Making a Swap: Network Formation with Increasing Marginal Costs*

Evan Sadler†

May 15, 2023

Abstract

I propose a simple theory of strategic network formation that accounts for many empirical patterns. The theory consists of three key parts: i) convex linking costs, ii) local linking benefits, and iii) swap-proofness, a refinement of pairwise stability. An acyclic preference condition—the *mutual favoritie property*—implies that a unique swap-proof stable graph generically exists. If players agree about who is a more desirable neighbor, then stability robustly begets homophily and clustering. With similar assumptions on players' desire for links, stable graphs take on structures—strong hierarchies or ordered overlapping cliques—that mirror real networks in different domains. I discuss several extensions, the relationship to matching, and a dynamic foundation for swap-proof stability.

*I dedicate this paper to my son, Percival Featherstone Sadler, whose impending birth provided a wellspring of motivation to quickly work out the main findings in this paper. I would like to thank Yeon-Koo Che, Ben Golub, Navin Kartik, Qingmin Liu, Alex Teytelboym, and Quitzé Valenzuela-Stookey for helpful conversations and references. I also thank seminar participants at Northwestern and Cornell for stimulating comments.

†Columbia University – es3668@columbia.edu.

1 Introduction

From friendships to coauthors to trading relationships and beyond, networks display striking regularities. Among the most reliable patterns are clustering, homophily, and the "small-worlds" property. One's neighbors are often neighbors of each other, one's neighbors are similar to oneself on multiple dimensions, and typical distances between two individuals are small relatitve to population size.[1] Often, networks also feature status hierarchies, though details differ across domains. Financial and trade networks display a tiered, core-periphery structure, in which larger firms or countries are much better connected than small ones (Soramaki et al., 2007; Craig and von Peter, 2014; Akerman and Seim, 2014). In contrast, social networks within small to medium peer groups (e.g., in workplaces and schools) break into separate cliques with a clear status ranking (Homans, 1950; Adler and Adler, 1995; Gest et al., 2007). What accounts for these recurring observations?

I propose a simple theory of strategic network formation that yields sharp predictions consistent with these patterns. The theory combines three key ingredients: i) increasing marginal costs of linking, ii) linking benefits that depend (only) on the two players' characteristics, and iii) a refinement of pairwise stability. A network formation game comprises a finite set of players with payoffs that depend on a graph that forms amongst them. In a pairwise stable graph, no player benefits from unilaterally severing a link, and no pair benefits from jointly forming a link. *Swap-proofness* asks for a minimal refinement: when evaluating an addition, each player in the pair may simultaneously sever one link. Despite a large multiplicity of pairwise stable graphs, swap-proofness generically selects a unique outcome.

[1]For instance, Barabási and Albert (2002) document that coauthorship networks across different academic fields exhibit far higher clustering coefficients than similarly dense random graphs, and Ugander et al. (2011) and Myers et al. (2014) show that many online social networks are also highly clustered. See McPherson et al. (2001) for a comprehensive review of evidence on homophily. Milgram (1967) demonstrated that seemingly distant people are often connected through relatively short network paths, coining the term "six degrees of separation." More recently, Dodds et al. (2003) and Backstrom et al. (2012) have replicated this famous experiment—if anything, distances are now shorter.

Homophily and clustering emerge as robust features of swap-proof stable graphs. Moreover, under different assumptions on link benefits, these graphs mirror both the hierarchies seen in trade networks and the ranked cliques seen in peer groups.

Beyond these qualitative predictions, the analysis helps organize our broader understanding of strategic incentives and network formation. Despite being nearly three decades old, this literature contains few results that characterize pairwise stable graphs in large classes of games. Moreover, the results we do have often depend on knife-edge properties.[2] This testifies to the difficulty of determining stable structures outside of the simplest examples. Following Sadler and Golub (2022), I focus on ordinal properties of players' payoffs, deriving results that do not rely on symmetry or any particular functional forms. This rewards us with simple—and more importantly, interpretable—conditions that ensure existence, uniqueness, and meaningful restrictions on the graphs that emerge.

For transparency, I focus on a restricted class of games with additively separable linking benefits—in the online Appendix, I identify more primitive payoff properties under which the analysis carries through. Each neighbor delivers a benefit that depends on the two players' types, and players incur linking costs that are increasing and convex in their degrees. A simple example illustrates an unnatural multiplicity of pairwise stable outcomes, motivating our refinement. Because swap-proof stable graphs need not exist in general, our next step is to identify conditions under which they do. This brings us to the paper's first contribution: the *mutual favorite property*.

Given a set of possible links E, I say $ij \in E$ is a *mutual favorite* if i prefers j as a neighbor over all k such that $ik \in E$, and j prefers i as a neighbor over all ℓ such that $j\ell \in E$. The

[2]In their seminal paper, Jackson and Wolinsky (1996) analyze two simple examples, the "connections model" and the "coauthor model." Subsequent work largely follows this approach, working out specific models rather than sytematically studying the relationship between payoffs and stable graphs. Some more recent work has started in this direction, though results typically depend on players having symmetric payoffs, which precludes any ex-ante heterogeneity that is independent of the network that forms. I discuss this in more detail under "Related Work."

mutual favorite property holds if *every* set E contains a mutual favorite. This property helps ensure existence because it rules out cycles of improving swaps. Several classes of games satisfy this condition—for instance, if players share a common ranking of potential neighbors, or if linking benefits are symmetric. Together with no indifferences, the mutual favorite property guarantees that a unique swap-proof stable graph exists.

I next study stable network structures using ordinal payoff properties. Player i is *more desirable* than player j if every other player gains more from a link with i than from a link with j. If this yields a total order on the set of players, then our existence result applies, and stable graphs have predictable features. Players limit their connections due to increasing marginal costs, so those who are most desirable quickly run through their budgets linking amongst each other, and we get a dense core at the top of the ranking. If the graph contains multiple components, each contains its own clustered core comprising its most attractive members. Moreover, a bound on player degrees translates into a bound on the distance between any linked pair in the common preference ranking: players assortatively match with those who are similarly ranked.[3] Note, however, that this result does not preclude core-periphery graphs, which often display *negative* assortativity—core vertices have high degrees in these graphs, and with a large degree bound, the distance between two neighbors in the rank order can become large. Nevertheless, the distance bound does not depend on the total number of players, so assortativity grows more pronounced in large populations.[4]

A common ranking of potential partners induces a natural status hierarchy, but the precise structures that emerge depend on how a player's attractiveness as a neighbor relates to her desire for links. If more attractive players also want more links, then stable graphs are *strong hierarchies*. In a strong hierarchy, a player i who ranks above j has (weakly)

[3]Although the degree of any player is endogenous, we can readily construct bounds. If the marginal cost of link $K+1$ is higher than the maximum possible benefit, no player can have more than K neighbors in a stable graph.

[4]This broadly agrees with empirical findings. For instance, Currarini et al. (2010) document that racial homophily in high school friendship networks is greater in larger high schools.

more neighbors than j, and these neighbors rank (weakly) higher than those of j. Strong hierarchies include the more familiar nested split graphs, a rigid structure in which neighborhoods are ordered by set inclusion.[5] In line with previous studies, nested split graphs appear if marginal costs are constant, but with increasing marginal costs, those at the top may find it prohibitively expensive to maintain so many links. In contrast, if more attractive players want fewer links, then stable graphs consist of ordered overlapping cliques: every player's neighborhood is an interval in the common rank order, with endpoints increasing in a player's own position. Although this agrees with earlier work assuming constant marginal costs (Sadler and Golub, 2022), we arrive through a different path. With constant marginal costs, all pairwise stable graphs consist of ordered overlapping cliques, but with increasing marginal costs, refinement via swap-proofness is crucial. Note that both of these characterizations are tight: any graph that is a strong hierarchy or ordered overlapping cliques can arise as the unique stable graph in an appropriate game.

These results not only replicate key features of real networks, they also pinpoint why particular structures appear. Prior work explains patterns in financial and trade networks as a consequence of complementarities—core-periphery structures emerge because large entities are more attractive partners, and they benefit more from additional links (König et al., 2014; Sadler and Golub, 2022). Our findings highlight an important implicit assumption: marginal linking costs are essentially constant. Moreover, to the extent this assumption fails in practice, we gain some insight into how real trading networks might differ from these stark predictions. At the boundary between our two cases, we find another pattern commonly seen in peer groups. If more attractive players desire neither more nor fewer links, then stable graphs partition the players into separate cliques, with more attractive players in larger groups.[6] Familiar schoolyard dynamics derive from a common desire for connection together

[5] In a nested split graph, if i ranks above j, then every neighbor of j is a neighbor of i.

[6] This is the only structure that is both a strong hierarchy and consists of ordered overlapping cliques.

with an agreed ranking of who is a desirable friend.

Implicit in our solution concept is that players choose precisely with whom they link, selecting among all others in the group. Particularly in large communities, opportunities to interact are typically more limited than this. To expand the applicability of our analysis to such settings, Section 5 introduces a hybrid model, augmenting our game with a set of *feasible links* that constrains network formation. In practice, feasible links might result from random meetings, geographic proximity, or in some cases, design choices. In contrast with random graphs or search models, players here actively select which links to form, given their constraints. Importantly, the existence and uniqueness result still applies, enabling a new approach for studying strategic network formation in larger populations.

Another potential concern in applications is that key results could fail with a little bit of noise in players' preferences. To address this, Section 6 highlights a model in which linking benefits include a noise term that is idiosyncratic to each pair. I show that the mutual favorite property still holds in this setting, ensuring a unique stable graph. Not only does this offer hope that we can estimate linking preferences from network data, it can also help stable graphs better match features of real networks. A small bound on player degrees, together with our earlier order conditions, precludes small-world graphs. However, because distances in a network rapidly shrink with a few random links (e.g. Watts and Strogatz, 1998), large enough noise terms should ameliorate this issue.

Towards the end of the paper, I discuss broader implications of the analysis along with extensions. If linking incentives lead to strong hierarchies, a result from Sadler (2022) implies that equilibrium actions in an unrelated network game of strategic complements follow the same ordering, highlighting a mechanism through which networks can reinforce existing inequalities. I also show that, if the mutual favorite property holds, better-response dynamics robustly converge to the unique swap-proof stable graph. In the online appendix, I adapt the solution concept from Sadler and Golub (2022) and note that our structural predictions

continue to hold if we replace exogenous types with endogenous actions. If players who take higher actions are more attractive neighbors, then we get either strong hierarchies or ordered overlapping cliques depending on whether actions and links are complements or substitutes.

This paper makes two main contributions. First, it characterizes stable graphs in a much larger class of games than any considered previously, linking predictions to qualitative features of players' payoffs. The mutual favorite property ensures existence and uniqueness in an even larger class—including games with arbitrary link constraints and noisy benefits—providing a key tool for future work. Second, it identifies a canonical model that explains several features of real networks. While random graphs can readily fit desired structures, here we tie these features to incentive properties like how steeply marginal linking costs increase and how one's desirability as a neighbor relates to one's desire for links. The analysis thus sheds light on when and why certain structures appear, identifying precisely what payoff assumptions we need. In addition to these contributions, the notion of an improving swap highlights the connection between network formation and matching—improving swaps are blocking pairs by another name. As I discuss at various points, results here offer a unified perspective that can help extend earlier findings. A companion project makes this connection much more explicit (Sadler, 2023).

Related Work

Many non-strategic network formation models generate realistic networks—we can build-in key features like homophily, clustering, and small-worlds through an exogenous random process.[7] For instance, Watts and Strogatz (1998) develop a canonical model that starts from a tightly clustered lattice and randomly rewires a small percentage of links. This procedure reliably generates highly clustered graphs with short path lengths. More recent efforts are tailored for estimation, serving as inputs to applied research (e.g. Jackson and Rogers, 2007;

[7]See van der Hofstad (2017) for a technical survey.

Breza et al., 2020; Chandrasekhar and Jackson, 2021). While good for fitting data, these models have limited ability to illuminate the mechanisms behind regularities. In contrast, strategic models provide a basis to understand how and why network features vary across contexts, which in turn can help predict the impact of efforts to change networks.

Previous work, taking a middle path between random graphs and fully strategic network formation, addresses possible sources of homophily. In Currarini et al. (2009, 2010), agents decide how intensively to search for partners, but subject to this effort choice, random meetings determine which links form. Homophily arises from both a preference for own-type links and bias in the meeting process. Preference for own-type links leads members of larger groups to invest more effort in friendships, while bias in the meeting process mechanically generates homophily. My results highlight a different mechanism that can explain similar patterns as the result of assortative matching, and I discuss later how one might attempt to distinguish between these two explanations.

König et al. (2014) take a different hybrid approach, studying a dynamic process in which links are continually added and deleted. They select agents to add and delete links based on an exogenous random process, but which links an agent adds or deletes depend on incentives. This combination of payoffs and random link changes specifies a Markov chain, and the authors study its steady-state distribution. The payoffs imply that, without noise in link selection, we get a nested split graph at every step of the process—this is a type of core-periphery network. Moreover, the authors tune parameters of the random process to fit other features, like the degree distribution, observed in trade and financial networks. As with many random graph models, it remains unclear exactly how these features of the steady state depend on various modeling choices. The present paper sheds light here, identifying precisely what properties of the payoff function we need to obtain nested split graphs—in particular, we get nested split graphs if more desirable partners also desire more links, and the marginal cost of additional links is roughly constant. In Section 7.1, I also discuss a

preference based explanation for heavy-tailed degree distributions.

Thus far, the strategic network formation literature contains precious few results that characterize stable structures in families of games. Early papers focus on specific examples that lead to simple structures. For instance, in their seminal work, Jackson and Wolinsky (1996) study the "connections model" and the "coauthor model." In the former, pairwise stable graphs are either complete, empty, or stars. In the latter, stable graphs always partition players into cliques with distinct sizes. Other early examples include the "spatial connections model" of Johnson and Gilles (2000), which leads to ordered overlapping cliques—though for different reasons than in the present paper—and models of trade and market sharing agreements in Furusawa and Konishi (2007) and Belleflamme and Bloch (2004) respectively, in which stable graphs partition players into cliques.[8]

Moving beyond examples, Goyal and Joshi (2006) connect families of payoff functions to particular network structures. In "playing the field games," a player's marginal payoff from a link depends on how many connections she has and how many other links are in the graph. Depending on whether own and others' links exert positive or negative spillovers, stable graphs can be complete, empty, stars, or have the dominant group architecture.[9] In "local spillover games," the value of a link depends on how many connections a player has and how many the potential neighbor has. Again, depending on the nature of spillovers, pairwise stable graphs fall into a simple taxonomy: complete graphs, empty graphs, dominant group architecture, interlinked stars, or exclusive groups.[10] Allowing a larger class of payoffs, Hellmann (2021) shows that if own and others' links exert positive spillovers, stable graphs are nested split graphs. Aside from stark predictions, symmetric payoffs are a key limitation

[8]There is also a significant literature in which linking decisions are unilateral (e.g. Bala and Goyal, 2000; Galeotti and Goyal, 2010; Herskovic and Ramos, 2020). While the lack of mutual consent is an important difference, much of this work similarly focuses on specific examples with very simple equilibrium structures.

[9]In the dominant group architecture, there is a single clique, and all other players are isolated.

[10]Interlinked stars partition players into a core and a periphery. Those in the core link with all others, and those in the periphery link only with the core. Exclusive groups means that every component is a clique.

in both papers—results do not extend to settings with heterogeneous players. In contrast, the present paper *relies* on heterogeneity to ensure unique predictions.

Sadler and Golub (2022) present the most closely related analysis, introducing a model of network games together with network formation. In the special case with degenerate action sets, their model reduces to that in the present paper, assuming constant marginal costs. Under analogous order conditions, pairwise stable graphs are either nested split graphs, where we get strong hierarchies, or ordered overlapping cliques. As already discussed, increasing marginal costs may preclude nested splits graphs because they require some players to maintain too many links, and swap-proofness is necessary to ensure ordered overlapping cliques in our setting. Moreover, with minimal effort we can translate our structural results to games on endogenous networks. Extending the framework of Sadler and Golub (2022), Section A.3 shows that if higher actions make players both more attractive as neighbors and more desiring of links, then stable outcomes are strong hierarchies. Likewise, if higher actions make players more attractive, but reduce own linking incentives, then stable outcomes entail ordered overlapping cliques.

Even if linking incentives have no relation to payoffs in a network game, the network that forms can have a profound impact on equilibrium behavior. In Section 7.2, I highlight particularly stark implications when linking incentives produce strong hierarchies. These graphs belong to a larger class identified in Sadler (2022) that ensure robust predictions in games of strategic complements. All else equal, equilibrium actions and payoffs align with players' positions in the network hierarchy. This illuminates a potentially important mechanism through which a status ranking in one domain can reproduce itself elsewhere, reinforcing and exacerbating existing inequalities.[11]

Although I focus on a refinement of pairwise stability, I nevertheless contribute to the

[11] Joshi et al. (2020) highlight a different mechanism that is similar in spirit. They study a network formation model in which one set of links is given exogenously (e.g., inherited from older family connections), and complementarities cause newly formed links to replicate the exogenous hierarchy.

conversation about when pairwise stable graphs exist. Because pairwise stability entails robustness to coalitional deviations—no two players can both benefit from forming the missing link between them—existence is difficult to ensure. Moreover, known results are often difficult to check—e.g., a pairwise stable graph exists if the game has a potential (Jackson and Watts, 2001; Chakrabarti and Gilles, 2007)—or require very strong assumptions—e.g., adding any link to the graph weakly increases the marginal value of all other links (Hellmann, 2013). Although my existence result applies less broadly than those based on potentials, the assumptions are straightforward to check and easy to interpret. In fact, increasing marginal costs often produce the opposite problem, a large multiplicity, which motivates refinement.

In general, swap-proofness is strictly weaker than other refinements. Goyal and Vega-Redondo (2007) define "bilateral equilibrium," which allows a pair of players to form a link and potentially sever many links at the same time—with increasing marginal costs, this turns out to be equivalent to swap-proof stability. Strong stability (Jackson and van den Nouweland, 2005) requires robustness against arbitrary deviations by coalitions of any size—the coalition can unilaterally sever any link attached to one of its members and add any link between members of the coalition. The core for cooperative games is closely related, though to define the value of a coalition, we must make a choice about how to deal with links between the coalition and those outside it. If we assume there are no links to those outside the coalition when determining its value, then Jackson and van den Nouweland (2005) show that the core is equivalent to the strongly stable set under a broad class of payoffs. Taking a different approach, Herings et al. (2009) define the farsightedly stable set, a concept that is non-nested with pairwise stability. Farsighted stability introduces additional deviations, as players compare the status quo to the endpoint of a sequence of changes, but it can also eliminate myopically beneficial deviations if anticipated future changes deter the first move. On net, this leads to a similar refinement in our setting, but as we see in Section 7.3, swap-proof stability has the advantage of being a steady state for natural learning dynamics.

The existence and uniqueness result in Section A.4 unifies and extends several findings in the matching literature. Improving swaps are analogous to blocking pairs, and strategic network formation presents a one-sided, many-to-many matching problem. However, through our choice of linking costs and benefits, we can encompass any combination of one or two-sided and one-to-one, one-to-many, or many-to-many matching. In two-sided matching, acyclic preference conditions ensure a unique stable matching in both one-to-one (Romero-Medina and Triossi, 2013) and many-to-many (Romero-Medina and Triossi, 2021) markets. The mutual favorite property extends this logic to one-sided markets—the earlier conditions imply it in the two-sided case. Gutin et al. (2022) show that a related condition is both necessary and sufficient for a unique stable matching in two-sided, one-to-one markets. Sadler (2023) generalizes this characterization to two-sided, many-to-many matching markets. Echenique and Oviedo (2006) establish the existence of stable matchings in two-sided, many-to-many markets under weak conditions. With two sides, we can take advantage of opposed interests and appeal to Tarksi's theorem for existence, so their argument applies to a broader class of payoffs. Using somewhat stronger assumptions, we get uniqueness in addition to existence, and we include one-sided markets.

2 Network Formation Games

A network formation game is a finite set of players $N = \{1, 2, ..., n\}$ together with a payoff function $u_i(G)$ for each player i. The payoff u_i takes as input a simple undirected graph G with vertex set N. Given a graph G, I write G_i for the set of i's neighbors and $d_i = |G_i|$ for player i's degree. I also write $G + E$ and $G - E$ for the graph G with the edges E respectively added and removed. Later sections provide more general results, but for now I restrict attention to the following class of games: each player i has an observable type t_i

taking values in an arbitrary set T, and payoffs take the form

$$u_i(G) = \sum_{j \in G_i} g(t_i, t_j) - c(d_i), \tag{1}$$

in which the cost c is increasing and convex. A link to j generates a benefit $g(t_i, t_j)$, and the marginal cost of a link $c(d_i + 1) - c(d_i)$ is increasing in d_i.[12]

The central question I ask is: what kinds of graphs G will the players form? The most widely used solution concept in these games is *pairwise stability*. Intuitively, a graph G is pairwise stable if no player is better off unilaterally deleting a link, and no two players are both better off forming a link between them. As a first contribution, I introduce a refinement of pairwise stability called *swap-proofness*.

Definition 1. *A graph is* ***pairwise stable*** *if there is no $ij \in G$ such that $u_i(G-ij) > u_i(G)$ and no $ij \notin G$ such that both $u_i(G+ij) \geq u_i(G)$ and $u_j(G+ij) \geq u_j(G)$ with at least one strict inequality.*

An ***improving swap*** *is a link $ij \notin G$ together with two players $k, \ell \in N$ such that both $u_i(G + ij - ik) \geq u_i(G)$ and $u_j(G + ij - j\ell) \geq u_j(G)$ with at least one strict inequality.*[13] *A pairwise stable graph G is* ***swap proof*** *if there is no improving swap, and G is then a* ***swap-proof stable*** *graph.*

A pairwise stable graph is swap-proof if no two players i and j are both better off if they form a link and each (possibly) delete a different link at the same time. If the cost function c is linear, or concave, then swap-proofness has no bite. In this case, the marginal value to i of a link with j is either constant or increasing as we add links to the graph. Hence, if an improving swap exists, the players involved are also better off simply adding the link without

[12] One can write equivalent payoff functions with concave benefits and constant costs, but this choice simplifies the presentation. Note that "types" here are public information, not private information.

[13] In this definition, I interpret $k = i$ as the case in which i simply adds the link to j without deleting any other link, and similarly $\ell = j$ means j simply adds the link to i without deleting any other link.

any deletions. However, if c is strictly convex, then pairwise stability allows unappealing outcomes: two players i and j might both benefit more from linking with one another than from existing links. For instance, imagine there are three players 1, 2, and 3, the cost of d links is $c(d) = 2d^2$, and every other player values a link to player i at $i+2$. This implies that a single link to any other player adds value, but a second link is always too costly. Hence, any graph with a single link is pairwise stable, including the one in which players 1 and 2 are linked. Intuitively, player 2 should drop her link to player 1 and form the more valuable link with player 3, and this is exactly what swap-proofness predicts—the unique swap-proof stable graph has exactly one link between players 2 and 3.

While we could imagine other deviations, I quite intentionally focus on a *minimal* refinement. As we shall see, this already yields a unique prediction in a large class of games—considering a larger set of deviations can only further refine outcomes, so our structural results still hold. One might worry about refining away existence, but at least in our setting, swap-proof stability precludes a much larger set of profitable deviations. With increasing marginal costs, pairwise stability is equivalent to pairwise Nash stability, a stronger solution concept requiring that no player can profit from unilaterally deleting any subset of her links. Moreover, since linking benefits depend on a potential partner's attributes, and not on her set of neighbors, players need not worry about what else a new neighbor might do when deciding to form a link. Hence, with payoffs of the form (1), swap-proof stability exactly captures robustness to *any* deviation that a coalition of size two could implement.

Improving swaps are analogous to blocking pairs in matching models: two players are better off dropping existing matches for one another. In fact, we can view many matching problems as special cases of network formation. To obtain a two-sided matching problem, partition types into two groups and make the benefit term $g(t_i, t_j)$ negative whenever t_i and t_j are in the same group. To impose a fixed link capacity, take $c(d) = 0$ for all sufficiently small d, and make subsequent cost increments larger than any benefit. Because I do not

insist on a partition of players into two sides (e.g., buyers and sellers, schools and students), any player can match with any other, and we have a challenge to ensure that swap-proof stable graphs exist. In two-sided matching, the two sides' implicitly opposed interests allow us to apply Tarski's theorem, yielding best and worst stable matchings for each side. In general, no such analysis is possible here. Indeed, pairwise stable graphs need not exist, and even when they do, the game may not have any that are swap-proof.

A three player example neatly illustrates the difficulties with existence. Suppose there are three players 0, 1, and 2, and the linking benefits are $g(t_1, t_0) = g(t_2, t_1) = g(t_0, t_2) = 1$, and $g(t_1, t_2) = g(t_2, t_0) = g(t_0, t_1) = 3$—player k earns 1 from linking to player $k - 1$ (mod 3) and 3 from linking to player $k + 1$ (mod 3). If the first link is free but the second costs 2, so $c(0) = c(1) = 0$ and $c(2) = 2$, then there is no pairwise stable graph. To see this, note the empty graph is not stable because any pair has an incentive to link, and any player with 2 links should unilaterally delete the less valuable one. Moreover, a graph with a single link is never stable because one of the two players is linked with her less preferred partner, so adding a link to the isolated player gains $3 - 2 = 1$.

Suppose instead that $c(2) = 4$, so a second link is always prohibitively expensive on the margin. Now any graph with a single link is pairwise stable, but any such graph has an improving swap—the player linked to her less preferred neighbor can swap this neighbor for the isolated player. Figures 1 and 2 illustrate. We clearly need some further assumptions. In the next section, I introduce the *mutual favorite property*, a novel acyclicity condition ensuring that a unique swap-proof stable graph exists.

3 The Mutual Favorite Property

The benefit $g(t_i, t_j)$ defines a preference order for each player over potential neighbors. I write $j \succeq_i k$ if $g(t_i, t_j) \geq g(t_i, t_k)$, indicating that player i prefers j as a neighbor over k.

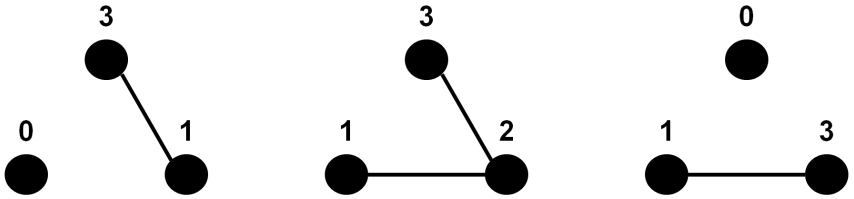

Figure 1: Payoffs with $c(2) = 2$: no pairwise stable outcome exists.

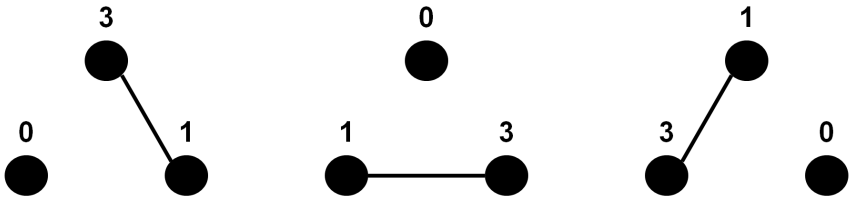

Figure 2: Payoffs with $c(2) = 4$: all graphs are pairwise stable, but all have improving swaps.

Given a set of edges E, I say that $ij \in E$ is a **mutual favorite** for E if j maximizes $\succeq_i$ among all k such that $ik \in E$, and likewise i maximizes $\succeq_j$ among all ℓ such that $j\ell \in E$. The game has the *mutual favorite property* if every set of edges contains a mutual favorite.

Definition 2. *Given a set of edges E, link $ij \in E$ is a **mutual favorite** for E if $j \succeq_i k$ for each k with $ik \in E$ and $i \succeq_j \ell$ for each ℓ with $j\ell \in E$. A network formation game has the **mutual favorite property** if every set of edges E contains a mutual favorite.*

Intuitively, the mutual favorite property is an acyclicity condition. Since each collection of edges $E = \{i_1i_2, i_2i_3, ..., i_Ki_1\}$ contains a mutual favorite, we can never have a cycle of improving swaps. Moreover, if the mutual favorite property fails, then we can find a cycle i_1i_2, i_2i_3,...,$i_Ki_{K+1} = i_Ki_1$ in which $i_{k+1} \succeq_{i_k} i_{k-1}$ for every $k = 1, 2, ..., K$.

At first glance, it might seem difficult to check whether the mutual favorite property ever holds, but at least two natural classes of payoffs always satisfy it. The first is central to our characterizations in the next section. Suppose players share a common ranking over potential neighbors—the orders $\succeq_i$ and $\succeq_j$ are the same for any i and j. Given payoffs of the form (1) and an ordered set of types, we might posit that the benefit $g(t, s)$ of a type s neighbor to a

type t player is increasing in s. To find a mutual favorite for any set of edges E, simply look for the highest type at the end of any edge in E and choose that player's most preferred edge. Alternatively, suppose payoffs take the form (1), and $g(t,s) = g(s,t)$ for all types $t, s \in T$. With symmetric benefits, the maximum over all links in E is clearly a mutual favorite. Such a structure arises if T is a metric space and the benefit from linking depends on the two players' distance from one another. Thus, we can readily cover settings in which players explicitly seek neighbors who are most similar to—or most different from—themselves.

In addition to the mutual favorite property, we need to assume there are no indifferences.

Definition 3. *Payoffs exhibit* ***no indifference*** *if* $g(t_i, t_j) \neq g(t_i, t_k)$ *for all* i, j, *and* k, *and* $u_i(G + ij) \neq u_i(G)$ *for all* i, j, *and* G.

This means that no player is ever indifferent between two potential neighbors, nor about adding a link. The following Theorem shows that the mutual favoritie property, together with no indifference, implies that a unique swap-proof stable graph exists. Moreover, the proof is constructive: we can find this graph using a greedy algorithm.

Theorem 1. *Suppose payoffs exhibit no indifference, and the game has the mutual favorite property. There exists a unique swap-proof stable graph.*

Proof. Using the mutual favorite property, order all potential links $i_1j_1, i_2j_2, i_3j_3, ...$ so that i_kj_k is a mutual favorite in the set $\{i_\ell j_\ell : \ell \geq k\}$.[14] Going in order, add link i_kj_k whenever doing so benefits both players, taking existing links as given. The resulting graph G is clearly pairwise stable as each new link benefits both players involved, and each existing link is even better than the latest addition. Any link $i_kj_k \notin G$ failed to benefit one of the two players when it was evaluated—without loss, suppose i_k preferred not adding link i_kj_k. Since costs are convex, and player i_k prefers partner j_k to all subsequent options on the list, we know

[14]The ability to order links in this way is in fact equivalent to the mutual favorite property—given such an ordering, the first link in any subset E is a mutual favorite for E.

that player i_k adds no further links. Since i_k prefers all existing partners to j_k, there can be no improving swap that adds link $i_k j_k$. We conclude that G is swap-proof.

Now suppose there are two distinct swap-proof stable graphs G and G'. Let $i_k j_k$ be the first link in our order that appears in one of the graphs but not the other—without loss, suppose $i_k j_k \in G$ and $i_k j_k \notin G'$. Write d_{i_k} and d_{j_k} for the two players' degrees in G, and write d'_{i_k} and d'_{j_k} for the two players' degrees in G'. I show that player i_k is willing to either add link $i_k j_k$ in graph G', or at least carry out a swap—a symmetric argument shows the same for player j_k. If $d_{i_k} > d'_{i_k}$, then since costs are convex, the value of a link to player j_k is higher than the marginal cost of that link in G': player i_k would willingly add link $i_k j_k$ in G'. Alternatively, if $d_{i_k} \leq d'_{i_k}$, then i_k has a neighbor j^* in G' that is not a neighbor in G. Since $i_k j_k$ is a mutual favorite for the set $\{i_\ell j_\ell : \ell \geq k\}$, and all previous links are either in both graphs or neither, we know that player i_k prefers j_k to j^*. Hence, player i_k has a neighbor she would swap for j_k. Since player j_k is similarly willing to either add link $i_k j_k$ to G', or swap another neighbor for i_k, we conclude that G' is not swap-proof.

□

Though tangential to my main goal—relating the structure of stable graphs to features of players' payoffs—Theorem 1 represents a significant contribution in its own right. Even simple network formation games often have a large multiplicity of pairwise stable graphs, but a natural refinement yields uniqueness in a wide range of settings. As I illustrate later in this section, this includes classic models of two-sided matching markets. With unique predictions, it becomes far easier to think about how to estimate linking preferences from network data. In Section 6, I show that the mutual favorite property holds in a model that incorporates noise into linking incentives, suggesting one potentially fruitful approach. In the Online Appendix, I present simple examples highlighting that no-indifference is important for *both* parts of this result, and I provide a detailed discussion of how the mutual favorite property relates to other acyclicity conditions in the matching literature.

4 Stable Network Structures

What do swap-proof stable graphs look like? This section highlights key structural features. If players have the same ranking of potential neighbors, then swap-proof stable graphs robustly exhibit homophily and clustering, and this ranking induces a clear status hierarchy. With further assumptions on players' desire for links, we obtain different architectures resembling real networks is certain contexts. If more desirable neighbors also want more links, then swap-proof stable graphs have a tiered structure similar to, but less rigid than, the core-periphery networks appearing in other models. If more desirable neighbors want fewer links, then players organize themselves into ordered cliques, potentially with some overlap.

Going forward, I assume the set of types T is linearly ordered, and the benefit $g(t, s)$ is strictly monotonic in its second argument. I say **desirability is increasing (decreasing) in type** if g is strictly increasing (decreasing) in its second argument. Since we can always reverse the type order, it is without loss to assume desirability is increasing in type. Substantively, this means that we can order players according to their attractiveness as neighbors, and every player agrees on this ordering. For convenience, I index the players in decreasing order, so $t_1 \geq t_2 \geq .... \geq t_n$, so lower indices are more desirable neighbors. Later results also require monotonicity in the first argument. I say **sociability is increasing (decreasing) in type** if g is strictly increasing (decreasing) in its first argument—if g is constant in the first argument, I say sociability is **constant in type**. Intuitively, desirability is increasing in type if higher types are more attractive neighbors, and sociability is increasing in type if higher types are more inclined to form links. As noted in the last section, if desirability is increasing in type, the game has the mutual favorite property, and Theorem 1 generically guarantees a unique swap-proof stable graph.

4.1 Homophily and Clustering

A first result bounds how far two neighbors are from one another in the type order, implying that swap-proof stable graphs entail significant homophily, particularly in large populations. Throughout this section, I assume types are distinct, and players are indexed in decreasing type order—we have $t_1 > t_2 > ... > t_n$.

Theorem 2. *Suppose payoffs take the form* (1)*, desirability is increasing in type, and G is a swap-proof stable graph. If there exists a bound $\overline{d}$ such that $1 \leq d_i \leq \overline{d}$ for each player i, then $|i-j| \leq \left\lceil \frac{\overline{d}}{2} \right\rceil \left(\left\lfloor \frac{\overline{d}}{2} \right\rfloor + 1 \right)$ for any $ij \in G$.*

Proof. I prove a slightly more general result that implies the theorem. Assume there is also a lower bound $\underline{d}$ so that $\underline{d} \leq d_i \leq \overline{d}$ for every player i, and two players $i < j$ are linked with one another—I show how to construct a bound on $|i-j|$ as a function of both $\underline{d}$ and $\overline{d}$.

I begin with a few preliminary observations. If there is a third player k with $i < k < j$ and $k \notin G_i$, then we must have $\ell < i$ for every $\ell \in G_k$—otherwise i should swap j for k, and k has a neighbor she would swap for i. Furthermore, we have $G_k \subseteq G_i$ because any $\ell \in G_k$ with $\ell \notin G_i$ would swap k for i, and i would swap j for ℓ. Finally, note that all players $\ell < i$ with a link to some $k > i$ must be linked with one another—if two were not linked, they would swap their higher indexed neighbors for each other.

Let S denote the set of players $\ell < i$ such that $k \in G_\ell$ for some k between i and j, with $k \notin G_i$, and define $r = |S|$. This group of players can form at most $r(\overline{d} - r)$ links with such players k—each member of S must spend $r-1$ links to other members of S and one additional link to player i. We necessarily have $r < \overline{d}$. Therefore, the total number of links available to players k between i and j, with $k \notin G_i$, is at most $\lfloor \frac{\overline{d}}{2} \rfloor \cdot \lceil \frac{\overline{d}}{2} \rceil$, which we can achieve taking $r = \lfloor \frac{\overline{d}}{2} \rfloor$. As r increases past this point, we decrease the number of available links, and we increase the number of players to whom i must link.

Suppose there are m_i players between i and j who are neighbors of i, and there are m_o

players between i and j, who are not neighobrs of i. The later group requires at least $m_o\underline{d}$ links, so we necessarily have

$$m_o\underline{d} \leq \left\lfloor \frac{\overline{d}}{2} \right\rfloor \cdot \left\lceil \frac{\overline{d}}{2} \right\rceil \quad \Longrightarrow \quad m_o \leq \frac{1}{\underline{d}} \cdot \left\lfloor \frac{\overline{d}}{2} \right\rfloor \cdot \left\lceil \frac{\overline{d}}{2} \right\rceil := C.$$

For each m_o below this bound, write $r(m_o)$ for the smallest value of r such that $r(\overline{d}-r) \geq m_o\underline{d}$. Player i has degree at least $m_i + r(m_o) + 1$ as there are m_i links to players between i and j, a total of $r(m_o)$ links to players $\ell < i$, and 1 link to player j. The maximum distance between i and j is therefore bounded by the solution to

$$\begin{aligned} \max_{m_i, m_o \geq 0} \quad & m_i + m_o + 1 \\ \text{s.t.} \quad & m_i + r(m_o) + 1 \leq \overline{d}. \end{aligned}$$

As long as $\underline{d} \leq \frac{\overline{d}}{2}$, the solution chooses $m_o = \lfloor C \rfloor$, $r(m_o) = \lfloor \frac{\overline{d}}{2} \rfloor$, and the optimal value is

$$m_i + m_o + 1 = \left(\overline{d} - \left\lfloor \frac{\overline{d}}{2} \right\rfloor - 1 \right) + \lfloor C \rfloor + 1 = \left\lceil \frac{\overline{d}}{2} \right\rceil + \lfloor C \rfloor.$$

Taking $\underline{d} = 1$ gives

$$\lfloor C \rfloor = \left\lfloor \frac{\overline{d}}{2} \right\rfloor \cdot \left\lceil \frac{\overline{d}}{2} \right\rceil,$$

and the bound becomes

$$m_i + m_o + 1 \leq \left\lceil \frac{\overline{d}}{2} \right\rceil \left(\left\lfloor \frac{\overline{d}}{2} \right\rfloor + 1 \right)$$

as desired. □

Theorem 2 constrains the distance between any two neighbors' types in the ordering. If $\overline{d}$ is small, the bound is very restrictive—if $\overline{d} = 1$, all neighbors are adjacent in the type order, and if $\overline{d} = 2$, neighbors' indices can differ by at most 2. As the bound grows quadratically,

it becomes much more permissive as $\overline{d}$ increases. Because the bound is independent of population size, its implications are more stark in larger groups. For instance, suppose T is a compact interval in $\mathbb{R}$, and types are i.i.d. draws from an atomless distribution on T—for n large enough, a player's neighbors come exclusively from an arbitrarily small ϵ-ball around her own type. This echoes empirical findings that racial homophily in high school friendship networks is more pronounced in larger schools (Currarini et al., 2010).

The underlying logic of this result is exactly the same as in classic models of assortative matching (Becker, 1973): people at the top of the ranking want to match with each other, leaving those below them to match amongst themselves. Theorem 2 extends this reasoning to a many-to-many matching market with only one side—any player can form a link with any other. Convex linking costs are key. If linking costs were linear, then as the population grows, each player should simply add more links. A type t player can have a type s neighbor, with s far away from t, as long as the gain $g(t,s)$ is larger than the constant marginal cost. Convexity forces players to be more discerning. If the marginal cost of a link $c(d+1)-c(d)$ grows without bound, then there must be some finite maximal degree $\overline{d}$ in any pairwise stable graph—if there exists $\overline{d}$ such that $c(\overline{d}+1)-c(\overline{d}) > g(t,s)$ for all types t and s, then no player can ever have more than $\overline{d}$ links in a stable outcome. Since having more than $\overline{d}$ neighbors is prohibitively expensive, each player seeks only the *best* neighbors who are willing to link with her. As the pool of potential neighbors expands, those with significantly higher types have better options, but a player need not look far below her own type to exhaust her budget.

Theorem 2 alone is insufficient to ensure clustering. Nevertheless, a simple argument shows that, at the top of the desirabiltiy order, we necessarily have a tightly connected core.

Proposition 1. *Suppose payoffs take the form* (1)*, desirability is increasing in type, and G is a swap-proof stable graph. Let K denote the largest integer such that $d_i \geq K-1$ for all $i \leq K$. Players $1,2,...,K$ form a clique in G.*

Proof. Suppose not. Then there exist players $i,j \leq K$ such that $ij \notin G$, and players $k,\ell > K$

such that $ik, j\ell \in G$. Adding ij to G while deleting ik and $j\ell$ is an improving swap. □

Although Proposition 1 only guarantees the existence of a single clique, the same holds within any connected component of G. Discarding the component with the most desirable players, the result applies immediately to the subgraph that remains. Moreover, the components themselves are necessarily ordered. If the component containing player 1 has m players in total, they are the m players with the highest types, assuming each has at least one neighbor—if not, we can readily construct an improving swap.[15] This outlines a general pattern: stable graphs have a strict hierarchy among their connected components, and each component contains a tightly connected core among its highest ranked members.

4.2 Strong Hierarchies and Ordered Overlapping Cliques

Previous work shows that order conditions on players' linking incentives imply strong restrictions on the kinds of network structures that can form (Sadler and Golub, 2022). These findings implicitly rely on linear (or concave) linking costs. While convex linking costs create new challenges, reviewing the structures that appear in earlier results provides a helpful reference point as we build towards a new characterization.

Definition 4. *A graph G is a **nested split graph** if we can parition the non-isolated vertices into sets $V_1, V_2, ..., V_K$ such that for each $i \in V_k$, we have*

$$G_i = \bigcup_{\ell=K+1-k}^{K} V_\ell \setminus \{i\}.$$

*A graph G consists of **ordered overlapping cliques** if we can order the vertices $\{1, 2, ..., n\}$ such that $G_i \cup \{i\}$ is an interval for each i, and the endpoints of this interval are weakly increasing in i.*

[15] If i is the lowest index not in this component, then one of the first $i-1$ players has a neighbor $j > i$ who she would swap for i, and i would swap any of her neighbors for this link.

Nested split graphs exhibit an extreme hierarchical structure—if i is in a lower partition element than j, then i's neighborhood is a strict subset of j's.[16] Assuming linear linking costs, the results of Sadler and Golub (2022) imply that, if sociability is increasing in type, pairwise stable graphs are nested split graphs. Because some vertices necessarily have a large number of neighbors in a nested split graph—those in the highest partition element link with all others—convex linking costs may preclude this structure. If sociability is decreasing in type, rather than increasing, the earlier findings predict ordered overlapping cliques. In these graphs, whenever two players i and j are linked, the set of players in between i and j form a clique. In contrast with nested split graphs, ordered overlapping cliques need not entail large neighborhoods—while this structure is consistent with a complete graph, it is also consistent with any finite bound on vertex degrees.

Because ordered overlapping cliques are consistent with low degrees, this prediction is robust to convex linking costs. In the opposite case, with sociability increasing in type, we require a more flexible class of graphs that I call *strong hierarchies*. Strong hierarchies sort players into ranked tiers much like nested split graphs, but they also permit arbitrary bounds on vertex degrees.

Definition 5. *A graph G is a* ***strong hierarchy*** *if we can order the vertices $\{1, 2, ..., n\}$ such that whenever $i < j$, we have $d_i \geq d_j$ and*

$$\max\{k \in G_i\} < \min\{\ell \in G_j \setminus \{i\} \,:\, \ell \notin G_i\}.$$

In a strong hierarchy, if i is ranked higher than j, then i has weakly more neighbors than j, and if j has a neighbor $\ell \notin G_i$, then every neighbor of i has higher rank than ℓ. Neighborhoods need not be ordered by set inclusion as long as a higher ranked vertex can match every neighbor of a lower ranked vertex with a higher ranked neighbor of its own.

[16] To be precise, this statement applies to i's self-inclusive neighborhood, meaning player i together with all of her neighbors.

Figure 3 illustrates a strong hierarchy and a comaprable nested split graph—both graphs feature 12 vertices with 4 distinct degree values. A nested split graph necessarily has one connected component, is always densely connected, and every vertex with a given degree has a symmetric network position. In contrast, strong hierarchies may have multiple connected components, can be far less dense, and feature more diverse local structures. Nevertheless, the strict ranking of vertices is still apparent—higher vertices are connected to other higher vertices, and they have more connections.

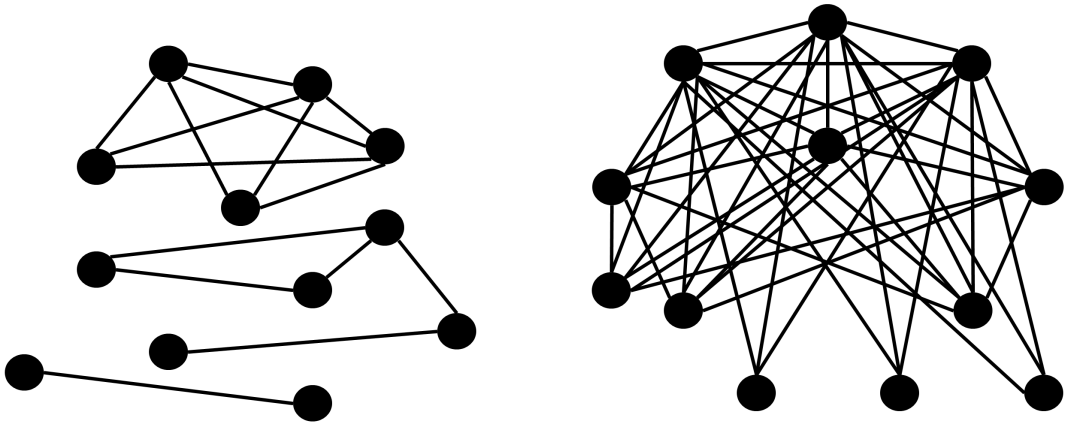

Figure 3: A strong hierarchy (left) and a nested split graph (right).

The main result in this section characterizes stable network structures when sociability is increasing or decreasing in type. In the former case, the graphs are precisely the strong hierarchies, and in the latter case, they are ordered overlapping cliques. The result is tight in that any graph within these classes can appear as a swap-proof stable graph in a network formation game satisfying the requisite order conditions.

Theorem 3. *Suppose payoffs take the form* (1)*, and desirability is increasing in type.*

(a) If sociability is increasing in type, then the unique swap-proof stable graph G is a strong hierarchy with respect to the player order. Moreover, for any strong hierarchy $\tilde{G}$, there exists a corresponding network formation game of this form such that $\tilde{G}$ is a swap-proof

stable graph.

(b) If sociability is decreasing in type, then the unique swap-proof stable graph G consists of ordered overlapping cliques with respect to the player order. Moreover, for any ordered overlapping cliques $\tilde{G}$, there exists a corresponding network formation game of this form such that $\tilde{G}$ is a swap-proof stable graph.

Proof. Suppose sociability is increasing in type. If the graph G is not a strong hierarchy with respect to the player order, then there exist players $i < j$, so $t_i > t_j$, such that either $d_j > d_i$, or there exist distinct players $k < \ell$ such that k is a neighbor of j but not i, and ℓ is a neighbor of i. In the latter case, player i would gladly swap ℓ for k, and k would gladly swap j for i, so the graph G is not swap-proof.

If $d_j > d_i$, there is some $k^* \in G_j$ with $k^* \notin G_i$. Since j finds it optimal to link with k^*, from $t_i > t_j$ we have

$$\begin{aligned} g(t_j, t_{k^*}) - (c(d_j) - c(d_j - 1)) \geq 0 &\implies g(t_i, t_{k^*}) - (c(d_j) - c(d_j - 1)) > 0 \\ &\implies g(t_i, t_{k^*}) - (c(d_i + 1) - c(d_i)) > 0, \end{aligned}$$

so player i wants to link with k^*. If k^* drops her link to j and exchanges it for the link to i, the change in utility is $g(t_{k^*}, t_i) - g(t_{k^*}, t_j) > 0$, so player k^* would gladly make this swap. Hence, the graph G is not swap-proof. We conclude that the graph G must be a strong hierarchy with respect to the player order.

Given any $\tilde{G}$ that is a strong hierarchy with respect to the player order, fix a strictly convex cost function c and construct benefits $g(t, s)$ as follows. For each degree d, let i_d denote the highest index such that player i_d has degree d in $\tilde{G}$—all players with higher indices have strictly fewer neighbors, and all with lower indices have weakly more. Let j_d denote the highest index among players in $\tilde{G}_{i_d}$, and define $g(t_{i_d}, t_{j_d}) = c(d) - c(d-1)$—the benefit is just high enough that player i_d is willing to link with j_d. For each $t \neq t_{j_d}$, define $g(t_{i_d}, t)$ to

be increasing in t, with $c(d+1)-c(d) > g(t_{i_d}, t) > c(d-1)-c(d-2)$ for all t. Then for each $i_{d+1} < i < i_d$, we can define $g(t_i, t)$ increasing in t so that $c(d)-c(d-1) > g(t_i, t) > g(t_{i_d}, t)$ whenever $t < t_{j_d}$ and $g(t_{i_{d+1}}, t) > g(t_i, t) > c(d)-c(d-1)$ whenever $t \geq t_{j_d}$—note that because $\tilde{G}$ is a strong hierarhcy, each such player i has exactly d neighbors, and these neighbors all have weakly higher types than player j_d. It is straightforward to check that the resulting benefits sustain $\tilde{G}$ as a swap-proof stable graph.

Now assume sociability is decreasing in type, and there are three players $i < j < k$ with $ik \in G$. I show that $ij \in G$—an analogous argument shows that $jk \in G$, thus establishing part (b). Given the ordering on types, we have $g(t_j, t_i) > g(t_j, t_k) > g(t_i, t_k)$ and $g(t_k, t_j) > g(t_i, t_j) > g(t_i, t_k)$. That is, players i and k find it at least as valuable to link with j as i finds it to link with k, and j finds it at least as valuable to link with k and i as i finds it to link with k. Note that i is always willing to swap k for j because j has a higher type. If $d_j < d_i$, then j benefits from linking with i because $g(t_j, t_i) > g(t_i, t_k)$ and costs are convex. It remains to show that if $d_j \geq d_i$, then j would benefit from either adding a link with i, or swapping an existing neighobr for i.

Suppose $d_j \geq d_i$ and $i \notin G_j$. Since $k \in G_i$, player j has at least one neighbor ℓ who is not a neighbor of i—if k were a neighobr of j, player j would swap k for i. If $\ell < i$, then i should swap k for ℓ, and ℓ should swap j for i—the graph cannot be swap-proof. If $\ell > i$, then j would swap ℓ for i. Hence, player j either benefits from linking with i, or has a neighbor she would swap for i. We conclude that if G is swap-proof and $ik \in G$, then we must have $ij \in G$ as well. The claim that $jk \in G$ follows from a similar argument.

Given any $\tilde{G}$ that consists of ordered overlapping cliques with respect to the player order, fix a strictly convex cost function c and construct benefits $g(t, s)$ as follows. For each player i, define $\overline{j}(i) = \max\{i, j \in \tilde{G}_i\}$, and take $g(t_i, t_{\overline{j}(i)}) = c(d_i) - c(d_i - 1)$—the benefit is just high enough so that i is willing to link with $\overline{j}(i)$. Because $\overline{j}(i)$ is weakly increasing in i, this is consistent with a benefit function g that is decreasing in its first argument and increasing

in its second, and $\tilde{G}$ is a swap-proof stable graph for any such g.

□

The strong hierarchies in case (a) are less stark than nested split graphs, but they retain a tiered structure with important behavioral implications. Strong hierarchies are a subset of the overlapping hierarchies first identified in Sadler (2022). The earlier paper shows that overlapping hierarchies are exactly those graphs that permit a robust ordering of players' actions in network games of strategic complements—all else equal, higher ranked players in the network hierarchy must take higher equilibrium actions. Hence, linking incentives that are unrelated to subsequent strategic interactions can nevertheless become an important determinant of behavior in such games. Section 7.2 discusses this in more detail.

The ordered overlapping cliques in case (b) imply a level of homophily and clustering beyond what Theorem 2 and Proposition 1 guarantee. If no player has more than $\overline{d}$ neighobrs, the structure immediately implies a more stringent bound on the distance between neighbors' types: we must have $|i - j| \leq \overline{d} + 1$ for any $ij \in G$. Moreover, any player with a neighbor at distance 2 or greater has a positive local clustering coefficient. While the qualitative prediction here is the same as with linear or concave linking costs, the underlying mechanism is distinct. With linear costs, whenever two players are linked there is a strict incentive to *add* all links to players in between. Here, the same logic implies the benefit of these links is greater than the existing one, but convex costs may preclude outright addition. Instead, we find that if such links are not part of the graph, then an improving swap must exist. If costs are linear, then every pairwise stable graph consists of ordered overlapping cliques, but that is *not* true if costs are convex—the swap-proof refinement is essential.

The contrast between convex and linear costs is particularly stark at the boundary between cases (a) and (b). If sociability is constant in type, and linking costs are linear or concave, then any pairwise stable graph is a nested split graph that consists of ordered overlapping cliques. The only compatible structure, the dominant group architecture, is ex-

tremely rigid—these graphs contain a single connected component that is a clique, possibly with some isolated vertices. With convex linking costs, a swap-proof stable graph is a strong hierarchy that consists of ordered overlapping cliques. These graphs partition vertices into potentially multiple cliques, with higher ranked vertices in weakly larger cliques.

Corollary 1. *Suppose payoffs take the form* (1)*, desirability is increasing in type, and G is a swap-proof stable graph. If sociability is constant in type, then every component in G is a clique, and higher types are in weakly larger cliques.*

Proof. Let k denote the highest index neighbor of player 1. Since the graph consists of ordered overlapping cliques, players 1 through k form a clique. Since the graph is a strong hierarchy, no member of this clique can have more than $k-1$ neighbors, so the first k players form a clique that is isolated from all other players. Iterating this argument for each subsequent component proves the claim. □

The graphs that emerge in this special case approximate friendship networks that emerge in schools (Adler and Adler, 1995; Gest et al., 2007). Homophily and clustering are especially pronounced. Corollary 1 also provides a microfoundation for group matching models, in which networks are *assumed* to have this structure, and each player chooses which clique to join.[17] Even if players could choose to interact outside their cliques, this result highlights natural conditions under which the group matching assumption is without loss.[18]

4.3 Explaining Different Networks

Our results so far highlight two dimensions along which linking incentives can vary: How quickly do marginal costs increase, and are more desirable neighbors more or less sociable? Differences in these incentives help explain different network structures. Whether

[17] See, for instance, Baccara and Yariv (2013) and Chade and Eeckhout (2018).

[18] Sadler and Golub (2022) provide an alternative microfoundation that depends on having natural divisions in the set of types.

the marginal cost of linking increases rapidly, slowly, or not at all dictates whether or not we can find individuals with very high degrees. Theorem 2 then tells us that a tighter bound translates into greater assortativity. When desirability and sociability are aligned, stable graphs display a stark hierarchy that one can identify from links alone. When desirability and sociability are opposed, status differences are less apparent from the network structure itself, and we get far more clustering.

Although the argument in Theorem 2 relies on assortative matching, this need not imply aggregate assortativity along either types or degrees. With a common ranking over potential neighbors, each player links with the highest types who are willing to reciprocate. This can lead to positive or negative assortativity depending on other payoff features. Negative assortativity arises if high types have many connections and low types have few, leading to core-periphery graphs in which low types link only with high types. As König et al. (2014) show, such graphs provide a good fit for trade networks, with large firms and countries occupying more central positions. In light of our findings, this suggests that large firms or countries are both more desirable partners and find links more profitable, and that linking costs are not very convex. In contrast, the cliques that appear in many social networks suggest more rapidly increasing marginal costs and less alignment between desirability and sociability. Taken together, our results help delineate when we should expect assortativity and when we should not, telling us what features of players' incentives are important.

5 A Hybrid Model for Large Networks

Within smaller communities—schools, neighborhoods, workplaces—it makes sense to think about people selecting precisely with whom they link, but this assumption becomes tenuous in large populations. For this reason, many researchers turn to random graph models or, more ambitiously, to search models (e.g. Currarini et al., 2009, 2010). While adept at

fitting key features of large graphs, such models offer limited explanatory power. Although I explicitly focus on smaller networks, an easy corollary of Theorem 1 opens up a new approach for studying larger ones.

Suppose we augment our players and payoffs with an exogenous set of *feasible links* F. I call the triple $(N, F, \{u_i\}_{i \in N})$ a **network formation game with link constraints**. Players form a graph using only links in F, but importantly, and in constrast with search models, they still actively choose *which links* to form. We can straightforwardly adapt our definitions of pairwise stability and swap-proof stability.

Definition 6. *In a network formation game with link constraints, a graph G is **pairwise stable** if it contains only links in F, there is no $ij \in G$ such that $u_i(G - ij) > u_i(G)$, and there is no $ij \in F$, with $ij \notin G$, such that both $u_i(G + ij) \geq u_i(G)$ and $u_j(G + ij) \geq u_j(G)$ with at least one strict inequality.*

*A pairwise stable graph G is **swap proof** if there is no improving swap that involves adding a link from F, and G is then a **swap-proof stable** graph.*

The only difference here from Definition 1 is that for link additions and swaps, we only consider links in the feasible set F. There are several ways to interpret feasible links. They might represent random meetings, with players then choosing whether or not to invest in a relationship. Alternatively, geographic proximity might determine which links are feasible (e.g., living in the same neighborhood). The feasible set could also be a design choice in some settings—students are assigned to classrooms, workers to teams, etc. In any event, the mutual favorite property, appropriately adjusted to consider only subsets of F, ensures a unique swap-proof stable graph via exactly the same argument as before.

Definition 7. *A network formation game with link constraints has the **mutual favorite property** if every subset of feasible edges $E \subseteq F$ contains a mutual favorite.*

Corollary 2. *In a network formation game with link constraints, suppose payoffs exhibit no indifference, and the game has the mutual favorite property. There exists a unique swap-proof stable graph.*

Proof. The result follows from an identical argument to that of Theorem 1, except at the start we only enumerate the links in F. □

As Corollary 2 illustrates, the existence and uniqueness of swap-proof stable graphs is robust to a wide range of added constraints. Our methods apply regardless of how feasible links arise, providing a new tool to study strategic network formation at scale. Unlike random graphs or search models, here players can direct their linking efforts towards specific other individuals, and we can readily incorporate realistic constraints through the feasible set F. Understanding how incentives and feasible links interact to determine network structure is a wide open frontier for further work. Although this problem requires its own literature to address, I can offer a few observations that should prove helpful.

6 A Model with Noise

One might worry that Theorem 1 breaks down when faced with the noisy reality of human behavior—real preferences are unlikely to strictly adhere to our order conditions. To address this concern, I present a model in which linking benefits depend on players' types as well as on noise that is idiosyncratic to each pair. These payoffs are consistent with any realized graph, and more importanty, they satisfy the mutual favorite property.

Suppose payoffs take the form

$$u_i(G) = \sum_{j \in G_i} (v_i + w_j + \epsilon_{ij}) - c(d_i), \tag{2}$$

in which $\epsilon_{ij} = \epsilon_{ji}$ is an idiosyncratic error term specific to each pair. The value v_i captures

player i's overall desire for links, the value w_i captures player i's attractiveness to others, and the error term captures match specific benefits or costs—this could depend on both individual attributes as well as random noise. Assuming distinct types, these payoffs specialize (1). In empirical applications, we would likely write v_i and w_i in terms of a coarse set of observable traits, and we would model the error terms via a parameterized distribution.

What is important from a theoretical standpoint is that the match specific term is common to both players: this gives us the mutual favorite property. If the errors ϵ_{ij} are drawn independently from a continuous distribution, then we get no indifference with probability one, and a unique swap-proof stable graph must exist.

Proposition 2. *If a network formation game has payoffs of the form* (2)*, it has the mutual favorite property.*

Proof. We proceed by contradiction. If the game does not have the mutual favorite property, then there exists a collection of edges $E = \{i_1i_2, i_2i_3, i_3i_4, ..., i_{K-1}i_K, i_Ki_1\}$ such that each player i strictly prefers the link to player $i+1$ over the link to player $i-1$ (mod K). This gives us the inequality

$$v_i + w_{i-1} + \epsilon_{i-1,i} < v_i + w_{i+1} + \epsilon_{i,i+1} \quad \Longrightarrow \quad \epsilon_{i-1,i} < w_{i+1} - w_{i-1} + \epsilon_{i,i+1}.$$

Start with $i = 1$ and proceed iteratively to obtain

$$\begin{aligned}\epsilon_{K1} &< w_2 - w_K + \epsilon_{12} < w_2 - w_K + w_3 - w_1 + \epsilon_{23} \\ &< \sum_{j=2}^{\ell} (w_j - w_{j-2}) + \epsilon_{\ell-1,\ell} < w_{K-1} - w_1 + \epsilon_{K-1,K} < \epsilon_{K1},\end{aligned}$$

in which the last line collapses the telescoping sum. Since we cannot have $\epsilon_{K1} < \epsilon_{K1}$, we conclude that no such collection E exists, and the game has the mutual favorite property.

□

Proposition 2 further illustrates the power of the mutual favorite property. Beyond its robustness to link constraints, the property persists even with a quite general form of noise in the linking benefits. This should prove helpful for empirical research that seeks to infer link preferences from network data.

7 Discussion

The analysis in this this paper touches on an unusually broad set of related questions. In this section I discuss in turn how to capture other empirical regularities, games played on the networks that form, and dynamic foundations.

7.1 Other Features of Real Networks

Particularly in large graphs, we find important regularities beyond homophily and clustering. Two of the most widespread phenomena are "small worlds" and heavy-tailed degree distributions. Small worlds refers to the short distances between typical individuals in a network.[19] Although our model is tailored for smaller graphs, we can still ask whether it would replicate this feature in a large population.

The answer clearly depends on payoffs—any graph can be stable if exactly those links are valuable to the players. However, the graphs we found in Section 4 need not feature small worlds. With a common desirability ranking, and a bound on player degrees, Theorem 2 implies that the average distance between two players grows linearly in population size. Small worlds are only possible if some players have very high degrees. However, long distances in a network are highly sensitive to noise, and a few random links quickly shrink the typical distance between a pair of vertices.[20] Looking at the model in Section 6, if the noise terms

[19] Milgram (1967) coined the term "six degrees of separation" to describe this feature. More recent work suggests distances in human social networks have become shorter over time (Dodds et al., 2003; Backstrom et al., 2012).

[20] For instance, Watts and Strogatz (1998) study a random graph model that starts from a highly clustered

are random and large enough, the stable graph will feature small worlds even if payoffs satisfy the order conditions without noise.

Heavy-tailed degree distributions are another prominent regularity in real networks—the vertices with the highest degrees typically have an order of magnitude more connections than average.[21] As with small worlds, this pattern is most stark in large networks, but it appears even in moderately sized ones. For instance, looking at the network of coauthorships among academic economists during the 1990s, Goyal et al. (2006) document that the average degree is 1.67 while the top hundred authors have an average degree over 25. We can readily generate such networks through appropriate type distributions, and at least in this example, doing so seems reasonable. Some researchers write few papers that are mostly single-authored, while others write many papers, each with multiple coauthors. The ubiquity of heavy-tailed distributions in other domains (e.g., the top few taxpayers account for a large share of overall tax revenue, a firm's top customers account for a large fraction of overall sales, etc.) suggests this phenomenon is not specific to network formation, so we should seek more broadly applicable explanations.

Nevertheless, there is a long history of trying to explain heavy tails in terms of the network formation process. The most popular approach is preferential attachment. In these dynamic models, vertices arrive one at a time and form links with those already present. If new links are biased towards well-connected vertices, then this bias reinforces itself after each new arrival, and we end up with a small number of very high degree nodes. This mechanism depends heavily on both sequential arrivals and permanent links. In contrast, predictions based on pairwise or swap-proof stability implicitly allow links to change at any

ring structure—nodes are arranged in a circle and linked to their k nearest neighbors for some k—and then rewires a small fraction of links. Their main results show that a small percentage of random links significantly reduces average distances without meaningful changes in clustering.

[21] There is an active debate around whether degree distributions follow power-laws—see Broido and Clauset (2019) and Voitalov et al. (2019) for recent contributions. Regardless of where one falls in this debate, it remains undisputed that real degree distributions have far heavier tails than what the simplest random graph models generate (e.g., Erdös-Rényi graphs).

time following strategic incentives—these solution concepts yield fixed points for natural adjustment dynamics (see Section 7.3). While each approach will have more or less appeal depending on context, strategic models can capture a similar effect using different incentives. If the benefit from linking with neighbor i is larger when i has more neighbors, then stable graphs should resemble preferential attachment networks. I explicitly rule out such link externalities here, but exploring other payoff specifications is a natural and important next step for this research agenda.

7.2 Reinforcing Inequality

If higher types are both more desirable and more sociable, the ensuing hierarchical networks tend to reproduce the type ordering in unrelated interactions. Imagine, for instance, that extraverted people are both more desired as friends and make more effort to form friendships. Separately, suppose people enjoy community sports leagues more if their friends join—athletic recreation exhibits complementarities. Even if introverts and extraverts have similar preferences over athletic activities, the social networks that form will induce more extraverts to participate, leading to inequality in health outcomes. Similar effects are likely in other settings like education, career choice, or migration.

To formalize this idea, suppose players form a strong hierarchy due to exogenous characteristics, and they subsequently play a network game of strategic complements. In the network game, actions take values in a compact set $S \subseteq \mathbb{R}$, and player i earns a payoff

$$u_i(\mathbf{s}) = v(s_i) + \sum_{j \in G_i} g(s_i, s_j),$$

in which $g \geq 0$ is twice continuously differentiable, is increasing in its second argument, and has a positive cross partial. Fixing the graph, such games always have minimal and maximal Nash equilibria. Moreover, symmetry ensures that any differences in equilibrium actions and

payoffs arise due to players' positions in the graph, not from idiosyncratic preferences.

Sadler (2022) introduces "weak centrality," a measure that robustly predicts the order of equilibrium actions in network games of strategic complements. In general, this centrality measure only partially orders the vertices of a graph, but the order is total in overlapping hierarchies—I reproduce the definition here.

Definition 8. *Given a graph $G = (V, E)$ and an ordered partition $\mathcal{P} = \{V_1, V_2, ..., V_K\}$ of V, the subset $S \subseteq V$ dominates $S' \subseteq V$ with respect to $\mathcal{P}$ if for each $k = 1, 2, ..., K$, we have*

$$\left|S \cap \left(\bigcup_{\ell=k}^{K} V_\ell\right)\right| \geq \left|S' \cap \left(\bigcup_{\ell=k}^{K} V_\ell\right)\right|.$$

*That is, subset S contains at least as many vertices in set V_k or higher for every k. The graph G is an **overlapping hierarchy** if we can find a partition $\mathcal{P}$ such that $G_i \cup \{i\}$ dominates $G_j \cup \{j\}$ with respect to $\mathcal{P}$ whenever i is in a higher partition element than j.*

A strong hierarchy is an overlapping hierarchy in which the corresponding partition $\mathcal{P}$ is the collection of all singletons. Hence, weak centrality yields a total order on the set of players that aligns with the hierarchy. Theorem 3 from Sadler (2022) then implies that, in the highest and lowest Nash equilibria, players who are higher in the ranking *must* take higher equilibrium actions. Since neighbors' actions exert positive externalities, players who are higher in the ranking also enjoy higher equilibrium payoffs. This finding highlights one reason why hierarchies in one domain reproduce themselves in others. By identifying incentives that bring about these networks, Theorem 3 in this paper can help pinpoint conditions under which endogenous networks reinforce inequality.

7.3 Dynamic Foundations

When the mutual favorite property holds, swap-proof stability enjoys a particularly robust learning foundation. In this section, I formalize two classes of better-response dynamics, one

in which opportunities to change particular links arise over time, and one in which players can periodically adjust their links. In either case, the mutual favorite property ensures convergence to the unique swap-proof stable graph with probability one.

In the **link adjustment dynamics**, we draw a pair ij at random in each discrete period t according to an arbitrary distribution with full support. If G_t is the graph at the start of the period, and $ij \in G_t$, then $G_{t+1} = G_t - ij$ if either $u_i(G_t - ij) > u_i(G_t)$ or $u_j(G_t - ij) > u_j(G_t)$—we remove ij if doing so benefits either of the two players. If $ij \notin G_t$, we add it to the graph if doing so benefits both i and j, at least one strictly, allowing either player to sever one other link if this is necessary to realize a benefit.[22] Under these dynamics, we get convergence to the unique swap-proof stable graph from any starting point.

Proposition 3. *Suppose a network formation game has payoffs of the form* (1) *that satisfy the mutual favorite property and no indifference. Given any initial graph G_0, the link adjustment dynamics always converge to the unique swap-proof stable graph.*

Proof. Under our assumptions, Theorem 1 applies, so there is a unique swap-proof stable graph, and by definition this is the only absorbing state of the Markov chain induced through the link adjustment dynamics. Hence, we need only show that there is a positive probability of getting to this state from any starting point. In the constructive proof of Theorem 1, we enumerated the possible links $i_1j_1, i_2j_2, ...$ so that i_kj_k was a mutual favorite among all subsequent links. Starting from any G_0, there is some positive probability that links are selected in exactly this order. Following the construction in the earlier result, it should be clear that, when going in this order, each link will be added to the graph if and only if it is in the unique swap-proof stable graph. When we consider link i_kj_k, the only difference from our original construction is that the graph may already contain some later links from the list, but since i_kj_k is a mutual favorite in this set, if adding it is beneficial without those later

[22] More precisely, if $u_i(G_t + ij) \geq u_i(G_t)$, player i does not sever another link—player i severs a link to the third player k only if $u_i(G_t + ij) < u_i(G_t)$ but $u_i(G_t + ij - ik) \geq u_i(G_t)$. If there are multiple such k, we make an arbitrary selection of which link to sever.

links, we can still add it through an improving swap that deletes some later links. Hence, there is a positive probability ϵ of getting to the absorbing state from any starting point, and after $K^{\frac{n(n-1)}{2}}$ periods, the probability that we are not yet at the absorbing state is no more than $(1-\epsilon)^K$. □

If players choose which links to adjust, then some links may not have any chance to change in a given period because there are better improvements available to the players involved. Nevertheless, a similar argument shows that the following **player adjustment dynamics** always converge. Instead of choosing a link in each period, we now choose a player i in each period t, drawn from an arbitrary distribution with full support. This player makes the most profitable adjustment available to her: player i either severs a link, proposes a link to another player j, or proposes a swap that adds a link to j and severs her link with k. If i proposes either an addition or a swap to j, we assume that j accepts as long as she weakly benefits—either from adding the link or from swapping a neighbor of her own—and we assume that i only makes proposals to players who will accept.

Proposition 4. *Suppose a network formation game has payoffs of the form* (1) *that satisfy the mutual favorite property and no indifference. Given any initial graph G_0, the player adjustment dynamics always converge to the unique swap-proof stable graph.*

Proof. Again, the unique swap-proof stable graph is the only absorbing state for these dynamics, so we need only show that we have a positive probability of getting there from any starting point. Enumerating the possible links $i_1j_1, i_2j_2, ...$ as before, we can see that if the graph G_t ever matches the swap-proof stable graph G^* on the first K links, it will continue to match on the first K links in all subsequent periods—if G_t matches G^* on the first K links, then by construction there is no way for the players to benefit from adding link i_kj_k for $k \leq K$ if this link is not in the graph, and if a player i could benefit from deleting one of these links, this is only because she has later links that would be more beneficial to delete.

Moreover, after selecting players i_{K+1} and j_{K+1} enough times, the graph G_t must eventually match G^* on link $K+1$—if this link is missing in G^*, one of the two players will delete it after deleting all links that appear later in the list, and if this link is present in G^* it will be the first link the two players add after exhausting more profitiable deletions. Hence, the graph G_t must eventually match G^* on all links.

□

8 Final Remarks

The literature on network formation thus far lacks a systematic treatment of how different payoff assumptions affect stable graphs. This paper's analysis based on separable benefits and ordinal rankings is a necessary first step. Even in this simple setting, a small refinement of pairwise stability can explain several stylized facts—homophily and clustering reliably appear—and we obtain sharp predictions about network structures under interpretable conditions. Beyond these specific conditions, the mutual favorite property offers a powerful tool. We get unique predictions in a large family of games covering many natural examples, and this result is robust to variations on the basic model. Moreover, swap-proofness highlights the neglected connection between network formation and matching, suggesting a more unified approach.

The limitations of my analysis present opportunities for future authors. Allowing externalities across links is clearly important in many settings—for instance, friends socialize in groups, which might make it easier to maintain relationships supported through mutual friends—and we might hope to find conditions other than separability that permit a clean characterization of stable graphs. A theorist might wonder exactly how tight the existence and uniqueness result is, and how far it applies beyond separable payoffs. As already noted in Section 5, combining strategic network formation with constraints on what links are fea-

sible may allow us to apply our tools to a much wider range of settings. Finding ways to infer preferences from network is data is important for practical applications, and the model in Section 6 suggests a place to start. I am hopeful that at least some of these directions will pique others' curiosity.

References

Adler, Patricia A and Peter Adler (1995), "Dynamics of inclusion and exclusion in preadolescent cliques." *Social Psychology Quarterly*, 58, 145–162.

Akerman, Anders and Anna Seim (2014), "The Global Arms Trade Network 1950–2007." *Journal of Comparative Economics*, 42, 535–551.

Baccara, Mariagiovanna and Leeat Yariv (2013), "Homophily in Peer Groups." *American Economic Journal: Microeconomics*, 5, 69–96.

Backstrom, Lars, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna (2012), "Four Degrees of Separation." In *Proceedings of the 4th ACM Web Science Conference*, 33–42.

Bala, Venkatesh and Sanjeev Goyal (2000), "A Noncooperative Model of Network Formation." *Econometrica*, 68, 1181–1229.

Barabási, A. and R. Albert (2002), "Statistical mechanics of complex networks." *Reviews of Modern Physics*, 74, 47–97.

Becker, Gary (1973), "A Theory of Marriage: Part I." *Journal of Political Economy*, 81, 813–846.

Belleflamme, Paul and Francis Bloch (2004), "Market Sharing Agreements and Collusive Networks." *International Economic Review*, 45, 387–411.

Breza, Emily, Arun Chandrasekhar, Tyler McCormick, and Mengjie Pan (2020), "Using Aggregated Relational Data to Feasibly Identify Network structure without Network Data." *American Economic Review*, 110, 2454–2484.

Broido, Anna and Aaron Clauset (2019), "Scale-Free Networks are Rare." *Nature Communications*, 10, 1017.

Chade, Hector and Jan Eeckhout (2018), "Homophily in Peer Groups." *Theoretical Economics*, 13, 377–414.

Chakrabarti, Subhadip and Robert Gilles (2007), "Network Potentials." *Review of Economic Design*, 11, 13–52.

Chandrasekhar, Arun and Matthew Jackson (2021), "A Network Formation Model Based on Subgraphs." Working Paper.

Craig, Ben and Goetz von Peter (2014), "Interbank Tiering and Money Center Banks." *Journal of Financial Intermediation*, 23, 322–347.

Currarini, S., Matthew Jackson, and Pablo Pin (2010), "Identifying the Roles of Race-Based Choice and Chance in High School Friendship Network Formation." *Proceedings of the National Academy of Sciences*, 107, 4857–4861.

Currarini, Sergio, Matthew Jackson, and Pablo Pin (2009), "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica*, 77, 1003–1045.

Dodds, Peter Sheridan, Roby Muhamad, and Duncan Watts (2003), "An Experimental Study of Search in Global Social Networks." *Science*, 301, 827–829.

Echenique, Federico and Jorge Oviedo (2006), "A Theory of Stability in Many-to-Many Matching Markets." *Theoretical Economics*, 1, 233–273.

Furusawa, Taiji and Hideo Konishi (2007), "Free Trade Networks." *Journal of International Economics*, 72, 310–335.

Galeotti, Andrea and Sanjeev Goyal (2010), "The Law of the Few." *American Economic Review*, 100, 1468–1492.

Gest, Scott D, Alice J Davidson, Kelly L Rulison, James Moody, and Janet A Welsh (2007), "Features of groups and status hierarchies in girls' and boys' early adolescent peer networks." *New Directions for Child and Adolescent Development*, 2007, 43–60.

Goyal, Sanjeev and Sumit Joshi (2006), "Unequal Connections." *International Journal of Game Theory*, 34, 319–349.

Goyal, Sanjeev, Marco van der Leij, and José Luis Moraga-González (2006), "Economics: An Emerging Small World." *Journal of Political Economy*, 114, 403–412.

Goyal, Sanjeev and Fernando Vega-Redondo (2007), "Structural Holes in Networks." *Journal of Economic Theory*, 137, 460–492.

Gutin, Gregory, Philip Neary, and Anders Yeo (2022), "Unique Stable Matchings." Working Paper.

Hellmann, Tim (2013), "On the Existence and Uniqueness of Pairwise Stable Networks." *International Journal of Game Theory*, 42, 211–237.

Hellmann, Tim (2021), "Pairwise Stable Networks in Homogeneous Societies with Weak Link Externalities." *European Journal of Operational Research*, 291, 1164–1179.

Herings, Jean-Jacques, Ana Mauleon, and Vincent Vannetelbosch (2009), "Farsightedly Stable Networks." *Games and Economic Behavior*, 67, 526–541.

Herskovic, Bernard and João Ramos (2020), "Acquiring Information through Peers." *American Economic Review*, 110, 2128–2152.

Homans, George (1950), *The Human Group*. Harcourt, New York.

Jackson, Matthew and Brian Rogers (2007), "Meeting Strangers and Friends of Friends: How Random are Social Networks?" *American Economic Review*, 97.

Jackson, Matthew and Anne van den Nouweland (2005), "Strongly Stable Networks." *Games and Economic Behavior*, 51, 420–444.

Jackson, Matthew and Alison Watts (2001), "The Existence of Pairwise Stable Networks." *Seoul Journal of Economics*, 14, 299–322.

Jackson, Matthew and Asher Wolinsky (1996), "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory*, 71, 44–74.

Johnson, Cathleen and Robert Gilles (2000), "Spatial Social Networks." *Review of Economic Design*, 5, 273–299.

Joshi, Sumit, Ahmed Saber Mahmud, and Sudipta Sarangi (2020), "Network Formation with Multigraphs and Strategic Complementarities." *Journal of Economic Theory*, 188, 105033.

König, Michael, Caludio Tessone, and Yves Zenou (2014), "Nestedness in Networks: A Theoretical Model and Some Applications." *Theoretical Economics*, 9, 695–752.

McPherson, M., L. Smith-Lovin, and J. Cook (2001), "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*, 27, 415–444.

Milgram, Stanley (1967), "The Small-World Problem." *Psychology Today*, 1, 61–67.

Myers, Seth, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin (2014), "Information Network or Social Network? The Structure of the Twitter Follow Graph." In *Proceedings of the 23rd International Conference on World Wide Web*, 493–498.

Romero-Medina, Antonio and Matteo Triossi (2013), "Acyclicity and Singleton Cores in Matching Markets." *Economics Letters*, 118, 237–239.

Romero-Medina, Antonio and Matteo Triossi (2021), "Two-Sided Strategy-Proofness in Many-to-Many Matching Markets." *International Journal of Game Theory*, 50, 105–118.

Sadler, Evan (2022), "Ordinal Centrality." *Journal of Political Economy*, 130, 926–955.

Sadler, Evan (2023), "A Unified Approach to Strategic Network Formation and Classical Matching Theory." Working Paper.

Sadler, Evan and Ben Golub (2022), "Games on Endogenous Networks." Working Paper.

Soramaki, Kimmo, Morten Bech, Jeffrey Arnold, Robert Glass, and Walter Beyeler (2007), "The Topology of Interbank Payment Flows." *Physica A: Statistical Mechanics and Its Applications*, 379, 317–333.

Ugander, J., B. Karrer, L. Backstrom, and C. Marlow (2011), "The Anatomy of the Facebook Social Graph." Working Paper.

van der Hofstad, Remco (2017), *Random Graphs and Complex Networks*. Cambridge University Press.

Voitalov, Ivan, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov (2019), "Scale-Free Networks Well Done." *Physical Review Research*, 1, 033034.

Watts, Duncan and S. Strogatz (1998), "Collective dynamics of small-world networks." *Nature*, 393, 440–442.

A Online Appendix

A.1 The Importance of No Indifference

While symmetry often simplifies network formation models, the opposite is true here. The no-indifference condition, which implies distinct types and strict preferences, is essential for *both* parts of Theorem 1. Suppose we have 3 identical players and every link yields one unit of benefit to each player. If a single link is always free, but a second link costs 2 on the margin, then no swap-proof stable graph exists. To see why, first note that no graph with two or more links is pairwise stable since any player with two links has a strict incentive to delete one. Similarly, the empty graph cannot be stable since any pair has a strict incentive to form a link. A graph with a single link is pairwise stable, but it cannot be swap-proof: one of the two linked players can always drop her neighbor and add a link to the isolated player. The player who severs a link is indifferent, but the formerly isolated player strictly gains, so this is an improving swap. Figure 4 illustrates.

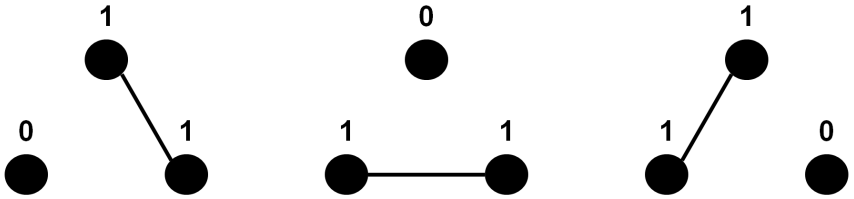

Figure 4: Payoffs for three identical players with $g(t,t) = 1$ and $c(0) = c(1) = 0 < 2 = c(2)$. Each graph can be obtained from the other two via an improving swap.

Even if existence is not an issue, we need strict preferences for uniqueness. Looking at the same game with 4 identical players (see Figure 5), we find that any graph in which players are linked in two pairs is pairwise stable and swap-proof. Indifferences create multiplicity.

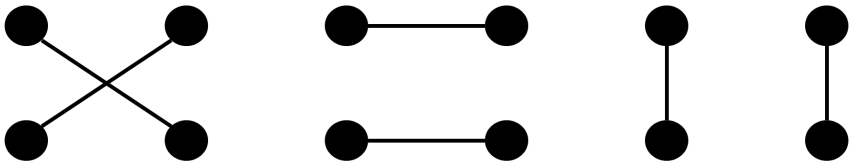

Figure 5: Four identical players with $g(t,t) = 1$ and $c(1) = 0 < 2 = c(2)$. All three graphs are pairwise stable and swap-proof.

A.2 Relationship to Other Acyclicity Conditions

The matching literature has identified acyclicity condtions that guarantee unique stable matchings in two-sided markets. In a marriage market, Romero-Medina and Triossi (2013) show that the absence of "simultaneous cycles" ensures a unique stable matching. A simultaneous cycle is a set of men $m_1, m_2, ..., m_K$ together with a set of women $w_1, w_2, ..., w_K$ such that each man k prefers woman k over woman $k-1$, and each woman k prefers man $k+1$ over man k (mod K). To represent a two-sided market in our setting, partition types into two sets, and make links between types in the same set have negative value. If we adjust Definition 2 to require a mutual favorite only in sets E such that every edge includes one player from each side of the market, our condition is equivalent to no simultaneous cycles.

The analogous condition for a one-sided market forbids cycles of players i_1, i_2, ..., $i_K = i_1$ such that each player i_k prefers linking with i_{k+1} over i_{k-1} (mod K). This is exactly the mutual favorite property. If such a cycle exists, then the corresponding set of links fails to include a mutual favorite. Conversely, if the mutual favorite property fails for a set of edges E, we can readily construct such a cycle from edges in E.

Romero-Medina and Triossi (2021) extend their own work to two-sided, many-to-many matching markets. In particular, they show that if one side of the market has acyclic preferences over the other side, then there is a unique stable matching. Partitioning players into firms $f_1, f_2, ...$ and workers $w_1, w_2, ...$, they require that there is no cycle of workers

$w_1, w_2, ..., w_K, w_{K+1} = w_1$, together with a corresponding set of firms $f_1, f_2, ..., f_K$, such that firm k always prefers worker $k+1$ (mod K) over worker k. If we again modify the mutual favorite property, applying it only to sets of edges that link firms to workers, then this one-sided acyclicity implies the mutual favorite property.

To see why, take any set of edges E that link firms to workers, and we can find a mutual favorite as follows. Pick any firm f_1 at the end of an edge in E, and find the worker w_1 that f_1 prefers most among all options within the set of edges E. That is, worker w_1 maximizes the firm's preferences among the set $\{w : f_1 w \in E\}$. If f_1 similarly maximizes the preferences of w_1, then we are done. If not, write f_2 for the firm that similarly maximizes the preferences of w_1. Again, if w_1 is the favorite of f_2 among options in E, then we are done, and otherwise we keep going. At some point, the sequence $f_1, w_1, f_2, w_2, ...$ must loop back on itself. If the length of this cycle is greater than 2, then we have a cycle that violates the one-sided acyclicity condition. The only remaining possibility is that we have found a mutual favorite.

A.3 Games on Endogenous Networks

The structural results from Section 4 easily adapt to network games with network formation, in which players simultaneously form a network and take strategic actions, and payoffs hinge on the interaction between the two (Sadler and Golub, 2022). Consider a finite set of players N, a set of actions S_i for each player $i \in N$, and a payoff function $u_i(G, \mathbf{s})$ for each player $i \in N$. Payoffs take as input both a graph G with vertex set N and a profile of actions $\mathbf{s} \in \prod_{i \in N} S_i$ for the players. Following the earlier paper, an outcome $(G, \mathbf{s})$ is pairwise stable if $\mathbf{s}$ is a Nash equilibrium, holding G fixed, and G is pairwise stable, holding $\mathbf{s}$ fixed. Applying definition 1 from the present paper, I call a pairwise stable outcome $(G, \mathbf{s})$ swap proof if G is swap proof, holding $\mathbf{s}$ fixed.

Suppose a network game with network formation has payoffs

$$u_i(G, \mathbf{s}) = v_i(\mathbf{s}) + \sum_{j \in G_i} g(s_i, s_j) - c(d_i), \tag{3}$$

in which c is increasing and convex. The term $v_i(\mathbf{s})$ describes idiosyncratic action incentives, and the common g determines linking incentives. Ignoring the first term, this is exactly the payoffs (1), with actions replacing types, and our earlier results apply without modification.

The desirability and sociability orders now have natural interpretations. Following Sadler and Golub (2022), I say the game has *positive (negative) spillovers* if higher actions make for more (less) attractive neighbors—that is, if g is increasing (decreasing) in its second argument. Similarly, the game has *action–link complements (substitutes)* if taking a higher action makes a player more (less) inclined to form links—that is, if g is increasing (decreasing) in its first argument. Positive spillovers and link–action complements, or negative spillovers and link–action substitutes, imply that a swap-proof stable outcome involves a strong hiearchy. In the other two cases, stability entails ordered overlapping cliques.

While Theorem 1 yields a unique swap-proof stable graph given a generic action profile, ensuring the existence of a swap-proof stable outcome $(G, \mathbf{s})$ is more delicate for two reasons. First, the interplay between actions and links could produce an improving cycle (e.g., a change in actions creates a profitable deviation in links, which in turn creates a profitable deviation in actions, etc.). Second, multiple players might take the same equilibrium action, creating troublesome indifferences. We can likely manage the latter through assumptions on idiosyncratic incentives, or a small adjustment to the solution concept, but the first issue demands an in-depth treatment.

A.4 Primitive Payoff Properties

While expositionally convenient, the payoffs (1) assume more than we need for the main findings to hold. I present two new definitions to capture the essence of increasing marginal costs and bilateral linking incentives. Analogs of each of our Theorems then follow from essentially the same arguments. Write

$$\Delta_{ij} u_i(G) = u_i(G + ij) - u_i(G - ij)$$

for player i's marginal value of a link with player j—note that $\Delta_{ij} u_i(G) = \Delta_{ij} u_i(G + ij) = \Delta_{ij} u_i(G - ij)$. In the following, I write a graph G as (G_i, G_{-i}) to distinguish the neighborhood G_i of player i from the collection of all edges G_{-i} that do not involve player i.

Definition 9. *In a network formation game, payoffs exhibit* ***no externalities*** *if $u_i(G_i, G_{-i})$ does not depend on G_{-i}. Payoffs are* ***quasi-concave in own links*** *if*

$$\Delta_{ij} u_i(G_i, G_{-i}) \geq (>) 0 \quad \Longrightarrow \quad \Delta_{ij} u_i(G_i', G_{-i}) \geq (>) 0$$

whenever $G_i' \subseteq G_i$. Payoffs are ***rank-consistent*** *if for every player i, whenever there exists a pair j, k, and a graph G with $j \notin G_i$ and $k \in G_i$, such that*

$$u_i(G + ij - ik) > u_i(G),$$

we also have

$$u_i(G' + ij - ik) > u_i(G')$$

for every other graph G' with $j \notin G_i'$ and $k \in G_i'$.

No externalities means that a player's payoff depends only on the links she has, not on other links in the network. Concavity in own-links says that additional links have decreasing

marginal returns.[23] Rank-consistency ensures that a player's ranking of potential partners does not change with the graph. Together with no externalities, this means that linking benefits are fundamentally bilateral: the value of a link to each player depends only on attributes of those two players.[24] The payoffs (1) clearly satisfy both conditions, but other natural formulations fit as well.[25] Throughout this subsection, I assume payoffs are quasi-concave in own-links and self-consistent.

Rank-consistency implies that each player i has a fixed preference order $\succeq_i$ over potential neighbors—I write $j \succ_i k$ if we ever have $u_i(G + ij - ik) > u_i(G)$ for some G with $j \notin G_i$ and $k \in G_i$.[26] Following the analysis of Section 3, I say that $ij \in E$ is a **mutual favorite** for E if j maximizes $\succeq_i$ among all k such that $ik \in E$, and likewise i maximizes $\succeq_j$ among all ℓ such that $j\ell \in E$. As before, the game has the mutual favorite property if every set of edges E contains a mutual favorite. As in Section 3, the mutual favorite property together with no indifference ensures the existence of a unique swap-proof stable graph.[27]

Proposition 5. *Suppose payoffs exhibit no externalities, rank-consistency, and quasi-concavity in own links. If there are no indifferences, and the game has the mutual favorite property, then there exists a unique swap-proof stable graph.*

Proof. The proof is substantively identical to that given for Theorem 1, and I omit it. □

We can similarly extend Theorems 2 and 3 using analogous order conditions.

[23] An equivalent definition appears in Hellmann (2013).

[24] This does exclude some classic examples like the "connections model" and the "coauthor model" of Jackson and Wolinsky (1996), in which linking incentives explicitly depend on more complex interactions.

[25] e.g., benefits that are a concave function of a linear aggregate together with either linear or convex costs.

[26] Note the definition precludes having both $j \succ_i k$ and $k \succ_i j$, and we can write $j \sim_i k$ if neither of these holds. Moreover, it should be clear that this relation is transitive—given a cycle $j_1 \succ_i j_2 \succ_i ... \succ_i j_K \succ_i j_1$, consider the graph that is empty except for one link ij_k, and a cycle of improving swaps then implies that $u_i(G) > u_i(G)$ for some graph, a contradiction.

[27] No indifference here means that $\Delta_{ij} u_i(G) \neq 0$ for all i, j, and G, and that $u_i(G + ij - ik) \neq u_i(G)$ for all i, j, k and all graphs G with $j \notin G_i$ and $k \in G_i$.

Definition 10. *Payoffs are* ***consistent*** *if, whenever there exists three players i, k, and ℓ, and a graph G with $k \notin G_i$ and $\ell \in G_i$ such that*

$$u_i(G + ik - i\ell) > u_i(G)$$

then we also have

$$u_j(G' + jk - j\ell) > u_j(G')$$

for any player j and any graph G' with $k \notin G'$ and $\ell \in G'$. That is, if any player ever benefits from swapping k for ℓ, then all players always benefit from swapping k for ℓ, and we say k is ***more desirable*** *than ℓ.*

Consistency ensures that players share a common ranking of who is a desirable neighbor—the orders $\succeq_i$ are identical for all players. Under this assumption, the conclusion of Theorem 2 holds via the same argument: there is a bound on the distance between any pair of neighbors, and this bound is invariant to the population size. Although the argument is the same, this meaningfully extends our earlier findings. The payoffs (1) require every player to have the same convex cost function, but linking costs that vary across players would still result in payoffs that satisfy definition 10.

Proposition 6. *Suppose payoffs exhibit no externalities, no indifference, consistency, and quasi-concavity in own links, and G is a swap-proof stable graph. If there exists a bound $\overline{d}$ such that $1 \leq d_i \leq \overline{d}$ for each player i, then $|i - j| \leq \left\lceil \frac{\overline{d}}{2} \right\rceil \left(\left\lfloor \frac{\overline{d}}{2} \right\rfloor + 1 \right)$ for any $ij \in G$.*

Proof. The argument is substantively identical to that of Theorem 2, and I omit it. □

The structures that Theorem 3 predicts require the strongest assumptions, depending on both consistent payoffs and a sociability order. The following definition provides the last ingredient we need to state our extension.

Definition 11. *Player i is **more sociable** than player j if for any third player k, and any graph G with $jk, ik \notin G$, we have*

$$\Delta_{jk} u_j(G) \geq (>) 0 \quad \implies \quad \Delta_{ik} u_i(G) \geq (>) 0$$

whenever $d_i \leq d_j$.

Proposition 7. *Suppose payoffs exhibit no externalities, no indifference, consistency, and quasi-concavity in own links, and G is a swap-proof stable graph. If i is more sociable than j whenever i is more desirable than j, then G is a strong hierarchy. If j is more sociable than i whenever i is more desirable than j, then G consists of ordered overlapping cliques.*

Proof. This argument is substantively identical to that of Theorem 3, and I omit it. □