of labor as a function of group size. *Behavioral Ecology and Sociobiology*, 62(2):289–298.

121. Jeanson, R. and Weidenmüller, A. (2014). Interindividual variability in social insects - proximate causes and ultimate consequences. *Biological Reviews*, 89(3):671–687.

122. Jennings, W., Bevan, S., Timmermans, A., Breeman, G., Brouard, S., Chaqués-Bonafont, L., Green-Pedersen, C., John, P., Mortensen, P. B., and Palau, A. M. (2011). Effects of the core functions of government on the diversity of executive agendas. *Comparative Political Studies*, 44(8):1001–1030.

123. Jolles, J. W., Boogert, N. J., Sridhar, V. H., Couzin, I. D., and Manica, A. (2017). Consistent individual differences drive collective behavior and group functioning of schooling fish. *Current Biology*, 27:1–7.

124. Jolles, J. W., King, A. J., and Killen, S. S. (2020). The role of individual heterogeneity in collective animal behaviour. *Trends in Ecology and Evolution*, 35(3):278–291.

125. Jones, M. I., Sirianni, A. D., and Fu, F. (2022). Polarization, abstention, and the median voter theorem. *Humanities and Social Sciences Communications*, 9(1):1–12.

126. Judd, C. M. and Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, 100(1):109–128.

127. Kaptein, N., Billen, J., and Gobin, B. (2005). Larval begging for food enhances reproductive options in the ponerine ant *Gnamptogenys striatula*. *Animal Behaviour*, 69(2):293–299.

128. Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, 26(5):594–604.

129. Kaushik, S., Katoch, B., and Nanjundiah, V. (2006). Social behaviour in genetically heterogeneous groups of *Dictyostelium giganteum*. *Behavioral Ecology and Sociobiology*, 59(4):521–530.

130. Kay, A. and Rissing, S. W. (2005). Division of foraging labor in ants can mediate demands for food and safety. *Behavioral Ecology and Sociobiology*, 58(2):165–174.

131. Keith, B., Magleby, D., Nelson, C., Orr, E., and Westlye, M. (1992). *The Myth of the Independent Voter*. University of California Press.

132. Kessinger, T. A. and Plotkin, J. B. (2022). Indirect reciprocity in populations with group structure. *arXiv*, 2204.10811. Accessed 25 Apr 2022.

133. Khuong, A., Gautrais, J., Perna, A., Sbaï, C., Combe, M., Kuntz, P., Jost, C., and Theraulaz, G. (2016). Stigmergic construction and topochemical information shape ant nest architecture. *Proceedings of the National Academy of Sciences*, 113(5):1303–1308.

134. King, A. J. and Sueur, C. (2011). Where next? Group coordination and collective decision making by primates. *International Journal of Primatology*, 32(6):1245–1267.

135. König, M. D. and Tessone, C. J. (2011). Network evolution based on centrality. *Physical Review E*, 84(5):056108.

136. König, M. D., Tessone, C. J., and Zenou, Y. (2014). Nestedness in networks: A theoretical model and some applications. *Theoretical Economics*, 9(3):695–752.

137. Kozlowski, A. C. and Murphy, J. P. (2021). Issue alignment and partisanship in the American public: Revisiting the 'partisans without constraint' thesis. *Social Science Research*, 94:102498.

138. Krause, S. M., Peixoto, T. P., and Bornholdt, S. (2013). Spontaneous centralization of control in a network of company ownerships. *PLOS ONE*, 8(12):e80303.

139. Kronauer, D. J. C., Pierce, N. E., and Keller, L. (2012). Asexual reproduction in introduced and native populations of the ant *Cerapachys biroi*. *Molecular Ecology*, 21(21):5221–5235.

140. Kunegis, J. (2013). KONECT: The Koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 1343–1350, New York, NY, USA. Association for Computing Machinery.

141. Kwapich, C. L., Valentini, G., and Hölldobler, B. (2018). The non-additive effects of body size on nest architecture in a polymorphic ant. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1753):20170235.

142. Leimar, O. and Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1468):745–753.

143. Lenz, G. S. (2013). *Follow the Leader?: How Voters Respond to Politicians' Policies and Performance*. University of Chicago Press.

144. Leonard, N. E., Lipsitz, K., Bizyaeva, A., Franci, A., and Lelkes, Y. (2021). The nonlinear feedback dynamics of asymmetric political polarization. *Proceedings of the National Academy of Sciences*, 118(50):e2102149118.

145. Leonard, N. E., Shen, T., Nabet, B., Scardovi, L., Couzin, I. D., and Levin, S. A. (2012). Decision versus compromise for animal groups in motion. *Proceedings of the National Academy of Sciences*, 109(1):227–232.

146. Levendusky, M. S. (2010). Clearer cues, more consistent voters: A benefit of elite polarization. *Political Behavior*, 32(1):111–131.

147. Levin, S. A. (2002). Complex adaptive systems: Exploring the known, the unknown and the unknowable. *Bulletin of the American Mathematical Society*, 40(1):3–19.

148. Liao, H., Mariani, M. S., Medo, M., Zhang, Y.-C., and Zhou, M.-Y. (2017). Ranking in evolving complex networks. *Physics Reports*, 689:1–54.

149. Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., and Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*, 650:1–63.

150. Macrae, C. N. and Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, 51(1):93–120.

151. Madison, J. (1787). The Federalist Papers: No. 10. In *The Avalon Project*. Yale Law School, New Haven, CT. https://avalon.law.yale.edu/18th_century/fed10.asp. Accessed 2 Jul 2021.

152. Maisonnasse, A., Lenoir, J.-C., Costagliola, G., Beslay, D., Choteau, F., Crauser, D., Becard, J.-M., Plettner, E., and Le Conte, Y. (2009). A scientific note on E-β-ocimene, a new volatile primer pheromone that inhibits worker ovary development in honey bees. *Apidologie*, 40(5):562–564.

153. Malka, A., Soto, C. J., Inzlicht, M., and Lelkes, Y. (2014). Do needs for security and certainty predict cultural and economic conservatism? A cross-national analysis. *Journal of Personality and Social Psychology*, 106(6):1031–1051.

154. Mann, T. and Ornstein, N. (2016). *It's Even Worse than It Looks: How the American Constitutional System Collided with the New Politics of Extremism*. Basic Books.

155. Mariani, M. S. and Lü, L. (2020). Network-based ranking in social systems: Three challenges. *Journal of Physics: Complexity*, 1(1):011001.

156. Martin, D., Hutchison, J., Slessor, G., Urquhart, J., Cunningham, S. J., and Smith, K. (2014). The spontaneous formation of stereotypes via cumulative cultural evolution. *Psychological Science*, 25(9):1777–1786.

157. Mas, F. and Kölliker, M. (2008). Maternal care and offspring begging in social insects: Chemical signalling, hormonal regulation and evolution. *Animal Behaviour*, 76(4):1121–1131.

158. Masuda, N. and Ohtsuki, H. (2007). Tag-based indirect reciprocity by incomplete social information. *Proceedings of the Royal Society B: Biological Sciences*, 274(1610):689–695.

159. May, R. M. (1972). Will a large complex system be stable? *Nature*, 238(5364):413–414.

160. McCarty, N. (2019). *Polarization: What Everyone Needs to Know®*. Oxford University Press.

161. McCarty, N., Poole, K., and Rosenthal, H. (2016). *Polarized America: The Dance of Ideology and Unequal Riches*. Walras-Pareto Lectures. MIT Press, second edition.

162. McCombs, M. and Zhu, J.-H. (1995). Capacity, diversity, and volatility of the public agenda: Trends from 1954 to 1994. *Public Opinion Quarterly*, 59(4):495–525.

163. Mehta, P. H. and Prasad, S. (2015). The dual-hormone hypothesis: A brief review and future research agenda. *Current Opinion in Behavioral Sciences*, 3:163–168.

164. Melke, P., Sahlin, P., Levchenko, A., and Jönsson, H. (2010). A cell-based model for quorum sensing in heterogeneous bacterial colonies. *PLOS Computational Biology*, 6(6):1–12.

165. Merling, M., Eisenmann, S., and Bloch, G. (2020). Body size but not age influences phototaxis in bumble bee (*Bombus terrestris*, L.) workers. *Apidologie*, 51(5):763–776.

166. Mersch, D. P., Crespi, A., and Keller, L. (2013). Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science*, 340(6136):1090–1093.

167. Mertl, A. L. and Traniello, J. F. A. (2009). Behavioral evolution in the major worker subcaste of twig-nesting *Pheidole* (Hymenoptera: Formicidae): Does morphological specialization influence task plasticity? *Behavioral Ecology and Sociobiology*, 63(10):1411–1426.

168. Milosh, M., Painter, M., Sonin, K., Dijcke, D. V., and Wright, A. L. (2020). Political polarisation impedes the public policy response to COVID-19. https://voxeu.org/article/political-polarisation-impedes-public-policy-response-covid-19. Accessed 27 Feb 2021.

169. Miyaguchi, T., Miki, T., and Hamada, R. (2020). Piecewise linear model of self-organized hierarchy formation. *Physical Review E*, 102(3):032213.

170. Morgan, A. C., Economou, D. J., Way, S. F., and Clauset, A. (2018). Prestige drives epistemic inequality in the diffusion of scientific ideas. *EPJ Data Science*, 7(1):40.

171. Myerscough, M. R. and Oldroyd, B. P. (2004). Simulation models of the role of genetic variability in social insect task allocation. *Insectes Sociaux*, 51(2):146–152.

172. Nadell, C. D., Bucci, V., Drescher, K., Levin, S. A., Bassler, B. L., and Xavier, J. B. (2013). Cutting through the complexity of cell collectives. *Proceedings of the Royal Society B: Biological Sciences*, 280(1755):20122770.

173. Naug, D. (2008). Structure of the social network and its influence on transmission dynamics in a honeybee colony. *Behavioral Ecology and Sociobiology*, 62(11):1719–1725.

174. Naug, D. and Gadagkar, R. (1998). The role of age in temporal polyethism in a primitively eusocial wasp. *Behavioral Ecology and Sociobiology*, 42(1):37–47.

175. Newcomb, T. M. (1961). The acquaintance process as a prototype of human interaction. In *The Acquaintance Process*, pages 259–261. Holt, Rinehart & Winston, New York, NY, US.

176. Nordlie, P. G. (1958). *A longitudinal study of interpersonal attraction in a natural group setting*. PhD thesis, University of Michigan, Ann Arbor, MI, United States.

177. North Dakota State University Department of Mathematics (2003). The Mathematics Genealogy Project. http://www.genealogy.math.ndsu.nodak.edu/. Accessed 18 Jul 2020.

178. Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

179. Nowak, M. A., Sasaki, A., Taylor, C., and Fudenberg, D. (2004). Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428(6983):646–650.

180. Nowak, M. A. and Sigmund, K. (1998a). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, 194(4):561–574.

181. Nowak, M. A. and Sigmund, K. (1998b). Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577.

182. Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

183. Nowak, M. A., Tarnita, C. E., and Antal, T. (2010). Evolutionary dynamics in structured populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):19–30.

184. O'Donnell, S. and Jeanne, R. L. (1992). Forager success increases with experience in *Polybia occidentalis* (Hymenoptera: Vespidae). *Insectes Sociaux*, 39(4):451–454.

185. Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–5.

186. Ohtsuki, H. and Iwasa, Y. (2004). How should we define goodness? - Reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology*, 231(1):107–120.

187. Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of theoretical biology*, 239(4):435–44.

188. Ohtsuki, H., Nowak, M. A., and Pacheco, J. M. (2007). Breaking the symmetry between interaction and replacement in evolutionary dynamics on graphs. *Physical Review Letters*, 98(10):108106.

189. Okada, I., Sasaki, T., and Nakai, Y. (2017). Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. *Scientific Reports*, 7(1):9737.

190. Okada, I., Sasaki, T., and Nakai, Y. (2018). A solution for private assessment in indirect reciprocity using solitary observation. *Journal of Theoretical Biology*, 455:7–15.

191. Oldroyd, B. P., Wossler, T., and Ratnieks, F. (2001). Regulation of ovary activation in worker honey-bees (*Apis mellifera*): Larval signal production and adult response thresholds differ between anarchistic and wild-type bees. *Behavioral Ecology and Sociobiology*, 50(4):366–370.

192. Overgoor, J., Benson, A. R., and Ugander, J. (2019). Choosing to grow a graph: Modeling network formation as discrete choice. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, pages 1409–1420, New York, New York, USA. Association for Computing Machinery.

193. Oxley, P. R., Ji, L., Fetter-Pruneda, I., McKenzie, S. K., Li, C., Hu, H., Zhang, G., and Kronauer, D. J. (2014). The genome of the clonal raider ant *Cerapachys biroi*. *Current Biology*, 24(4):451–458.

194. Pacala, S., Gordon, D., and Godfray, H. (1996). Effects of social group size on information transfer and task allocation. *Evolutionary Ecology*, 10(2):127–165.

195. Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. Techreport, Stanford InfoLab.

196. Page, R. E. and Mitchell, S. D. (1998). Self-organization and the evolution of division of labor. *Apidologie*, 29(1-2):171–190.

197. Pamminger, T., Foitzik, S., Kaufmann, K. C., Schützler, N., and Menzel, F. (2014). Worker personality and its association with spatially structured division of labor. *PLoS ONE*, 9(1):e79616.

198. Pande, S. and Velicer, G. J. (2018). Chimeric synergy in natural social groups of a cooperative microbe. *Current Biology*, 28(2):262–267.

199. Pankiw, T. and Page, R. E. (2000). Response thresholds to sucrose predict foraging division of labor in honeybees. *Behavioral Ecology and Sociobiology*, 47(4):265–267.

200. Pankiw, T. and Page Jr., R. E. (1999). The effect of genotype, age, sex, and caste on response thresholds to sucrose and foraging behavior of honey bees (*Apis mellifera* L.). *Journal of Comparative Physiology A*, 185(2):207–213.

201. Pankiw, T., Page Jr, R. E., and Kim Fondrk, M. (1998). Brood pheromone stimulates pollen foraging in honey bees (*Apis mellifera*). *Behavioral Ecology and Sociobiology*, 44(3):193–198.

202. Park, S., Bizyaeva, A., Kawakatsu, M., Franci, A., and Leonard, N. E. (2022). Tuning cooperative behavior in games with nonlinear opinion dynamics. *IEEE Control Systems Letters*, 6:2030–2035.

203. Perna, A. and Theraulaz, G. (2017). When social behaviour is moulded in clay: On growth and form of social insect nests. *Journal of Experimental Biology*, 220(1):83–91.

204. Pettit, B., Flack, A., Freeman, R., Guilford, T., and Biro, D. (2013). Not just passengers: Pigeons, *Columba livia*, can learn homing routes while flying with a more experienced conspecific. *Proceedings of the Royal Society B: Biological Sciences*, 280(1750):20122160.

205. Pierson, P. and Schickler, E. (2020). Madison's constitution under stress: A developmental analysis of political polarization. *Annual Review of Political Science*, 23(1):37–58.

206. Pinter-Wollman, N., Hobson, E. A., Smith, J. E., Edelman, A. J., Shizuka, D., de Silva, S., Waters, J. S., Prager, S. D., Sasaki, T., Wittemyer, G., Fewell, J., and McDonald, D. B. (2014). The dynamics of animal social networks: Analytical, conceptual, and theoretical advances. *Behavioral Ecology*, 25(2):242–255.

207. Porter, M. A. (2020). Nonlinearity + networks: A 2020 vision. In Kevrekidis, P. G., Cuevas-Maraver, J., and Saxena, A., editors, *Emerging Frontiers in Nonlinear Science*, Nonlinear Systems and Complexity, pages 131–159. Springer International Publishing, Cham, Germany.

208. Pósfai, M. and D'Souza, R. M. (2018). Talent and experience shape competitive social hierarchies. *Physical Review E*, 98(2):020302.

209. R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

210. Radzvilavicius, A. L., Kessinger, T. A., and Plotkin, J. B. (2021). Adherence to public institutions that foster cooperation. *Nature Communications*, 12(1):3567.

211. Radzvilavicius, A. L., Stewart, A. J., and Plotkin, J. B. (2019). Evolution of empathetic moral evaluation. *eLife*, 8:e44269.

212. Rand, D. G., Pfeiffer, T., Dreber, A., Sheketoff, R. W., Wernerfelt, N. C., and Benkler, Y. (2009). Dynamic remodeling of in-group bias during the 2008 presidential election. *Proceedings of the National Academy of Sciences*, 106(15):6187–6191.

213. Ravary, F., Jahyny, B., and Jaisson, P. (2006). Brood stimulation controls the phasic reproductive cycle of the parthenogenetic ant *Cerapachys biroi*. *Insectes Sociaux*, 53(1):20–26.

214. Ravary, F. and Jaisson, P. (2002). The reproductive cycle of thelytokous colonies of *Cerapachys biroi* Forel (Formicidae, Cerapachyinae). *Insectes Sociaux*, 49(2):114–119.

215. Ravary, F. and Jaisson, P. (2004). Absence of individual sterility in thelytokous colonies of the ant *Cerapachys biroi* Forel (Formicidae, Cerapachyinae). *Insectes Sociaux*, 51(1):67–73.

216. Ravary, F., Lecoutey, E., Kaminski, G., Châline, N., and Jaisson, P. (2007). Individual experience alone can generate lasting division of labor in ants. *Current Biology*, 17(15):1308–1312.

217. Read, A. F. and Taylor, L. H. (2001). The ecology of genetically diverse infections. *Science*, 292(5519):1099–1102.

218. Riolo, R. L., Cohen, M. D., and Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, 414(6862):441–443.

219. Roberts, G. and Sherratt, T. N. (2002). Does similarity breed cooperation? *Nature*, 418(6897):499–500.

220. Robinson, G. (1992). Regulation of division of labor in insect societies. *Annual Review of Entomology*, 37(1):637–665.

221. Robinson, G. E. (1987). Regulation of honey bee age polyethism by juvenile hormone. *Behavioral Ecology and Sociobiology*, 20(5):329–338.

222. Rosenthal, S. B., Twomey, C. R., Hartnett, A. T., Wu, H. S., and Couzin, I. D. (2015). Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *Proceedings of the National Academy of Sciences*, 112(15):4690–4695.

223. Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

224. Sánchez-Tójar, A., Schroeder, J., and Farine, D. R. (2018). A practical guide for inferring reliable dominance hierarchies and estimating their uncertainty. *Journal of Animal Ecology*, 87(3):594–608.

225. Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3490–4.

226. Santos, F. P., Santos, F. C., and Pacheco, J. M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695):242–245.

227. Sasaki, T., Okada, I., and Nakai, Y. (2017). The evolution of conditional moral assessment in indirect reciprocity. *Scientific Reports*, 7(1):1–8.

228. Sayama, H., Pestov, I., Schmidt, J., Bush, B. J., Wong, C., Yamanoi, J., and Gross, T. (2013). Modeling complex systems with adaptive networks. *Computers & Mathematics with Applications*, 65(10):1645–1664.

229. Schmid, L., Shati, P., Hilbe, C., and Chatterjee, K. (2021). The evolution of indirect reciprocity under action and assessment generosity. *Scientific Reports*, 11(1):17443.

230. Seeley, T. (2010). *Honeybee Democracy*. Princeton University Press.

231. Seeley, T. D. (1982). Adaptive significance of the age polyethism schedule in honeybee colonies. *Behavioral Ecology and Sociobiology*, 11(4):287–293.

232. Seeley, T. D., Visscher, P. K., Schlegel, T., Hogan, P. M., Franks, N. R., and Marshall, J. A. R. (2012). Stop signals provide cross inhibition in collective decision-making by honeybee swarms. *Science*, 335(6064):108–111.

233. Sendova-Franks, A. B. and Franks, N. R. (1995). Spatial relationships within nests of the ant *Leptothorax unifasciatus* (Latr.) and their implications for the division of labour. *Animal Behaviour*, 50(1):121–136.

234. Settle, J. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge University Press.

235. Shizuka, D. and McDonald, D. B. (2015). The network motif architecture of dominance hierarchies. *Journal of The Royal Society Interface*, 12(105):20150080.

236. Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482.

237. Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., and Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences*, 104(44):17435–17440.

238. Spaethe, J. and Weidenmüller, A. (2002). Size variation and foraging rate in bumblebees (*Bombus terrestris*). *Insectes Sociaux*, 49(2):142–146.

239. Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., and Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature*, 573(7772):117–121.

240. Strassmann, J. E., Zhu, Y., and Queller, D. C. (2000). Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature*, 408(6815):965–967.

241. Stroeymeyt, N., Casillas-Pérez, B., and Cremer, S. (2014). Organisational immunity in social insects. *Current Opinion in Insect Science*, 5:1–15.

242. Sumpter, D. J. T. (2006). The principles of collective animal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1465):5–22.

243. Sunstein, C. (2001). *Republic.com*. Princeton University Press.

244. Taagepera, R. and Grofman, B. (1985). Rethinking duverger's law: Predicting the effective number of parties in plurality and pr systems - parties minus issues equals one*. *European Journal of Political Research*, 13(4):341–352.

245. Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, 33(1):1–39.

246. Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1(2):149–178.

247. Tarnita, C. E., Antal, T., Ohtsuki, H., and Nowak, M. A. (2009a). Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21):8601–8604.

248. Tarnita, C. E., Ohtsuki, H., Antal, T., Fu, F., and Nowak, M. A. (2009b). Strategy selection in structured populations. *Journal of Theoretical Biology*, 259(3):570–581.

249. Taylor, C., Fudenberg, D., Sasaki, A., and Nowak, M. A. (2004). Evolutionary game dynamics in finite populations. *Bulletin of Mathematical Biology*, 66(6):1621–1644.

250. Taylor, D., Meyer, S. A., Clauset, A., Porter, M. A., and Mucha, P. J. (2017a). Data from: PhD Exchange in the Mathematics Genealogy Project. https://sites.google.com/site/danetaylorresearch/data. Accessed 18 Jul 2020.

251. Taylor, D., Myers, S. A., Clauset, A., Porter, M. A., and Mucha, P. J. (2017b). Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574.

252. Taylor, D., Porter, M. A., and Mucha, P. J. (2019). Supracentrality analysis of temporal networks with directed interlayer coupling. In Holme, P. and Saramäki, J., editors, *Temporal Network Theory*, Computational Social Sciences, pages 325–344. Springer International Publishing, Cham, Germany.

253. Taylor, P. D. and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1):145–156.

254. Teseo, S., Châline, N., Jaisson, P., and Kronauer, D. J. C. (2014). Epistasis between adults and larvae underlies caste fate and fitness in a clonal ant. *Nature Communications*, 5(1):3363.

255. Theraulaz, G., Bonabeau, E., and Deneubourg, J. L. (1998). Response threshold reinforcement and division of labour in insect societies. *Proceedings of the Royal Society B: Biological Sciences*, 265(1393):327–332.

256. Tokita, C. K. and Tarnita, C. E. (2020). Social influence and interaction bias can drive emergent behavioural specialization and modular social networks across systems. *Journal of The Royal Society Interface*, 17(162):20190564.

257. Tomasello, M. and Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology*, 64(1):231–255.

258. Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

259. Traulsen, A., Claussen, J. C., and Hauert, C. (2006). Coevolutionary dynamics in large, but finite populations. *Physical Review E*, 74(1):011901.

260. Traulsen, A. and Nowak, M. A. (2007). Chromodynamics of cooperation in finite populations. *PLOS ONE*, 2(3):e270.

261. Traulsen, A., Pacheco, J. M., and Nowak, M. A. (2007). Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology*, 246(3):522–529.

262. Treier, S. and Hillygus, D. S. (2009). The nature of political ideology in the contemporary electorate. *Public Opinion Quarterly*, 73(4):679–703.

263. Trible, W., McKenzie, S. K., and Kronauer, D. J. C. (2020). Globally invasive populations of the clonal raider ant are derived from Bangladesh. *Biology Letters*, 16(6):20200105.

264. Tripet, F. and Nonacs, P. (2004). Foraging for work and age-based polyethism: The roles of age and previous experience on task choice in ants. *Ethology*, 110(11):863–877.

265. Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57.

266. Tsuji, K. and Yamauchi, K. (1995). Production of females by parthenogenesis in the ant, *Cerapachys biroi*. *Insectes Sociaux*, 42:333–336.

267. Uchida, S. (2010). Effect of private information on indirect reciprocity. *Physical Review E*, 82(3):036111.

268. Ulrich, Y., Burns, D., Libbrecht, R., and Kronauer, D. J. (2016). Ant larvae regulate worker foraging behavior and ovarian activity in a dose-dependent manner. *Behavioral Ecology and Sociobiology*, 70(7):1011–1018.

269. Ulrich, Y., Saragosti, J., Tokita, C. K., Tarnita, C. E., and Kronauer, D. J. C. (2018). Fitness benefits and emergent division of labor at the onset of group-living. *Nature*, 560:635–638.

270. van de Waal, E., Borgeaud, C., and Whiten, A. (2013). Potent social learning and conformity shape a wild primate's foraging decisions. *Science*, 340(6131):483–485.

271. Vehrencamp, S. L. (1983). A model for the evolution of despotic versus egalitarian societies. *Animal Behaviour*, 31(3):667–682.

272. Waibel, M., Floreano, D., Magnenat, S., and Keller, L. (2006). Division of labour and colony efficiency in social insects: Effects of interactions between genetic architecture, colony kin structure and rate of perturbations. *Proceedings of the Royal Society B: Biological Sciences*, 273(1595):1815–1823.

273. Wang, X., Sirianni, A. D., Tang, S., Zheng, Z., and Fu, F. (2020). Public discourse and social network echo chambers driven by socio-cognitive biases. *Physical Review X*, 10(4):041042.

274. Ward, A. J. W., Sumpter, D. J. T., Couzin, I. D., Hart, P. J. B., and Krause, J. (2008). Quorum decision-making facilitates information transfer in fish shoals. *Proceedings of the National Academy of Sciences*, 105(19):6948–6953.

275. Webster, S. W. and Abramowitz, A. I. (2017). The ideological foundations of affective polarization in the U.S. electorate. *American Politics Research*, 45(4):621–647.

276. Weidenmüller, A. (2004). The control of nest climate in bumblebee (*Bombus terrestris*) colonies: Interindividual variability and self reinforcement in fanning response. *Behavioral Ecology*, 15(1):120–128.

277. Weidenmüller, A., Chen, R., and Meyer, B. (2019). Reconsidering response threshold models—short-term response patterns in thermoregulating bumblebees. *Behavioral Ecology and Sociobiology*, 73(8):1–13.

278. Werfel, J., Petersen, K., and Nagpal, R. (2014). Designing collective behavior in a termite-inspired robot construction team. *Science*, 343(6172):754–758.

279. Westley, P. A. H., Berdahl, A. M., Torney, C. J., and Biro, D. (2018). Collective movement in ecology: From emerging technologies to conservation and management. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746):20170004.

280. Wetterer, J. (1999). The ecology and evolution of worker size-distribution in leaf-cutting ants (Hymenoptera: Formicidae). *Sociobiology*, 34:119–144.

281. Wilson, E. O. (1980). Caste and division of labor in leaf-cutter ants (Hymenoptera: Formicidae: *Atta*). *Behavioral Ecology and Sociobiology*, 7(2):157–165.

282. Yamagishi, T. and Mifune, N. (2008). Does shared group membership promote altruism?: Fear, greed, and reputation. *Rationality and Society*, 20(1):5–30.