Evolutionarily stable networks[*]

Péter Bayer[†]

March 13, 2024

Abstract

This paper studies the evolution of behavior governing strategic network formation. I first propose a general framework of evolutionary selection in non-cooperative games played in heterogeneous groups under assortative matching. I show that evolution selects strategies that (i) execute altruistic actions towards others in the interaction group with rate of altruism equal to the rate of assortative matching and (ii) are stable against pairwise coalitional deviations under two qualifications: pairs successfully coordinate their deviations with probability equaling the rate of assortative matching and externalities are taken into account with the same weight. I then restrict the domain of interaction games to strategic network formation and define a new stability concept for networks called 'evolutionarily stable networks'. The concept fuses ideas of solution concepts used by evolutionary game theory and network formation games. In a game of communication, evolutionarily stable networks prescribe equal information access. In the classic co-authorship game only the least efficient network, the complete network, is evolutionarily stable. Finally, I present an evolutionary model of homophilistic network formation between identity groups and show that extreme high degrees of homophily may persist even in groups with virtually no preference for it; thus societies may struggle to eliminate segregation between identity groups despite becoming increasingly tolerant.

Keywords: Networks, evolution, relatedness, stability, homophily.

JEL classification: C73, D85.

[*] Job market paper. Funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d'Avenir, grant ANR-17-EURE-0010) and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 789111 - ERC EvolvingEconomics awarded to Ingela Alger) are gratefully acknowledged. I thank Ingela Alger, Matteo Bizzarri, Francis Bloch, Yann Bramoullé, Ani Guerdjikova, Benjamin Golub, Jean-Jacques Herings, Marion Hoffman, Matthew Jackson, Marijn Keijzer, Maria Kleshnina, Christoph Kuzmics, Mánuel Mágó, Jonathan Newton, Ronald Peeters, Jorge Peña, Jérôme Renault, Alexandros Rigos, Paul Seabright, and Anastas Tenev for comments and discussions.

[†] Toulouse School of Economics, CNRS, University of Toulouse Capitole, 1 Esplanade de l'université 31080 Toulouse. E-mail: peter.bayer@tse-fr.eu

1 Introduction

The economic study of networks relies in large part on the *homo economicus* paradigm. Models of endogenously arising social and economic networks assume that individuals divide their budgets (time, attention, material resources) to form and sustain costly ties in anticipation of material benefits (information access, trade opportunities, risk management) or hedonistic rewards (leisure opportunities, belonging) that may depend intricately on the network structure. The theory of such *network formation games* (Jackson and Wolinsky, 1996; Bala and Goyal, 2000) have provided highly valuable insights in understanding the formation of endogenously formed ties between rational individuals. Interactions in networks determine a wide range of individual economic outcomes, including on the labor market (Myers and Shultz, 1951; Granovetter, 1973; Bayer et al., 2008; Beaman, 2012), educational attainment (Calvó-Armengol et al., 2009; Patacchini et al., 2017), and criminal behavior (Patacchini and Zenou, 2008, 2012; Lee et al., 2021), making networks an important field of study in mainstream economics. However, a comprehensive evolutionary approach to strategic network formation, connecting the field to the biological realities of the *homo sapiens*, has not yet been attempted.

Human tendencies to organize into social networks, like all behavior pertaining to social organization, is a product of Darwinian evolution by natural selection. While various fields of social science have put forward theories about the evolutionary mechanics and functions of social ties between non-kin (Hamilton et al., 2007; Apicella et al., 2012; Christakis and Fowler, 2014; Boyd and Richerson, 2022), there seems to be little disagreement that networks matter for evolutionary success. Thus, behavioral traits responsible for the creation of social ties are subject to evolutionary forces. An evolutionary framework of network formation is further motivated by the general view of behavioral genetics that genes account for up to half of individual variation in a wide array of behavioral traits,[1] as well as specific evidence on the genetic heritability traits responsible for creating social networks. In particular, three social network characteristics have been shown to be heritable: in-degree, betweenness centrality, and clustering (Fowler et al., 2009).[2] Inherited traits therefore play a role in determining the size of individuals' social network, their position within it, and its density around them.

In this paper, I propose an evolutionary model of endogenous network formation. Following

[1] A succinct summary is the first of "Turkheimer's three laws of behavior genetics", stating "All behavioral traits are heritable" (Turkheimer, 2000).

[2] As common in behavioral genetics, Fowler et al. (2009) uses a twin study design in which the similarity of social networks of monozygotic (identical) twins is compared to that of same-sex dizygotic (fraternal) twins. As both sets of twins share a common environment, the higher similarity of the three measured social network characteristics found in the former group is attributed to higher genetic relatedness.

the logic of evolutionary game theory, strategies represent inherited predispositions and behavioral traits rather than conscious decisions. In the network formation setting, actions represent behavioral patterns that form bilateral ties between players, which is transmitted genetically to future generations. The networks that form are thus not results of rational actions or biased learning but outcomes of Darwinian forces acting on genes over evolutionary time.

To build the model, I rely on the theory of static evolutionary games, using evolutionarily stable strategy (ESS) as a solution concept (Maynard Smith and Price, 1973; Maynard Smith, 1982). The model has two key elements: (i) kin selection (Hamilton, 1964) modeled through assortative matching of strategies (Bergstrom, 2003; Jensen and Rigos, 2018) and (ii) heterogeneous interaction modeled through types of a Bayesian game (Harsanyi, 1967). Assortative matching ensures that possibilities for coordination necessary for the formation of bilateral links arises naturally in the model. Types encode the possible positions that players may take in an interaction game; in network formation games, they represent the nodes of the forming network. The direct antecedent of the model is Alger and Weibull (2013) which features heterogeneous pairwise interaction and Alger and Weibull (2016) which features homogeneous N-player interaction, both under assortative matching. Crucially, however, this paper only tackles the case of strategy evolution.

In the base model, which is not specific to network formation, I define two generalizations of the ESS concept, called 1-ESS and 2-ESS. While the classic ESS concept contains strategies that are stable against any mutation, the 1-ESS and 2-ESS prescribe stability against mutations that deviate from the status quo strategy in one and two positions, respectively. The two stability concepts are used in the definition of evolutionarily stable networks under one-sided and two-sided link formation, respectively. As in past applications of evolutionary game theory within economics (Güth and Yaari, 1992; Bester and Güth, 1998; Sethi and Somanathan, 2001; Alger and Weibull, 2013), persistent deviations from *homo economicus* behavior arise endogenously. Characterizing the 1-ESS and the 2-ESS gives two such results: In the 1-ESS, evolution selects for strategies that execute altruistic actions towards the rest of the interaction group with rate of altruism equaling the strength of assortative matching (Proposition 1). In 2-ESS, evolution further selects for strategies that are stable against pairwise coalitional deviations where the pair coordinates with probability equaling the strength of assortative matching and externalities are taken into account by the same rate (Proposition 2). The evolution of altruistic behavior through kin selection is a known phenomenon, going back to Hamilton (1963, 1964); the result I obtain amounts to a form of Hamilton's rule, stated for general N-player games with heterogeneous interaction. The evolution of coalitional thinking through kin selection is, to my knowledge, a new result, as previous work on the evolution of coalitions rely on coalitional behavior forming at random (Newton, 2012) or an exogenous capacity for shared intentionality (Newton and Angus, 2015; Newton, 2017).

I then restrict the base model to strategies that correspond to a network formation game. On the broadest level, this contributes to the theory of evolutionary foundations of economic behavior which included equilibrium behavior in games (Samuelson, 1988), intertemporal preferences (Fudenberg and Maskin, 1990), and risk attitudes (Robson, 1996). In the most specific level, this work adds to the theory of equilibrium concepts in network formation games (Jackson and Watts, 2001; Jackson and Van den Nouweland, 2005; Bloch and Jackson, 2006, 2007; Sadler, 2022) by defining evolutionarily stable networks, combining equilibrium concepts in the network formation games and evolutionary game theory literatures. For network formation under one-sided link formation, evolutionarily stable networks are defined through the 1-ESS and obtain as Nash networks under altruism, in keeping with Proposition 1. For two-sided link formation I use the 2-ESS to build an equilbrium concept that incorporates the pairwise coalitional thinking of Proposition 2. Proposition 3 characterizes such 'pair-evolutionarily stable networks' whose properties are reminiscent of pairwise stable networks (Jackson and Wolinsky, 1996) and pairwise stable networks with transfers (Bloch and Jackson, 2006) under altruism.

The evolutionary approach unlocks unique insights into network formation behavior. When building and severing connections, individuals consider the consequences of their actions on their would-be (former) partners, as well as the rest of the interaction group. Recently, the literature began incorporating exogenously given altruistic preferences in games on networks (Bourlès et al., 2017, 2021); by the evolutionary paradigm, such behavior arises endogenously. To showcase specific properties and insights, this paper presents three applications of the model: a communication game, the classic co-authorship game, and a model of homophilistic linking between identity groups.

In a game of non-rival communication with one-sided link formation, evolutionarily stable networks prescribe the formation of regular networks if assortativity is weak (Proposition 4). Evolved altruistic behavior leads to nodes sponsoring links to those who are worst off, leading to equality of information access in equilibrium. If assortativity is strong, evolutionarily stable networks are identical in structure to Nash networks, allowing for irregular structures such as multi-centered stars. However, the direction of links is reversed and the source of information inequalities is different. In Nash networks, degree inequalities arise from the indifference of *homo economicus* players in whom they link to, thus some nodes end up with more links, more information access than others, and end up not needing to sponsor links of their own.

In evolutionarily stable networks under strong assortativity, it is the pure sponsors who create degree inequalities; compelled by altruism to satisfy the needs of others, they initiate more than the optimal number of links while receiving none themselves (Proposition 5).

Under two-sided link formation, evolutionarily stable networks being stable against coalitionary deviations of pairs provides an evolutionary foundation for cooperative models of strategic network formation and stability concepts (Jackson and Wolinsky, 1996; Jackson and Van den Nouweland, 2005). The specific predictions, however, are different. In Jackson and Wolinsky (1996)'s classic co-author game, coalitional formation may destabilize pairwise stable structures. Inefficiencies arise from two sources: (i) coalitional formation of links compels individuals to agree to too many collaborations, and (ii) if assortativity is weak, altruism is not strong enough to deter the formation of links that harms the pairs' neighbors. Instead of the pairwise stable structure, complete components of different sizes, the complete network forms with at most a single "monogamous" pair staying out. However, the efficient structure, "full monogamy" becomes the least unstable (requiring the lowest degree of assortativity to be made stable) of all other structures (Proposition 6).

Finally, I apply the model to capture homophilistic network formation between identity groups. Homophily, the tendency to preferentially link to similar individuals, is a stumbling block of information exchange and finding consensus (Golub and Jackson, 2012) as well as a driver of economic inequality, immobility, and inefficiencies (Jackson, 2021). Structural models have proposed individual bias as sources of homophily (Currarini et al., 2009; Mele, 2022)[3] while evolutionary models of homophily posit exogenous fitness advantage of associating with in-group members (Fu et al., 2012). Here, I introduce a model in which individuals oppose all cross-group linking (e.g., a preference for "apartheid" or "anti-miscegenation"), a model producing identical behavior as models of in-group bias under the *homo economicus* paradigm. Under the evolutionary paradigm, however, individuals partially internalize the external effects of their linking choices. As a result, even individuals who do not have a preference for in-group neighbors display preferential in-group linking as long as the interaction group contains individuals who do (Proposition 7). As groups get large and the prevalence of bias in the population disappears, there is a stark divergence of predictions by the two paradigms: the *homo economicus* paradigm prescribes a vanishing homophily, while under the evolutionary paradigm, homophily converges to its maximum possible value (Proposition 8). The evolutionary lens thus explains why even a marginal prevalence of bias in the population can create strongly homophilistic networks. More

[3] Bramoullé et al. (2012) and Currarini et al. (2016) use a similar channel: searching for in-group friends is less costly than out-group ones, hence meeting opportunities are more frequent between members of the same identity group.

broadly, it provides an explanation for the persistence of biased linking and the sluggishness to eliminate segregation even in increasingly tolerant societies.

This paper proceeds as follows: Section 2 introduces general evolutionary games where interactions occur in heterogeneous groups and derives two equilibrium concepts based on the ESS, the 1-ESS and the 2-ESS. Section 3 restricts attention to network formation games and defines evolutionarily stable networks in games with one-sided and two-sided link formation. Section 4 presents three applications, a game of communication, the co-authorship game, and homophily. Section 5 concludes.

2 Evolutionary games on heterogeneous groups

In this section I introduce a multi-player evolutionary game played in heterogeneous groups under assortative matching of strategies, define evolutionarily stable strategies, and characterize two generalizations, the 1-ESS and the 2-ESS.

2.1 Interaction in heterogeneous groups

Consider a countably infinite population of individuals who live for a single period. At birth, the individuals, also called players, are sorted into groups of $N > 1$. Groups represent typical interaction situations of the population, such as a village, a tribe, a school class, or an office floor. Each group plays the same one-shot non-cooperative *interaction game*. Individuals are born with a strategy which determines their play, while the payoffs received from the interaction game determine their *evolutionary fitness*. The fitness of a strategy, determining which strategies may be selected by the evolutionary process, is the average fitness of all individuals in the population with that strategy.

Group interaction is heterogeneous. Each individual in each group is assigned a *position* $i \in I = \{1, \ldots, N\}$ uniformly at random. Positions represent differences in the interaction circumstances of players be they in individual characteristics not related to network formation, social hierarchy, economic situation, or geographic location. In game theory terms, positions describe the possible asymmetry of the interaction game (e.g., 'sender' and a 'receiver' in an ultimatum game). In every group, the individual in position $i \in I$ has *action set* X_i. Let $X = \prod_{i \in I} X_i$ denote the action space of the interaction game. The profile of actions $(x_1, \ldots, x_N)$ with $x_i \in X_i$ for each $i \in I$ determines the payoffs (fitness) of the players participating in the interaction with the player in position i receiving $u_i(x_1, \dots, x_N)$, for $u_i \colon X \to \mathbb{R}$.[4][5]

As each individual may end up in any position, their *strategy*, "what an individual will do in any situation in which it may find itself" (Maynard Smith, 1982), specifies for each position i which action the individual will play if they are assigned that position. The set of strategies of an individual is an N-vector of the action sets for each position which is mathematically identical to the action space of the interaction game, X. A typical element of X is called a strategy, denoted by x. A player born with strategy x is called an 'x-player'. Figure 1 depicts the components of the model for interaction games of $N = 4$ positions.

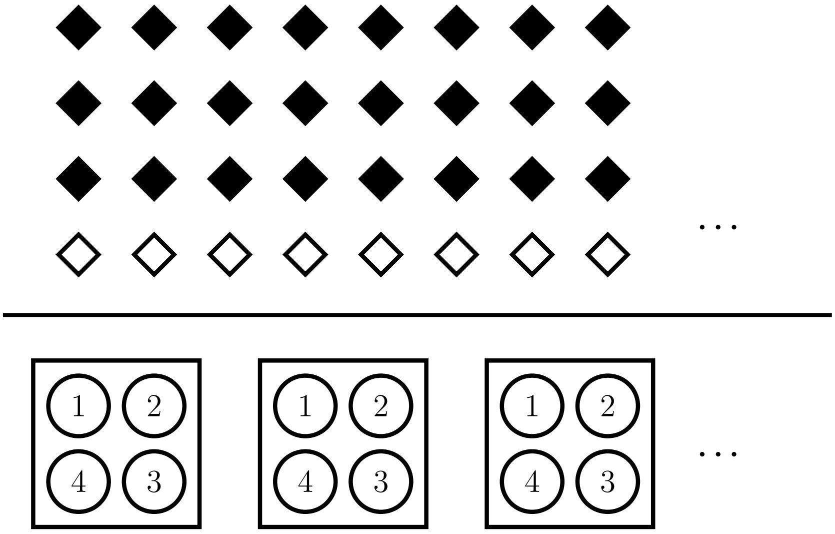

Figure 1: The model components, players (diamonds), positions (numbered circles), and interaction groups (squares) with $N = 4$ positions per interaction group. In this example, two strategies are represented in the population, black and white, with frequencies 0.75 and 0.25.

To define evolutionarily stable strategies, it is sufficient to consider the case where two strategies are represented in the population, a resident (typically with frequency close to 1), and a mutant (typically with frequency close to 0). For two strategies represented in the population x an y, let $q = (q_x, q_y) \in [0, 1]^2$ with $q_x + q_y = 1$ denote the vector of *frequencies*, the relative proportion of players with strategies x and y, respectively. Figure 1 shows two strategies, black and white, with frequencies 0.75 and 0.25.

I assume that individuals are matched into interaction groups assortatively based on their strategies. Following Bergstrom (2003), let $r \in [0, 1]$ be given as the rate (or index) of assortativity. For represented strategies x an y and positions $i, j \in I$, conditional on the event that

[4] Positions correspond to the "roles" in Alger and Weibull (2013)'s asymmetric interactions. Formally, they are payoff-types of a Bayesian game (Harsanyi, 1967) with the common prior being the uniform distribution over all bijections between the N players sorted into the group and the N possible positions.

[5] In this paper, positions are not heritable, though the model could be extended to include social inertia.

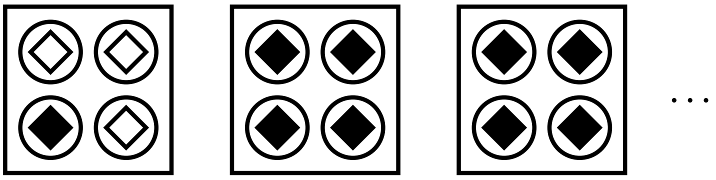

Figure 2: Players are assigned to positions assortatively, making interaction between individuals of the same strategy more likely than it would be in a well-mixed population. The strategies determine how each position will behave in the interaction game.

position i of a group is assigned to an x-player, the probability that position j of the same group is also assigned to an x-player, given frequency vector q, is

$$P_q(x|x) = r + (1-r)q_x.$$

Conditional on the same event, the probability that position j of the group is assigned to a y-player is

$$P_q(y|x) = (1-r)q_y.$$

The interpretation is the following: If $r = 0$, the population is well-mixed and there is no assortative matching of strategies. If $r = 1$, there is full assortative matching and the probability that an individual encounters any strategy other than their own is 0. For in-between values, r is the rate by which an individual meets others with the same strategy and $1 - r$ is the rate by which they meet a random strategy drawn from a well-mixed population. The mechanics of the matching process is independent of the players' strategies and their position in the interaction game. Furthermore, the matching process displays *conditional independence* (Alger and Weibull, 2016); the probability of encountering an x-player or a y-player at position j, conditional on the strategy in position i is independent of the strategy in position $k \neq j$. Section 2.3.6. of Jensen and Rigos (2018) shows the existence of such a matching process. As typical in the literature, I assume that the matching process is exogenous with assortativity arising due to local interaction and local dispersion of offspring; see Alger et al. (2020) for a model of endogenous assortativity arising in an island model (Frank, 1998; Rousset, 2004) featuring migration between homogenous interaction groups. Figure 2 shows a realization of the assortative matching process with most black strategies matched to majority black interaction groups and white strategies matched to majority white ones.

Once individuals are matched into interaction groups and positions are assigned, players execute the actions prescribed by their strategies in their positions: an x-player in position $i \in I$ executes x_i. Interaction payoffs are awarded. Figure 3 shows an example of network formation if black and white strategies are different linking strategies.

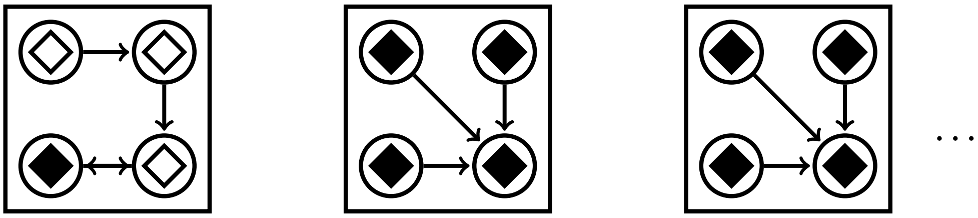

Figure 3: The interaction game is played in each interaction group. The strategies encode what actions the players execute in each position, hence the actions are both position- and strategy-dependent. The focus is on interaction games being network formation games, hence the figure showcases an example where actions represent linking decisions. In this figure, positions with white strategies link to the position one number greater than they are, while positions with black strategies link to position 3 no matter what position they are in. The expected interaction payoff of each strategy determines its evolutionary fitness.

The *evolutionary fitness* of a strategy is given by the average interaction payoff of that strategy across all positions and all groups. Given represented strategies x and y and frequency vector q, the fitness of strategy x is thus given by

$$V_q(x) = \frac{1}{N}\sum_{i\in I}\sum_{(y^j)_{j\in I\setminus\{i\}}\in\{x,y\}^{I\setminus\{i\}}}\left(\prod_{j\in I\setminus\{i\}} P_q(y^j|x)\right) u_i(x_i,(y^j_j)_{j\in I\setminus\{i\}}).\text{[6]} \tag{1}$$

The interpretation of (1) is as follows: The payoff of strategy x is averaged over all positions $i \in I$, multiplied by the probability of getting assigned that position, $1/N$, with each position's value evaluated against all action profiles they can encounter in the interaction game, multiplied by the probability of encountering the strategy profiles in the interaction game that produce the action profile.

2.2 Mutation and evolutionary stability

Evolutionary stability is the standard solution concept of evolutionary game theory. While it is a static concept, not requiring an explicit definition of evolutionary dynamics, it is typically interpreted as an outcome of Darwinian selection acting on competing strategies over a long period of time. A strategy is called *resident* if it is played by almost every individual in the population. Such a strategy can only be an outcome of an evolutionary process if, in fitness terms, it outperforms any *mutant* strategy appearing in the population at random at low frequencies. If, under such conditions, the mutant's fitness is higher than the resident's, it *invades* the resident. As a result of such an invasion, the mutant may replace the resident, co-exist with the resident, or open the possibility for other mutations to spread; in any case, the resident existing by itself

[6] In the formula, the strategy of the player sorted into position $j \in I \setminus \{i\}$ is denoted by y^j. The dual index y^j_j indicates that the player executes the action corresponding to this position, the jth component of y^j.

is not a stable outcome of an evolutionary process. Formally, stability against invasion is defined as follows:

Definition 1. For represented strategies, x an y and let $q(\varepsilon)$ denote the frequency distribution such that $q_x(\varepsilon) = 1 - \varepsilon$ and $q_y(\varepsilon) = \varepsilon$. Then, x (the resident) is called *evolutionarily stable against y* (the mutant) if there exists $\delta > 0$ such that for every $\varepsilon \in (0, \delta)$ we have

$$V_{q(\varepsilon)}(x) > V_{q(\varepsilon)}(y), \tag{2}$$

and *neutrally stable against y* if (2) is satisfied with weak inequality.

By the fact that fitness is a continuous function of ε, the conditions that ensure that x is stable against y imply a property in the limit as ε goes to zero. Then, x is stable against y if

$$V_{q(0)}(x) > V_{q(0)}(y), \tag{3}$$

or if

$$\begin{gathered} V_{q(0)}(x) = V_{q(0)}(y), \text{ and} \\ \frac{\partial}{\partial \varepsilon} V_{q(\varepsilon)}(x) \bigg|_{\varepsilon=0} > \frac{\partial}{\partial \varepsilon} V_{q(\varepsilon)}(y) \bigg|_{\varepsilon=0}, \end{gathered} \tag{4}$$

and neutrally stable under the same conditions with weak inequality in (4).

Mutations occur naturally; in genetic evolution, they occur due to copying errors during mitosis. Most mutations only effect a marginal change upon status quo strategy as larger differences require more errors to occur independently. Thus, mutant strategies may be highly similar to the resident it mutated from. In this paper I use an abstract, reduced-form object called a *mutation protocol* to capture which strategies can mutate into which. Formally, a mutation protocol is a correspondence $\mathcal{M} \colon X \rightrightarrows X$ such that $x \notin \mathcal{M}(x)$. The set $\mathcal{M}(x)$ contains the set of strategies that can obtain from strategy $x \in X$ through mutation. I define evolutionary stability in relation to the possible mutation protocols.

Definition 2. A strategy x is called an *evolutionarily stable strategy (ESS) under mutation protocol $\mathcal{M}$* if it is stable against all possible mutants $y \in \mathcal{M}(x)$. It is called *neutrally evolutionarily stable (nESS) under $\mathcal{M}$* if it is neutrally stable against the same set of strategies.

Naturally, if a strategy is an (n)ESS under a given protocol, it is also one under a more restrictive protocol which allows for fewer mutations, as summarized by the following Lemma.

Lemma 1. *Let mutation protocols $\mathcal{M}$ and $\mathcal{M}'$ be given such that for all $x \in X$ we have $\mathcal{M}(x) \subseteq \mathcal{M}'(x)$. Then, if x^* is an (n)ESS under protocol $\mathcal{M}'$, it is an (n)ESS under protocol $\mathcal{M}$.*

To obtain a metric, I take the Hamming distance on the strategy space X and thus measure the distance between two strategies x and y as the number of positions where they prescribe different actions for the player. Then, for $m \in \{1, \ldots, N\}$, the mutation protocol $\mathcal{H}_m$ is defined as

$$\mathcal{H}_m(x) = \{y \colon H(x, y) \leq m\} \setminus \{x\} \tag{5}$$

for every $x \in X$, where $H(x, y)$ denotes the Hamming distance of x and y, that is, the cardinality of the largest set $M \subseteq I$ for which $y_i \neq x_i$ holds for all $i \in M$.

A subset of evolutionary game theory literature assumes that mutations are small; this would correspond to $\mathcal{M}(x)$ being an ε-ball around x for some metric on X. This assumption reflects the biological reality that genetic mutation of quantitative traits such as body size is slow and gradual. In the present paper, I take a different approach, accounting for the possibility of genetic mutation of qualitative traits measured on discrete scales, cultural evolution (in which mutations amount to norm-breaking or experimentation), as well as the discrete nature of my primary application, network formation.

For simplicity, I call y an m-mutation of x if $y \in \mathcal{H}_m(x)$. I call strategy x an m-ESS if it is stable under protocol $\mathcal{H}_m$, that is, stable against all m-mutations. From Lemma 1, it follows that the set of m-ESSs are nested, with 1-ESS being the most permissive and the N-ESS corresponding to the strictest possible evolutionary stability notion; in the latter, $\mathcal{H}_m(x) = X \setminus \{x\}$ for all x, which amounts to the original definition of the ESS. Of special interest for an analysis focused on network formation are the 1-ESS and the 2-ESS. The former is intended to capture situations in which link formation is a one-sided decision by individual nodes, not requiring consent of the links' recipients, while the latter captures situations where bilateral agreement between nodes is necessary for link formation.

2.3 The evolution of altruism

In static evolutionary games featuring assortative matching, ESSs are typically different from the Nash equilibria of the underlying non-cooperative interaction game. In particular, while in a fitness game, the Nash benchmark features pure self-regard, the ESS concept is known to select strategies that are consistent with special classes of other-regarding preferences such as altruism and spite (Bester and Güth, 1998), reciprocity (Sethi and Somanathan, 2001), and moral behavior (Alger and Weibull, 2013).

This paper's first result shows that if mutations change an individual's play in a single position, then the set of evolutionarily stable strategies prescribe altruistic behavior with rate of altruism equaling the strength of assortative matching in the population.

Proposition 1 (Altruistic actions in 1-ESS)**.** *Given the interaction game u, a strategy $x \in X$ is an ESS of the induced evolutionary game V under the mutation protocol $\mathcal{H}_1$, that is, x is a* 1*-ESS, if and only if*

$$u_i(x) - u_i(y) + r \sum_{j \in I \setminus \{i\}} \Big(u_j(x) - u_j(y) \Big) > 0, \tag{6}$$

for all $i \in I$ and every $y = (y_i, x_{-i})$ with $y_i \neq x_i$.

Furthermore, x is a 1*-nESS if (6) holds with a weak inequality.*

Letting $\tilde{u}_i(x) = u_i(x) + r \sum_{j \neq i} u_j(x)$ denote the game derived from the interaction game u played by altruistic players with rate of altruism equaling r, Proposition 1 shows an equivalence between 1-ESS strategies and strict Nash equilibrium profiles of $\tilde{u}$, as well as 1-nESS strategies and Nash equilibrium profiles of the same game.

To provide intuition in understanding this result and upcoming results in this section, I show how the condition (6) obtains from (3). The rest of the formal proof, showing why (4) ends up not mattering for the 1-ESS is moved to the Appendix which contains all other proofs of this paper.

To obtain the first part of the result, one needs to calculate the payoff (dis)advantage of a y-player appearing in low frequencies in a resident population of x-players. At the limit, with the mutant frequency nearing zero, the average x-player only interacts with other x-players due to

$$P_{q(0)}(x|x) = r + (1-r) \cdot 1 = 1.$$

Thus, the expected payoff of an x-player sorted into any position j is simply $u_j(x)$. The fitness of strategy x is then $1/N \sum_{j \in I} u_j(x)$.

Now turn to y-players. As the mutation affects the action in position i only, if a focal y-player is sorted into position i, they will execute action y_i, while the players in every other position $j \neq i$ execute x_j regardless of whether they are residents or mutants, so the action profile $y = (y_i, x_{-i})$ is played: the mutant's payoff in position i is thus $u_i(y)$. If the y-player is sorted into a position $j \neq i$, they will execute the action x_j as a resident would. Then, by the rules of assortative matching, the probability that position i is assigned to a mutant is

$$P_{q(0)}(y|y) = r + (1-r) \cdot 0 = r.$$

If another y-player is assigned to position i, they execute action y_i and the action profile y is played, giving the focal y-player $u_j(y)$. Otherwise, if position i is assigned to an x-player, they execute action x_i and the action profile x is played, giving the focal y-player $u_j(x)$. The payoff of the focal mutant does not depend on whether residents or mutants are assigned to all other positions $k \neq i$ as both execute the same action, x_k. Thus, the expected payoff of a y-player sorted to position $j \neq i$ equals $ru_j(y)+(1-r)u_j(x)$. The fitness of a focal mutant is the expected interaction fitness across all positions, amounting to $\big(u_i(y) + \sum_{j\neq i}(ru_j(y) + (1-r)u_j(x))\big)/N$. Then, the fitness of the resident is strictly larger than that of the mutant if and only if

$$u_i(x) - u_i(y) + r\Big(\sum_{j\in I\setminus\{i\}} u_j(x) - u_j(y)\Big) > 0,$$

completing this part of the proof.

By Proposition 1, the set of 1-ESSs equals the set of Nash equilibria of a game where individuals internalize part of the externalities of their actions on the interaction group, valuing others' payoffs with a weight equaling the measure of assortativity. Intuitively, a mutant strategy y that compels an individual in position i to take action $y_i \neq x_i$ not only affects the fitness of that individual, but that of every other individual in the interaction group. By the rules of assortative matching, the probability that a player in any other position $j \neq i$ carries the same mutation y, is r. This is despite the fact that every other individual in the interaction group, mutant or resident, behaves exactly like a resident would (due to y being a 1-mutation, i.e., $y_j = x_j$ for $j \neq i$). Hence, such a mutation can only compete with x if, in expectation, it does no more harm to its clones in other positions than the advantage it bestows to the position in which it causes a change of action. Even stronger, a 1-mutation may invade even if it causes a payoff loss to the position in which it switches action, as long as the external effects are positive and large enough.

Proposition 1 generalizes existing results of first-order conditions of evolutionary stability (e.g., Proposition 3 of Alger et al. (2020)) for heterogeneous interactions. As the sets of m-ESSs are nested, every m-ESS is a Nash equilibrium of the game $\tilde{u}$, thus the result is a necessary condition of evolutionary stability. On its own, if mutations arise with likelihood inversely proportional to Hamming distance, 1-ESS strategies are the ones stable against the most likely sets of mutations.

2.4 The evolution of coalitional action

In network formation games with two-sided link formation, ties form between pairs of players by bilateral action. To be able to define the stability of a network against the appearance of links, that is, against strategies that form links between nodes, the next result characterizes the set of 2-ESSs.

Proposition 2 (Coalitional action in 2-ESS). *A strategy x is a 2-ESS, that is, evolutionarily stable under the $\mathcal{H}_2$ protocol, if and only if for every pair of positions $i, j \in I$ and every 2-mutation $y = (y_i, y_j, x_{-ij})$ we have*

$$\begin{aligned} &r\Big(u_i(x) - u_i(y) + u_j(x) - u_j(y)\Big) + (1-r)\Big(u_i(x) - u_i(y^{(i)}) + u_j(x) - u_j(y^{(j)})\Big) + \\ &r \sum_{k \in I \setminus \{i,j\}} \Big(r\big(u_k(x) - u_k(y)\big) + (1-r)\big(2u_k(x) - u_k(y^{(i)}) - u_k(y^{(j)})\big)\Big) > 0, \end{aligned} \tag{7}$$

or (7) holds with equality and

$$\sum_{k \in \{i,j\}} \Big(u_k(y^{(i)}) + u_k(y^{(j)}) - u_k(x) - u_k(y)\Big) + 2r \sum_{k \in I \setminus \{i,j\}} \Big(u_k(y^{(i)}) + u_k(y^{(j)}) - u_k(x) - u_k(y)\Big) > 0, \tag{8}$$

where $y^{(i)} = (y_i, x_{-i})$ and $y^{(j)} = (y_j, x_{-j})$.

As y differs from x in two positions, i, and j, both positions must be assigned to mutants to implement the action profile y. If only one position is assigned to a mutant while the other is assigned to a resident, one of the two action profiles $y^{(i)}$ or $y^{(j)}$, both 1-mutations of x, will be played. If neither position is assigned to a mutant, the action profile x will be played. Condition (7) identifies four terms determining the success of a 2-mutant: (1) the two positions' net benefits of reaching y rather than x, multiplied by the probability of a successful coordination between the positions, r, (2) the two positions' net benefits while failing to coordinate the two-sided move and thus effecting a 1-mutation instead, multiplied by the probability of a coordination failure, $1-r$, (3) the total externalities of implementing y, multiplied by the probability that both i and j are assigned to mutants r^2 (from the point of view of a mutant in position $k \neq i, j$), and (4), the total externalities of implementing a 1-mutation, multiplied by the probability that exactly one of positions i and j is assigned to a mutant, $r(1-r)$. If y is a 1-mutation (e.g., if $y = y^{(i)}$ and $y^{(j)} = x$), then (7) reduces to (6) and (8) cannot be true.

The conditions in Proposition 2 obtain in terms of the total gains/losses for the deviating positions i and j, and externalities for all other positions. As such, if the expected total gains of a mutation, accounting for externalities, are positive for a 2-mutation, it will invade, even if it comes with a loss for one or both deviating positions. In behavioral terms, 2-ESS strategies are stable against coalitional deviations from pairs of positions, as well as individual deviations by single positions.

Remark 1 (Kantian morality, altruism, and coalitional action). Assume $N = 2$ for the remainder of this section. Consider an interaction game played between two players, with one individual, called player A, being an x-player and another, player B, being a y-player for general strategies,

$x = (x_1, x_2)$ and $y = (y_1, y_2)$. Then, the expected (ex ante) payoff of player A of interacting with B, depending only on the players' strategies and thus denoted as the (Bayesian) game $\pi(x, y)$, is given as

$$\pi(x, y) = \frac{1}{2}\left(u_1(x_1, y_2) + u_2(y_1, x_2)\right).$$

The first component is A's interaction payoff if A is assigned to position 1 and thus executing action x_1 with B executing y_2, and the second is the payoff if A is assigned position to 2, thus executing action x_2 with B executing action y_1.

Expressing payoffs in terms of the Bayesian game π, Proposition 1 of Alger and Weibull (2013) states that a strategy x is an ESS if and only if

$$\pi(x, x) > \pi(y, x) + r\big(\pi(y, y) - \pi(y, x)\big), \tag{9}$$

or

$$\begin{aligned} &\pi(x, x) = \pi(y, x) + r\big(\pi(y, y) - \pi(y, x)\big), \text{ and} \\ &\pi(x, y) > \pi(y, y) + r\big(\pi(y, y) - \pi(y, x)\big). \end{aligned} \tag{10}$$

for all $y \in X$.

The authors interpret strategies satisfying condition (9) as equilibria of games played by decision makers adhering to Kantian morality. The strategies satisfying the condition obtain as equilibria of the Bayesian game given by

$$\widehat{\pi}(x, y) = (1 - r)\pi(x, y) + r\pi(x, x),$$

with individuals placing weight $1 - r$ on their actual (ex ante) expected payoffs under strategy profile (x, y) and weight r on the expected payoff they would get if their opponent had the same strategy. The latter component captures individuals' Kantian concern, that is, how much they value strategy x if their interaction partner also adopts that strategy.

If we constrain y to be a 1-mutation, that is, if we take $y = (y_1, x_2)$, then (9) reduces to

$$u_1(x_1, x_2) - u_2(y_1, x_2) + r\big(u_2(x_1, x_2) - u_2(y_1, x_2)\big) > 0, \tag{11}$$

which is the $N = 2$ case of (6). Thus, for x to be an ESS in Alger and Weibull (2013), it must be an equilibrium of the interaction game given by $\tilde{u}_i(x) = u_i(x) + ru_{-i}(x)$. The equation and inequality in (10) become mutually exclusive: the former will prescribe that (11) be met with equality and the other that it be met as it is, with strict inequality.[7]

[7] Akdeniz et al. (2023) shows an example where altruistic individuals outperform moral ones à la Alger and Weibull (2013). As this remark shows, under the direct evolutionary model (strategy evolution), all evolutionarily stable strategies will be moral (in the ex ante game) and they will perform altruistic actions (in the ex post game).

If y is any mutant strategy (y_1, y_2), which, with $N = 2$, amounts to a 2-mutation, then (9) becomes

$$u_1(x_1, x_2) + u_2(x_1, x_2) - r\big(u_1(y_1, y_2) + u_2(y_1, y_2)\big) - (1 - r)\big(u_1(y_1, x_2) + u_2(x_1, y_2)\big) > 0, \quad (12)$$

which is the $N = 2$ case of (7). Similarly, condition (10) obtains from (8).

3 Evolutionary network formation games

In this section I focus on interaction games whose action spaces correspond to link formation decisions. The set of positions corresponds to the set of nodes (or vertices) of the forming network. For all nodes $i \in I$, the action set X_i contains the possible lists of other nodes that i could link to: $X_i = \{0, 1\}^{I \setminus \{i\}}$. It is convenient to use a matrix notation: For a pair $i, j \in I$, the component x_{ij} of player i's action $x_i \in X_i$ encodes whether or not node i links to node j; if $x_{ij} = 1$, then I say that i links to j, if $x_{ij} = 0$, then i does not link to j. As x defines a directed graph on the set of nodes I, I also use the notation $\vec{ij} \in x$ and $\vec{ij} \notin x$ to denote $x_{ij} = 1$ and $x_{ij} = 0$, respectively.

The network formation literature focuses most of its attention on two types of link formation, one-sided and two-sided. In one-sided link formation processes, an individual's action to link to another creates a link independently of the recipient's action. If a link between nodes i and j is formed through i's action, I call i the *sponsor* of the link, while j is its *recipient.* The link's cost is borne entirely by its sponsor. The network that forms bestows benefits to all nodes. In *one-way flow* models, the direction of a link (which node sponsored it) matters for benefits, for instance, valuable information may flow from the recipient to the sponsor but not the other way. In *two-way flow* models, the direction of the link does not matter for benefits, only for costs. Formally, a network formation game with one-sided link formation is given as follows:

Definition 3. Let $X_i = \{0, 1\}^{I \setminus \{i\}}$ for all i. The interaction game u is a *network formation game with one-sided link formation* if u_i obtains as

$$u_i(x) = b_i(f(x)) - c_i\Big(\sum_{j \in I \setminus \{i\}} x_{ij}\Big), \qquad (13)$$

with *flow function* $f\colon X \to X$, benefit functions $b_i\colon X \to \mathbb{R}$, and cost parameters $c_i \in \mathbb{R}_+$ for all i, . The game has *one-way flow* of benefits if f is the identity function, and *two-way flow* if $f(x)$ is the underlying graph of x.[8]

[8] The underlying graph of a directed graph is the undirected graph obtained by replacing all directed edges by undirected ones, that is, $(f(x))_{ij} = \max\{x_{ij}, x_{ji}\}$.

One sided link formation processes go back to Bala and Goyal (2000). The strand of the literature that originates from this paper uses the (strict) Nash equilibrium as its primary solution concept; a network is considered to be in equilibrium if no node is able to rewire the links that it initiates in a way that would improve its payoff (or, in case of strict Nash, every rewiring would leave the node with a lower payoff). This principle exactly aligns with that of the 1-ESS. Hence, under one-sided network formation, the directed network defined by strategy x will be called evolutionarily stable for game u if the strategy x is a 1-ESS. That is, if

$$b_i\big(f(x)\big) - b_i\big(f(y_i, x_{-i}\big) + r \sum_{j \in I \setminus \{i\}} \Big(b_k\big(f(x)\big) - b_k\big(f(y_i, x_{-i})\big)\Big) > c_i \cdot \left(\sum_{j \in I \setminus \{i\}} x_{ij} - \sum_{j \in I \setminus \{i\}} y_{ij} \right)$$

for all $i \in I$ and all $y_i \neq x_i$.

In case of two-sided link formation, an undirected link forms between two nodes if and only if both participants link to each other, otherwise, no link forms. Let G denote the set of undirected networks on set of nodes I. For $g \in G$, $ij \in g$ ($ij \notin g$) denotes that a link between i and j exists (does not exist) in network g. As usual in the literature, the notation $g + ij$ ($g - ij$) is used to denote the network obtained from g by adding (removing) the link ij. Let the *two-sided link formation process* $g\colon X \to G$ be defined as

$$g_{ij}(x) = g_{ji}(x) = x_{ij} x_{ji}$$

for $x \in X$ and $i, j \in I$.

The formal definition of network formation games with such link formation process is the following:

Definition 4. Let $X_i = \{0,1\}^{I \setminus \{i\}}$ for all i. The interaction game u is a *network formation game with two-sided link formation* if u_i is given as

$$u_i(x) = b_i(g(x)) - c_i \cdot \sum_{j \in I \setminus \{i\}} x_{ij}(1 - (1-\rho)(1 - x_{ji})), \tag{14}$$

with benefit functions $b_i\colon G \to \mathbb{R}$, cost parameters $c_i \in \mathbb{R}$ and reciprocity parameter $\rho \in [0,1]$ for all i.

Nodes collect benefits from the (now undirected) network that forms and pay costs for linking attempts. For simplicity of presentation, I assume that costs are additively separable along links and players pay a proportional cost for initiating an unreciprocated linking attempt that does not lead to link formation. The interpretation is that a node i pays c_i for each link formed in the network that involve i and ρc_i for each unreciprocated linking attempt. In the analysis, I will focus on the case of $\rho > 0$ to ensure that strategies that extend unreciprocated links are never stable.

To account for stability against the formation of potentially profitable links, the network formation literature with two-sided link formation typically uses equilibrium concepts that guarantee stability against two-sided deviations. Notably, a network is called pairwise stable (Jackson and Wolinsky, 1996) if (i) there is no link in the network whose removal, leaving every other part of the network the same, would increase the payoff of at least one of the two participating nodes (stability against individual severance) and, (ii) for every pair of nodes who do not have a link running between them, adding the link would decrease the payoff for at least one of them (stability against pairwise addition).

Next, I define an evolutionary equivalent of pairwise stability through the evolutionary stability of the strategy that forms the network under the appropriate mutation protocol. As pairwise stability amounts to withstanding changes along individual links, the chosen mutation protocol will follow this principle as well. The mutations of this protocol are collectively called pair-mutations and networks stable against such mutations are called pair-evolutionarily stable. The formal definition of the notion is as follows:

Definition 5. A network g is called (neutrally) *pair-evolutionarily stable*, if there exists a strategy $x \in X$ such that $g(x) = g$ and x is (n)ESS with respect to the mutation protocol $\mathcal{P}$ given as follows:

$$\mathcal{P}(x) = \{y \in X \colon \exists i, j \in I \text{ such that } y_{i'j'} = x_{i'j'} \ \forall (i', j') \text{ such that } \{i', j'\} \neq \{i, j\}\},$$

for $x \in X$.

Given a strategy x, mutation protocol $\mathcal{P}(x)$ is the set of all action changes that affect a single pair of nodes; a mutant strategy y is identical to the resident except for the two components y_{ij} and y_{ji}. As such, for every x and every $y \in \mathcal{P}(x)$ we either have $g(y) = g(x)$, $g(y) = g(x) + ij$, or $g(y) = g(x) - ij$ for some pair i, j. Note that $\mathcal{P}(x) \subseteq \mathcal{H}_2(x)$ for all $x \in X$.

The following result characterizes pair-evolutionarily stable networks:

Proposition 3. *For $\rho > 0$, a network g is pair-evolutionarily stable if for every pair of positions $i, j \in I$*

- *$ij \notin g$ implies*
  - $b_i(g+ij) - b_i(g) + b_j(g+ij) - b_j(g) + r\sum_{k \in I \setminus \{i,j\}} \Big(b_k(g+ij) - b_k(g)\Big) < \frac{r+(1-r)\rho}{r}(c_i + c_j)$ *(stability vs pairwise formation).*

- *$ij \in g$ implies*
  - $b_i(g) - b_i(g-ij) + r\sum_{k\in I\setminus\{i\}}\Big(b_k(g) - b_k(g-ij)\Big) > c_i + r(1-\rho)c_j$ *(stability vs severance by i),*
  - $b_j(g) - b_j(g-ij) + r\sum_{k\in I\setminus\{j\}}\Big(b_k(g) - b_k(g-ij)\Big) > c_j + r(1-\rho)c_i$ *(stability vs severance by j),*
  - $b_i(g) - b_i(g-ij) + b_j(g) - b_j(g-ij) + r(2-r)\sum_{k\in I\setminus\{i,j\}}\Big(b_k(g) - b_k(g-ij)\Big) > c_i + c_j$ *(stability vs pairwise severance).*

If a network g is pair-evolutionarily stable, the same conditions hold with weak inequalities.

In the Appendix, I use Proposition 2 to provide an exact characterization of pair-evolutionarily stable networks. Proposition 3 only uses the first condition of a 2-ESS (7) to obtain a more succinct partial characterization, omitting the second conditions derived from (8). These additional conditions become relevant if the conditions stated in the Proposition happen to be met with equality. If u and r are given such that the statements are never met with equality (which is the case for a generic calibration of the game's primitives), the characterization stated here becomes exact.

The interpretation of stability against individual severance of an existing link in Proposition 3 is reminiscent to its counterpart in pairwise stability. The difference is that nodes act as if they partially internalize the payoff-effect of severing a link incurred by other nodes; both the other endpoint of the severed link and all other nodes in the network. As r tends to 0, the evolutionary and the non-evolutionary conditions become identical.

Stability against pairwise link formation acts differently in the evolutionary setting. The pair forming the link behaves as if the nodes evaluate the gains and losses of the link jointly, comparing total benefits, including externalities with weight r, against the total costs, accounting for the possibility of discoordination. In the case where such discoordination is costless, that is, as ρ goes to zero, the condition becomes

$$b_i(g+ij) - b_i(g) + b_j(g+ij) - b_j(g) + r\sum_{k\in I\setminus\{i,j\}}\Big(b_k(g+ij) - b_k(g)\Big) < c_i + c_j,$$

so the link forms if and only if total benefits of the link for the two nodes, partially inclusive of externalities are larger than the total costs of forming the link. As r tends to zero, this condition becomes a condition of stability against link formation while allowing for side payments (Bloch and Jackson, 2006).[9] Note that there is no component of the model that would allow for commitment to such side payments, nor are such direct fitness transfers possible in most

[9] Under link formation with side payments, a link between two nodes that benefits one and harms the other may still form as long as the node that benefits can compensate the losing node for its loss.

evolutionary applications, yet the tendency to form the links *as if it was possible* still evolves. This is due to the fact that the fitness advantage of a strategy adding a link takes into account both endpoints of the link equally.

Finally, stability against pairwise severance is a new condition not present in the original notion of pairwise stable networks. It is possible that a network is stable against individual severance of a link ij but not against a strategy that severs the link from both endpoints. As r tends to zero, this condition ceases to have any bite as it follows from the individual severance conditions. Furthermore, if total externalities are non-negative, meaning that the addition of a link by a pair weakly increases the total interaction payoffs of individuals outside of the pair, pairwise severance is once again implied by the two individual severance conditions, as the following Corollary shows:

Corollary 1. *Let a strategy $x \in X$ and $i, j \in I$ be such that $x_{ij} = x_{ji} = 1$, and let the strategies $y^{(i)}$, $y^{(j)}$, and y be given as follows:*

- $y^{(i)}_{ij} = 0$, $y_{k\ell} = x_{k\ell}$ *otherwise (unilateral severance by i),*
- $y^{(j)}_{ji} = 0$, $y_{k\ell} = x_{k\ell}$ *otherwise (unilateral severance by j),*
- $y_{ij} = y_{ji} = 0$, $y_{k\ell} = x_{k\ell}$ *otherwise (bilateral severance).*

Then, if $\sum_{k \neq i,j} \Big(b_k(g + ij) - b_k(g) \Big) \geq 0$, and if x is stable against $y^{(i)}$ and $y^{(j)}$, x is also stable against y.

Corollary 1 follows by adding up the unilateral severance conditions for player i and player j. Due to the total externalities of the link being non-negative, the bilateral severance condition follows for any value of ρ.

For a generic, positive r, pairwise severance produces results that are counterintuitive from a *homo economicus* standpoint. The pair attaches a strictly larger weight on the link's externalities in the pairwise severance condition than in the pairwise formation condition, $r(2 - r)$ instead of r. At the same time, the cost evaluation of the link, for ρ small, is almost equal. As a result, if the total externalities of a link are negative, it is possible a link not in the network to be formed via a pair-mutation that creates the link from both endpoints, whereas if the link is in the network, it would be dissolved by another pair-mutation that severs the link from both endpoints. Such a phenomenon is common in evolutionary game theory and not unique to pair-evolutionarily stable networks, nor is it the only feature of the equilibrium concept that precludes its existence (as pairwise stable networks are also known to lack existence in general network formation games).

Another feature unique to the evolutionary framework is the clash between pairwise formation and individual severance. This clash precludes general existence even for as few as two nodes: The empty network is invaded by the strategy that forms the link if its total benefits are higher than total costs. However, if the link decreases the payoff of one of the nodes, for low enough r, the strategy that cuts the link also invades, and neither the existence nor the absence of the link between the two nodes is evolutionarily stable.

The value of pair-evolutionarily stable networks rests in the minimality of requirements needed to define and characterize them. Just as it was for pairwise stable networks for non-cooperative network formation games, the evolutionary equivalent is the most permissive point-valued solution concept that fits with the philosophy of two-sided link formation in evolutionary network formation games. Yet, to obtain an evolutionary equilibrium notion that guarantees existence, one has to consider set valued solution concepts in the spirit of the evolutionarily stable set, which is an interesting direction for future work on evolutionary network formation games.

4 Applications

In this section I present three applications of evolutionarily stable networks, a communication game with one-sided link formation, the classic co-authorship game with two-sided link formation, and a model of homophily and in-group bias between two identity groups.

4.1 Communication networks

Consider a game of communication with players initiating costly links in order to access non-rival information from co-players. Linking costs are assumed to be homogeneous across nodes with $c_i = c$ for all $i \in I$. Assume that the benefits that each node receives is a concave function of the number of its direct neighbors.

Linking is one-sided and only node initiating the link incurs the cost c, whereas benefits flow both ways on each link. Given the directed network formed through action profile x, let $\bar{x}$ denote its underlying graph, that is, $ij \in \bar{x}$ if either $\vec{ij} \in x$ or $\vec{ji} \in x$. For node i, let $d_i^{out}(x) = |\{j \colon \vec{ij} \in x\}|$ denote node i's *out-degree*, the number of links it initiates in action profile x, $d_i^{in}(x) = |\{j \colon \vec{ji} \in x\}|$ denote its *in-degree*, the number of links targeting i, and $d_i(x) = |\{j \colon ij \in \bar{x}\}|$ its *degree*, the number of links formed involving i, whether they are sponsored by i or another node in action profile x.

Formally, the interaction game is then given by

$$u_i(x) = b\big(d_i(x)\big) - c \cdot d_i^{out}(x), \tag{15}$$

with $b\colon \mathbb{N} \to \mathbb{R}_+$ increasing and concave, and $c > 0$. Furthermore, assume that there exists $d^* > 0$ such that $b(d^*) - b(d^* - 1) < c < b(d^* + 1) - b(d^*)$.

A specific game that obtains in the form (15) is a case of non-rival information exchange between neighbors. Suppose that each node may receive a signal independently with probability p. If a node has not received the signal, it may still receive it from one of its network neighbors; if a link of any direction exists between nodes i and j and if i as received the signal, it will communicate it to j with probability q independently on each link. The signal is not transferred to neighbors of neighbors. If the value of receiving the signal is 1 either by the random draw or from a network neighbor, and not receiving it is worth 0, then, the benefit function obtains as

$$b(d) = p + (1-p)\Big(1 - (1-pq)^d\Big),$$

for $i \in I$, while for $d \geq 1$ we have $b(d) - b(d-1) = (1-p)(1-pq)^{d-1}pq$, which decreases in d, hence b is concave.

Under these assumptions, the best response behavior of any given node i, given the linking actions of all other nodes is as follows: If other nodes sponsor at least d^* links to i, then, as the marginal benefits of any more direct connections are below the marginal costs of initiating a link, i initiates no additional links. If other nodes sponsor $d < d^*$ links, i initiates $d^* - d$ links, stopping when marginal benefits fall below marginal costs. Furthermore, i is indifferent in which non-neighboring nodes to link to. The characterizations of Nash equilibria and strict Nash equilibria follow in a straightforward way.

Lemma 2. *1. The directed graph $x \in X$ is a Nash equilibrium of a communication network formation game satisfying (15) if and only if*

- *x is oriented,*[10]
- *$d_i(x) \geq d^*$ for all $i \in I$,*
- *if $\vec{ij} \in x$, then $d_i(x) = d^*$.*

2. The action profile $x \in X$ is a strict Nash equilibrium of the same game if and only if it is oriented, $d^* \geq N-1$, and $\bar{x}$ is the complete network. If $d^* < N-1$, there does not exist a strict Nash equilibrium.

[10] An *oriented* directed graph is a directed graph without reciprocated links: $\vec{ij} \in x \Rightarrow \vec{ji} \notin x$.

By Lemma 2, in every Nash equilibrium, there exist two types of nodes; those with degree exactly d^* may be sending and receiving links, while those whose degree exceeds d^* are pure receivers. For $d^* < N-1$, there does not exist a strict Nash equilibrium; this is due to the fact that nodes are indifferent in who they link to, meaning that unless a node links to every other, which only obtains in equilibrium if $d^* \geq N-1$, any node can swap an existing neighbor to a non-neighbor without a change in payoffs. The formal proof is shown in the Appendix.

The evolutionarily stable networks are markedly different from Nash networks. Due to individuals behaving altruistically towards their co-players, sponsors of links are no longer indifferent between targets; while the benefits and costs of sending a single link remain the same, and there are no linking externalities on any player other than the target, the target is positively affected by receiving an additional link. Due to the concavity of b, the rise in payoffs of an additional link received is a decreasing function of degree. Hence, for any $r > 0$, individuals will preferentially link to low-degree nodes. For low r, individuals' optimal degree in the game with other-regarding preferences, $\tilde{u}$ is the same as in u, and only regular and "almost regular networks", where difference between the maximum and the minimum degree is no greater than 1, turn out to be neutrally evolutionarily stable. Further restricting to evolutionarily stable networks eliminates even the almost regular networks.

Proposition 4. *There exists $\underline{r} \in (0,1)$ such that for every $r \in (0,\underline{r})$ the following statements hold:*

1. *Strategy $x \in X$ is a 1-nESS of the evolutionary game induced by a communication game in the shape of (15) if and only if*
   - *x is oriented,*
   - *$d_i(x) \in \{d^*, d^*+1\}$,*
   - *if $\vec{ij} \in x$, then $d_i(x) = d^*$,*
   - *if $d_i(x) = d^*+1$, then $d_i^{out}(x) = 0$.*
2. *Strategy $x \in X$ is a 1-ESS of the evolutionary game induced by the same game if and only if x is oriented and $\bar{x}$ is a d^*-regular network. If no d^* regular network exists with N nodes, there does not exist a 1-ESS.*

Proposition 4 shows that the notion of evolutionary stability produces starkly different equilibrium predictions for communication games even for low r. If the nodes are inhabited by *homo economicus* agents, there is a vast multiplicity of Nash equilibria ranging from d^*-regular networks to d^*-centered stars with sponsors in the periphery and receivers in the center. As the best-response correspondences are never single-valued in the generic case of $d^* < N - 1$, restricting to the strict Nash equilibrium does not help with selection as all Nash equilibria are weak.

If, instead, nodes are inhabited by individuals who are products of evolution by natural selection then, under assortative matching of strategies, altruistic behavior evolves. If r is low enough, then altruism does not change the optimal number of links of the nodes, but sponsors will preferentially target nodes with the lowest degree in the network, as if to maximize the marginal benefits of linking conferred to the *target* of the link. Regular and almost regular network featuring sponsors with degree d^* and receivers with in-degree d^*+1 are neutrally stable. Restricting to evolutionary stability throws out even the almost regular networks, leaving only the regular networks where the best-response correspondences under altruistic preferences are all single-valued. Figure 4 showcases a d^*-centered star, the most irregular Nash network structure, an almost regular network, the most irregular 1-nESS structure, and a regular network, the only 1-ESS structure for $N = 10$ and $d^* = 2$.

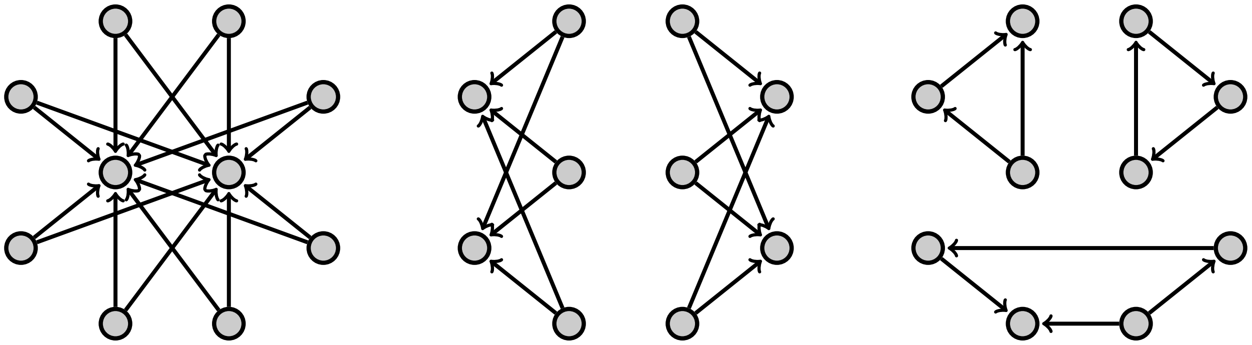

Figure 4: For low assortativity, evolutionary stability selects regular networks in one-sided communication network formation games: a d^*-centered, periphery-sponsored star network, the most irregular Nash equilibrium structure (left); a network of sponsors (with d^* out-links) and receivers (with d^*+1 in-links), the most irregular 1-nESS structure (middle); and a d^*-regular network, the only 1-ESS structure (right), for $N = 10$ and $d^* = 2$.

The surviving network structure is egalitarian in terms of the nodes' information access but payoff-inequalities may arise due to them sponsoring different numbers of links. This outcome is utilitarian efficient if b flattens out above d^*; namely, if $2\big(b(d^*+1) - b(d^*)\big) < c$, then equilibria are more efficient the fewer links there are. If b is steep above d^*, it is even possible for this outcome to be the least utilitarian efficient of all Nash equilibria; if $2\big(b(N-1) - b(N-2)\big) > c$, then every additional link is welfare-increasing, making the regular network the least utilitarian efficient of all Nash equilibria, d^*-centered stars the most utilitarian efficient Nash equilibria, and the complete network the most utilitarian efficient network.

To provide an example with large assortativity, consider a case with the benefit function b flattening out above d^* and r close to 1. While d^*-regular networks are also evolutionarily stable in this case, any network is evolutionarily stable where links target nodes with degree d^*. Most notably, d^*-centered stars again become stable; however, instead of being periphery-sponsored such as in the case of the Nash equilibria, they are center-sponsored.

Proposition 5. *Let b be given such that $b(d^* + 1) - b(d^*) = 0$. Then, there exists $\bar{r}$ such that for all $r \in (\bar{r}, 1)$, every 1-ESS network x satisfies the following properties:*

- *x is oriented,*
- *$d_i(x) \geq d^*$ for all $i \in I$,*
- *if $\vec{ij} \in x$, then $d_j(x) = d^*$.*

A comparison between Lemmas 2 and Propositions 4, 5 shows the importance of r in determining evolutionarily stable network structures and how they compare to Nash structures. For low levels of assortativity, when the level of evolved altruism is low, evolutionarily stable networks select the most regular Nash equilibria with the least number of links, independently of whether these structures are more or less utilitarian efficient than irregular networks. While altruism changes equilibrium structures through preferential linking, it does not cause nodes to initiate links that reduce their own payoffs in the interaction game u. As r gets higher, this changes and nodes initiate links that lower their interaction payoffs but increases the fitness of their strategies. Yet, instead of stabilizing efficient networks only, for the case of disappearing benefits above d^*, a high degree of evolved altruism also stabilizes non-Nash, highly inefficient network architectures, particularly, center sponsored d^*-stars. In these networks, groups of " martyrs" sponsor too many links to their own detriment, and to that of the whole interaction group, all while acting out of almost pure altruism (5).

4.2 Co-authorship networks

To showcase equilibrium selection in two-sided network formation, I consider the classic co-authorship model of Jackson and Wolinsky (1996). In this model, links represent pairwise collaboration agreements between researchers. The value of each collaboration between any two nodes i and j is the sum of three factors: how much time i has for the project, how much time j has for it, and a synergistic component depending on how much time they have for each other. Each new link decreases the time that the participants have for their other projects. For

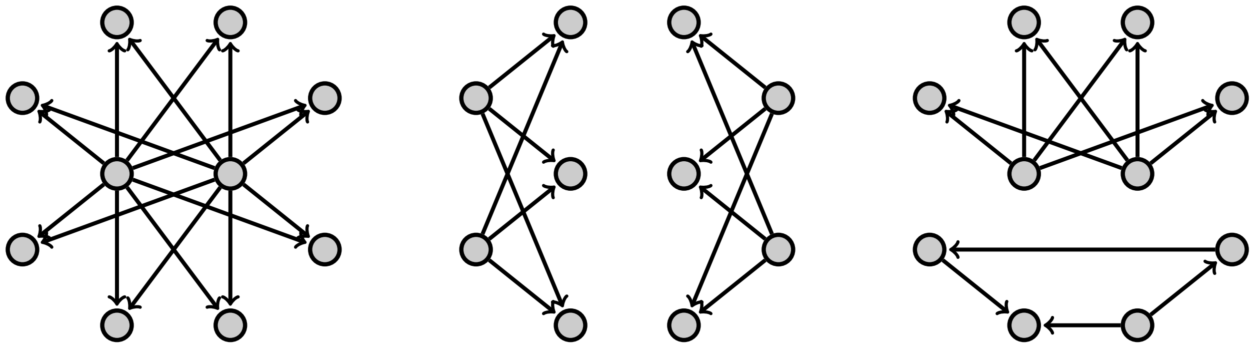

Figure 5: For high assortativity and disappearing benefits above d^*, non-Nash networks become evolutionarily stable with groups of "martyrs" altruistically linking to a larger than optimal number of nodes: a d^*-centered, center-sponsored star network (left); a network of two d^*-centered star components (middle); and a network with a d^*-centered star and a regular component (right), for $N = 10$ and $d^* = 2$.

$i \in I$ and $x \in X$, let $d_i(x)$ denote player i's degree in network $g(x)$. Network benefits in the co-authorship game are then given as

$$b_i(x) = \sum_{j:ij \in g(x)} \left(\frac{1}{d_i(x)} + \frac{1}{d_j(x)} + \frac{1}{d_i(x)d_j(x)} \right), \tag{16}$$

while I assume costs c_i to be small enough to never influence invasion conditions.[11]

Jackson and Wolinsky (1996)'s Proposition 4 characterizes the efficient and pairwise stable networks of this game: If N is even, "monogamy", nodes forming $N/2$ pairs joined by one link per pair, is the unique efficient structure. In pairwise stable networks, complete components of different sizes form with any two components of sizes $n, m \in \{1, \ldots, N\}$ having to satisfy $n^2 > m$. Notably, the efficient network structure is not pairwise stable.

As the next proposition shows, under the evolutionary paradigm, stable networks have starkly different properties. I showcase this through a characterization of evolutionarily stable networks that can be decomposed into regular connected components.

Proposition 6. *Let g be a network of regular connected components $C_1, \ldots, C_K$ for $K < N$ and let $n_1, \ldots, n_K$ denote the degrees of the components. Then, g is pair-evolutionarily stable of the co-authorship game given by (16) if and only if*

1. $2\frac{n}{n+1}\frac{m}{m+1}\frac{n+m+3}{n+m} - 1 < r$ *for all* $n, m \in \{n_1, \ldots, n_K \setminus \{N-1\}\}$ *such that* $ij \notin g$ *exists with* $d_i(x) = n$ *and* $d_j(x) = m$.

2. $1 - \frac{\sqrt{2}}{\sqrt{n+1}} > r$ *for* $n = \min\{\{n_1, \ldots, n_K\} \setminus \{1\}\}$.

[11] While there are no explicit linking costs in the original co-authors game, here, costs are still positive in order to deter invasions by strategies with unreciprocated linking attempts.

n/m	1	2	3	4	5	6	7	8	9	10	...	∞
1	0.25	0.33	0.31	0.28	0.25	0.22	0.20	0.19	0.17	0.16		0
2		0.56	0.60	0.60	0.59	0.57	0.56	0.54	0.53	0.52		0.33
3			0.69	0.71	0.73	0.71	0.71	0.70	0.69	0.68		0.50
4				0.76	0.78	0.78	0.78	0.78	0.77	0.77		0.60
5					0.81	0.82	0.82	0.82	0.82	0.82		0.67
6						0.84	0.85	0.85	0.85	0.85		0.71
7							0.86	0.87	0.87	0.87		0.75
8								0.88	0.88	0.89		0.78
9									0.89	0.89		0.80
10										0.90		0.82

Table 1: $L(n, m)$, the lower bound of assortativity required to prevent linking between regular with degrees n and m for $n \leq m$.

The first condition of Proposition 6 provides a lower limit to the rate of assortativity sustaining the candidate stable network. If the rate is lower than this threshold, the evolved altruism is too weak to stop an unlinked pair from forming a new link between components or within them (for $n = m$). Let

$$L(n, m) = 2\frac{n}{n+1}\frac{m}{m+1}\frac{n+m+3}{n+m} - 1$$

denote the left hand side of the condition. For a fixed n, $L(n, m)$ is non-monotonic in m; it increases between 1 and $\lfloor (n + \sqrt{3}\sqrt{n(n+2)})/2 \rfloor$, and decreases after $\lceil (n + \sqrt{3}\sqrt{n(n+2)})/2 \rceil$, converging to $(n-1)/(n+1)$ as m goes to infinity. For instance, the level of altruism that prevents a 2-regular component from creating a link with a 1-regular component (a linked pair) is the same as the level that prevents it from creating a link with a very large component, 1/3, but it would take a rate of 5/9 to prevent it from linking with another 2-regular component (or a new link forming within the same component) and 3/5 to prevent it from linking with a 3-regular or a 4-regular component. I report values of $L(n, m)$ for $n, m \in \{1, \ldots, 10\}$ as well as the limits in Table 1.

The second condition gives an upper bound; if assortativity is higher, altruism is so strong that pairs within existing components sever links to bring benefits to their neighbors. This condition is simpler to interpret; as the left hand side increases with n, the more connected a component is, the easier it is to sustain it against such altruistic severance.

However, the two conditions are difficult to satisfy at once. In fact, no n-regular network is evolutionarily stable, except for "monogamy" where the second condition is empty, and the complete network where the first condition is empty. This is obtained as an immediate corollary of Proposition 6.

Corollary 2. *An n-regular network is pair-evolutionarily stable of the co-authorship game given by (16) if and only if*

1. $n = 1$ *and* $r > 0.25$,
2. $n = N - 1$ *and* $r < 1 - \sqrt{2}/\sqrt{N}$.

To extend Proposition 6, a reading of Table 1 shows that, as a rule, preventing linking between components requires relatively high degrees of assortativity. Smaller components survive more easily than large ones and components of similar sizes merge more readily than components with a larger degree difference. For $r < 0.25$, the lower bound guaranteeing no linking between two disjoint pairs, the only evolutionarily stable networks are the complete network and a single pair and a large complete component of at least six nodes.

The case of the co-authorship game further showcases the differences between predictions under the *homo economicus* and the evolutionary paradigms. The pairwise stable structures characterized by Jackson and Wolinsky (1996) disappear due to pairwise formation of links. Links between components of different sizes form, even though it is not profitable for the nodes in the smaller components, due to evolution compelling the pair to form links as if they evaluated the move coalitionally. For r low the only stable networks with regular components feature a large complete component forms with at most a single pair staying out. This is formalized in the following Corollary, obtained directly from Proposition 6 and the values reported in Table 1.

Corollary 3. *Let g be a network of regular connected components $C_1, \ldots, C_K$ for $K < N$ and let $n_1, \ldots, n_K$ denote the degrees of the components. Then, there exists $\underline{r} \in (0,1)$ such that for every $r \in (0, \underline{r})$ it holds that g is pair-evolutionarily stable of the co-authorship game given by (16) if and only if one of the following holds:*

- *g is the complete network,*
- *g has two complete components, a linked pair and the complete network of the remaining $N - 2$ players.*

However, the evolved altruism stabilizes less connected components more easily than more connected ones, resulting in the efficient structure becoming stable for a broader range of the assortativity parameter than any other structure without a large complete component. Notably, the efficient network of $N/2$ linked pairs for N even is stable for r larger than 0.25.

4.3 In-group preference and the strength of homophily

Homophily is the tendency of individuals to link to others similar to them. To clarify the difference between assortative matching of strategies and homophily: in this paper, the former is an exogenous property of interactions between strategies arising due to local interaction and dispersion, while the latter is an endogenous outcome of the network formation process arising due to the selection of strategies. Accordingly, I consider homophily between identity groups that are assumed to be social constructs and have no genetic interpretation.

Formally, identity groups are modeled by partitioning the positions of the interaction group into pre-determined identity affiliations. Individuals are born with a strategy, are assortatively matched into interaction groups of N based on their strategies, then are sorted into positions which determines their identity. The variables of interest in this section is the number of links formed within and between identity groups.

Consider a two-sided link formation game in an interaction group with two non-overlapping sub-groups called *clans* (identity groups) A and B, denoted by $I = \{I_A, I_B\}$. Let N_A and N_B denote the sizes of the clans, let $C(i)$ denote the clan of node i and $-C(i)$ denote the other clan. Assume that for two nodes i and j, forming the link ij confers benefit w_{ij} and that benefits are additively separable across links. For simplicity I assume reciprocity of benefits, expressed as $w_{ij} = w_{ji}$, but a very similar analysis would go through for non-reciprocal benefits between pairs. Each value w_{ij} is drawn from an i.i.d. distribution with complementary cumulative distribution function (cdf) $\overline{F}$ before the interaction game is played: individuals are thus perfectly informed of the w_{ij} values.[12] All links have identical cost c paid by both participating nodes. Assume that ρ is close to 0.

A simple model of in-group preferences is to consider interaction payoffs of the following form:

$$u_i(x) = \sum_{j:\, ij \in g(x)} (w_{ij} - \phi_i \mathbb{1}_{\{C(i) \neq C(j)\}}) - c \sum_{j \in I \setminus \{i\}} x_{ij}(1 - (1-\rho)(1 - x_{ji})). \tag{17}$$

In the *individual model* of (17), nodes display a preference for links in their own clan with the strength of the preference given by the parameter ϕ_i, measuring individual distaste for the other clan.

While this model is attractive for its simplicity and captures the spirit of models deriving homophily from preferential linking to in-group members it does not account for one of the most historically prevalent forms of in-group bias: a preference against any inter-group mixing. Notably, individuals may not only seek to limit their own association with other clans, they oppose

[12] Given a cdf F, the complementary cdf $\overline{F}$ is given as $\overline{F}(z) = 1 - F(z)$ for $z \in \mathbb{R}$.

any association between clan members. In the present context, this is represented by a negative payoff-component for all links formed between clans, captured by the following formulation:

$$u_i(x) = \sum_{j\colon ij \in g(x)} w_{ij} - \phi_i |E_{AB}(g(x))| - c \sum_{j \in I \setminus \{i\}} x_{ij}(1 - (1-\rho)(1-x_{ji})), \tag{18}$$

with $E_{AB}(g) = \{ij \in g\colon C(i) \neq C(j)\}$ denoting the set of inter-clan links in network g.[13]

In the *apartheid model* of (18), node i suffers a payoff penalty of ϕ_i for every link formed between clans. An evolutionary model producing fitness values consistent with (18) relies on inter-clan conflicts. I do not model these conflicts; they may include clashes over differing sharing practices, cultural misunderstandings, or fears of a changing power balance between the clans (such as cultural assimilation for a minority or a loss of privilege for a majority) that may spark from inter-clan links but not from intra-clan links.

The following Lemma characterizes pairwise stable networks in the two models.

Lemma 3. *For both the individual model of (17) and the apartheid model of (18), there exists a unique pairwise stable network g such that $ij \in g$ if and only if*

$$w_{ij} > c + \max\{\phi_i, \phi_j\} \mathbb{1}_{\{C(i) \neq C(j)\}}.[14]$$

Crucially, under the *homo economicus* paradigm, it is impossible to distinguish between the two models as they prescribe the same best-response and equilibrium behavior. This is reflected by Lemma 3 as well.

To obtain a group-level measure of clan bias, assume, for simplicity that $\phi_i \in \{0, \phi\}$. Individuals with $\phi_i = \phi$ participate in the inter-clan conflict and/or are unable to escape it, while those with $\phi_i = 0$ do not participate and are unaffected. I will call the former type *radicals* and the latter type *moderates* based on the types of preferences induced by these payoff functions; radicals oppose inter-group linking, moderates do not. Let R_A, R_B, and M_A, M_B denote the number of radicals and moderates in the two clans, and R and M denote their total number in the interaction group. Under these assumptions, the frequency of radicals, R/N, is a direct measure of in-group bias in the interaction group, while R_A/N_A and R_B/N_B measure the strength of in-group bias of clans A and B, respectively.

Lemma 4. *Under pairwise stability, the ex ante probability of a link ij appearing in the network is*

[13] A similar analysis would go through if, instead of individuals suffering a payoff penalty for between-clans links, they would receive rewards for links forming within their own clan.

[14] As in the rest of the paper, I ignore the cases of indifference. If the distribution defined by F has a density function, the probability of any pair being indifferent is 0.

- $\overline{F}(c)$ *if* $C(i) = C(j)$ *or if* $C(i) \neq C(j)$ *and* $i, j \in M$,

- $\overline{F}(c + \phi)$ *if* $C(i) \neq C(j)$ *and either* $i \in R$ *or* $j \in R$.

As Lemma 4, under the *homo economicus* paradigm, homophily of both moderates and radicals is driven entirely by presence of radicals in the group. If node i is moderate, it will be linked to a fraction $\overline{F}(c)$ of its own clan and a fraction $\overline{F}(c)M_{-C(i)}/N_{-C(i)} + \overline{F}(c+\phi)R_{-C(i)}/N_{-C(i)}$ of the opposite clan in expectation. If i is a radical, it will be linked to a fraction $\overline{F}(c)$ of its own clan and a fraction $\overline{F}(c + \phi)$ of the opposite clan in expectation.

The next Lemma states the conditions under which a link ij forms under the evolutionary paradigm.

Lemma 5. *Let* $x \in X$ *such that* $x_{ij} = x_{ji} = 0$. *Then, there exists* $\bar{\rho} \in (0, 1)$ *such that for all* $\rho \in (0, \bar{\rho})$, x *is not stable against the strategy* y *given as* $y_{ij} = y_{ji} = 1$ *and* $y_{k\ell} = x_{k\ell}$ *if*

$$w_{ij} > c + \mathbb{1}_{C(i) \neq C(j)} \frac{1}{2} \left(\sum_{k \in \{i,j\}} \phi_k + r \sum_{k \in I \setminus \{i,j\}} \phi_k \right). \tag{19}$$

Furthermore, if $C(i) = C(j)$ *and* x *is not stable against* y, *then* y *is stable against* x *and the strategies* $y^{(i)}$ *and* $y^{(j)}$.

Lemma 5 shows the conditions under which a strategy that forms link ij invades a strategy without the link. In the remainder of this section I will refer to this condition as the evolutionary pairwise formation rule. Furthermore, if a link ij forms under this rule, and i and j belong to the same clan, it is stable against invasion from any strategy that severs it. In contrast, if i and j do not belong to the same clan but ij forms under this rule, strategies that sever the link, either unilaterally or bilaterally, may invade. As such, calculating the strength of homophily under this rule gives a lower bound of the strength of homophily.

Proposition 7. *If link formation is governed by the evolutionary pairwise formation rule of (19), the ex ante probability that a link forms between* i *and* j *is given as*

- $\overline{F}(c)$ *if* $C(i) = C(j)$,

- $\overline{F}(c + r\phi R/2)$ *if* $C(i) \neq C(j)$ *and both* i *and* j *are moderates,*

- $\overline{F}(c + \phi/2 + r\phi(R - 1)/2)$ *if* $C(i) \neq C(j)$ *and one of* i *and* j *is moderate and the other is radical,*

- $\overline{F}(c + \phi + r\phi(R - 2)/2)$ *if* $C(i) \neq C(j)$ *and both* i *and* j *are radicals.*

Proposition 7 shows that the evolutionary paradigm allows for a more nuanced view of homophily and its dependence on the strength of in-group bias. The differences arise in the frequency of between-clans links due to evolved other-regarding preferences. Since every link formed between clans is subject to the distaste of radicals of both groups and since individuals partially internalize linking externalities under the evolutionary formation rule, all linking outcomes are affected, even those between moderates of different clans. This effect is proportional to the absolute number of both clans' radicals, R. As in Lemma 4, radicals create yet fewer between-clans links even with the other clans' moderates. However, uniquely in the evolutionary case, linking between radicals of different clans is more infrequent than linking between radicals and moderates of different clans.

To calculate the strength of homophily in the population, I use the following measure:

$$\eta = 1 - \frac{[\text{Frequency of between-clans links}]}{[\text{Frequency of within-clans links}]}.$$

A value of $\eta = 1$ shows full homophily as every link is within-clans, a value of $\eta = 0$ shows no preference for the own clan, and negative values indicate heterophily.[15]

By Lemma 4, the probability that two individuals, drawn uniformly from different clans are linked under the pairwise stability forming rule is

$$\frac{M_1 M_2}{N_1 N_2} \overline{F}(c) + \left(1 - \frac{M_1 M_2}{N_1 N_2}\right) \overline{F}(c + \phi),$$

and the strength of homophily obtains as

$$\eta^{PS} = \left(1 - \frac{M_1 M_2}{N_1 N_2}\right)\left(1 + \frac{\overline{F}(c+\phi)}{\overline{F}(c)}\right). \tag{20}$$

By Proposition 7, the probability that two individuals, drawn uniformly from different clans are linked under the evolutionary forming rule is

$$\frac{M_1 M_2}{N_1 N_2} \overline{F}(c + r\phi R/2) + \frac{R_1 M_2 + M_1 R_2}{N_1 N_2} \overline{F}(c + \phi(1/2 + r(R-1)/2)) + \frac{R_1 R_2}{N_1 N_2} \overline{F}(c + \phi(1 + r(R-2)2)).$$

and the strength of homophily obtains as

$$\begin{aligned}\eta^{EV} =& 1 - \frac{M_1 M_2}{N_1 N_2} \frac{\overline{F}(c + r\phi R/2)}{\overline{F}(c)} + \frac{R_1 M_2 + M_1 R_2}{N_1 N_2} \frac{\overline{F}(c + \phi(1/2 + r(R-1)/2))}{\overline{F}(c)} \\ &+ \frac{R_1 R_2}{N_1 N_2} \frac{\overline{F}(c + \phi(1 + r(R-2)/2))}{\overline{F}(c)}.\end{aligned} \tag{21}$$

[15] Frequency of between- and within-clan links is the ex ante probability that two nodes drawn uniformly at random from different clans and from the same clan, respectively, are linked. The literature tends to focus on two individual-level homophily measures: The "baseline homophily" index measures the ratio of a node's within-clan links over all its links, while Coleman's "inbreeding homophily" index (Coleman, 1958) which corrects for a natural linking bias towards overrepresented groups through a normalization. I present results in terms of the group-level measure of η using within-clan links in the denominator rather than all links due its relative simplicity, but very similar results go through for the other measures.

A key difference between (20) and (21) is the nature by which homophily depends on the number of radicals in the population. If link formation is by the logic of pairwise stability, the strength of homophily depends on the *frequency* of radicals. If link formation follows the evolutionary principles of bilateral link formation, the strength of homophily depends both on the *frequency* and the *density* (the absolute number) of radicals in the interaction group.

In recent history, populations and interaction groups have gotten larger, while stated preferences against inter-group mixing across identity groups have been declining in most developed countries, though they persist to various degrees.[16] The following Proposition shows the stark divergence of the strength of homophily under the two paradigms in the asymptotic limit.

Proposition 8. *For fixed r and ϕ we have*

1. $\lim_{N,R\to\infty,R/N\to 0} \eta^{PS} = 0$,
2. $\lim_{N,R\to\infty,R/N\to 0} \eta^{EV} = 1$.

Proposition 8's first point states that as populations get large and as the population's in-group bias disappears, the homophily of the pairwise stable network approaches 0. This is consistent with the frequency-dependent strength of homophily in (20): as most of the population consists of moderates and as moderates do not discriminate (Lemma 4), the frequencies of within-group and between-group links are almost identical. The second point, stating that homophily converges to its extreme level under the evolutionary paradigm even as the population's in-group bias approaches zero, follows from the density-dependence of homophily in (21): as the number of radicals grows, discrimination of all groups, moderates and radicals rises and the frequency of between-group links forming goes to zero. This effect is driven by moderates internalizing the preferences of a "loud minority" that increases in size even as it is decreasing in prevalence.

A key assumption driving this result is that population dynamics have changed too quickly for evolved behavior to adapt. As a result, linking actions follow rules that adapted to a fixed environment over evolutionary time, specifically, the assortativity of the population. These linking actions then drive homophily in today's modern environment. While these assumptions are strong and more nuanced studies in the evolution of homophilistic network formation ought to be mindful of the co-evolution of assortativity and group size, Propositions 7 and the second point of 8 are congruent with observed patterns of high homophily in present-day societies.

[16] For instance, 2021 Gallup poll measured a 94% approval rate of interracial marriages in the US while the 2019 Eurobarometer report on discrimination states that 53% of EU citizens would approve of their children having a romantic relationship with a Muslim person, both showing rising tendencies.

5 Concluding remarks

In this paper I propose an evolutionary model of interaction in heterogeneous groups. The model relies on two key elements, assortative matching of strategies modeled by a constant index of assortativity and heterogeneous interaction modeled by the interaction game being asymmetric across positions. A strategy is called an m-ESS if it is evolutionarily stable against mutant strategies that execute different actions in no more than m positions. In each position, 1-ESS strategies execute actions as if driven by altruism towards the interaction group at rate equaling the assortative matching of the population. 2-ESS strategies are further stable against pairwise coalitional moves as if pairs of positions evaluate such moves jointly but taking into account the possibility of miscoordination as well as externalities. While the evolution of altruistic action in 1-ESS follows from classical inclusive fitness arguments, the evolution of shared intentionality in 2-ESS through kin selection is, to my knowledge, a new result.

I then apply the model to strategic network formation. To my knowledge, this is the first evolutionary model that can be directly applied to general classes of network formation games à la Jackson and Wolinsky (1996) and Bala and Goyal (2000). The exercise is motivated by the evolutionary history of the *homo sapiens* and the recent demand in the networks literature for non-*homo economicus* models. In 1-ESS – a natural solution concept for evolutionary network formation games with one-sided link formation – a node sponsoring a link to another takes into account the payoff changes induced to the recipient of the link as well as the externalities induced on the rest of the network. For games with two-sided link formation, I introduce pair-evolutionarily stable networks as a simple and natural solution concept. As it accounting for stability against mutations that form or sever individual links, it may be viewed as an evolutionary equivalent of pairwise stability. Since the concept obtains as a variant of the 2-ESS, any candidate stable network must withstand coalitional link formation and severance where the pair forming or severing the link acts as if the members evaluate the move jointly rather than individually. As such, the notion is also reminiscent of pairwise stability with transfers.

Evolutionarily stable networks open the way for the study of social and economic networks where individuals' propensity to form ties is guided by Darwinian principles in line with the evolutionary history of the *homo sapiens.* This paper thus constitutes a move away from the *homo economicus* paradigm in networks, reinforcing the movement in the networks literature that aims for a closer connection with behavioral models. Even stronger, the evolutionary mindset has the potential to motivate, identify, and systematize such deviations.

The analysis in this paper uses a single feature from evolutionary game theory, assortative matching by strategies, to obtain its results under the theory's first, static equilibrium notion, evolutionarily stable strategies. The base model is readily extendable to include additional features such as group selection, other solution concepts such as convergence stability, and the rich types of evolutionary dynamics that the field is known for.

While in this paper I focused on the evolutionary equivalent of pairwise stability, other solution concepts in network formation games lend themselves to the same analysis. For instance, extending the pair-mutation protocol to include the unilateral severance of any number of links, a 1-mutation, would produce the evolutionary equivalent of the pairwise Nash equilibrium (Bloch and Jackson, 2006), merging the properties of 1-ESS networks and pair-evolutionarily stable networks. As another example, allowing for swaps, a 2-mutation, would produce the evolutionary equivalent of swap-proof graphs (Sadler, 2022). Both obtain as straightforward refinements of pair-evolutionarily stable networks that can be characterized through the general 1-ESS and 2-ESS results. The evolutionary equivalent of networks stable against coalitional moves with more than two players obtain from higher order m-ESS results. For instance, strongly stable networks (Jackson and Van den Nouweland, 2005) would correspond to the N-ESS.

If evolutionarily stable networks fail to exist, one may turn to evolutionarily stable sets of networks to obtain a prediction, obtaining as evolutionary equivalents to the von Neumann-Morgenstern stable set. Alternatively, evolutionarily stable strategies of the mixed extension of the network formation game, if they exist, would produce stable distributions of networks. While the analysis of this paper deliberately focused on point-valued solutions in pure strategies, these directions are natural next steps in the study of evolutionary network formation games.

Coalitional moves for more than two positions are of interest for the evolutionary study of coalition formation. As the 2-ESS result shows, coalitional moves are subject to discoordination between deviating positions, a problem that becomes more pronounced with more deviating positions. If such discoordination is costless, however, as in the case of link formation with a low cost of initiating an unreciprocated link, evolution favors the coalition if its total gains plus partial externalities are positive. The model introduced in Section 2 thus lends itself to be used in the evolutionary foundations of TU-games and solution concepts.

Directions for future work also include extending the pool of applications in Section 4. Promising candidates include the 'connections game' (Jackson and Wolinsky, 1996), risk-sharing games (Bramoullé and Kranton, 2007), and general classes of communication games (Bala and Goyal, 2000). These applications are attractive as the network ties constitute bonds of trust, affection, or friendship between individuals, necessarily subject to evolutionary forces. I conjecture that network formation between institutions operating on a free market, such as R&

D networks (Goyal and Moraga-Gonzalez, 2001) and oligopolies (Goyal and Joshi, 2003) are arguably less likely to be guided by Darwinian forces, though, ultimately, such ties are also ties between (groups of) individuals. Finally, a complete model of networked behavior includes not only network formation games but also games on networks. A holistic evolutionary model should include both components à la Galeotti and Goyal (2010), mindful of possible differences between the speed and nature of evolutionary dynamics between link formation and other activities.

References

Akdeniz, A., Graser, C., and van Veelen, M. (2023). Homo moralis and regular altruists ii. Technical report, Tinbergen Institute.

Alger, I. and Weibull, J. W. (2013). Homo moralis–preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

Alger, I. and Weibull, J. W. (2016). Evolution and kantian morality. *Games and Economic Behavior*, 98:56–67.

Alger, I., Weibull, J. W., and Lehmann, L. (2020). Evolution of preferences in structured populations: genes, guns, and culture. *Journal of Economic Theory*, 185:104951.

Apicella, C. L., Marlowe, F. W., Fowler, J. H., and Christakis, N. A. (2012). Social networks and cooperation in hunter-gatherers. *Nature*, 481(7382):497–501.

Bala, V. and Goyal, S. (2000). A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229.

Bayer, P., Ross, S. L., and Topa, G. (2008). Place of work and place of residence: Informal hiring networks and labor market outcomes. *Journal of Political Economy*, 116(6):1150–1196.

Beaman, L. A. (2012). Social networks and the dynamics of labour market outcomes: Evidence from refugees resettled in the us. *The Review of Economic Studies*, 79(1):128–161.

Bergstrom, T. C. (2003). The algebra of assortative encounters and the evolution of cooperation. *International Game Theory Review*, 5(03):211–228.

Bester, H. and Güth, W. (1998). Is altruism evolutionarily stable? *Journal of Economic Behavior & Organization*, 34(2):193–209.

Bloch, F. and Jackson, M. O. (2006). Definitions of equilibrium in network formation games. *International Journal of Game Theory*, 34(3):305–318.

Bloch, F. and Jackson, M. O. (2007). The formation of networks with transfers among players. *Journal of Economic Theory*, 133(1):83–110.

Bourlès, R., Bramoullé, Y., and Perez-Richet, E. (2017). Altruism in networks. *Econometrica*, 85(2):675–689.

Bourlès, R., Bramoullé, Y., and Perez-Richet, E. (2021). Altruism and risk sharing in networks. *Journal of the European Economic Association*, 19(3):1488–1521.

Boyd, R. and Richerson, P. J. (2022). Large-scale cooperation in small-scale foraging societies. *Evolutionary Anthropology: Issues, News, and Reviews*, 31(4):175–198.

Bramoullé, Y., Currarini, S., Jackson, M. O., Pin, P., and Rogers, B. W. (2012). Homophily and long-run integration in social networks. *Journal of Economic Theory*, 147(5):1754–1786.

Bramoullé, Y. and Kranton, R. (2007). Risk-sharing networks. *Journal of Economic Behavior & Organization*, 64(3-4):275–294.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267.

Christakis, N. A. and Fowler, J. H. (2014). Friendship and natural selection. *Proceedings of the National Academy of Sciences*, 111(3):10796–10801.

Coleman, J. (1958). Relational analysis: The study of social organizations with survey methods. *Human organization*, 17(4):28–36.

Currarini, S., Jackson, M. O., and Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045.

Currarini, S., Matheson, J., and Vega-Redondo, F. (2016). A simple model of homophily in social networks. *European Economic Review*, 90:18–39.

Fowler, J. H., Dawes, C. T., and Christakis, N. A. (2009). Model of genetic variation in human social networks. *Proceedings of the National Academy of Sciences*, 106(6):1720–1724.

Frank, S. A. (1998). *Foundations of social evolution*, volume 2. Princeton University Press.

Fu, F., Nowak, M. A., Christakis, N. A., and Fowler, J. H. (2012). The evolution of homophily. *Scientific reports*, 2(1):845.

Fudenberg, D. and Maskin, E. (1990). Evolution and cooperation in noisy repeated games. *The American Economic Review*, 80(2):274–279.

Galeotti, A. and Goyal, S. (2010). The law of the few. *American Economic Review*, 100(4):1468–1492.

Golub, B. and Jackson, M. O. (2012). How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338.

Goyal, S. and Joshi, S. (2003). Networks of collaboration in oligopoly. *Games and Economic Behavior*, 43(1):57–85.

Goyal, S. and Moraga-Gonzalez, J. L. (2001). R&d networks. *Rand Journal of Economics*, pages 686–707.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Güth, W. and Yaari, M. (1992). An evolutionary approach to explain reciprocal behavior in a simple strategic game. *U. Witt. Explaining Process and Change–Approaches to Evolutionary Economics. Ann Arbor*, pages 23–34.

Hamilton, M. J., Milne, B. T., Walker, R. S., Burger, O., and Brown, J. H. (2007). The complex structure of hunter–gatherer social networks. *Proceedings of the Royal Society B: Biological Sciences*, 274(1622):2195–2203.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *The American Naturalist*, 97(896):354–356.

Hamilton, W. D. (1964). The genetical evolution of social behaviour, parts i and ii. *Journal of Theoretical Biology*, 7(1):17–52.

Harsanyi, J. C. (1967). Games with incomplete information played by “bayesian” players, i–iii part i. the basic model. *Management science*, 14(3):159–182.

Jackson, M. O. (2021). Inequality's economic and social roots: The role of social networks and homophily. *SSRN 3795626*.

Jackson, M. O. and Van den Nouweland, A. (2005). Strongly stable networks. *Games and Economic Behavior*, 51(2):420–444.

Jackson, M. O. and Watts, A. (2001). The existence of pairwise stable networks. *Seoul Journal of Economics*, 14.

Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of Economic Theory*, 71:44–74.

Jensen, M. K. and Rigos, A. (2018). Evolutionary games and matching rules. *International Journal of Game Theory*, 47(3):707–735.

Lee, L.-F., Liu, X., Patacchini, E., and Zenou, Y. (2021). Who is the key player? a network analysis of juvenile delinquency. *Journal of Business & Economic Statistics*, 39(3):849–857.

Maynard Smith, J. (1982). *Evolution and the Theory of Games.* Cambridge University Press.

Maynard Smith, J. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246(5427):15–18.

Mele, A. (2022). A structural model of homophily and clustering in social networks. *Journal of Business & Economic Statistics*, 40(3):1377–1389.

Myers, C. A. and Shultz, G. P. (1951). The dynamics of a labor market: a study of the impact of employment changes on labor mobility, job satisfactions, and company and union policies. *(No Title)*.

Newton, J. (2012). Coalitional stochastic stability. *Games and Economic Behavior*, 75(2):842–854.

Newton, J. (2017). Shared intentions: The evolution of collaboration. *Games and Economic Behavior*, 104:517–534.

Newton, J. and Angus, S. D. (2015). Coalitions, tipping points and the speed of evolution. *Journal of Economic Theory*, 157:172–187.

Patacchini, E., Rainone, E., and Zenou, Y. (2017). Heterogeneous peer effects in education. *Journal of Economic Behavior & Organization*, 134:190–227.

Patacchini, E. and Zenou, Y. (2008). The strength of weak ties in crime. *European Economic Review*, 52(2):209–236.

Patacchini, E. and Zenou, Y. (2012). Juvenile delinquency and conformism. *The Journal of Law, Economics, & Organization*, 28(1):1–31.

Robson, A. J. (1996). A biological basis for expected and non-expected utility. *Journal of Economic Theory*, 68(2):397–424.

Rousset, F. (2004). *Genetic structure and selection in subdivided populations*, volume 40. Princeton University Press.

Sadler, E. (2022). Making a swap: network formation with increasing marginal costs. *Available at SSRN 4294169*.

Samuelson, L. (1988). Evolutionary foundations of solution concepts for finite, two-player, normal-form games. In *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 211–225.

Sethi, R. and Somanathan, E. (2001). Preference evolution and reciprocity. *Journal of Economic Theory*, 97(2):273–297.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current directions in psychological science*, 9(5):160–164.

A Appendix

Proofs of Section 2

Proposition 1.

The main body of the text showcases how (6) obtains from (3), so here I only show how the two conditions provided by (4) are impossible to meet if the inequality is strict and reduce to just the equality if the inequality is weak. As a result, the 1-ESS only obtains through (6) as the strict Nash equilibrium of $\tilde{u}$, whereas the 1-nESS obtains as the set of Nash equilibria of the same game.

The equality of the two conditions in (3) obtains as

$$u_i(x) - u_i(y) + r\Big(\sum_{j\in I\setminus\{i\}} u_j(x) - u_j(y)\Big) = 0$$

through similar steps as (6). To obtain the inequality, notice that the fitness of the resident x as a function of ε obtains as

$$V_{q(\varepsilon)}(x) = \frac{1}{N}\left(u_i(x) + \sum_{j\neq i}\Big((1-r)\varepsilon u_j(y) + \big(r + (1-r)(1-\varepsilon)\big)u_j(x)\Big)\right).$$

Differentiating with respect to ε and multiplying by n gives $(1-r)\sum_{j\neq i}\big(u_j(y) - u_j(x)\big)$.

The fitness of the resident y as a function of ε obtains as

$$V_{q(\varepsilon)}(y) = \frac{1}{N}\left(u_i(y) + \sum_{j\neq i}\Big(\big(r + (1-r)\varepsilon\big)u_j(y) + (1-r)(1-\varepsilon)u_j(x)\Big)\right).$$

Differentiating with respect to ε and multiplying by n gives $(1-r)\sum_{j\neq i}\big(u_j(y) - u_j(x)\big)$ as before, hence we have

$$\frac{\partial}{\partial\varepsilon}V_{q(\varepsilon)}(x) = \frac{\partial}{\partial\varepsilon}V_{q(\varepsilon)}(y),$$

meaning that it is impossible to have strict inequality between the two, while the weak inequality is automatically satisfied.

Proposition 2.

For the first condition, as in the case of (1), assume that $\varepsilon = 0$. Then, I calculate the expected interaction payoff of a focal y-player in each position. For such a player in position i, two cases may obtain. If position j is assigned to a mutant, they obtain the interaction payoff $u_i(y)$, if it is assigned to a resident, they obtain $u_i(y^{(i)})$ with $y^{(i)} = (y_i, x_{-i})$. The expected payoff in position i is thus $ru_i(y) + (1-r)u_i(y^{(i)})$.

Similarly, the expected payoff in position j amounts to $ru_j(y) + (1-r)u_j(y^{(j)})$ for $y^{(j)} = (y_j, x_{-j})$.

A focal y-player in position $k \neq i, j$ will receive (i) interaction payoff $u_k(y)$ if both positions i and j are assigned to mutants, which happens with probability r^2; (ii) payoff $u_k(y^{(i)})$ if position i is assigned to a mutant but j is assigned to a resident, occurring with probability $r(1-r)$; (iii) payoff $u_k(y^{(j)})$ if position j is assigned to a mutant and i is assigned to a resident, happening with the same probability; (iv) and payoff $u_k(x)$ if both positions i and j are assigned to residents, happening with probability $(1-r)^2$. The expected payoff in position k thus amounts to $r^2u_k(y) + r(1-r)(u_k(y^{(i)}) + u_k(y^{(j)})) + (1-r)^2u_k(x)$.

The fitness of the mutant y-player is thus

$$\begin{aligned} V_{q(0)}(y) = \frac{1}{N}\Bigg(& r\big(u_i(y) + u_j(y)\big) + (1-r)\big(u_i(y^{(i)}) + u_j(y^{(j)})\big) \\ & + \sum_{k \in I \setminus \{i,j\}} \Big(r^2 u_k(y) + r(1-r)\big(u_k(y^{(i)}) + u_k(y^{(j)})\big) + (1-r)^2 u_k(x)\Big)\Bigg). \end{aligned}$$

As in the case of the 1-ESS, the expected fitness of the resident x-player is $\sum_{j\in I} u_j(x)/N$. Subtracting the fitness of y and multiplying by N gives

$$\begin{aligned} & N\Big(V_{q(0)}(x) - V_{q(0)}(y)\Big) = \\ & \qquad r\big(u_i(x) - u_i(y) + u_j(x) - u_j(y)\big) + (1-r)\big(u_i(x) - u_i(y^{(i)}) + u_j(x) - u_j(y^{(j)})\big) \\ & \qquad + r \sum_{k \in I \setminus \{i,j\}} \Bigg(r\big(u_k(x) - u_i(y)\big) + (1-r)\big(2u_k(x) - u_k(y^{(i)}) - u_k(y^{(j)})\big)\Bigg). \end{aligned}$$

Thus, the first condition of ESS, (3) is satisfied if the above expression is positive, as stated.

To obtain (4), calculating $V_{q(\varepsilon)}(x)$ and $V_{q(\varepsilon)}(y)$ by similar steps gives

$$\begin{aligned} & V_{q(\varepsilon)}(x) = \big(u_i(x) + u_j(x)\big)(r + (1-r)(1-\varepsilon)) + \big(u_i(y^{(j)}) + u_j(y^{(i)})\big)(1-r)\varepsilon + \\ & \sum_{k \in I \setminus \{i,j\}} u_k(x)(r + (1-r)(1-\varepsilon))^2 + \big(u_k(y^{(i)}) + u_k(y^{(j)})\big)(1-r)\varepsilon(r + (1-r)(1-\varepsilon)) + u_k(y)(1-r)^2\varepsilon^2, \end{aligned}$$

$$\begin{aligned} & V_{q(\varepsilon)}(y) = \big(u_i(y) + u_j(x)\big)(r + (1-r)\varepsilon) + \big(u_i(y^{(i)}) + u_j(y^{(j)})\big)(1-r)(1-\varepsilon) + \\ & \sum_{k \in I \setminus \{i,j\}} u_k(x)(1-r)^2(1-\varepsilon)^2 + \big(u_k(y^{(i)}) + u_k(y^{(j)})\big)(1-r)(1-\varepsilon)(r + (1-r)\varepsilon) + u_k(y)(r + (1-r)\varepsilon)^2. \end{aligned}$$

Taking derivatives with respect to ε gives

$$\frac{\partial}{\partial\varepsilon}V_{q(\varepsilon)}(x) = (1-r)\big(u_i(y^{(j)}) - u_i(x) - u_j(y^{(i)}) - u_j(x)\big) + \\ \sum_{k\in I\setminus\{i,j\}} \Bigg(u_k(x)\big(-2r(1-r) - 2(1-r)^2(1-\varepsilon)\big) \\ +\big(u_k(y^{(i)}) + u_k(y^{(j)})\big)\big((1-r)r + (1-r)^2(1-2\varepsilon)\big) \\ +u_k(y)\big(2(1-r)\varepsilon\big)\Bigg),$$

$$\frac{\partial}{\partial\varepsilon}V_{q(\varepsilon)}(y) = (1-r)\big(u_i(y) - u_i(y^{(i)}) + u_j(y) - u_j(y^{(j)})\big) + \\ \sum_{k\in I\setminus\{i,j\}} \Bigg(u_k(x)\big(-2(1-r)^2(1-\varepsilon)\big) \\ +\big(u_k(y^{(i)}) + u_k(y^{(j)})\big)\big(-(1-r)r + (1-r)^2(1-2\varepsilon)\big) \\ +u_k(y)\big(2r(1-r) + 2(1-r)^2\varepsilon\big)\Bigg).$$

Dividing by $(1-r)$, substituting $\varepsilon = 0$ and taking the difference $V_{q(\varepsilon)}(x) - V_{q(\varepsilon)}(y)$ gives

$$\left[\frac{\partial}{\partial\varepsilon}V_{q(\varepsilon)}(x)\right]_{\varepsilon=0} - \left[\frac{\partial}{\partial\varepsilon}V_{q(\varepsilon)}(y)\right]_{\varepsilon=0} > 0 \Leftrightarrow \\ \sum_{k\in\{i,j\}}\Big(u_k(y^{(i)}) + u_k(y^{(j)}) - u_k(x) - u_k(y)\Big) + 2r\sum_{k\in I\setminus\{i,j\}}\Big(u_k(y^{(i)}) + u_k(y^{(j)}) - u_k(x) - u_k(y)\Big) > 0.$$

Proofs of Section 3

Proposition 3

As stated in the main body of the text, I provide an exact characterization of pair-evolutionarily stable networks, from which the simpler characterization as stated by the Proposition obtains in a straightforward way.

As $\rho > 0$, no strategy x with $x_{ij} = 1$ and $x_{ji} = 0$ for $i, j \in I$ is stable against the strategy defined by $y_{ij} = 0$, $y_{ji} = 0$ and $y_{k\ell} = x_{k\ell}$ otherwise. Hence, a strategy x is evolutionarily stable under the protocol $\mathcal{P}$ only if x does not include unreciprocated linking attempts.

Given that x does not include unreciprocated linking attempts, I proceed through the three types of mutations that can occur on any pair i, j; unilateral severance of an existing link between i and j, pairwise formation of a new link between i and j, and pairwise severance of an existing link between i and j.

First, I consider the unilateral severance of a link:

Lemma 6. *Let $x \in X$ with $x_{ij} = x_{ji} = 1$. Then, x is evolutionarily stable against the strategy y given by $y_{ij} = 0$ and $y_{i'j'} = x_{i'j'}$ otherwise if and only if*

$$b_i\big(g(x)\big) - b_i\big(g(x) - ij\big) + r \sum_{k \in I \setminus \{i\}} \Big(b_k\big(g(x)\big) - b_k\big(g(x) - ij\big)\Big) > c_i + r(1-\rho)c_j \tag{22}$$

Proof. The statement follows from a direct application of Proposition 1 on the game given by (14). One just needs to substitute $u_i(x) - u_i(y) = b_i(x) - b_i(x - ij) + c_i$, $u_j(x) - u_j(y) = b_i(x) - b(x - ij) + (1-\rho)c_j$, and $u_k(x) - u_k(y) = b_k(x) - b_k(x - ij)$ for $k \neq i, j$. ■

Next, consider stability against the bilateral addition of a link.

Lemma 7. *Let $x \in X$ with $x_{ij} = x_{ji} = 0$. Then, x is evolutionarily stable against the strategy y given by $y_{ij} = y_{ji} = 1$ and $y_{i'j'} = x_{i'j'}$ otherwise if and only if*

$$\begin{aligned} b_i\big(g(x) + ij\big) - b_i\big(g(x)\big) + b_j\big(g(x) + ij\big) - b_j\big(g(x)\big) + r \sum_{k \in I \setminus \{i,j\}} \Big(b_k\big(g(x) + ij\big) - b_k\big(g(x)\big)\Big) \\ < \frac{r + (1-r)\rho}{r}(c_i + c_j) \end{aligned} \tag{23}$$

or if (23) holds with equality and we have

$$r^2 \sum_{k \in I \setminus \{i,j\}} \big(b_k(g(x)) - b_k(g(x) + ij)\big) > c_i + c_j. \tag{24}$$

Proof. Condition (23) obtains from (7) by substituting $u_i(x) - u_i(y) = b_i(g(x)) - b_i(g(x) + ij) + c_i$, $u_j(x) - u_j(y) = b_j(g(x)) - b_j(g(x) + ij) + c_j$, $u_i(x) - u_i(y^{(i)}) = \rho c_i$, $u_j(x) - u_j(y^{(j)}) = \rho c_j$, as well as $u_k(x) - u_k(y) = b_k(g(x)) - b_k(g(x) + ij)$, $u_k(x) - u_k(y^{(i)}) = 0$, and $u_k(x) - u_k(y^{(j)}) = 0$ for $k \neq i, j$, taking the negative, and dividing by r.

To (24), assume that (23) is met with equality. Substituting $u_i(y^j) - u_i(x) = 0$, $u_i(y^i) - u_i(y) = b_i(g(x)) - b_i(g(x) + ij) + c_i(1-\rho)$ and $u_j(y^i) - u_j(x) = 0$, $u_j(y^j) - u_j(y) = b_j(g(x)) - b_j(g(x) + ij) + c_j(1-\rho)$ into the first term of (8) and $u_k(y^{(i)}) - u_k(x) = 0$ and $u_k(y^{(j)}) - u_k(y) = b_k(g(x)) - b_k(g(x) + ij)$ into its second term returns

$$b_i(g(x)) - b_i(g(x)+ij) + b_j(g(x)) - b_j(g(x)+ij) + 2r \sum_{k \in I \setminus \{i,j\}} \big(b_k(g(x)) - b_k(g(x)+ij)\big) > -(1-\rho)(c_i + c_j).$$

With (23) holding with equality, adding both sides to the respective side of the above equation and multiplying by r returns (24). ■

Finally, consider stability against bilateral severance.

Lemma 8. *Let $x \in X$ with $x_{ij} = x_{ji} = 1$. Then, x is evolutionarily stable against the strategy y given by $y_{ij} = y_{ji} = 0$ and $y_{i'j'} = x_{i'j'}$ otherwise if and only if*

$$b_i\big(g(x)\big) - b_i\big(g(x) - ij\big) + b_j\big(g(x)\big) - b_j\big(g(x) - ij\big) + r(2-r) \sum_{k \in I \setminus \{i,j\}} \Big(b_k\big(g(x)\big) - b_k\big(g(x) - ij\big)\Big)$$
$$> c_i + c_j \tag{25}$$

or if (25) is met with equality and

$$r^2 \sum_{k \in I \setminus \{i,j\}} \big(b_k(g(x) - ij) - b_k(g(x))\big) > \rho(c_i + c_j) \tag{26}$$

Proof. Condition (25) also obtains from (7) by substituting $u_i(x) - u_i(y) = u_i(x) - u_i(y^{(i)}) = b_i(g(x)) - b_i(g(x) - ij) - c_i$, $u_j(x) - u_j(y) = u_j(x) - u_j(y^{(j)}) = b_j(g(x)) - b_j(g(x) - ij) - c_j$, as well as $u_k(x) - u_k(y) = u_k(x) - u_k(y^{(i)}) = u_k(x) - u_k(y^{(j)}) = b_k(g(x)) - b_k(g(x) - ij)$ for $k \neq i, j$.

To get the second condition, assume that (25) is met with equality. Substituting $u_i(y^{(i)}) - u_i(x) = b_i(g(x) - ij) - b_i(g(x)) + c_i$, $u_i(y^{(j)}) - u_i(y) = -\rho c_i$ and $u_j(y^{(j)}) - u_i(x) = b_j(g(x) - ij) - b_j(g(x)) + c_j$, $u_j(y^{(i)}) - u_j(y) = -\rho c_j$ into the first term of 8 and $u_k(x) - u_k(y^{(i)}) = b_k(g(x)) - b_k(g(x) - ij)$, $u_k(y^j) - u_k(y) = 0$ into the second gives

$$b_i(g(x)-ij)-b_i(g(x))+b_j(g(x)-ij)-b_j(g(x))+2r \sum_{k \in I \setminus \{i,j\}} \big(b_k(g(x)-ij)-b_k(g(x))\big) > -(c_i+c_j)(1-\rho)$$

As (25) holds with equality, adding both sides to the respective sides of the above equation returns (26). ■

Together, Lemmas 6, 7, and 8 characterize evolutionarily stable networks: If $ij \in g$, then the strategy x forming g must be stable against strategies that remove the link (Lemmas 6 and Lemma 8), while if $ij \notin g$, it must be stable against strategies that add it (Lemma 7). Proposition 3 follows.

Section 4

Lemma 2

It is straightforward to see that point 1 contains the sufficient and necessary conditions of best response behavior by node i.

For point 2, start by noting that in a strict Nash equilibrium, any two nodes who sponsor links must have a link running between them. If i and j are sponsors of links but are not themselves linked to each other, i could delete a link to another node, say k, and link to j

instead, which would keep its payoff the same. It follows that if x is a strict Nash equilibrium, the network $\bar{x}$ is complete. Next assume that we have a strict Nash equilibrium, but two nodes, i and j are not linked; it follows that both are pure receivers with at least d^* links. Thus, the network $\bar{x}$ contains a complete subgraph $J \subseteq I \setminus \{i,j\}$ of sponsors. Since no sponsor can have degree larger than d^*, we must have $|J| \leq d^* + 1$. However, node i must have at least degree d^*, so $|J| \geq d^*$ must hold as well and every member of J must sponsor a link to i. It follows that $|J| = d^*$, meaning that each member of J is has a directed link with every other member of J and sends a link to i, amounting to degree d^*. Then, however, j's in-degree and its degree must be zero, which, due to $d^* > 0$, contradicts that x is a Nash equilibrium.

Proposition 4

Let $\underline{r} \in (0,1)$ be given such that for every $d, d' \in \{0,\ldots,N-2\}$ we have $b(d+1) - b(d) < c \leftrightarrow b(d+1) - b(d) + \underline{r}\big(b(d'+1) - b(d')\big) < c$. Then, the degree of altruism in $\tilde{u}$ is low enough that it does not change the number of links that nodes initiate. The only effect of altruism is a preferential linking to low-degree nodes. Then, every Nash equilibrium of $\tilde{u}$ is a Nash equilibrium of u.

1. From Proposition 1, we know that every 1-nESS is a Nash equilibrium of the game $\tilde{u}$. It is easy to see that the proposed networks are Nash equilibria of $\tilde{u}$ as no sponsor is motivated to sever links, create new links, or rearrange existing links, while no pure receiver is motivated to initiate links, completing the 'if' direction.

For the 'only if' direction, consider $x \in X$, a Nash equilibrium of u, with a node i such that $d_i^{in}(x) \geq d^* + 2$. Then, there must exist $d^* + 2$ disjoint nodes who each send a link to i. As each sponsor has degree exactly d^*, there exist two such sponsors j and k who are not linked. Then, either of them, say j could delete its link to i and link to k instead, and doing so would raise $\tilde{u}$ due to the concavity of b. Hence, x is not a Nash equilibrium of u.

2. By the same argument as point 1, x is a 1-ESS only if it is a strict Nash equilibrium of $\tilde{u}$ which are all strict Nash equilibria of u due to the condition placed on r. In every d^*-regular network, every node is playing a strict best response in the game $\tilde{u}$, completing the 'if' direction. For the 'only if' direction, we only have to consider the almost regular networks of point 1. In such a network, if there exists a node i receiving $d^* + 1$ links, then there must exist two sponsors j and k who link to i but do not link to each other. Then, either of them, say j could delete its link to i and link to k instead, which would leave $\tilde{u}_j$ unchanged, hence almost regular networks cannot be strict Nash equilibria of $\tilde{u}$.

Proposition 5

It is clear that x must be oriented and $d_i(x) \geq d^*$ for all 1-ESS networks, independently of the value of r. To show that for large r, every link must point to a node with exactly degree d^*, assume that there exists $ij \in x$ with $d_j(x) > d^*$. Then, we must have $d_i(x) = d^*$, otherwise the link would add benefit zero. Now, consider k who is a neighbor of j but not of i. Then, $d_k(x) = d^*$ must also hold, otherwise the link between j and k would add benefit zero. Then, i could replace the link ij with the link ik, which would leave its payoffs unchanged in the game $\tilde{u}$, hence x cannot be a 1-ESS.

Proposition 6

If g is as described by the statement, then pairwise mutations take one of three forms.

1. A link ij bilaterally forms with $d_i(g) = n$ and $d_i(g) = m$, and with i being in an n-regular component and j being in an m-regular one (possibly with $n = m$), for $n, m \in \{n_1, \dots, n_K\}$.
2. A link ij breaks with $d_i(g) = d_j(g) = n$ through unilateral severance, and with i and j both being in an n-regular component for some $n \in \{n_1, \dots, n_K\}$.
3. A link ij is bilaterally severed with $d_i(g) = d_j(g) = n$, and with i and j both being in the same n-regular component, for $n \in \{n_1, \dots, n_K\}$.

I show that the conditions given by the Proposition ensure g against all such invasions.

1. The payoff of node i prior to adding the link ij equals

$$n\left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n^2}\right) = 2 + \frac{1}{n}.$$

Its payoff after adding the link ij is

$$n\left(\frac{1}{n} + \frac{1}{n+1} + \frac{1}{n(n+1)}\right) + \frac{1}{n+1} + \frac{1}{m+1} + \frac{1}{(n+1)(m+1)} = 2 + \frac{n+m+3}{(n+1)(m+1)},$$

so the gain of i is

$$\frac{n+m+3}{(n+1)(m+1)} - \frac{1}{n}.$$

The payoff gain of a neighbor of i, node $k \neq j$ by the addition of the link ij equals

$$-\left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n^2}\right) + \left(\frac{1}{n} + \frac{1}{n+1} + \frac{1}{n(n+1)}\right) = -\frac{1}{n^2}.$$

As i has n neighbors in g, the total payoff gain in its neighborhood amounts to $-1/n$.

Through similar calculations, the payoff gain of node j equals

$$\frac{n+m+3}{(n+1)(m+1)} - \frac{1}{m},$$

while the total payoff gain in j's neighborhood equals $-1/m$.

By Lemma 7, g is stable against the formation of the link ij if and only if

$$2\frac{n+m+3}{(n+1)(m+1)} - \left(\frac{1}{n} + \frac{1}{m}\right)(1+r) < 0.$$

Rearranging gives

$$2\frac{n+m+3}{n+m}\frac{n}{n+1}\frac{m}{m+1} - 1 < r.$$

2. The payoff gain of node i through the severance of link ij equals

$$-\sum_{k:\, ik\in g}\left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n^2}\right) + \sum_{k\neq j:\, ik\in g}\left(\frac{1}{n-1} + \frac{1}{n} + \frac{1}{n(n-1)}\right) =$$
$$n\left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n^2}\right) + (n-1)\left(\frac{1}{n-1} + \frac{1}{n} + \frac{1}{n(n-1)}\right) = -\frac{1}{n}$$

The payoff gain of node j is similarly $-1/n$. For $n = 1$, severing this link does not benefit any other node $k \neq i, j$, hence the condition ensuring against unilateral severance is vacuous.

For $n \geq 2$, the payoff gain of a neighbor of i, node $k \neq j$ from the link ik equals

$$-\left(\frac{1}{n} + \frac{1}{n} + \frac{1}{n^2}\right) + \left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n(n-1)}\right) = \frac{n+1}{n^2(n-1)},$$

while the payoff gain of a neighbor of j, node $\ell \neq i$, from the link $j\ell$ is similarly $(n+1)/\big(n^2(n-1)\big)$. As both i and j have exactly $n-1$ other neighbors, the total externalities of severing the link amount to $2(n+1)/n^2$.

Thus, by Proposition 1, g is stable against the mutation severing ij from the side of i if and only if

$$-\frac{1}{n}(1+r) + \frac{2(n+1)}{n^2}r < 0.$$

Simplifying gives $r < n/(n+2)$. This condition is stricter the smaller n is, hence if it holds for the minimum degree node that has at least two neighbors, it holds for all nodes.

3. As in point 1, the gains of players i and j from severing the link ij amount to $-2/n$. For $n = 1$, severing this link does not benefit any other node $k \neq i, j$, hence the condition ensuring against unilateral severance is vacuous. For $n \geq 2$, the total gains amount to $2(n+1)/n^2$, again, as in point 1. By Lemma 8, g is stable against the bilateral severance of the link ij if and only if

$$-\frac{1}{n} + r(2-r)\frac{n+1}{n^2} < 0.$$

Rearranging gives $n/(n+1) > r(2-r)$, and solving for r under the condition that $r \in [0,1]$ one gets $1 - \sqrt{2}/\sqrt{n+1} > r$. This condition is stricter the smaller n is, hence if it holds for the minimum degree nodes with at least two neighbors, it holds for all nodes. As $1 - \sqrt{2}/\sqrt{n+1} < n/(n+2)$ for all n, if g is stable against bilateral severance, it is also stable against unilateral severance, hence only the conditions derived in points 1 and 3 end up mattering.

Corollary 2

The 'if' direction is a straightforward consequence of Proposition 6. For the 'only if' direction, it is enough to show that the conditions guaranteeing the stability of an n-regular network against bilateral link formation and link severance contradict each other for $n \in \{2, \ldots, N-2\}$. By Proposition 6, the former needs

$$\frac{n^2+n-1}{(n+1)^2} > r,$$

while the latter needs

$$1 - \frac{\sqrt{2}}{\sqrt{n+1}} < r.$$

The second inequality implies $1 - 2/(n+1) < r$, however $1 - 2/(n+1) < (n^2+n-1)/(n+1)^2$, so the set of assortativity parameters that guarantee the stability of an n-regular network for $n \in \{2, \ldots, N-2\}$ is empty.

Lemmas 3 and 4

In pairwise stable networks, the link ij forms if and only if its benefits are larger than the costs for both players. The cost of every link is c while the benefit of the link ij for player i is $w_{ij} - \phi_i \mathbb{1}_{C(i)\neq C(j)}$ in both models. Due to reciprocity ($w_{ij} = w_{ji}$), benefits are larger than the costs for both players if and only if it holds for the player with the larger distaste parameter. Thus, taking the maximum for both players gives the formulation stated the Lemma 3.

As w_{ij} is drawn from the distribution characterized by $\overline{F}$, the probability that the network benefit exceeds the cost for both players obtain as stated by Lemma 4.

Lemma 5 and Proposition 7

Let $\rho = 0$. Then, substituting $b_i(g(x)+ij) - b_i(x) = w_{ij} - \phi_i \mathbb{1}_{C(i)\neq C(j)}$, $b_j(g(x)+ij) - b_j(x) = w_{ij} - \phi_j \mathbb{1}_{C(i)\neq C(j)}$, and $b_k(g(x)+ij) - b_k(g(x)) = -\phi_k \mathbb{1}_{C(i)\neq C(j)}$ into the complement of (23) (the pairwise formation condition of Lemma 7) and dividing by two gives the condition stated by (19).

If $C(i) = C(j)$, (19) reduces to $w_{ij} > c$. Furthermore, $b_i(g(y)) - b_i(g(y) - ij) = b_j(g(y)) - b_j(g(y) - ij) = w_{ij}$ and there are no externalities. Then, by the first pairwise severance condition of Lemma 8, (8), y is stable against x if $w_{ij} > c$, while by the individual severance conditions of Lemma 6, y is stable against $y^{(i)}$ and $y^{(j)}$ if $w_{ij}(1 + c) > c(1 + r)$. Hence, if x is not stable agaisnt y, then y is stable against x, $y^{(i)}$, and $y^{(j)}$.

Proposition 7 follows by substituting the ϕ_i values for all four cases and finding the probability that w_{ij} is large enough to satisfy (19).

Proposition 8

1. As the second factor is constant and the first approaches 0 as the number of radicals goes to 0, η^{PS} converges to 0 as well. 2. As $\lim_{z\to\infty} \overline{F}(z) = 0$, while all other factors in the fractions are bounded or constant, all fractions converge to 0.