Social groups and social network formation[☆]

Bassel Tarbush[a,*], Alexander Teytelboym[b]

[a] *Department of Economics and Merton College, University of Oxford*
[b] *Institute for New Economic Thinking, University of Oxford*

Abstract

We present a dynamic model of social network formation in which a fixed number of agents interact in overlapping social groups. We derive several results on the formation of links in such networks, including results on the degree distribution, on comparative statics relating degree and group size, and on the dynamics of homophily. In particular, we derive comparative statics showing that degree is typically positively related to social group size but negatively related to the size of the overlap across multiple social groups. This is supported by evidence from a Facebook dataset. We also show that homophily over an agent's lifespan in the network can be non-monotonic, reaching a global maximum in some period before eventually decreasing.

Keywords: social groups, dynamic network formation, social networks, homophily
JEL: D85, A14, Z13

1. Introduction

Friendships are an essential part of economic life. Friendships result in peer effects, which impact educational performance (Sacerdote, 2001), health (Kremer and Levy, 2008), group lending (Banerjee et al., 2012), and productivity at work (Falk and Ichino, 2006). The structure of friendships can be described by a social network.[1] How friendships form is key to understanding the properties of social networks and "one of the most important areas of network research [is] developing richer, but still tractable, models of network formation." (Jackson, 2014, p. 17).

[☆] An earlier version of this paper was circulated under the title "Friending".

[*] Corresponding author
Email addresses: bassel.tarbush@economics.ox.ac.uk (Bassel Tarbush), alexander.teytelboym@inet.ox.ac.uk (Alexander Teytelboym)

[1] The best recent summaries of applications of networks in the social sciences are by Jackson (2008), Goyal (2009), Easley and Kleinberg (2010), and Newman (2010).

The theoretical model presented in this paper is a new dynamic network formation process in which a fixed number of agents interact in overlapping social groups. In every period, every agent interacts with others with a probability that depends on mutual social group sizes and on the size of their overlaps (as well as on a set of network-level parameters). When interacting with others in a social group, an agent forms a friendship with another agent chosen at random from among those in the group who are not yet his friends. An example we have in mind is the formation of friendships among college students. A college freshman interacts with students in his class and his dorm. The sizes of these two social groups and the size of their overlap (number of students who are both in his class and in his dorm) determine his chance of becoming friends with students either from his class or his dorm or both.

In this paper, we derive several properties of social networks that arise from our network formation process, including results on the degree distribution, on comparative statics relating degree and group size, and on the dynamics of degree and of homophily – the propensity of agents to be friends with others who are similar to themselves.

Our dynamic network formation process has two key features: a fixed number of agents and overlapping social groups. The fixed number of agents in our model drives a number of features of the resulting link formation.[2] For example, we find that agents make friends at a decreasing rate over time because they gradually exhaust the pool of potential friends in each social group. Let us return to our college freshman: after numerous interactions with students in his dorm, he will have become friends with everyone from that social group, and although he continues to interact with students in his dorm, he no longer spends this time making *new* friends, which thus reduces his overall rate of friendship formation. This explanation for the concavity of degree over time contrasts with those given in the literature. In a model with an infinite population of agents who are matched according to probabilities that depend on their types, Currarini et al. (2009) offer one alternative: agents have a decreasing marginal utility of friendships and stop making friends when the marginal cost of a friend exceeds the marginal benefit. The concavity of degree over time is also a feature of many *growing* random network models, in which new agents enter in each period and form links with pre-existing agents, who are chosen according to a specific stochastic process, which may depend on the number of links of the pre-existing agents (Barabási and Albert, 1999, Jackson and Rogers, 2007) or on their characteristics (Bramoullé et al., 2012). In these models, agents make friends

[2] Watts and Strogatz (1998) examine dynamics on a network of fixed size but their linking process does not depend on agent characteristics, as it does here.

at a decreasing rate because they are less likely to receive a link from the incoming agent as the population grows over time. This paradigm is well suited for the analysis of social networks in which the growth of the network is important in capturing features of link formation, whereas focusing on a fixed network size, as we do here, may better capture situations in which there is relatively low volatility in the network growth relative to the rate of link formation.

Our second key feature – overlapping social groups – allows us to derive nuanced comparative statics on the relationship between an agent's degree and the size of the social groups that the agent belongs to. For example, if our freshman is studying for a degree in economics and is resident of a particular dorm, then one might ask what the effect of being in a larger dorm would be, *ceteris paribus*, on his number of friends. The problem is that the ceteris is not paribus in a network (or college) of fixed size. If the freshman's dorm is now larger, then the size of some other social group must have changed for the total number of students to remain unchanged. In other words, one must keep track of what new students joined the dorm. If they were previously in the freshman's economics classes, then their joining the dorm increases the overlap across the freshman's social groups, which, as we show, has a negative impact on his expected number of friends. On the other hand, if non-economists join the dorm, then this positively impacts the freshman's number of friends. Our comparative statics on varying the size of *overlapping* social groups are novel in the literature. Currarini et al. (2009) show that agents belonging to larger groups have higher degrees. However, in their model, there is only one group per agent (e.g. race), so the interaction across social groups cannot be studied. de Marti and Zenou (2011) and Iijima and Kamada (2014) consider strategic network formation models in which the costs and benefits of link formation depend on agents' social groups.[3] de Marti and Zenou (2011) study segregation patterns that arise in stable networks as a function of relative costs and benefits of link formation between and across social groups, whereas Iijima and Kamada (2014) show how properties of stable networks (such as clustering and average path length) depend on "social distance" parameters. However, these papers examine neither dynamics of link formation nor the effect of varying social group sizes and their overlaps.

We also derive results on the dynamics of homophily. Homophily is a commonly observed empirical phenomenon (Kandel, 1978, Shrum et al., 1988, McPherson et al., 2001, Moody, 2001, Mouw and Entwisle, 2006, Mayer and Puller, 2008, Currarini et al.,

[3]These models are in the spirit of Jackson and Wolinsky (1996) and Bala and Goyal (2000), but these earlier contributions did not consider the effect of social groups.

2009, Wimmer and Lewis, 2010), and most empirical studies of homophily have used surveys of close friendships. For example, Shrum et al. (1988) find that for school children homophily in gender falls over time, but that homophily in race increases over time. In a theoretical paper, Bramoullé et al. (2012) derive a negative relationship between homophily and time and find some empirical support for their prediction in physics citation networks. In contrast to previous work, we show that in our model, homophily is not necessarily monotonic in an agent's degree or in the amount of time that the agent has spent making friends. We provide sufficient conditions on the effective social group sizes for (i) homophily to monotonically decrease over time, and (ii) to increase up to peak and eventually fall over time.

One interpretation of our network formation process – that we use in a running example throughout the paper – is to consider it as a model of friendship formation in online social networks, such as Facebook. "Friending" – recording friendships on online social networking platforms – is different from maintaining real-world friendships. Indeed, different types of friendships – close, distant, romantic or online – generate remarkably dissimilar social networks (Jackson, 2008, Newman, 2010). Most people have few close friends and even fewer lovers. Many platforms, such as Facebook and LinkedIn, provide a record of its users' real-life meetings – an online Rolodex. Typically, after meeting each other, people "send friend requests" in order to record the meeting on Facebook and maintain a "Facebook friendship". For this reason, many Facebook users have more Facebook friends than friends with whom they interact daily.[4] Our model complements growing random network models for understanding the process governing friendship formation in online social networks. Firstly, we retain one standard feature of growing random network models that captures the Rolodex aspect of many online social networks – that agents do not break friendships. Secondly, we focus on friendship formation in networks of fixed size since it is link formation rather than network growth that is the phenomenon of interest in some online social networks. Finally, in our model, agents may initiate and receive multiple friendships throughout the friendship formation process whereas in growing random networks agents who have already entered the network do not make friendships among themselves.

Although the focus of this paper is theoretical, we use Facebook data to provide supporting empirical evidence for our main comparative static results. The data represent a September 2005 cross-section of the complete structures of social connections on

[4]Therefore, rather than reflect close real-world relationships, many Facebook friendships represent "weak ties", which play an important role in economic and social outcomes (Granovetter, 1973, 2005).

www.facebook.com within (but not across) the first ten American colleges and universities that joined Facebook. We support our findings so far as it is possible with the *cross-sectional* data at our disposal, and the empirical evidence presented lends support to our results.

Section 2 presents the model, and we use a mean-field approximation to derive baseline results regarding agents' friendship rate over time and the degree distribution in Section 3. The main results regarding the relationship between degree and social group size, as well as supporting empirical evidence using the Facebook data, are given in Section 4. The dynamics of homophily and its implications are discussed in Section 5. Comparisons of simulations of the model against our mean-field approximation are presented in Section 6, and Section 7 concludes. All proofs, further discussion of the model, and a data summary are in the Appendix.

2. Model

In this section, we first introduce the social structure of our model (Section 2.1), and then describe our complex stochastic network formation process (Section 2.2). In the rest of the paper (starting from Section 3), we employ the mean-field approximation method in order to get a handle on this process analytically. According to this method, we assume that the realization of a random variable in any period is its expected value. Hence, the dynamic system generated by our model is not seen as evolving stochastically, but rather deterministically at the rate proportional to the expected change. The method has been adopted by the economics literature from statistical physics, and our analysis here is similar to the one carried out in Jackson and Rogers (2007).

2.1. Social groups and social categories

Let $[K^1, ..., K^R]$ be a finite ordered list of *social categories*. For each $r \in \mathcal{R} = \{1, ..., R\}$ the element K^r is the r^{th} category and an element of K^r is a *characteristic* within that category. The *type* of an agent $i \in N$ is represented by a vector of characteristics $k_i = (k_i^1, ..., k_i^R) \in \mathcal{K}$. Denote by $\mathcal{K}$ the set of all distinct vectors $k = (k^1, ..., k^R)$ where $k^r \in K^r$ for each $r \in \mathcal{R}$, so the number of all possible types is $|\mathcal{K}| = \prod_{r\in\mathcal{R}} |K^r|$. For any type $k \in \mathcal{K}$ and any agent $i \in N$, denote by $N_i(k)$ the set of all agents other than i who are of type k. For each $r \in \mathcal{R}$ and agent $i \in N$, define a *social group* $\Gamma_i^r = \cup\{N_i(k) : k^r = k_i^r\}$,[5] which is the set of all agents (other than i) that share the characteristic k_i^r within the social category r with i. Additionally, define the social

[5]For any sequence of sets $X_1, ..., X_n$, $\cup\{X_1, ..., X_n\} = \cup_{i=1}^n X_i$.

group $\Gamma_i^{\emptyset} = N \setminus \{i\}$ as the set of all agents other than i. For each $S \subseteq \mathcal{R}$ we can also define the *social subgroup* $\pi_i(S)$ as the set of agents (other than i) that share *only* the characteristics within the set of categories indexed by S with i. That is,

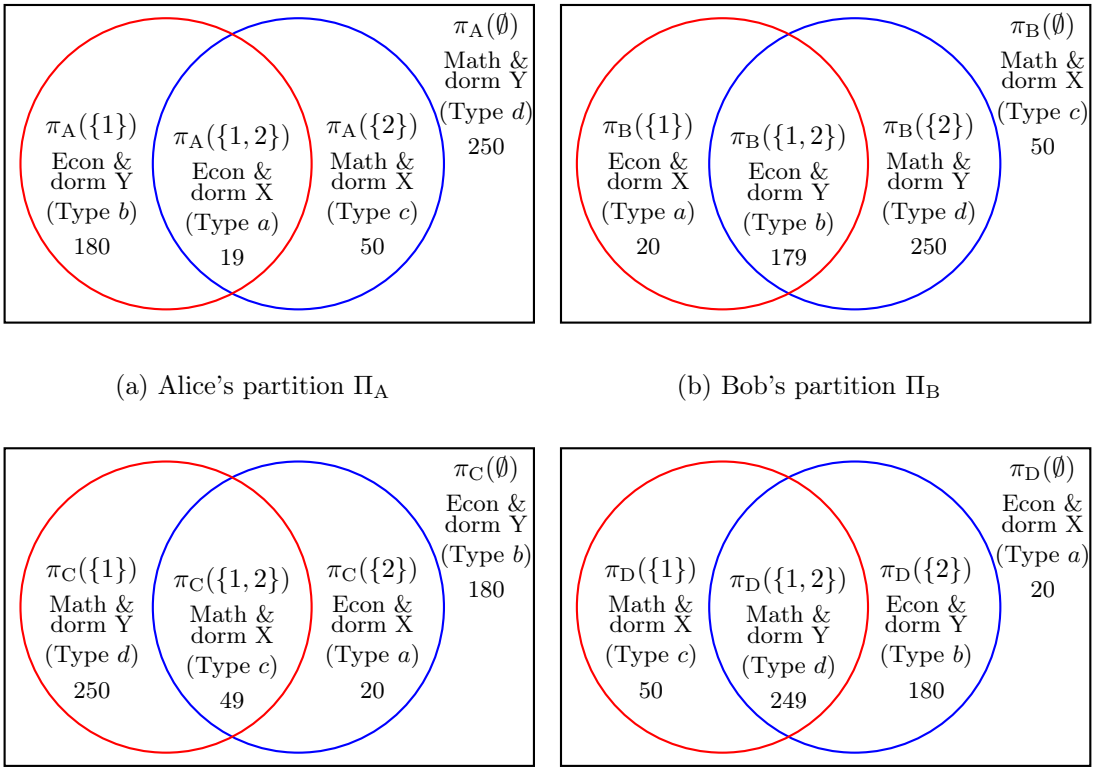
$$\pi_i(S) = \cup\{N_i(k) : k^r = k_i^r \text{ for each } r \in S, \text{ and } k^r \neq k_i^r \text{ for each } r \in \mathcal{R} \setminus S\} \quad (1)$$

Naturally, the subgroup $\pi_i(\emptyset)$, which we refer to as the $\emptyset$-*subgroup*, denotes the set of agents other than i who share no characteristics with i. We refer to $\pi_i(\mathcal{R})$ as the *core subgroup*, and for any $r \in \mathcal{R}$, $\pi_i(\{r\})$ is referred to as a *singleton subgroup*. The set $\Pi_i = \{\pi_i(S) : S \subseteq \mathcal{R}\}$ induces a partition on $N \setminus \{i\}$.

We assume throughout that there are at least two social categories ($|\mathcal{R}| > 1$) and that for every agent $i \in N$, every set $N_i(k)$ contains at least one agent.

Example. Suppose that the list of social categories at a university is given by $[K^1, K^2] =$ [class, dorm]. The "class" social category, K^1, is given by {Econ, Math}, and the "dorm" social category, K^2, is given by {dorm X, dorm Y}. There are therefore four possible types of students, namely a = (Econ, dorm X), b = (Econ, dorm Y), c = (Math, dorm X), and d = (Math, dorm Y). That is, $\mathcal{K} = \{a, b, c, d\}$. To be concrete, suppose there are 20 students of type a, 180 of type b, 50 of type c, and 250 of type d. Furthermore, suppose that four particular students Alice, Bob, Charlie, and Diana, are students of type a, b, c, and d respectively.[6] For example, Figure 1a shows Alice's partition Π_{A} induced over all the other students at the university. The singleton subgroup $\pi_{\text{A}}(\{1\})$ includes all students other than Alice who are in her class *only* (and thus share no other characteristic with Alice). The cardinality of $\pi_{\text{A}}(\{1\})$ is therefore 180. The $\emptyset$-subgroup includes all students other than Alice who are neither in her class (Econ) nor in her dorm (dorm X). The "class" social group $\Gamma_{\text{A}}^1 = \pi_{\text{A}}(\{1\}) \cup \pi_{\text{A}}(\{1,2\})$ includes all students other than Alice who are in her class (red circle on the left). Note that since the core subgroup $\pi_{\text{A}}(\{1,2\})$ includes all students of type a other than Alice, the cardinality of this set is 19, and the cardinality of Γ_{A}^1 is therefore 199. The "dorm" social group $\Gamma_{\text{A}}^2 = \pi_{\text{A}}(\{2\}) \cup \pi_{\text{A}}(\{1,2\})$ includes all students other than Alice who are in her dorm (blue circle on the right), and finally the social group $\Gamma_{\text{A}}^{\emptyset} = N \setminus \{\text{Alice}\}$ includes all students other than Alice (the entire rectangle). Note that for any other student i of type a, the corresponding elements in Π_i and Π_{A} have the same cardinality. The analogous partitions for Bob, Charlie, and Diana are represented in Figures 1b-1d respectively. ■

[6] For these four students, every mathematical expression involving a student is subscripted by the first letter of their name.



(a) Alice's partition Π_A

(b) Bob's partition Π_B

(c) Charlie's partition Π_C

(d) Diana's partition Π_D

Figure 1: The partition Π_j for students j of each type $k \in \mathcal{K}$ in the Example

2.2. The network formation process

The network formation process is as follows: At time period $t = 0$ all agents are *active* and have no friends. An agent is active in period $t > 0$ if the agent sends friend requests and accepts all friend requests in period t. An agent is *idle* in period $t > 0$ if the agent does not send friend requests and cannot receive friend requests in period t. Let $\mathbf{q} = (q^{\emptyset}, q^1, ..., q^R)$ and $\sum_{r \in \mathcal{R} \cup \{\emptyset\}} q^r = 1$. In each period $t \in \{1, 2, 3, ...\}$ every active agent i *interacts* with other agents as follows: Agent i selects the social group Γ_i^r with probability $q^r > 0$ for $r \in \mathcal{R} \cup \{\emptyset\}$ and for each $N_i(k) \subseteq \Gamma_i^r$, selects the group of agents of type k with probability $\frac{|N_i(k)|}{|\Gamma_i^r|}$. Agent i then sends a friend request to an agent j selected uniformly at random from among the active agents in $N_i(k)$ who are not yet i's friends. If such an agent j can be found (i.e. the set of active agents in $N_i(k)$ who are not yet i's friends is not empty), then the request is immediately accepted (since j would be

active) and agents i and j become friends at t.[7] Otherwise, agent i does not initiate any friendships in period t. At the end of every period $t \geq 0$, agent i remains active with probability $p \in (0,1)$ until the following period and becomes idle with probability $1-p$. If agent i becomes idle, i retains all his friendships, but no longer forms any new links with other agents in all subsequent periods.[8]

To be clear, when we say that agents i and j become friends, we mean that an undirected link forms between them, and such a link is formed in period t if and only if i sends a friend request to j at t that is immediately accepted, or j sends a friend request to i at t that is immediately accepted. Note that it is therefore irrelevant for the formation of the link ij whether it was i or j who initiated the friendship by sending the friend request. Once formed, a link cannot disappear. That is, we do not allow agents to delete links or "unfriend" each other.[9]

Example. *(cont.)* Let us imagine that Alice, Bob, Charlie and Diana are users of an online social networking platform, such as Facebook. Alice interacts with students in her class Γ_{A}^1, or in her dorm Γ_{A}^2, or with everyone $\Gamma_{\mathrm{A}}^{\emptyset}$, with probabilities q^1, q^2, and $q^{\emptyset}$ respectively. Suppose that in period t, Alice interacts with students in dorm X (that is, with students of type a or c). This event occurs with probability q^2. Conditional on interacting with students in her dorm, Alice interacts with type c students with probability $\frac{|N_{\mathrm{A}}(c)|}{|\Gamma_{\mathrm{A}}^2|} = \frac{|\pi_{\mathrm{A}}(\{2\})|}{|\Gamma_{\mathrm{A}}^2|} = \frac{50}{69}$ (Our running example is particularly simple because each subgroup $\pi \in \Pi_i$ for any student i contains students of only one type. So $\pi_{\mathrm{A}}(\{2\})$ is precisely the set of type c students in this case).[10] Alice then sends a friend request to a student j selected uniformly at random from among the type c students in her dorm who are active users of the online social network and who are not yet her friends. If such a student, say Charlie, can be found, the request is immediately accepted and Alice and Charlie become (online) friends. On the other hand, if the relevant set of students is empty, Alice initiates no new friendships in period t.[11] Alice keeps making friends in this

[7] We effectively assume that friend requests are accepted with probability 1, but one could generalize this to an exogenous probability $0 < m^{kk'} \leq 1$ that a friend request from an agent i of type k is accepted by an agent j of type k'.

[8] There are several ways of interpreting $1-p$, the probability of becoming idle: There must be reasons, *other than having linked with every user in the network*, for why people stop adding new friends (online) such as reaching a cognitive capacity for social interaction, losing interest, finding an alternative (online) social network and so on. Including all these explanations would require a much richer model, so we simply capture them as a random process with the idleness probability $1-p$.

[9] As noted in the Introduction, this feature captures the Rolodex aspect of many online social networks.

[10] If there were an additional subject, say Computer Science, then $\pi_{\mathrm{A}}(\{2\})$ would contain dorm X students studying Econ or CS. That is, we would have $\pi_{\mathrm{A}}(\{2\}) = N_{\mathrm{A}}((\text{dorm X, Econ})) \cup N_{\mathrm{A}}((\text{dorm X, CS}))$.

[11] The probability that Alice sends a friend request to any *particular* type c student conditional on

manner until she becomes idle (that is, stops being an active user of the online social network). This process happens simultaneously for all students in every period. The implications of this are worth highlighting: If Bob (of type b) interacts with students in his dorm in period t, then he would be interacting with students in dorm Y in that period, whereas if Alice interacts with students in her dorm in the same period, she would be interacting with students in dorm X (and could thus send friend requests only to students in that dorm at t). But in the same period t she may receive a friend request from students who are not in her dorm. For example, Bob could be interacting with students in his class in period t and could send a friend request to Alice in that period (since Alice and Bob are in the same class). In this context, we interpret the vector $\mathbf{q}$ as students' propensity to browse (or students' allocation of time spent browsing) through the online profiles of students of a particular social category. The probability $q^{\emptyset}$ is then interpreted as the propensity to browse anyone in the network.[12] ■

The example highlights an important distinction between the probability of *interacting* with other agents and the probability of making a link with other agents. In our model, an agent interacts with a *set* of agents. That is, when an agent i interacts with a set of agents in period t we mean that this *set* is selected from among possible sets of agents at t. The probability of interacting with a set of agents of a particular type is constant and depends only on the number of agents of each type and on $\mathbf{q}$. *Conditional* on interacting with a set of agents of a particular type, agent i sends a friend request to an active potential friend in this set. If such an agent can be found a link is established with probability 1. So even though i's probability of interacting with a set of agents of a particular type is constant, i's probability of establishing a link with a *particular* agent j from the set is increasing since the subset of active potential friends becomes smaller over time.[13]

Our model is conceptually related to affiliation networks introduced in sociology by Breiger (1974) and Feld (1981). An affiliation network is described by a set of agents and a set of *memberships*, such as clubs, online fora, research topics, or social groups (Newman

interacting with type c students increases over time: She sends a friend request to a student selected uniformly at random from among the active type c students who are not yet her friends, but this pool of students becomes smaller as Alice befriends type c students and as they become idle.

[12] In the Online Appendix, we describe an alternative version of the model in which $\mathbf{q}$ can be interpreted as the fraction of time that students allocate *physically* to being in a particular social group.

[13] Currarini et al. (2009, 2010) distinguish between the probability of agents of particular types "meeting" and a separate probability of agents forming a link together conditional on a meeting. Both of these probabilities are constant over time. Our notion of interacting is similar to their notion of "meeting", but we also keep track of the time-varying probabilities of linking between agents of particular types.

et al., 2002). Some affiliation network models have found wide-spread application in online social networks (Botha and Kroon, 2010, Kumar et al., 2010, Xiang et al., 2010). In more recent evolving models of affiliation networks, new memberships may emerge over time, and the likelihood of meeting new agents can depend on their memberships (Lattanzi and Sivakumar, 2009, Zheleva et al., 2009). However, these models typically contain a large number of parameters and most, such as those by Leskovec et al. (2005, 2008), rely entirely on simulations.[14]

3. Baseline results

This section presents our baseline results. All the analytical results presented in this paper are derived using the mean-field approximation to the stochastic network formation model described in Section 2.2.[15] In Section 6, we show that the mean-field approximation performs well against simulations of the model. Appendix I contains proofs of the analytical results.

Lemma 1. *The probability with which agent i interacts with agents from a subgroup $\pi_i(S)$ such that $S \subseteq \mathcal{R}$ is given by*

$$q^{\pi_i(S)} = |\pi_i(S)| \left[\sum_{r \in S \cup \{\emptyset\}} \frac{q^r}{|\Gamma_i^r|} \right] \tag{2}$$

and by definition $\sum_{\pi \in \Pi_i} q^\pi = \sum_{S \subseteq \mathcal{R}} q^{\pi_i(S)} = 1$.

Example. *(cont.)* Equation (2) above highlights the role of overlaps across social groups in our model. Suppose that q^1 is small, so that Alice allocates little time specifically to

[14] Within our framework, let $K^\emptyset$ be a social category containing a single characteristic (which implies it is shared by all agents). Then the set of all memberships would be $\{k^r \in K^r : r \in \mathcal{R} \cup \{\emptyset\}\}$ and a link between an agent i and a membership $k^r \in K^r$ is given the weight q^r for all $i \in N$. New links form over time via a form of *focal closure* (Easley and Kleinberg, 2010, p. 97): in every period, every agent i is assigned a membership $k^r \in K^r$ at random according to $\mathbf{q}$. Agent i then selects a type of agent from among those agents (other than i) that have a link with k^r at random according to their relative proportions, and then creates an undirected link with an agent j chosen uniformly at random from among the remaining active agents of the selected type.

[15] This mean-field analysis highlights technical differences between our model and several other models of network formation. Currarini et al. (2009) derive the steady-state of a process in which pools of agents are matched at random to meet and strike friendships, whereas our analysis essentially derives the most likely outcome of our process (which may be thought of as a finite Markov chain on a finite state space). This also differs from all growing random network models (Barabási and Albert, 1999, Jackson and Rogers, 2007, Bramoullé et al., 2012, for example) which have an infinite number of states, or which are ergodic (Fosco et al., 2010).

interacting with students from her class. If we increase q^2, she would be more likely to interact with those students in her dorm who are also in her class. ■

Let $d_i(t)$ denote the *degree* (or number of friends) of agent i in period t. Analogously, let $d_i^\pi(t)$ denote the number of friends i has in period t with agents in the subgroup $\pi \in \Pi_i$. Finally, let $\Delta d_i^\pi(t)$ denote the number of new friends i makes in period t with agents in the subgroup $\pi \in \Pi_i$.

Proposition 1. *For any agent $i \in N$ and any $\pi \in \Pi_i$, the function $\Delta d_i^\pi(t)$ is given by*

$$\Delta d_i^\pi(t) = 2q^\pi \mathbf{1}(t \leq T^\pi) \tag{3}$$

where for each $\pi \in \Pi_i$, the "expected depletion time" T^π, which denotes the expected number of periods it takes i to make a link with every other active agent in π, is given by

$$T^\pi = \frac{\ln\left(\frac{2q^\pi p}{2q^\pi p + (1-p)|\pi|}\right)}{\ln(p)} \tag{4}$$

This proposition states that agent i makes new friends at a rate of $2q^\pi$ in the subgroup $\pi \in \Pi_i$ for every period $t \leq T^\pi$. We now provide some intuition for this result. As shown in Lemma 1, agent i interacts with agents in the subgroup π with probability q^π. Conditional on interacting with agents in this subgroup, agent i sends an immediately accepted friend request to some active agent in π who is not yet i's friend (such an agent can be found in every period $t \leq T^\pi$). Therefore, for any period $t \leq T^\pi$, agent i initiates an expected q^π links with agents in π. It remains for us to determine the expected number of links initiated by agents in π that agent i receives in any period $t \leq T^\pi$. To do this, we must determine the probability with which each agent $j \in \pi$ interacts with agents of i's type (which will depend on the size of the subgroups in Π_j and is not necessarily equal to q^π). Conditional on j interacting with agents of i's type, we must also determine the probability that j selects agent i specifically as the agent to whom the friend request is sent. But j is selecting uniformly at random from among the active potential friends of i's type, and the size of this pool of agents varies with time as j makes friends. Hence the probability that j selects i is also time-varying. Nevertheless, in Appendix I we show that the expected number of links that agent i makes in a period $t \leq T^\pi$ that are initiated by agents in $\pi \in \Pi_i$ is constant and also equal to q^π.

Example. *(cont.)* Recall that all the students in $\pi_{\mathrm{A}}(\{2\})$ are of type $c =$ (Math, dorm X). In every period, Alice interacts with them with probability $q^{\pi_{\mathrm{A}}(\{2\})} = q^2 \frac{50}{69} + q^\emptyset \frac{50}{499}$. (She can interact with them via interacting with students in her dorm or via interacting with

everyone.) Conditional on interacting with them, Alice sends an immediately accepted friend request to some active student in $\pi_A(\{2\})$ who is not yet her friend. Such a student can be found if $t \leq T^{\pi_A(\{2\})}$. Alice therefore initiates an expected $q^{\pi_A(\{2\})}$ links with students in $\pi_A(\{2\})$ in a period $t \leq T^{\pi_A(\{2\})}$.

Now, we will determine the expected number of friend requests that Alice receives (and immediately accepts) from students in $\pi_A(\{2\})$. Denote by $Z^{\pi_A(\{2\})}(t)$ the number of active students in $\pi_A(\{2\})$ who are not Alice's friends in period t, and let Charlie be one such student. In every period, Charlie interacts with students of type a with probability $q^{\pi_C(\{2\})} = q^2 \frac{20}{69} + q^{\emptyset} \frac{20}{499}$. Conditional on interacting with type a students in period t, he selects one to send a friend request to from among the remaining active students in $\pi_C(\{2\})$ who are not yet his friends in period t. Since there are $Z^{\pi_C(\{2\})}(t)$ such students, Charlie selects Alice in period t with probability $\frac{1}{Z^{\pi_C(\{2\})}(t)}$. Therefore, the expected number of friend requests that Alice receives in period t is given by

$$q^{\pi_C(\{2\})} \frac{Z^{\pi_A(\{2\})}(t)}{Z^{\pi_C(\{2\})}(t)} \tag{5}$$

Following the reasoning outlined in Appendix I, Equation (5) is equal to $q^{\pi_A(\{2\})}$ in this example. The reason is that $Z^{\pi_A(\{2\})}(t)$ and $Z^{\pi_C(\{2\})}(t)$ have the same growth rate and therefore the fraction $\frac{Z^{\pi_A(\{2\})}(t)}{Z^{\pi_C(\{2\})}(t)}$ remains fixed at the initial value $\frac{Z^{\pi_A(\{2\})}(0)}{Z^{\pi_C(\{2\})}(0)} = \frac{50}{20}$. Therefore, Alice makes friends at a rate of $2q^{\pi_A(\{2\})}$ with students in $\pi_A(\{2\})$. ■

Corollary 1. *The degree of agent i as a function of time t is a continuous, increasing, piecewise linear concave function, and is given by*

$$d_i(t) = \sum_{\pi \in \Pi_i} d_i^{\pi}(t) = 2 \sum_{\pi \in \Pi_i} q^{\pi} \left[t \mathbf{1}(t \leq T^{\pi}) + T^{\pi} \mathbf{1}(t > T^{\pi}) \right] \tag{6}$$

This corollary shows that agents make friends at a decreasing rate over time. Essentially, in every period, each agent $i \in N$ makes new friends at a rate of $2q^{\pi}$ in each subgroup $\pi \in \Pi_i$, and therefore at an overall rate which corresponds to the sum of the rates in each subgroup. As time passes, the agents in a subgroup $\pi \in \Pi_i$ either become idle or become friends with i – thus leaving i with fewer agents in π to strike new friendships with over time. Eventually, a period T^{π} will be reached at which every agent in π will either have become idle or will already be a friend of i, therefore any time spent interacting with agents in the subgroup π after period T^{π} no longer adds to i's degree. The overall rate at which i makes new friends therefore diminishes by $2q^{\pi}$ after the period T^{π}. Note that this explanation for making new friends at a decreasing rate is purely

due to the fact that as times passes, agents deplete the pools of potential friends, and the expected depletion times are finite because the pools are finite. This contrasts with models in which decreasing rates of friendship formation are due to decreasing marginal utility of friendships (Currarini et al., 2009).

The following lemma provides us with a useful characterization of the expected depletion times.

Lemma 2. *For every $i \in N$ and any $S, S' \subseteq \mathcal{R}$, if $S' \subseteq S$, then $T^{\pi_i(S')} \geq T^{\pi_i(S)}$.*

This result indicates that any agent i will deplete the subgroups of agents that share all their characteristics with i first and then will deplete subgroups of agents that share subsets of those characteristics with i. The last subgroup to be depleted is the $\emptyset$-subgroup (consisting of agents sharing no characteristics with i). Note that *within a social group* this partial ordering over the expected depletion times of its subgroups is expressed, somewhat remarkably, in terms of the *number* of characteristics only and it holds independently of social group sizes and of the vector $\mathbf{q}$.

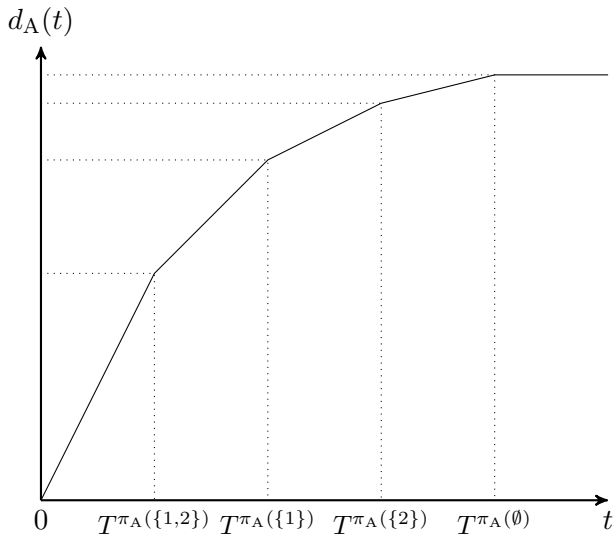


Figure 2: A sketch of the function $d_A(t)$

Example. *(cont.)* We sketch the function $d_A(t)$ in Figure 2. By Lemma 2, Alice must first deplete the core subgroup, followed by the singleton subgroups (and we assume here that $\pi_A(\{1\})$ is depleted before $\pi_A(\{2\})$, although the reverse could also hold), and finally followed by the $\emptyset$-subgroup. We therefore have that $0 \leq T^{\pi_A(\{1,2\})} \leq T^{\pi_A(\{1\})} \leq$

$T^{\pi_A(\{2\})} \leq T^{\pi_A(\emptyset)}$. Furthermore, by Corollary 1 the slope of $d_A(t)$ is given by

$$\begin{cases} 2(q^{\pi_A(\emptyset)} + q^{\pi_A(\{2\})} + q^{\pi_A(\{1\})} + q^{\pi_A(\{1,2\})}) & \text{if } t \in (0, T^{\pi_A(\{1,2\})}] \\ 2(q^{\pi_A(\emptyset)} + q^{\pi_A(\{2\})} + q^{\pi_A(\{1\})}) & \text{if } t \in (T^{\pi_A(\{1,2\})}, T^{\pi_A(\{1\})}] \\ 2(q^{\pi_A(\emptyset)} + q^{\pi_A(\{2\})}) & \text{if } t \in (T^{\pi_A(\{1\})}, T^{\pi_A(\{2\})}] \\ 2(q^{\pi_A(\emptyset)}) & \text{if } t \in (T^{\pi_A(\{2\})}, T^{\pi_A(\emptyset)}] \\ 0 & \text{if } t > T^{\pi_A(\emptyset)} \end{cases}$$

That is, Alice starts off making friends at a rapid rate of 2 (since $\sum_{S \subseteq \mathcal{R}} q^{\pi_A(S)} = 1$). Eventually, she becomes friends with everyone who is both in her class and her dorm (or some of them become idle). Having depleted the core subgroup, the rate at which she makes new friends drops by $2q^{\pi_A(\{1,2\})}$, since any period that she now spends interacting with agents in the core subgroup will not be spent making new friends. Alice subsequently depletes the subgroup of agents who are in her class only, and then depletes the subgroup of agents who are in her dorm only. Her rate of making new friends drops each time. Finally, she is left with the $\emptyset$-subgroup. Once that is depleted, she can no longer make any new friends.

Note that the function $d_C(t)$ for Charlie (who is of type c) will differ from Alice's since the slope and the expected depletion times depend on the sizes of the subgroups in Π_C and these differ from the corresponding ones for Alice. ■

There are two sources of heterogeneity in degree that are admitted by our mean-field approximation. Firstly, the function $d_i(t)$ for an agent i depends only on p, $\mathbf{q}$, and the size of the subgroups in the partition Π_i. If we consider agents i and j of different types, the sizes of the subgroups in their partitions Π_i and Π_j will typically be different from each other. Hence $d_i(t)$ and $d_j(t)$ will differ. That is, the mean-field approximation preserves the heterogeneity of predicted degree trajectories for agents of different types (with $|\mathcal{K}|$ types there are generically $|\mathcal{K}|$ distinct predicted trajectories for degree over time). Secondly, the function $d_i(t)$ provides us with the expected degree of an agent i in period t, *provided that the agent actually remains active until period* t. Therefore even if i and j are of the same type (so that $d_i(t) = d_j(t)$ for all t), the *realized* degrees of these agents will typically differ since agents i and j may become idle at different times. This highlights the fact that to derive the degree distribution in the population, we must keep track of the point at which an agent becomes idle and thus the point at which that agent's degree must be measured. This point is determined stochastically according to the probability of idleness $1 - p$. The following proposition provides an expression for

the resulting degree distribution.

Proposition 2. *The overall cumulative distribution function for degree is given by*

$$G(d) = \frac{1}{|N|} \sum_{i \in N} \left(1 - p^{t_i(d)+1}\right) \tag{7}$$

where $t_i(d)$ is the inverse of $d_i(t)$ for each agent $i \in N$.

Proposition 2 shows that with our network formation process, the resulting distribution is roughly geometric (for which the continuous analogue is the exponential distribution). Jackson (2008) observes that empirically "more purely social networks" tend to have degree distributions that are close to exponential.[16] Other social networks tend to follow a power law distribution. Such degree distributions can be generated by a link formation process exhibiting *preferential attachment* in growing random networks. In these models, every new agent links to existing agents with a probability that is proportional to the degree of these agents (Price, 1976, Barabási and Albert, 1999, Jackson and Rogers, 2007). However, we show in Appendix II that none of the results in this paper would change if our link formation process were governed by preferential attachment.

4. Comparative statics on degree and group size

There are many ways of defining "an increase in the size of a social group" when social groups can be overlapping and the number of agents is finite. In particular, it is crucial to keep track of how the overlap across social groups changes as we expand a particular social group. In some cases, we will show that expanding the size of an agent's group increases that agent's expected degree, but in other cases, it may decrease the agent's expected degree.

Before proceeding to the main definitions and results of this section, it will be useful to impose a mild restriction on our model.

Assumption 1. *For every $i \in N$ and any $S, S' \subseteq \mathcal{R}$, if $S' \neq S$, then $T^{\pi_i(S')} \neq T^{\pi_i(S)}$.*

This assumption allows us to impose a strict order on the expected depletion times, which is relatively weak and considerably simplifies the proofs. It will also be useful to

[16] Quoting Jackson (2008, p. 65): "some of the more purely social networks have parameters that indicate much higher levels of random link formation, which are very far from satisfying a power law. In fact, the degree distribution of the romance network among high school students is essentially the same as that of a purely random [growing] network [i.e. exponential]."

introduce the following definition: For any $r \in \mathcal{R}$ and any agent $i \in N$, we refer to $\frac{|\Gamma_i^r|}{q^r}$ as the *effective size* of the social group Γ_i^r (since the size of the group is normalized by the probability of interacting with that group). We refer to $\frac{|\Gamma_i^\emptyset|}{q^\emptyset}$ as the *effective network size.*

We consider two definitions of "subgroup expansion" below. In each case, we denote every variable after the expansion by that variable's name with an added "hat". For example, the core subgroup is denoted by $\pi_i(\mathcal{R})$ before an expansion (as usual), and by $\hat{\pi}_i(\mathcal{R})$ after the expansion.

Definition 1. *For any agent $i \in N$, there is a* ***singleton subgroup expansion*** *by $\delta > 0$ of a singleton subgroup $\pi_i(\{r\})$ if*

i. $|\hat{\pi}_i(\{r\})| = |\pi_i(\{r\})| + \delta$

ii. $|\hat{\pi}_i(\emptyset)| = |\pi_i(\emptyset)| - \delta$

and the cardinality of all other subgroups remains unchanged.

That is, there is a singleton subgroup expansion by δ of a singleton subgroup $\pi_i(\{r\})$ if $|\pi_i(\{r\})|$ increases by δ and $|\pi_i(\emptyset)|$ decreases by δ. Note that such an expansion implies that the cardinality of Γ_i^r increases by δ while leaving the cardinality of every other social group for agent i unchanged.

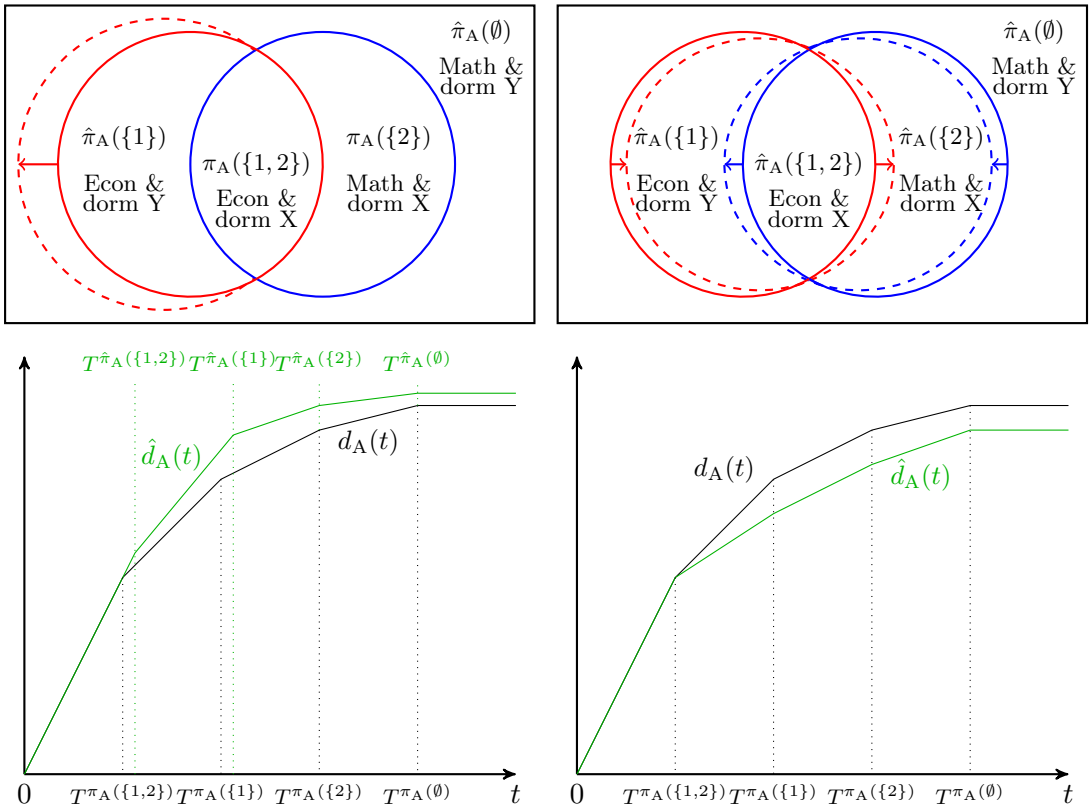
Proposition 3. *For any agent $i \in N$ and some $r \in \mathcal{R}$, suppose there is a singleton subgroup expansion by δ of $\pi_i(\{r\})$ and suppose Assumption 1 holds. There is a $\bar{\delta} > 0$ such that if $\frac{|\Gamma_i^\emptyset|}{q^\emptyset} \geq (|\Gamma_i^r| + \bar{\delta})\frac{|\Gamma_i^r|}{q^r}$ then, for all $\delta < \bar{\delta}$ and all t, $\hat{d}_i(t) \geq d_i(t)$.*

This proposition shows that the degree of agent i increases from $d_i(t)$ to $\hat{d}_i(t)$ for every t following a singleton subgroup expansion if the effective network size is large enough. A detailed explanation of singleton subgroup expansions, as well as the intuition for this result can be given with the help of our example.

Example. *(cont.)* On the left-hand side of Figure 3 (top of panel (a)) we illustrate the singleton subgroup expansion of $\pi_{\mathrm{A}}(\{1\})$ by δ for Alice, so that $|\hat{\pi}_{\mathrm{A}}(\{1\})| = |\pi_{\mathrm{A}}(\{1\})| + \delta$, and $|\hat{\pi}_{\mathrm{A}}(\emptyset)| = |\pi_{\mathrm{A}}(\emptyset)| - \delta$. All other subgroups remain unchanged. The expansion is equivalent to δ Math students living in dorm Y changing their "class" to Econ. The number of students other than Alice who are in the Econ class but not in her dorm is thus greater by δ.

Before the expansion, let us suppose that the expected depletion times satisfied

$$0 < T^{\pi_{\mathrm{A}}(\{1,2\})} < T^{\pi_{\mathrm{A}}(\{1\})} < T^{\pi_{\mathrm{A}}(\{2\})} < T^{\pi_{\mathrm{A}}(\emptyset)}$$



(a) A singleton subgroup expansion of $\pi_A(\{1\})$ for Alice in the Example

(b) A core subgroup expansion for Alice in the Example

Figure 3: An illustration of Propositions 3 and 4 using the Example

The strict inequalities follow from Assumption 1. If δ is sufficiently small ($\delta < \bar{\delta}$), the order of expected depletion times remains unchanged, so after the expansion they satisfy

$$0 < T^{\hat{\pi}_A(\{1,2\})} < T^{\hat{\pi}_A(\{1\})} < T^{\hat{\pi}_A(\{2\})} < T^{\hat{\pi}_A(\emptyset)}$$

On the left-hand side of Figure 3 (bottom of panel (a)) we represent degree as a function of time for Alice before the expansion by $d_A(t)$ (in black) and after the expansion by $\hat{d}_A(t)$ (in green). By the definition of expected depletion times, we can show that $T^{\pi_A(S)}$ for any $S \subseteq \mathcal{R}$ such that $1 \in S$ will be greater after the expansion, and will remain unchanged for any S such that $1 \notin S$. In other words, any subgroup that is a subset of the "class" social group will take longer to deplete. This shift in the expected depletion times is shown in the diagram. Interestingly, because the cardinality of $\pi_A(\{1\})$ and of Γ^1_A has increased, Alice's probability of interacting with students within the "class" social

group increases. Indeed, since the "class" social group is larger, Alice is now more likely to become friends with students from her class when she is interacting with students from *other* social groups that overlap with the "class" social group. From this we can show that for any $t \leq T^{\pi_A(\{1\})}$, the slope of $d_A(t)$ is smaller than that of $\hat{d}_A(t)$, from which it follows that $\hat{d}_A(t) \geq d_A(t)$ for all $t \leq T^{\pi_A(\{1\})}$.

Following the depletion of every subgroup that is a subset of the "class" social group, Alice is left with fewer students to strike new friendships with (since the $\emptyset$-subgroup is now smaller). Effectively, Alice is now more likely to interact with students that she is already friends with, and thus adds new friends at a slower rate. The slope of $\hat{d}_A(t)$ is therefore smaller than that of $d_A(t)$ for any $t \geq T^{\hat{\pi}_A(\{1\})}$. Nevertheless, if the effective network size is large enough the functions never cross, so $\hat{d}_A(t) \geq d_A(t)$ for all t.[17,18] ■

We now introduce a different type of expansion which is specific to *overlapping* groups.

Definition 2. *For any agent $i \in N$, there is a **core subgroup expansion** by $\delta > 0$ if*

i. $|\hat{\pi}_i(\mathcal{R})| = |\pi_i(\mathcal{R})| + \delta$

ii. $|\hat{\pi}_i(\{r\})| = |\pi_i(\{r\})| - \delta$, *for all* $r \in \mathcal{R}$

iii. $|\hat{\pi}_i(\emptyset)| = |\pi_i(\emptyset)| + (|\mathcal{R}| - 1)\delta$

and the cardinality of all other subgroups remains unchanged.

A core subgroup expansion by δ represents a situation in which the overlap of all of agent i's social groups increases by δ, while leaving *all* social group sizes unchanged. Let r^* denote the social category for which the singleton subgroup $\pi_i(\{r^*\})$ is the last to be depleted for agent i (before $\pi_i(\emptyset)$).

Proposition 4. *For any agent $i \in N$ suppose there is a core subgroup expansion by δ of $\pi_i(\mathcal{R})$ and suppose Assumption 1 holds. If $\frac{|\Gamma_i^{\emptyset}|}{q^{\emptyset}} \geq (|\mathcal{R}| - 2)\frac{|\Gamma_i^{r^*}|}{q^{r^*}}$, then for all t, $\hat{d}_i(t) \leq d_i(t)$.*

This proposition shows that the degree of agent i decreases from $d_i(t)$ to $\hat{d}_i(t)$ for every t following a core subgroup expansion if the effective network size is large enough.

[17] Note that $\frac{|\Gamma_i^{\emptyset}|}{q^{\emptyset}} \geq (|\Gamma_i^r| + \bar{\delta})\frac{|\Gamma_i^r|}{q^r}$ is a sufficient but not a necessary condition. For example, one can verify that $\lim_{p\to 1^-} \hat{d}_i(t) \geq \lim_{p\to 1^-} d_i(t)$ regardless of effective social group sizes.

[18] Note that if the network itself were allowed to expand by δ, that is, $|\hat{\pi}_i(\{r\})| = |\pi_i(\{r\})| + \delta$ and $|\hat{\pi}_i(\emptyset)| = |\pi_i(\emptyset)| + \delta$, then one could easily show that $\hat{d}_i(t)$ would lie above $d_i(t)$ for every t.

Note that when there are only two social categories ($|\mathcal{R}| = 2$), the sufficient condition holds trivially. A detailed explanation of core subgroup expansions and some intuition for this result is given by returning our running example.

Example. *(cont.)* On the right-hand side of Figure 3 (top of panel (b)) we illustrate a core subgroup expansion by δ for Alice. So $|\hat{\pi}_{\mathrm{A}}(\{1,2\})| = |\pi_{\mathrm{A}}(\{1,2\})| + \delta$, and $|\hat{\pi}_{\mathrm{A}}(\{1\})| = |\pi_{\mathrm{A}}(\{1\})| - \delta$, and $|\hat{\pi}_{\mathrm{A}}(\{2\})| = |\pi_{\mathrm{A}}(\{2\})| - \delta$, and $|\hat{\pi}_{\mathrm{A}}(\emptyset)| = |\pi_{\mathrm{A}}(\emptyset)| + \delta$. If there were any other subgroups, their cardinality would remain unchanged. Furthermore, note that the cardinality of *every* social group remains unchanged. That is, $|\hat{\Gamma}^r_{\mathrm{A}}| = |\Gamma^r_{\mathrm{A}}|$ for all $r \in \mathcal{R} \cup \{\emptyset\}$. Somewhat less abstractly, a core subgroup expansion in this case would be equivalent to δ Econ students from dorm Y switching to Alice's dorm (dorm X), and δ Math students from Alice's dorm switching to dorm Y.[19]

Now, since the cardinality of every social group Γ^r_{A} ($r \in \mathcal{R} \cup \{\emptyset\}$) is unchanged, the expected depletion times remain unchanged. On the right-hand side of Figure 3 (bottom of panel (b)) we represent degree as a function of time for Alice before the expansion by $d_{\mathrm{A}}(t)$ (in black) and after the expansion by $\hat{d}_{\mathrm{A}}(t)$ (in green). The slope of $\hat{d}_{\mathrm{A}}(t)$ is smaller than that of $d_{\mathrm{A}}(t)$ over some range because after having depleted the core subgroup $\pi_{\mathrm{A}}(\{1,2\})$, Alice is likely to deplete the singleton subgroups next, but these are now smaller. The slope of $\hat{d}_{\mathrm{A}}(t)$ eventually becomes larger than that of $d_{\mathrm{A}}(t)$ because the $\emptyset$-subgroup is larger after the expansion. However, we show that the functions never cross and so $\hat{d}_{\mathrm{A}}(t) \leq d_{\mathrm{A}}(t)$ for all t. ■

Remark 1. Under the relevant conditions, Proposition 3 (Proposition 4) shows that the function for degree over time is higher (lower) in every period following a singleton (core) subgroup expansion. Since the probability of idleness is unchanged from an expansion, it follows that the expectation for the *realized* degree is higher (lower) following a singleton (core) subgroup expansion. ■

4.1. Supporting evidence from Facebook data

We use Facebook data to provide some supporting evidence for the theoretical results of this section. Our data represent a September 2005 cross-section of the complete

[19] Note that the definition of core subgroup expansion allows the number of agents transferred from a singleton subgroup to the core subgroup to be different for different singleton subgroups. That is, we could transfer δ_1 units from $\pi_i(\{1\})$ to $\pi_i(\{1,2\})$, and $\delta_2 \neq \delta_1$ units from $\pi_i(\{2\})$ to $\pi_i(\{1,2\})$. If we then transfer δ_1 units from the $\emptyset$-subgroup to $\pi_i(\{1\})$ and δ_2 from the $\emptyset$-subgroup to $\pi_i(\{2\})$, the number of agents and the cardinality of all social groups remains unchanged. Note that this system of transfers is equivalent to the definition of core subgroup expansion when $\delta = \delta_1 + \delta_2$ and can easily be generalized to more than two social categories.

structures of social connections on www.facebook.com *within* (but not across) the first ten American colleges and universities that joined Facebook. The (anonymized) raw data contain over 130,000 nodes (users) and over 5.6 million links (friendships). We observe six social categories for each user: gender, year of graduation, major, minor, dorm, and high school. We cleaned the data as described in Appendix III. There are 27,454 users and 492,236 links in our cleaned data, consisting only of students graduating between 2006 and 2009, who have supplied all the relevant personal characteristics (except high school).[20] We provide a more complete description of the dataset in Appendix III.

Using the available information in our data, we define agents i and j to be in the same class if and only if they have the same year of graduation and major or have the same year of graduation and minor. We then let $[K^1, K^2] = [\text{class, dorm}]$ to match the running example of this paper. That is, we assumed that every student in our dataset interacts with other students in their "class" social group Γ_i^1 and with other students in their "dorm" social group Γ_i^2. Naturally, these groups often overlap.

Remark 2. To match our Example, suppose there are only two social categories ($|\mathcal{R}| = 2$). Also note that any two agents i and j are distinguished only by their type. That is, for a given p and $\mathbf{q}$, if i and j have social groups and overlaps across those social groups that are of the same size, their degrees and expected depletion times should be the same; the labels of the social groups within the social categories are therefore irrelevant, only their sizes matter. Proposition 4 can therefore be read as saying that if i and j have the same social group sizes but i's core social subgroup is larger, then i's expected degree will be smaller than j's. Similarly, under the relevant conditions, Proposition 3 can be read as saying that if only one of i's social groups is larger than j's (and i and j have core subgroups of the same size), then i's expected degree is larger than j's. ■

To test the implications in Remark 2, we ran the following regression for each college

$$d_i = \beta_0 + \beta_1 |\Gamma_i^1| + \beta_2 |\Gamma_i^2| + \beta_3 |\pi_i(\{1,2\})| + \epsilon_i \tag{8}$$

That is, we assumed that the degree of student i depends on the size of i's "class" social group, the size of i's "dorm" social group, and on the size of the intersection of these groups. The parameter β_1 is interpreted as the marginal effect on degree of increasing the size of the "class" social group *holding the size of the "dorm" social group, and of*

[20] Technically, we consider a non-random subsample of the data since there might be selection biases in data disclosure preferences. But while cleaning the data may affect the value of the parameter estimates, we do not expect their *sign* to be significantly affected, which is the only aspect that is relevant for testing our comparative static results.

	Dependent variable: agent's degree							
	$\lvert\Gamma_i^1\rvert$	s.e.	$\lvert\Gamma_i^2\rvert$	s.e.	$\lvert\pi_i(\{1,2\})\rvert$	s.e.	const.	N
Harvard	0.170***	(0.032)	0.239***	(0.034)	-0.273	(0.308)	9.002	1325
Columbia	0.149***	(0.022)	0.012	(0.012)	-0.627***	(0.141)	33.94	2663
Stanford	0.319***	(0.031)	0.071***	(0.021)	-1.995***	(0.429)	35.53	2254
Yale	0.035	(0.043)	0.056**	(0.023)	0.518	(0.458)	25.06	1431
Cornell	0.034**	(0.017)	0.002	(0.003)	-0.308***	(0.096)	21.07	2509
Dartmouth	0.200***	(0.037)	-0.035	(0.036)	-0.689	(0.512)	37.04	1612
UPenn	0.153***	(0.018)	-0.018***	(0.005)	-0.427***	(0.128)	37.55	3006
MIT	0.063**	(0.032)	-0.028	(0.018)	-0.328	(0.274)	42.38	1563
NYU	0.085***	(0.012)	0.020***	(0.003)	-0.218***	(0.061)	23.98	5581
Boston U.	0.091***	(0.011)	0.008***	(0.002)	-0.274***	(0.044)	26.02	5510

Comment: Standard OLS regression with robust standard errors in parentheses

***/**/* denote rejection of $H_0 : \beta = 0$ at the 1/5/10% significant level respectively

Table 1: Regression results

the intersection of the groups constant. Similarly, β_3 is interpreted as the marginal effect on degree of increasing the size of the core subgroup *holding the size of the social groups constant.* Given Remark 2, we should find β_1 and β_2 to be positive and β_3 to be negative.[21]

The evidence is reported in Table 1 and largely supports the results of this section, with β_1 and β_2 appearing as positive, and β_3 as negative. We therefore learn that the definition of a social group matters for the resulting comparative statics. If one were to define agents in $\pi_i(\{1,2\})$ as constituting their own "group", then one would be surprised to find that expanding that group has a negative effect on degree. On the other hand, defining the set of agents $\pi_i(\{1,2\})$ as the intersection of social groups yields the result established in our model.

We have only considered singleton subgroup and core subgroup expansions. While many other types of expansion exist, the results of this section suffice to highlight the importance of accounting for the overlap across social groups.

5. Homophily

We now turn to the dynamics of homophily in our model and show that they depend crucially on the relative sizes of social groups and on their overlaps. First of all, for any

[21] From Appendix III, note that the average dorm social group size in our dataset is 50.5 and the average class social group size is 62.8, but the network size $|\Gamma_i^\emptyset|$ is two orders of magnitude larger, so we can assume that the sufficient condition for a large enough effective network size in Proposition 3 is satisfied.

agent i, the *individual homophily index* in social category $r \in \mathcal{R}$ is given by[22]

$$\frac{\text{number of friends of } i \text{ that share characteristic } k_i^r \text{ with } i}{\text{number of friends of } i} \tag{9}$$

To express the individual homophily index within our model, let us define $\Pi_i^r = \{\pi_i(S) \in \Pi_i : S \subseteq \mathcal{R}, r \in S\}$. This is the set of subgroups that contain agents who share i's characteristic in social category r. Using Equation (9), the individual homophily index in social category r of agent i in period t is

$$H_i^r(t) = \frac{\sum_{\pi \in \Pi_i^r} d_i^\pi(t)}{\sum_{\pi \in \Pi_i} d_i^\pi(t)} = \frac{\sum_{\pi \in \Pi_i^r} d_i^\pi(t)}{d_i(t)} \tag{10}$$

Proposition 5. *For any agent $i \in N$ and any $r \in \mathcal{R}$, the function $H_i^r(t)$ is equal to the constant $\sum_{\pi \in \Pi_i^r} q^\pi$ for any $t \leq T^{\pi_i(\mathcal{R})}$, is decreasing for any $t \in (T^{\pi_i(\{r\})}, T^{\pi_i(\emptyset)}]$, and is equal to the constant $\frac{\sum_{\pi \in \Pi_i^r} q^\pi T^\pi}{\sum_{\pi \in \Pi_i} q^\pi T^\pi}$ for any $t > T^{\pi_i(\emptyset)}$.*

This result shows that the homophily index for agent i is a constant for any t up to the expected depletion time of the core subgroup $T^{\pi_i(\mathcal{R})}$, and the index then decreases from $T^{\pi_i(\{r\})}$ to some constant. The fact that homophily decreases over some range is intuitive: *Within a social group*, agents deplete subgroups from those that contain agents who share the largest number of characteristics with them to those that contain agents who share the smallest number of characteristics. Therefore, as time passes, the agents that are added as friends later in time from that social group are likely to be less similar to the agent, which reduces the homophily index. Note that this feature of homophily eventually decreasing over time is specific to our model, which captures the Rolodex aspect of online social networks. Indeed, rather than recording close friendships, such networks tend to become a repository of past acquaintances. This contrasts with "best" friendship networks reported in surveys in which we observe that homophily in certain social categories increases over time (e.g. in race among school students, see Shrum et al. 1988, McPherson et al. 2001).

A feature to highlight in Proposition 5 is that we cannot determine the shape of $H_i^r(t)$ in the range $(T^{\pi_i(\mathcal{R})}, T^{\pi_i(\{r\})}]$ without further restrictive assumptions. That is, while homophily must eventually decrease over time, it is possible for it to *increase over*

[22] Boucher (2012) and Iijima and Kamada (2014) analyze static network formation models in which agents have multi-dimensional characteristics. They focus their attention on a homophily index which measures the distance between any two vectors of characteristics. This differs somewhat from the usual measure typically encountered (McPherson et al., 2001, Currarini et al., 2009) and used in this paper.

some range. This non-monotonicity can have important implications, which we return to after Proposition 6 (a special case of Proposition 5 for two social categories).

Proposition 6. *Suppose that* $\mathcal{R} = \{r, r'\}$ *and that Assumption 1 holds. Then, for any agent* $i \in N$*,*

i. If $\frac{|\Gamma_i^r|}{q^r} < \frac{|\Gamma_i^{r'}|}{q^{r'}}$*, then* $H_i^r(t)$ *is decreasing for all* t*.*

ii. If $\frac{|\Gamma_i^r|}{q^r} > \frac{|\Gamma_i^{r'}|}{q^{r'}}$ *and* $|\Gamma_i^r \cap \Gamma_i^{r'}|$ *is sufficiently small, then* $H_i^r(t)$ *is non-monotonic and reaches a global maximum at* $T^{\pi_i(\{r\})} \gg 0$*.*

The proposition above states that when there are only two relevant social categories, the homophily index for the social category of the effectively smaller social group is decreasing over time. On the other hand, if the intersection of the social groups is sufficiently small, the homophily index for the social category of the effectively larger social group increases over some range until it reaches a global maximum at $T^{\pi_i(\{r\})}$ before decreasing down to a constant. (More specifically, the proof of the second part of Proposition 6 reveals that $H_i^r(t)$ is constant in the range $(0, T^{\pi_i(\mathcal{R})}]$, is decreasing in the range $(T^{\pi_i(\mathcal{R})}, T^{\pi_i(\{r'\})}]$, is increasing in the range $(T^{\pi_i(\{r'\})}, T^{\pi_i(\{r\})}]$, is decreasing in the range $(T^{\pi_i(\{r\})}, T^{\pi_i(\emptyset)}]$, and is finally constant for any $t > T^{\pi_i(\emptyset)}$).

We can provide some intuition for this result by returning to our running example.

Example. *(cont.)* Suppose we are interested in Alice's homophily index over time for the "class" social category, that is, $H_{\mathrm{A}}^1(t)$. We illustrate the first part of Proposition 6 on the left-hand side of Figure 4 (panel (a)), and the second part of Proposition 6 on the right-hand side of Figure 4 (panel (b)). Suppose $|\Gamma_{\mathrm{A}}^1| < \frac{q^1}{q^2}|\Gamma_{\mathrm{A}}^2|$. Then Alice will first deplete the core subgroup, followed by the "class" subgroup $\pi_{\mathrm{A}}(\{1\})$, followed by the "dorm" subgroup $\pi_{\mathrm{A}}(\{2\})$, and finally by the $\emptyset$-subgroup. Therefore, Alice makes friends with the students who share her characteristic within the "class" social category first, and therefore her homophily index decreases over time. On the other hand, suppose $|\Gamma_{\mathrm{A}}^1| > \frac{q^1}{q^2}|\Gamma_{\mathrm{A}}^2|$. Now, Alice will first deplete the core subgroup, followed by the "dorm" subgroup $\pi_{\mathrm{A}}(\{2\})$, followed by the "class" subgroup $\pi_{\mathrm{A}}(\{1\})$, and finally by the $\emptyset$-subgroup (Notice the switched expected depletion times in Figure 4). So, having depleted the core subgroup, Alice will make friends with students from within the "dorm" subgroup, which reduces her homophily index in the "class" social category. But, having depleted the "dorm" subgroup $\pi_{\mathrm{A}}(\{2\})$, Alice continues making friends from within the "class" subgroup $\pi_{\mathrm{A}}(\{1\})$ which will increase her homophily index. If, in addition, the size of the core subgroup $\pi_{\mathrm{A}}(\{1, 2\})$ is sufficiently small, this increase in her homophily

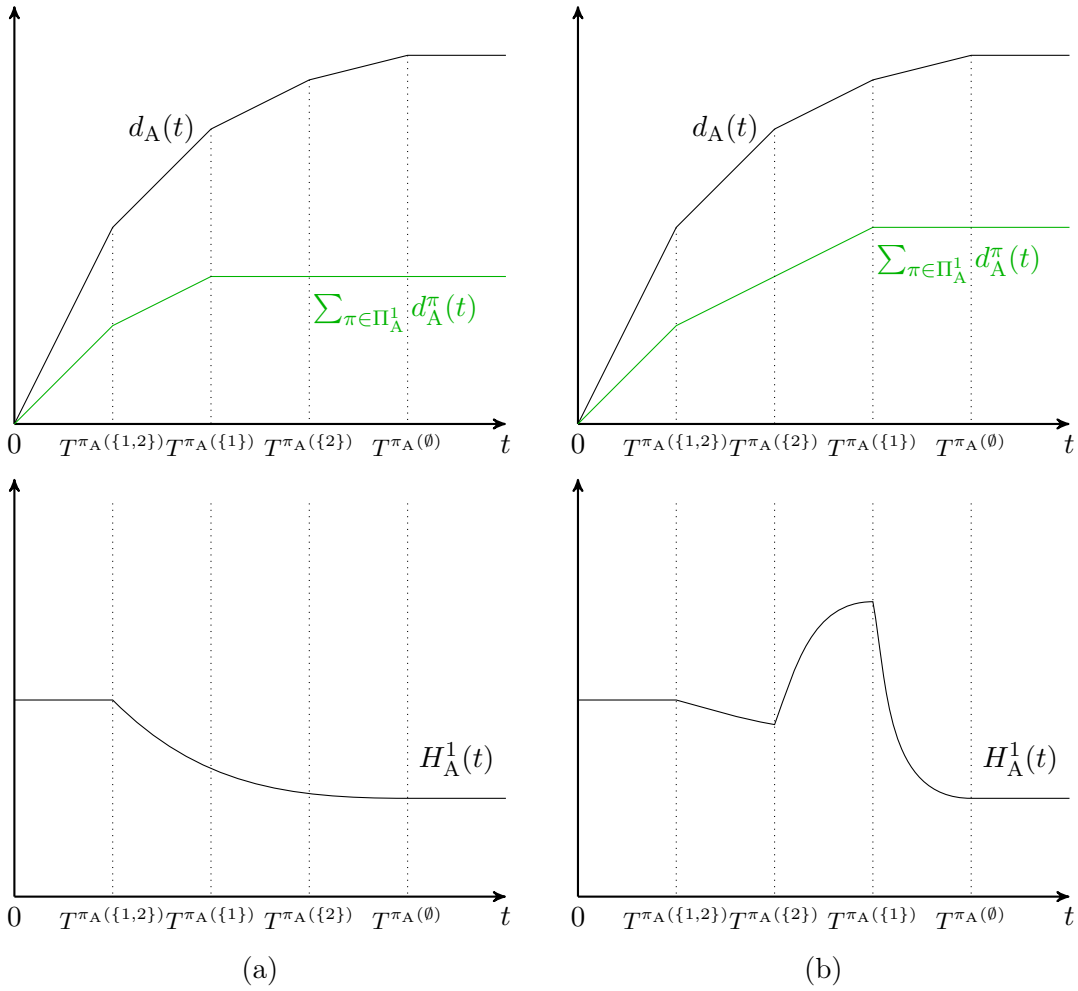


Figure 4: An illustration of Proposition 6 using the Example

index reaches a global maximum (as shown in panel (b) of Figure 4). The reason for this is that if the core subgroup is small (and since the "dorm" social group is smaller than the "class" social group), Alice's homophily index will have started off low in early
545 periods so the eventual addition of new friends who are similar to her in the relevant social category has a large effect on her homophily index.[23] ■

Bramoullé et al. (2012) derive a negative relationship between homophily and time (or

[23]Looking at panel (b) of Figure 4, while the slope of $d_{\mathrm{A}}(t)$ must be everywhere greater than the slope of $\sum_{\pi\in\Pi^1_{\mathrm{A}}} d^{\pi}_{\mathrm{A}}(t)$, it is not necessarily true that the *growth rate* of $d_{\mathrm{A}}(t)$ is everywhere greater than that of $\sum_{\pi\in\Pi^1_{\mathrm{A}}} d^{\pi}_{\mathrm{A}}(t)$, and that is what would be required for $H^1_{\mathrm{A}}(t)$ decreasing for all t.

indeed between homophily and degree). As Propositions 5 and 6 show, this relationship holds only under certain conditions in our model. The novel prediction of our analysis is that under the condition of part (ii) of Proposition 6, homophily increases over some range, which offers a way to distinguish empirically between the models.

The non-monotonicity of the homophily index over time may affect various dynamic economic processes that occur in social networks. For example, Golub and Jackson (2012) find that increasing homophily tends to reduce the speed of information transmission because the information will tend to circulate rapidly within but not across clusters of similar individuals. If homophily is indeed non-monotonic, as we demonstrate, then our results suggest that information will transmit quickest through the "older" nodes or through the "youngest" nodes in the network (for example, through seniors and freshmen respectively, in the context of an American university), but may transmit more slowly among "middle-aged" nodes.[24]

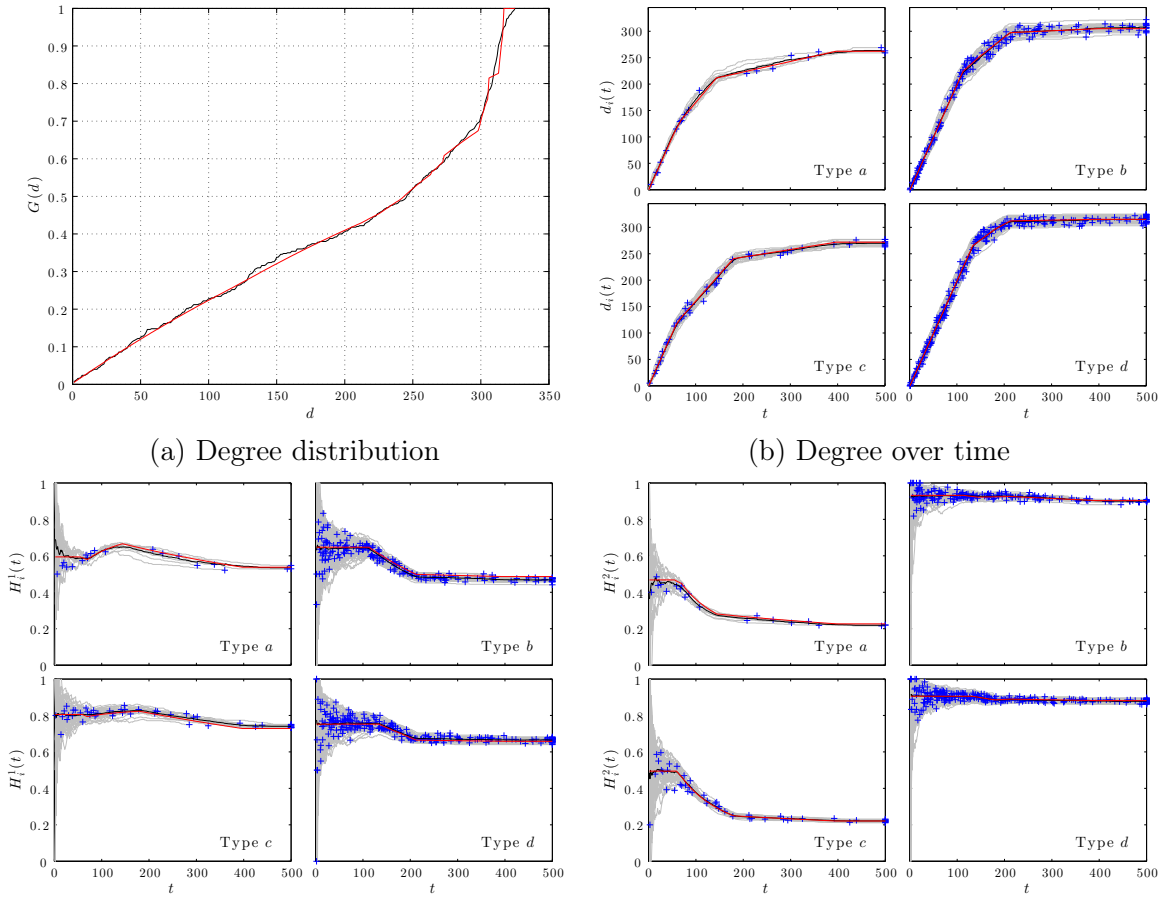
6. Simulations

Since we derived our analytical results using a mean-field approximation, it is fruitful to verify that the approximation is consistent with simulation results. We do this by returning to our running example.

Example. *(cont.)* In line with our running example, we have 20 type a students, 180 type b students, 50 type c students, and finally 250 type d students. We chose $q^1 = q^2 = 0.4$ and $p = 0.995$. In Figure 5 we can see the performance of our approximation against a *single* run of the simulation.[25] Panel (a) shows the degree distribution resulting from one run of the simulation in black against our analytical degree distribution, corresponding to Equation (7), in red. Panel (b) shows our predicted degree over time for each agent type (in red) against the simulation results. Grey lines trace the degree over time for each agent as long as they are active, and blue crosses indicate the point at which each agent becomes idle. The black line shows the average degree over time across all the active agents of a given type. Similarly, panels (c) and (d) show our predicted path for

[24] It is unfortunately not possible for us to test the non-monotonicity of the homophily index over time with our cross-sectional dataset. While it is possible for us to discard the time dimension by defining the composition function $h_i^\tau(d) = (H_i^\tau \circ t_i)(d)$ which expresses homophily as a function of degree, the function $h_i^\tau(d)$ will typically be different for different agents. Since we only have a single observation for degree and for the homophily index in a given category for each agent, we cannot test what $h_i^\tau(d)$ looks like empirically for a given agent.

[25] That is, the simulation was initialized with the chosen parameters and terminated only when every student became idle.



(a) Degree distribution

(b) Degree over time

(c) Homophily in social category 1

(d) Homophily in social category 2

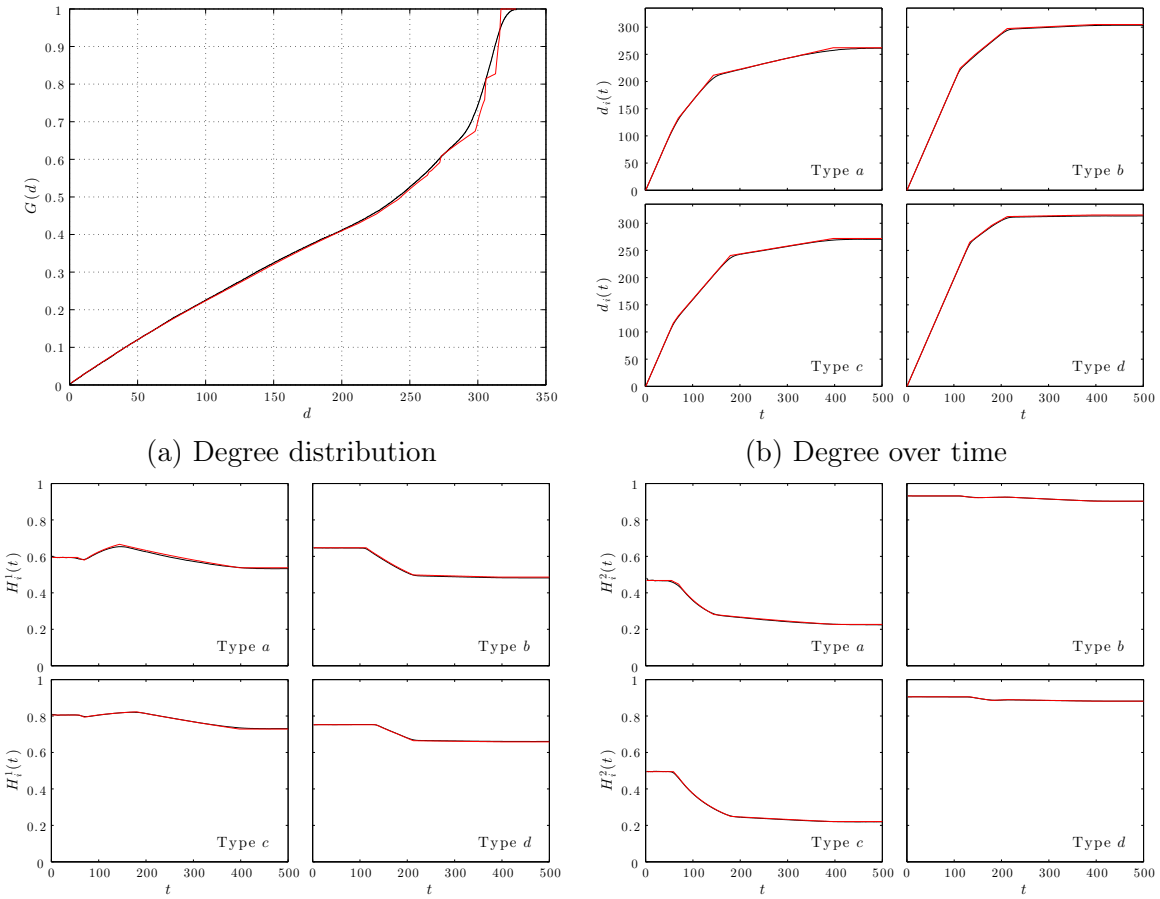
Figure 5: Performance of the mean-field approximation against one run of the simulation

homophily over time in red, in social categories 1 and 2 respectively, for each agent type. Grey lines trace the homophily index over time for each agent as long as they are active, and blue crosses indicate the point at which each agent becomes idle. Also, the black line shows the average homophily over time across all the active agents of a given type.

The parametrization used to generate Figure 5 was chosen to highlight the fact that homophily over time can be non-monotonic (see the top left of panel (c)), but Figure 5 is representative of how well our approximation matches the simulation results.

Figure 6 uses the same parametrization to show the performance of our approximation against the average of 100 runs of the simulation. Panel (a) shows degree distribution averaged across 100 runs of the simulation in black against our analytical degree distribution in red. In panels (b) to (d), the black lines show the average degree (or homophily) over time across all the active agents of a given type averaged across 100 runs of the

simulation. The red lines in those panels show our analytical predictions. Once again, Figure 6 is representative of how well our approximation matches the simulations. ■



(a) Degree distribution

(b) Degree over time

(c) Homophily in social category 1

(d) Homophily in social category 2

Figure 6: Performance of the mean-field approximation against 100 runs of the simulation

While our approximation matches the simulation output well, it is not indistinguishable from the output for the following reasons: (i) The expected depletion times are the same for any two agents of the same type. So in our mean-field approximation, agents of the same type all abruptly and simultaneously stop making new friends with others from a particular social group. Due to the randomness in the simulation however, agents of the same type may deplete their social groups at slightly different times. The implication is that the average degree over time across active agents will be slightly smoother than our

prediction for degree as a concave piecewise linear function.[26] (ii) In deriving the degree over time for an agent i, we added the expected number of friendships that i initiates with other agents and the expected number of friendships that are initiated by other agents j with i in every period t. However, we do not subtract the intersection: Namely, we ignore the possibility that i initiates a friendship with some agent j in period t while j simultaneously initiates a friendship with i in that period. The likelihood of such an event is small when the pools of active agents are large, but becomes more significant when the pools are close to being completely depleted.

In an Online Appendix, we present further simulations of the model (and of variants of the model) for different values of p and $\mathbf{q}$.[27]

7. Conclusion

We presented a new dynamic model of network formation over a fixed number of agents with overlapping social groups. We derived some comparative static results on the relationship between degree and group size, showing that degree should increase with social group size, but should decrease when the overlap of social groups is increased. We gave some supportive evidence using data from Facebook. We also showed that homophily can be non-monotonic, reaching a global maximum in some period before eventually falling. Future work could empirically investigate the testable implications for the dynamics of degree and homophily of this model.

8. Acknowledgments

We are grateful to the advisory editor and to two anonymous referees for valuable comments that have substantially improved this paper. We also thank Dan Beary, Vincent Crawford, Francis Dennig, Fulya Ersoy, Peter Eső, Marcel Fafchamps, Edo Gallo, Bernie Hogan, Matthew Jackson, Alan Kirman, Paul Klemperer, Rachel Kranton, Manuel Mueller-Frank, Raviv Murciano-Goroff, John Quah, Simon Quinn, Zaifu Yang and Peyton Young for their suggestions as well as seminar participants at the Workshop for Information in Networks 2012, WINE 2012, RES 2013 Conference, SAET 2013, EEA 2013 Conference, APET Workshop on Diversity and Public Policy, Université de Montréal, St Andrews, Leicester, George Mason, Baruch College and the Oxford

[26] Evidence of this can be seen in panel (a) of Figure 6: For degrees greater than approximately 275 and particularly for degrees between 300 and 325, our analytic degree distribution (in red) is visibly more "jagged" than the degree distribution generated from the simulation.

[27] The MATLAB code for running the simulations is available at http://users.ox.ac.uk/~scat3580/MATLABFriending.zip.

Internet Institute. Tarbush gratefully acknowledges the generous support of the Royal Economic Society.

Appendix I: Proofs

Baseline results

In this section, we first carry out the analysis in terms of pairs of agent types. We then extend the analysis to obtain the formulae in the main text which express the relevant variables (such as probabilities of interaction, expected depletion times, etc) in terms of an agent i and their corresponding partition Π_i.

For any agent $i \in N$ of type k and any type $k' \in \mathcal{K}$ denote by $q^{N_i(k')}$ the probability with which agent i interacts with agents of type k'. Also, for any set X, let $X_+ = X \cup \{\emptyset\}$.

Lemma 3. *Consider an agent i of type k and any type $k' \in \mathcal{K}$. Denote by S the set of indices for which the vectors k and k' are equal. Then, (i) $q^{N_i(k')} = \sum_{r \in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|}$, and (ii) for any other agent j of type k, $q^{N_i(k')} = q^{N_j(k')}$.*

Proof. (i) Agent i interacts with agents in $N_i(k') \subseteq \Gamma_i^r$ with positive probability only if $r \in S_+$. Furthermore, for any $r \in S_+$ agent i interacts with agents in the social group Γ_i^r with probability q^r, and conditional on interacting with agents in Γ_i^r, agent i interacts with those in $N_i(k') \subseteq \Gamma_i^r$ with probability $\frac{|N_i(k')|}{|\Gamma_i^r|}$. Therefore, an agent interacts with agents in $N_i(k')$ with probability

$$\sum_{r \in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|} \tag{11}$$

(ii) If agents i and j are of the same type, then for all $k' \in \mathcal{K}$, $|N_i(k')| = |N_j(k')|$, and for all $r \in S_+$, $|\Gamma_i^r| = |\Gamma_j^r|$. □

Since all agents of the same type have the same probability of interacting with agents of a given type (Lemma 3 part (ii)), for any types $k, k' \in \mathcal{K}$ we can let $\alpha^{kk'}$ denote the probability with which an agent of type k interacts with agents of type k'. That is for any $i \in N$ such that $k_i = k$, $\alpha^{kk'} = q^{N_i(k')}$.[28]

For any types $k, k' \in \mathcal{K}$, let $Z^{kk'}(t)$ denote the remaining number of agents of type k' that an agent of type k can still make a link with in period t. That is, for agent i of type k, $Z^{kk'}(t)$ is the number of agents of type k' who are active in period t and who have neither received a friend request from i (that they accepted) nor sent a friend request to i (that i accepted) before period t. Also denote by $T^{kk'}$ the period in which an agent of type k can no longer make any new links with agents of type k' (because they are all

[28] The probability $\alpha^{kk'}$ resembles what Currarini et al. (2009, 2010) refer to as the probability of two agent types "meeting".

either friends with the agent of type k or are idle). That is, $T^{kk'}$ is the smallest t for which $Z^{kk'}(t) = 0$.

Remark 3. For any $k, k' \in \mathcal{K}$, although it is not necessarily the case that $Z^{kk'}(t) = Z^{k'k}(t)$, it must be the case that $T^{kk'} = T^{k'k}$. Indeed, the initial number of type k agents, which is equal to $Z^{k'k}(0)$, may differ from the initial number of type k' agents, which is equal to $Z^{kk'}(0)$. The fact that $T^{kk'} = T^{k'k}$ simply follows from the fact that if $T^{kk'}$ is the period in which agents of type k can no longer make new links with agents of type k', then it also must be the period in which agents of type k' can no longer make new links with agents of type k. ■

Denote by $\Delta d_i^{kk'}(t)$ the number of friends that an agent i of type k makes with agents of type k' in period t. Note that the index "i" in $\Delta d_i^{kk'}(t)$ is in some sense redundant because any two agents of the same type are indistinguishable as long as they are active. Nevertheless, the notation is useful for tracking any two agents of the same type who may differ in their *realized* times of idleness, and thus in their *realized* degrees. The equations for $Z^{kk'}(t)$ and for $\Delta d_i^{kk'}(t)$, as well as their relationship, are derived below.

The equation for $\Delta d_i^{kk'}(t)$. We find the equation for $\Delta d_i^{kk'}(t)$ by deriving the expected number of (immediately accepted) friend requests that an agent i of type k sends to agent of type k' in period t and the expected number of friend requests that an agent i of type k receives (and immediately accepts) from agents of type k' in period t.

In period t an agent i of type k interacts with agents of type k' with probability $\alpha^{kk'}$. Provided that $t \leq T^{kk'}$ agent i sends an immediately accepted friend request to one of the remaining type k' agents. If $t > T^{kk'}$ agent i makes no new links with agents of type k' in period t. Therefore the expected number of immediately accepted friend requests that an agent i of type k sends to agents of type k' in period t is given by

$$\alpha^{kk'} \mathbf{1}(t \leq T^{kk'}) \tag{12}$$

Similarly, in period t, an agent j of type k' interacts with agents of type k with probability $\alpha^{k'k}$. Provided that $t \leq T^{k'k} = T^{kk'}$ (see Remark 3), and conditional on interacting with agents of type k, agent j (of type k') sends an immediately accepted friend request to some remaining agent of type k (who is not idle and is not yet friends with j). There are precisely $Z^{k'k}(t)$ agents of type k that an agent j of type k' can still make new links with in period t. Since j selects the specific agent that the friend request is sent to uniformly at random, the probability that it is specifically agent i who receives the friend request is therefore $\frac{1}{Z^{k'k}(t)}$. Finally, there are precisely $Z^{kk'}(t)$ agents j of type

k' from which an agent i of type k could accept a friend request from in period t. From the above, it follows the expected number of friend requests that an agent i of type k receives (and immediately accepts) from agents of type k' in period t is given by

$$\alpha^{k'k}\frac{Z^{kk'}(t)}{Z^{k'k}(t)}\mathbf{1}(t \leq T^{kk'}) \tag{13}$$

The total expected number of friends that an agent i of type k makes with agents of type k' in period t is therefore given by the sum of Equations (12) and (13). That is,

$$\Delta d_i^{kk'}(t) = \alpha^{kk'}\mathbf{1}(t \leq T^{kk'}) + \alpha^{k'k}\frac{Z^{kk'}(t)}{Z^{k'k}(t)}\mathbf{1}(t \leq T^{kk'}) \tag{14}$$

The equation for $Z^{kk'}(t)$. For any $t < T^{kk'}$ the equation for $Z^{kk'}(t)$ is given by

$$Z^{kk'}(t+1) = Z^{kk'}(t) - \left[\Delta d_i^{kk'}(t) + (1-p)Z^{kk'}(t) - (1-p)\Delta d_i^{kk'}(t)\right] \tag{15}$$

The interpretation of Equation (15) is straightforward. The number of remaining active agents of type k' that an agent i of type k can make a link with in period $t+1$ is the number of active agents of type k' at t less the number of such agents that have either become idle or that become friends with i at t. This includes the agents who became friends with i at t $[\Delta d_i^{kk'}(t)]$ and those who have become idle at t $[(1-p)Z^{kk'}(t)]$ and excludes the ones who became friends with i at t *and* have become idle at t $[(1-p)\Delta d_i^{kk'}(t)]$.

Lemma 4. *For any* $k, k' \in \mathcal{K}$*, the growth rates of* $Z^{kk'}(t)$ *and* $Z^{k'k}(t)$ *are equal for all* t*.*

Proof. For any $t \leq T^{kk'}$, Equation (15) for $Z^{kk'}(t)$ can be written as

$$\begin{aligned} Z^{kk'}(t+1) - Z^{kk'}(t) &= -\left[(1-p)Z^{kk'}(t) + p\Delta d_i^{kk'}(t)\right] &(16)\\ &= -\left[(1-p)Z^{kk'}(t) + p\left(\alpha^{kk'} + \alpha^{k'k}\frac{Z^{kk'}(t)}{Z^{k'k}(t)}\right)\right] &(17)\end{aligned}$$

where Equation (14) was used to obtain the second line. Equation (17) can then be re-arranged to obtain

$$\frac{Z^{kk'}(t+1) - Z^{kk'}(t)}{Z^{kk'}(t)} = -\left[(1-p) + p\left(\frac{\alpha^{kk'}}{Z^{kk'}(t)} + \frac{\alpha^{k'k}}{Z^{k'k}(t)}\right)\right] \tag{18}$$

Similarly, from the equation for $Z^{k'k}(t)$ one obtains

$$\frac{Z^{k'k}(t+1) - Z^{k'k}(t)}{Z^{k'k}(t)} = -\left[(1-p) + p\left(\frac{\alpha^{k'k}}{Z^{k'k}(t)} + \frac{\alpha^{kk'}}{Z^{kk'}(t)}\right)\right] \tag{19}$$

The right hand sides of Equations (18) and (19) are identical, from which it follows that their left hand sides are also equal. Therefore the growth rates of $Z^{kk'}(t)$ and $Z^{k'k}(t)$ are equal. □

Lemma 4 is key in rendering the mean-field approximation analytically tractable by allowing us to simplify the expression for $\Delta d_i^{kk'}(t)$ as shown in the following corollary.

Corollary 2. *Equation (14) for $\Delta d_i^{kk'}(t)$ can be written as $\Delta d_i^{kk'}(t) = 2\alpha^{kk'}\mathbf{1}(t \leq T^{kk'})$.*

Proof. Consider agents i and j of types k and k' respectively. It follows from Lemma 4 that $\frac{Z^{kk'}(t)}{Z^{k'k}(t)}$ is a constant and equal to $\frac{Z^{kk'}(0)}{Z^{k'k}(0)}$. Equation (14) can therefore be re-written as

$$\Delta d_i^{kk'}(t) = \alpha^{kk'}\mathbf{1}(t \leq T^{kk'}) + \alpha^{k'k}\frac{Z^{kk'}(0)}{Z^{k'k}(0)}\mathbf{1}(t \leq T^{kk'}) \tag{20}$$

$$= \sum_{r \in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|}\mathbf{1}(t \leq T^{kk'}) + \sum_{r \in S_+} q^r \frac{|N_j(k)|}{|\Gamma_j^r|}\frac{|N_i(k')|}{|N_j(k)|}\mathbf{1}(t \leq T^{kk'}) \tag{21}$$

$$= 2\sum_{r \in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|}\mathbf{1}(t \leq T^{kk'}) \tag{22}$$

The second line follows from the first by the definitions of $\alpha^{kk'}$ and $\alpha^{k'k}$ (S denotes the set of indices for which the vectors k and k' are equal), and by the fact that $Z^{kk'}(0)$ is simply the initial number of agents of type k' and is therefore equal to $|N_i(k')|$. Similarly, $Z^{k'k}(0)$ is the initial number of agents of type k and is equal to $|N_j(k)|$. The the third line follows from the second by the definition of social groups: Since $r \in S_+$, $i \in \Gamma_j^r$ and $j \in \Gamma_i^r$; from which it follows that $|\Gamma_i^r| = |\Gamma_j^r|$. □

Corollary 2 shows that the expected number of agents of type k' that an agent i of type k becomes friends with in period t is a constant and is equal to $2\alpha^{kk'}$ for any $t \leq T^{kk'}$. In fact, the expected number of agents of type k' that agent i sends an (immediately accepted) friend request to in period t is $\alpha^{kk'}$, and the expected number of friend requests that i receives (and immediately accepts) from agents of type k' in period t is also equal to $\alpha^{kk'}$.

Having carried out the analysis in terms of interacting pairs of types, we now aggregate our results over elements of Π_i for any agent i to obtain the formulae stated in

the main text. To do this, for any agent i and any $\pi \in \Pi_i$, it will be useful to define $\mathcal{K}^\pi = \{k' \in \mathcal{K} : N_i(k') \subseteq \pi\}$. That is, $\mathcal{K}^\pi$ is the set of types of all the agents in the subgroup $\pi \in \Pi_i$.

Proof of Lemma 1. Consider any subset of indices $S \subseteq \mathcal{R}$. The probability $q^{\pi_i(S)}$ with which an agent i of type k interacts with agents in the subgroup $\pi_i(S)$ is the sum of the probabilities with which i interacts with each type in $\mathcal{K}^{\pi_i(S)}$.[29] That is,

$$q^{\pi_i(S)} = \sum_{k' \in \mathcal{K}^{\pi_i(S)}} \alpha^{kk'} = \sum_{k' \in \mathcal{K}^{\pi_i(S)}} \sum_{r \in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|} \tag{23}$$

$$= \sum_{r \in S_+} q^r \frac{|\pi_i(S)|}{|\Gamma_i^r|} \tag{24}$$

The second line follows from the fact that $\sum_{k' \in \mathcal{K}^{\pi_i(S)}} |N_i(k')| = |\pi_i(S)|$. □

For an agent i of type k and any subgroup $\pi \in \Pi_i$, the expected time it takes for all the agents of type $k' \in \mathcal{K}^\pi$ to either become friends with i or to become idle is given by $T^{kk'}$. The following lemma shows that for any $k', k'' \in \mathcal{K}^\pi$, $T^{kk'}$ is the same as $T^{kk''}$ and their value is given by Equation (25).

Lemma 5. *Consider an agent i of type k. For any $S \subseteq \mathcal{R}$ and any $k' \in \mathcal{K}^{\pi_i(S)}$,*

$$T^{kk'} = \frac{\ln\left(\frac{2q^{\pi_i(S)}p}{2q^{\pi_i(S)}p + (1-p)|\pi_i(S)|}\right)}{\ln(p)} \tag{25}$$

Proof. By Lemma 2, for any $t \leq T^{kk'}$, we can re-write Equation (16) as

$$Z^{kk'}(t+1) - Z^{kk'}(t) = -\left[(1-p)Z^{kk'}(t) + p2\alpha^{kk'}\right] \tag{26}$$

Solving this difference equation with $Z^{kk'}(0) = |N_i(k')|$, we obtain

$$Z^{kk'}(t) = |N_i(k')|p^t + \frac{2\alpha^{kk'}p(p^t - 1)}{1-p} \tag{27}$$

Solving Equation (27) with $Z^{kk'}(T^{kk'}) = 0$ yields the expression for $T^{kk'}$ which is given

[29] Obviously, for any $k', k'' \in \mathcal{K}^{\pi_i(S)}$, the set of indices such that k and k' are equal is the set as the set of indices such that k and k'' are equal, and that set of indices is simply S.

by

$$T^{kk'} = \frac{\ln\left(\frac{2\alpha^{kk'}p}{2\alpha^{kk'}p+(1-p)|N_i(k')|}\right)}{\ln(p)} = \frac{\ln\left(\frac{2\sum_{r\in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|}p}{2\sum_{r\in S_+} q^r \frac{|N_i(k')|}{|\Gamma_i^r|}p+(1-p)|N_i(k')|}\right)}{\ln(p)} \tag{28}$$

$$= \frac{\ln\left(\frac{2\sum_{r\in S_+} q^r \frac{|\pi_i(S)|}{|\Gamma_i^r|}p}{2\sum_{r\in S_+} q^r \frac{|\pi_i(S)|}{|\Gamma_i^r|}p+(1-p)|\pi_i(S)|}\right)}{\ln(p)} = \frac{\ln\left(\frac{2q^{\pi_i(S)}p}{2q^{\pi_i(S)}p+(1-p)|\pi_i(S)|}\right)}{\ln(p)} \tag{29}$$

The first line simply uses the definition of $\alpha^{kk'}$ (noting that the set of indices S for which the vectors k and k' are equal is the same for any $k' \in \mathcal{K}^{\pi_i(S)}$). To obtain the second line, we simply replace $|N_i(k')|$ by $|\pi_i(S)|$ in both the top and bottom of the fraction that appears in the numerator, and then use the definition of $q^{\pi_i(S)}$. □

We are finally in a position to prove Proposition 1.

Proof of Proposition 1. Consider an agent i of type k and any $S \subseteq \mathcal{R}$. The expected depletion time $T^{\pi_i(S)}$ is the expected time it takes for every agent in $\pi_i(S)$ to either become friends with i or to become idle. Since $T^{kk'}$, the expected time it takes for every agent in $N_i(k') \subseteq \pi_i(S)$ to become friends with i or to become idle, is the same for each $k' \in \mathcal{K}^{\pi_i(S)}$ (see Lemma 5), it follows that $T^{\pi_i(S)} = T^{kk'}$. Therefore, for any $k' \in \mathcal{K}^{\pi_i(S)}$, $\Delta d_i^{kk'}(t) = 2\alpha^{kk'}\mathbf{1}(t \leq T^{\pi_i(S)})$. The number of friends i adds in period t from the set $\pi_i(S) \in \Pi_i$ is the sum of the friends that i adds in period t from each set $N_i(k') \subseteq \pi_i(S)$ for $k' \in \mathcal{K}^{\pi_i(S)}$. That is

$$\Delta d_i^{\pi_i(S)}(t) = \sum_{k'\in\mathcal{K}^{\pi_i(S)}} \Delta d_i^{kk'}(t) = \sum_{k'\in\mathcal{K}^{\pi_i(S)}} 2\alpha^{kk'}\mathbf{1}(t \leq T^{\pi_i(S)}) = 2q^{\pi_i(S)}\mathbf{1}(t \leq T^{\pi_i(S)}) \tag{30}$$

The last step follows from Equation (23). □

Proof of Corollary 1. In period $t = 0$, every agent i has no friends. Solving Equation (30) with the initial condition $d_i^{\pi_i(S)}(0) = 0$ gives

$$d_i^{\pi_i(S)}(t) = 2q^{\pi_i(S)}\left[t\mathbf{1}(t \leq T^{\pi_i(S)}) + T^{\pi_i(S)}\mathbf{1}(t > T^{\pi_i(S)})\right] \tag{31}$$

The degree of agent i in period t is therefore given by Equation (6) below

$$d_i(t) = \sum_{\pi\in\Pi_i} d_i^{\pi}(t) = 2\sum_{\pi\in\Pi_i} q^{\pi}\left[t\mathbf{1}(t \leq T^{\pi}) + T^{\pi}\mathbf{1}(t > T^{\pi})\right] \tag{32}$$

Note that $d_i(t)$ is a continuous concave piecewise linear function that is strictly increasing in the range $(0, \max_{\pi \in \Pi_i}\{T^\pi\}]$. □

715 Incidentally, it should be clear from the above proof that in every period $t \leq T^\pi$, agent i initiates an expected q^π links with agents in $\pi \in \Pi_i$, *and* an expected number q^π of i's links are initiated by agents in $\pi \in \Pi_i$.

Proof of Lemma 2. Using Equations (2) and (4), we obtain

$$T^{\pi_i(S)} = \frac{\ln\left(\frac{2p\left[\sum_{r\in S_+}\frac{q^r}{|\Gamma_i^r|}\right]}{2p\left[\sum_{r\in S_+}\frac{q^r}{|\Gamma_i^r|}\right]+(1-p)}\right)}{\ln(p)} \tag{33}$$

This equation immediately shows that if $S' \subseteq S$, then $\sum_{r\in S'_+}\frac{q^r}{|\Gamma_i^r|} \leq \sum_{r\in S_+}\frac{q^r}{|\Gamma_i^r|}$ and therefore $T^{\pi_i(S')} \geq T^{\pi_i(S)}$ (since $\ln(p) < 0$). □

Proof of Proposition 2. Since $d_i(t)$ is increasing, we can find its inverse in the range $(0, d_i(\max_{\pi\in\Pi_i}\{T^\pi\})]$, which is given by

$$d_i^{-1}(d) = t_i(d) = \frac{d - 2\sum_{\pi\in\Pi_i} q^\pi T^\pi \mathbf{1}(d > d_i(T^\pi))}{2\sum_{\pi\in\Pi_i} q^\pi \mathbf{1}(d \leq d_i(T^\pi))} \tag{34}$$

We now obtain $G_i(d)$ – the probability that agent i has degree at most d (degree distribution of agent i).

$$\Pr(d_i(t) \leq d) = \Pr(d_i^{-1}(d_i(t)) \leq d_i^{-1}(d)) = \Pr(t \leq t_i(d)) = G_i(d) \tag{35}$$

Since an agent i remains active exactly x periods with probability $p^x(1-p)$, we have that

$$\Pr(t \leq x) = \sum_{t=0}^{t=x} p^t(1-p) = 1 - p^{x+1} \tag{36}$$

Therefore, the degree distribution of agent i is given by

$$G_i(d) = \Pr(t \leq t_i(d)) = 1 - p^{t_i(d)+1} \tag{37}$$

Finally, the overall degree distribution $G(d)$ is the average of the degree distributions across all agents and is given by

$$G(d) = \frac{1}{|N|}\sum_{i\in N}\left(1 - p^{t_i(d)+1}\right) \tag{38}$$

□

Comparative statics on degree and group size

We introduce the following notation which will be useful for the rest of the proofs. Let $\mathcal{V} = \{1, ..., V\}$. For some sequence of distinct terms $S_0, S_1, S_2, ..., S_V$ such that $T^{\pi_i(S_0)} = 0$, and in which $S_1, ..., S_V$ are the subsets of $\mathcal{R}$ partially ordered by set inclusion,[30] Lemma 2 allows us to order the expected depletion times as

$$T^{\pi_i(S_0)} < T^{\pi_i(S_1)} < T^{\pi_i(S_2)} < \cdots < T^{\pi_i(S_V)} \tag{39}$$

Assumption 1 guarantees the inequalities to be strict. Note that $V = 2^{|\mathcal{R}|}$, and that $S_V = \emptyset$, and that $S_1 = \mathcal{R}$. For any $r \in \mathcal{R}$, let us define $\mathcal{V}_r = \{j \in \mathcal{V} : r \in S_j\}$ and $\mathcal{V}_{\neg r} = \{j \in \mathcal{V} : r \notin S_j\}$. This notation allows us to re-express Equation (6) for the degree of agent i over time as

$$\begin{aligned} d_i(t) &= 2 \sum_{\pi \in \Pi_i} q^{\pi} \left[t\mathbf{1}(t \leq T^{\pi}) + T^{\pi}\mathbf{1}(t > T^{\pi}) \right] \\ &= 2 \sum_{v=1}^{V} q^{\pi_i(S_v)} \left[t\mathbf{1}(t \leq T^{\pi_i(S_v)}) + T^{\pi_i(S_v)}\mathbf{1}(t > T^{\pi_i(S_v)}) \right] \end{aligned} \tag{40}$$

Using Equation (40) note that for any $v \in \mathcal{V}$, and any t in the interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$, the function $d_i(t)$ can be written as

$$d_i(t) = d_i\left(T^{\pi_i(S_{v-1})}\right) + 2\left(t - T^{\pi_i(S_{v-1})}\right) \sum_{j=v}^{V} q^{\pi_i(S_j)} \tag{41}$$

where $d_i(T^{\pi_i(S_0)}) = 0$. Notably, this implies that for any $v \in \mathcal{V}$ and any t in the interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$, the function $d_i(t)$ can be written as

$$d_i(t) = 2 \left[\sum_{j=1}^{v-1} q^{\pi_i(S_j)} T^{\pi_i(S_j)} + t \sum_{j=v}^{V} q^{\pi_i(S_j)} \right] \tag{42}$$

Proof of Proposition 3. In this proof, we show that the expected depletion times are shifted rightwards after the singleton subgroup expansion of $\pi_i(\{r\})$. Furthermore, the slope of the degree function over time is larger for every $t \leq T^{\pi_i(\{r\})}$ after the expansion, from which it follows that $\hat{d}_i(t) \geq d_i(t)$ for every $t \leq T^{\pi_i(\{r\})}$. In addition, the slope of

[30] For any S_v with $v > 1$, there is a set $S_{v'}$ with $v' < v$ $(v, v' \in \mathcal{V})$ such that $S_v \subseteq S_{v'}$.

the degree function over time is weakly smaller for every $t \in (T^{\pi_i\{r\}}, T^{\pi_i(\emptyset)}]$ after the expansion, but we can show that for any $t \geq T^{\pi_i(\emptyset)}$, $\hat{d}_i(t)$ and $d_i(t)$ are both flat and satisfy $\hat{d}_i(t) \geq d_i(t)$ when the effective size of $\Gamma_i^{\emptyset}$ is large compared to that of Γ_i^r. From this it follows that the functions never cross and therefore $\hat{d}_i(t) \geq d_i(t)$ for all t.

For reference, let us set $S_{v^*} = \{r\}$. Assumption 1 allows us to order the expected depletion times before the expansion as in Equation (39). From Equation (33) one can verify that $T^{\pi_i(S)}$ is increasing in the cardinality of Γ_i^r for any $S \subseteq \mathcal{R}$ such that $r \in S$. The singleton subgroup expansion of $\pi_i(\{r\})$ results in an increase of the cardinality of the social group Γ_i^r, but leaves the cardinality of any other social group unchanged. From this it follows that for any $v \in \mathcal{V}$ such that $v > v^*$, $T^{\pi_i(S_v)} = T^{\hat{\pi}_i(S_v)}$, and for any $v \leq v^*$, $T^{\pi_i(S_v)} \leq T^{\hat{\pi}_i(S_v)}$. Finally, $T^{\pi_i(S_0)} = T^{\hat{\pi}_i(S_0)} = 0$. Now, Equation (33) shows that expected depletion times vary continuously with social group sizes, so for a sufficiently small $\delta < \bar{\delta}$, the order of expected depletion times after the expansion becomes

$$T^{\hat{\pi}_i(S_0)} < T^{\hat{\pi}_i(S_1)} < T^{\hat{\pi}_i(S_2)} < \cdots < T^{\hat{\pi}_i(S_V)} \tag{43}$$

where the sequence of the sets $S_0, S_1, ..., S_V$ is unchanged.

From Equation (2), and by definition of the singleton subgroup expansion, one can verify that for any $v \in \mathcal{V}$,

$$q^{\hat{\pi}_i(S_v)} = \begin{cases} q^{\emptyset} \frac{|\pi_i(S_v)|-\delta}{|\Gamma_i^{\emptyset}|} & \text{if } v = V \\ q^{\emptyset} \frac{|\pi_i(S_v)|}{|\Gamma_i^{\emptyset}|} + q^r \frac{|\pi_i(S_v)|}{|\Gamma_i^r|+\delta} + \sum_{h \in S_v \setminus \{r\}} q^h \frac{|\pi_i(S_v)|}{|\Gamma_i^h|} & \text{if } v \in \mathcal{V}_r \setminus \{v^*\} \\ q^{\emptyset} \frac{|\pi_i(S_v)|+\delta}{|\Gamma_i^{\emptyset}|} + q^r \frac{|\pi_i(S_v)|+\delta}{|\Gamma_i^r|+\delta} & \text{if } v = v^* \\ q^{\emptyset} \frac{|\pi_i(S_v)|}{|\Gamma_i^{\emptyset}|} + \sum_{h \in S_v} q^h \frac{|\pi_i(S_v)|}{|\Gamma_i^h|} & \text{if } v \in \mathcal{V}_{\neg r} \setminus \{V\} \end{cases} \tag{44}$$

For any $v \in \mathcal{V}$ the slope of $\hat{d}_i(t)$ in the non-empty interval $(T^{\hat{\pi}_i(S_{v-1})}, T^{\hat{\pi}_i(S_v)}]$ is given by $2\sum_{j=v}^{V} q^{\hat{\pi}_i(S_j)}$ (see Equation (42)). We will compare this slope with the slope of $d_i(t)$ in the corresponding non-empty interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$, which is given by $2\sum_{j=v}^{V} q^{\pi_i(S_j)}$. That is, for any $v \in \mathcal{V}$ we need to evaluate the expression

$$2\sum_{j=v}^{V} \left(q^{\hat{\pi}_i(S_j)} - q^{\pi_i(S_j)} \right) \tag{45}$$

Now, note that for any $v \in \mathcal{V}_r$, $v \leq v^*$. That is, if we consider a $v \leq v^*$, then at least some set in $\{S_v, S_{v+1}, ..., S_{v^*}\}$ must contain r. But, for every $v > v^*$, $r \notin S_v$. Note also that since $|\mathcal{R}| > 1$, $v^* > 1$.

Suppose $v \leq v^*$. By Equation (44) one can verify that Equation (45) is smallest when $v = 1$, and in which case its value is zero. So for any $v \leq v^*$, $\hat{d}_i(t)$ has a steeper slope (at least weakly) in the interval $(T^{\hat{\pi}_i(S_{v-1})}, T^{\hat{\pi}_i(S_v)}]$ than $d_i(t)$ in the corresponding interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$. From the above, we have that $\hat{d}_i(t)$ lies above $d_i(t)$ for every $t \leq T^{\pi_i(\{r\})}$.

Suppose $v > v^*$. The summation in Equation (45) is therefore over sets S_v not containing r. In this case it follows that $2\sum_{j=v}^{V}\left(q^{\hat{\pi}_i(S_j)} - q^{\pi_i(S_j)}\right) = -2\frac{q^{\emptyset}\delta}{|\Gamma_i^{\emptyset}|}$ which is negative. So the slope of $d_i(t)$ over any interval in $(T^{\pi_i(\{r\})}, T^{\pi_i(\emptyset)}]$ must be greater than that of $\hat{d}_i(t)$ in the corresponding interval.

However, both $\hat{d}_i(t)$ and $d_i(t)$ are flat for any $t \geq T^{\hat{\pi}_i(\emptyset)} = T^{\pi_i(\emptyset)}$. And we now show that when the effective size of $\Gamma_i^{\emptyset}$ is large compared to that of Γ_i^r, $\hat{d}_i(T^{\pi_i(\emptyset)}) \geq d_i(T^{\pi_i(\emptyset)})$, from which it follows that the functions never cross, and therefore $\hat{d}_i(t)$ lies above $d_i(t)$ for every t. Using Equation (42) one can verify that $\hat{d}_i(T^{\pi_i(\emptyset)}) = 2\sum_{v=1}^{V} q^{\hat{\pi}_i(S_v)} T^{\hat{\pi}_i(S_v)}$ and that $d_i(T^{\pi_i(\emptyset)}) = 2\sum_{v=1}^{V} q^{\pi_i(S_v)} T^{\pi_i(S_v)}$. Since for every $v \in \mathcal{V}$, $T^{\hat{\pi}_i(S_v)} \geq T^{\pi_i(S_v)}$, for $\hat{d}_i(T^{\pi_i(\emptyset)}) \geq d_i(T^{\pi_i(\emptyset)})$ to hold it suffices to show that

$$\sum_{v=1}^{V} q^{\hat{\pi}_i(S_v)} T^{\pi_i(S_v)} \geq \sum_{v=1}^{V} q^{\pi_i(S_v)} T^{\pi_i(S_v)} \tag{46}$$

We can think of $\left(q^{\hat{\pi}_i(S_v)}\right)_{v\in\mathcal{V}}$ and of $\left(q^{\pi_i(S_v)}\right)_{v\in\mathcal{V}}$ as being probability distributions over $\left(T^{\pi_i(S_v)}\right)_{v\in\mathcal{V}}$. To show that the expectation of the former is greater than that of the latter, it suffices to show that the former second-order stochastically dominates the latter. That is, we must establish that for every $h \in \mathcal{V}$, $\sum_{j=1}^{h}\sum_{v=1}^{j}\left(q^{\hat{\pi}_i(S_v)} - q^{\pi_i(S_v)}\right) \leq 0$.

Using Equation (44), one can verify that $\sum_{v=1}^{j}\left(q^{\hat{\pi}_i(S_v)} - q^{\pi_i(S_v)}\right)$ is equal to

$$\begin{cases} -q^r \frac{\delta}{|\Gamma_i^r|(|\Gamma_i^r|+\delta)} \sum_{v\in\mathcal{V}_r\cap\{1,..,j\}} |\pi_i(S_v)| & \text{if } 1 \leq j < v^* \\ q^{\emptyset}\frac{\delta}{|\Gamma_i^{\emptyset}|} & \text{if } v^* \leq j < V \\ 0 & \text{if } j = V \end{cases} \tag{47}$$

Therefore $\sum_{j=1}^{h}\sum_{v=1}^{j}\left(q^{\hat{\pi}_i(S_v)} - q^{\pi_i(S_v)}\right) \leq 0$ is equal to

$$\begin{cases} -q^r \frac{\delta}{|\Gamma_i^r|(|\Gamma_i^r|+\delta)} \sum_{j=1}^{h}\sum_{v\in\mathcal{V}_r\cap\{1,..,j\}} |\pi_i(S_v)| & \text{if } 1 \leq h < v^* \\ (h-(v^*-1))q^{\emptyset}\frac{\delta}{|\Gamma_i^{\emptyset}|} - q^r \frac{\delta}{|\Gamma_i^r|(|\Gamma_i^r|+\delta)} \sum_{j=1}^{v^*-1}\sum_{v\in\mathcal{V}_r\cap\{1,..,j\}} |\pi_i(S_v)| & \text{if } v^* \leq h < V \\ (V-v^*)q^{\emptyset}\frac{\delta}{|\Gamma_i^{\emptyset}|} - q^r \frac{\delta}{|\Gamma_i^r|(|\Gamma_i^r|+\delta)} \sum_{j=1}^{v^*-1}\sum_{v\in\mathcal{V}_r\cap\{1,..,j\}} |\pi_i(S_v)| & \text{if } h = V \end{cases} \tag{48}$$

The top line is negative, and the second line is negative if the expression in the third line is negative. Therefore, we obtain the desired result if the third line is (weakly) negative. That is,

$$\frac{|\Gamma_i^{\emptyset}|}{q^{\emptyset}} \geq (|\Gamma_i^r| + \delta) \left[\frac{V - v^*}{\sum_{j=1}^{v^*-1} \sum_{v \in \mathcal{V}_r \cap \{1,..,j\}} |\pi_i(S_v)|} \right] \frac{|\Gamma_i^r|}{q^r} \tag{49}$$

We now show that the fraction in square brackets in Equation (49) is at most 1. The smallest value of v^* is $2^{|\mathcal{R}|-1}$, since it is the index of the first singleton set; and $V = 2^{|\mathcal{R}|}$. Therefore the largest value of the numerator is $2^{|\mathcal{R}|} - 2^{|\mathcal{R}|-1}$. Now consider the denominator: Since every set $N_i(k)$ contains at least one agent, every social subgroup contains at least one agent, so $|\pi_i(S_v)| \geq 1$ for each $v \in \mathcal{V}$. The sum in the denominator is equal to $(2^{|\mathcal{R}|-1} - 1)|\pi_i(\mathcal{R})|$ when $|\mathcal{R}| = 2$, and it is at least $(2^{|\mathcal{R}|-1} - 1)|\pi_i(\mathcal{R})|$ plus at least one other $|\pi_i(S)|$ for some S such that $r \in S$ when $|\mathcal{R}| > 2$. Therefore, the sum in the denominator is at least $2^{|\mathcal{R}|-1}$ when $|\mathcal{R}| > 2$. This therefore establishes that for *any* $|\mathcal{R}| > 1$, the fraction in square brackets in Equation (49) is at most 1. Finally, since $\delta < \bar{\delta}$, the condition in Equation (49) holds if $\frac{|\Gamma_i^{\emptyset}|}{q^{\emptyset}} \geq (|\Gamma_i^r| + \bar{\delta}) \frac{|\Gamma_i^r|}{q^r}$. □

Proof of Proposition 4. This proofs follows similar steps to the ones in the proof of Proposition 3. Assumption 1 allows us to order the expected depletion times before the expansion as in Equation (39). However, since the core subgroup expansion leaves the cardinality of *every* social group Γ_i^r ($r \in \mathcal{R}_+$) unchanged, it follows from Equation (33) that $T^{\pi_i(S)} = T^{\hat{\pi}_i(S)}$ for every $S \subseteq \mathcal{R}$.

From Equation (2), and by definition of the core subgroup expansion, one can verify that for any $v \in \mathcal{V}$,

$$q^{\hat{\pi}_i(S_v)} = \begin{cases} q^{\emptyset} \frac{|\pi_i(S_v)| + (|\mathcal{R}|-1)\delta}{|\Gamma_i^{\emptyset}|} & \text{if } S_v = \emptyset \\ q^{\emptyset} \frac{|\pi_i(S_v)| - \delta}{|\Gamma_i^{\emptyset}|} + q^r \frac{|\pi_i(S_v)| - \delta}{|\Gamma_i^r|} & \text{if } S_v = \{r\} \text{ for } r \in \mathcal{R} \\ q^{\emptyset} \frac{|\pi_i(S_v)| + \delta}{|\Gamma_i^{\emptyset}|} + \sum_{h \in \mathcal{R}} q^h \frac{|\pi_i(S_v)| + \delta}{|\Gamma_i^h|} & \text{if } S_v = \mathcal{R} \\ q^{\emptyset} \frac{|\pi_i(S_v)|}{|\Gamma_i^{\emptyset}|} + \sum_{h \in S_v} q^h \frac{|\pi_i(S_v)|}{|\Gamma_i^h|} & \text{otherwise} \end{cases} \tag{50}$$

For any $v \in \mathcal{V}$ the slope of $\hat{d}_i(t)$ in the non-empty interval $(T^{\hat{\pi}_i(S_{v-1})}, T^{\hat{\pi}_i(S_v)}]$ is given by $2\sum_{j=v}^{V} q^{\hat{\pi}_i(S_j)}$ (see Equation (42)). We will compare this slope with the slope of $d_i(t)$ in the corresponding non-empty interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$, which is given by $2\sum_{j=v}^{V} q^{\pi_i(S_j)}$. That is, for any $v \in \mathcal{V}$ we need to evaluate the expression given in Equation (45).

Recall that $S_1 = \mathcal{R}$ and $S_V = \emptyset$, and let $\tilde{\mathcal{V}}$ denote the set of indices v such that S_v is

a singleton. Let $\underline{v}$ be the smallest index and $\bar{v}$ be the largest index in $\tilde{\mathcal{V}}$, and note that $\bar{v} = V - 1$, and that $\underline{v}$ must be at least $2^{|\mathcal{R}|-1}$. Also, let $X_n = \tilde{\mathcal{V}} \cap \{n, ..., V\}$ denote the set of indices of singleton sets S_v for which $v \geq n$. Note finally that $|\tilde{\mathcal{V}}| = |\mathcal{R}|$. One can verify that $\sum_{j=v}^{V} \left(q^{\hat{\pi}_i(S_j)} - q^{\pi_i(S_j)}\right)$ is equal to

$$\begin{cases} 0 & \text{if } v = 1 \\ -\delta \frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|} - \delta \sum_{r \in \mathcal{R}} \frac{q^r}{|\Gamma_i^r|} & \text{if } 1 < v \leq \underline{v} \\ \delta \frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(|\mathcal{R}| - 1 - |X_v|) - \delta \sum_{r \in \cup_{h \in X_v} S_h} \frac{q^r}{|\Gamma_i^r|} & \text{if } \underline{v} < v \leq \bar{v} \\ \delta \frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(|\mathcal{R}| - 1) & \text{if } v = V \end{cases} \tag{51}$$

Equation (51) shows that the difference between the slope of $\hat{d}_i(t)$ and $d_i(t)$ is negative for all $t \in (T^{\pi_i(S_0)}, T^{\pi_i(S_{v^*})}]$, and switches to being positive for any $t > T^{\pi_i(S_{v^*})}$, for some $v^* > \underline{v}$. This shows that $\hat{d}_i(t)$ lies below $d_i(t)$ for every $t \leq T^{\pi_i(S_{v^*})}$, and that $\hat{d}_i(t)$ eventually has a steeper slope than $d_i(t)$. However, both $\hat{d}_i(t)$ and $d_i(t)$ are flat for any $t \geq T^{\hat{\pi}_i(\emptyset)} = T^{\pi_i(\emptyset)}$. And we now show that $\hat{d}_i(T^{\pi_i(\emptyset)}) \leq d_i(T^{\pi_i(\emptyset)})$, from which it follows that they never cross, and therefore $\hat{d}_i(t)$ lies below $d_i(t)$ for every t.

Using Equation (42), $\hat{d}_i(T^{\pi_i(\emptyset)}) \leq d_i(T^{\pi_i(\emptyset)})$ holds if

$$\sum_{v=1}^{V} q^{\hat{\pi}_i(S_v)} T^{\pi_i(S_v)} \leq \sum_{v=1}^{V} q^{\pi_i(S_v)} T^{\pi_i(S_v)} \tag{52}$$

Equation (52) holds if the distribution $\left(q^{\pi_i(S_v)}\right)_{v \in \mathcal{V}}$ second-order stochastically dominates $\left(q^{\hat{\pi}_i(S_v)}\right)_{v \in \mathcal{V}}$, and this is true if Equation (53) is positive for all $h \in \mathcal{V}$.

$$\sum_{j=1}^{h} \sum_{v=1}^{j} \left(q^{\hat{\pi}_i(S_v)} - q^{\pi_i(S_v)}\right) \tag{53}$$

Let $X^n = \tilde{\mathcal{V}} \cap \{1, ..., n\}$ denote the set of indices of singleton sets S_v for which $v \leq n$. For any $j \in \mathcal{V}$ we have that $\sum_{v=1}^{j} \left(q^{\hat{\pi}_i(S_v)} - q^{\pi_i(S_v)}\right)$ is equal to

$$\begin{cases} \delta \frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|} + \delta \sum_{r \in \mathcal{R}} \frac{q^r}{|\Gamma_i^r|} & \text{if } 1 \leq j < \underline{v} \\ -\delta \frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(|X^j| - 1) + \delta \sum_{r \in \mathcal{R} \setminus (\cup_{h \in X^j} S_h)} \frac{q^r}{|\Gamma_i^r|} & \text{if } \underline{v} \leq j < \bar{v} \\ -\delta \frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(|\mathcal{R}| - 1) & \text{if } j = \bar{v} \\ 0 & \text{if } j = V \end{cases} \tag{54}$$

The second line of Equation (54) is smallest for $j = \bar{v} - 1$, and is positive if

$$\frac{|\Gamma_i^{\emptyset}|}{q^{\emptyset}} \geq (|\mathcal{R}| - 2)\frac{|\Gamma_i^{r^*}|}{q^{r^*}} \tag{55}$$

where r^* satisfies $S_{\bar{v}} = \{r^*\}$. If Equation (55) holds, then the second line of Equation (54) is positive. One can then verify that for any $1 \leq h < \bar{v}$, Equation (53) is a sum of positive terms, and (by the first line of Equation (54)), must be at least $\delta\frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(\underline{v} - 1)$, and therefore must be at least $\delta\frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(2^{|\mathcal{R}|-1} - 1)$. At $h = \bar{v}$, we subtract $\delta\frac{q^{\emptyset}}{|\Gamma_i^{\emptyset}|}(|\mathcal{R}| - 1)$, but the difference remains positive for any positive $|\mathcal{R}|$. Therefore, if Equation (55) holds, Equation (53) is positive for all $h \in \mathcal{V}$, as desired. □

Homophily

Consider any sequence of expected depletion times $T^{\pi_i(S_0)}, T^{\pi_i(S_1)}, ..., T^{\pi_i(S_V)}$, ordered according to Equation (39). Similarly to Equation (42), for any Π_i^r induced by $r \in \mathcal{R}$, and for any t in the interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$ with $v \in \mathcal{V}$, the function $\sum_{\pi\in\Pi_i^r} d_i^{\pi}(t)$ can be written as

$$\sum_{\pi\in\Pi_i^r} d_i^{\pi}(t) = 2\left[\sum_{j\in\{1,...,v-1\}\cap\mathcal{V}_r} q^{\pi_i(S_j)}T^{\pi_i(S_j)} + t\sum_{j\in\{v,...,V\}\cap\mathcal{V}_r} q^{\pi_i(S_j)}\right] \tag{56}$$

Finally, from Equations (42) and (56), one can verify that for any $v \in \mathcal{V}$ and any t in the interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$ the slope of $H_i^r(t)$ is negative if and only if

$$\frac{\sum_{j\in\{v,...,V\}\cap\mathcal{V}_r} q^{\pi_i(S_j)}}{\sum_{j\in\{1,...,v-1\}\cap\mathcal{V}_r} q^{\pi_i(S_j)}T^{\pi_i(S_j)}} \leq \frac{\sum_{j\in\{v,...,V\}} q^{\pi_i(S_j)}}{\sum_{j\in\{1,...,v-1\}} q^{\pi_i(S_j)}T^{\pi_i(S_j)}} \tag{57}$$

Equation (57) shows that $H_i^r(t)$ is decreasing in the interval $(T^{\pi_i(S_{v-1})}, T^{\pi_i(S_v)}]$ if the growth rate of $d_i(t)$ is greater than the growth rate of $\sum_{\pi\in\Pi_i^r} d_i^{\pi}(t)$ in that interval.

Proof of Proposition 5. Using Equation (42), since $(T^{\pi_i(S_0)}, T^{\pi_i(S_1)}] = (0, T^{\pi_i(\mathcal{R})}]$, for any t in this interval, $d_i(t) = 2\sum_{\pi\in\Pi_i} q^{\pi}t$. But note that $\sum_{\pi\in\Pi_i} q^{\pi} = 1$. Similarly, using Equation (56), $\sum_{\pi\in\Pi_i^r} d_i^{\pi}(t) = 2\sum_{\pi\in\Pi_i^r} q^{\pi}t$, therefore $H_i^r(t)$ is equal to $\sum_{\pi\in\Pi_i^r} q^{\pi}$. Since $T^{\pi_i(S_V)} = T^{\pi_i(\emptyset)}$, using Equation (42), we find that for any $t \in (T^{\pi_i(\emptyset)}, \infty)$, $d_i(t) = 2\sum_{\pi\in\Pi_i} q^{\pi}T^{\pi}$. Similarly, $\sum_{\pi\in\Pi_i^r} d_i^{\pi}(t) = 2\sum_{\pi\in\Pi_i^r} q^{\pi}T^{\pi}$, therefore $H_i^r(t)$ is equal to $\frac{\sum_{\pi\in\Pi_i^r} q^{\pi}T^{\pi}}{\sum_{\pi\in\Pi_i} q^{\pi}T^{\pi}}$. Finally, the slope of $\sum_{\pi\in\Pi_i^r} d_i^{\pi}(t)$ is zero for any $t > T^{\pi_i(\{r\})}$, while the slope of $d_i(t)$ is positive for any $t \leq T^{\pi_i(\emptyset)}$. It follows that the slope of $H_i^r(t)$ is negative for any $t \in (T^{\pi_i(\{r\})}, T^{\pi_i(\emptyset)}]$. □

Proof of Proposition 6. Let $\mathcal{R} = \{r, r'\}$. From Equation (33) one can verify that $T^{\pi_i(\{r\})} < T^{\pi_i(\{r'\})}$ if and only if $\frac{|\Gamma_i^r|}{q^r} < \frac{|\Gamma_i^{r'}|}{q^{r'}}$. Supposing Assumption 1 holds, we obtain the following ordering of the expected depletion times for the first part of the proposition:

$$0 < T^{\pi_i(\{r,r'\})} < T^{\pi_i(\{r\})} < T^{\pi_i(\{r'\})} < T^{\pi_i(\emptyset)} \tag{58}$$

And, we obtain the following ordering for the second part of the proposition:

$$0 < T^{\pi_i(\{r,r'\})} < T^{\pi_i(\{r'\})} < T^{\pi_i(\{r\})} < T^{\pi_i(\emptyset)} \tag{59}$$

By Proposition 5, $H_i^r(t)$ is a constant for any $t \in (0, T^{\pi_i(\{r,r'\})}] \cup (T^{\pi_i(\emptyset)}, \infty)$, and it is decreasing in the range $(T^{\pi_i(\{r\})}, T^{\pi_i(\emptyset)}]$. Therefore it only remains for us to show that the relevant conditions hold in the interval $(T^{\pi_i(\{r,r'\})}, T^{\pi_i(\{r\})}]$ for the first part of Proposition 6, and in the intervals $(T^{\pi_i(\{r,r'\})}, T^{\pi_i(\{r'\})}]$ and $(T^{\pi_i(\{r'\})}, T^{\pi_i(\{r\})}]$ for the second part.

Let us focus on the first part of Proposition 6. We will verify that the slope of $H_i^r(t)$ is negative in the interval $(T^{\pi_i(S_1)}, T^{\pi_i(S_2)}] = (T^{\pi_i(\{r,r'\})}, T^{\pi_i(\{r\})}]$. Applying Equation (57) in this interval by setting $v = 2$, we obtain the following inequality which holds trivially,

$$\frac{q^{\pi_i(\{r\})}}{q^{\pi_i(\{r,r'\})} T^{\pi_i(\{r,r'\})}} \leq \frac{q^{\pi_i(\{r'\})} + q^{\pi_i(\{r\})} + q^{\pi_i(\emptyset)}}{q^{\pi_i(\{r,r'\})} T^{\pi_i(\{r,r'\})}} \tag{60}$$

Now, let us focus on the second part of Proposition 6. Firstly, we show that $H_i^r(t)$ is decreasing in the interval $(T^{\pi_i(S_1)}, T^{\pi_i(S_2)}] = (T^{\pi_i(\{r,r'\})}, T^{\pi_i(\{r'\})}]$. Applying Equation (57) in this interval, we once again obtain exactly Equation (60) which holds. Secondly, we show that $H_i^r(t)$ is increasing in the interval $(T^{\pi_i(S_2)}, T^{\pi_i(S_3)}] = (T^{\pi_i(\{r'\})}, T^{\pi_i(\{r\})}]$, and reaches a global maximum if $|\Gamma_i^r \cap \Gamma_i^{r'}| = |\pi_i(\{r, r'\})|$ is sufficiently small. Applying Equation (57) with $v = 3$, the slope of $H_i^r(t)$ in the interval $(T^{\pi_i(\{r'\})}, T^{\pi_i(\{r\})}]$ is positive if and only if

$$\begin{aligned} &\frac{q^{\pi_i(\{r\})}}{q^{\pi_i(\{r,r'\})} T^{\pi_i(\{r,r'\})}} > \frac{q^{\pi_i(\{r\})} + q^{\pi_i(\emptyset)}}{q^{\pi_i(\{r,r'\})} T^{\pi_i(\{r,r'\})} + q^{\pi_i(\{r'\})} T^{\pi_i(\{r'\})}} \\ \Leftrightarrow\ &q^{\pi_i(\{r\})} q^{\pi_i(\{r'\})} T^{\pi_i(\{r'\})} > q^{\pi_i(\emptyset)} q^{\pi_i(\{r,r'\})} T^{\pi_i(\{r,r'\})} \end{aligned} \tag{61}$$

Equation (61) holds when $|\pi_i(\{r, r'\})|$ is sufficiently small. Finally, we show that under this condition, $H_i^r(T^{\pi_i(\{r\})}) > H_i^r(T^{\pi_i(\{r,r'\})})$. Applying Equations (42) and (56), we

find that

$$H_i^r(T^{\pi_i(\{r\})}) = \frac{\sum_{\pi\in\Pi_i} d_i^\pi(T^{\pi_i(\{r\})})}{d_i(T^{\pi_i(\{r\})})}$$
$$= \frac{2\left[q^{\pi_i(\{r,r'\})}T^{\pi_i(\{r,r'\})} + q^{\pi_i(\{r\})}T^{\pi_i(\{r\})}\right]}{2\left[q^{\pi_i(\{r,r'\})}T^{\pi_i(\{r,r'\})} + q^{\pi_i(\{r'\})}T^{\pi_i(\{r'\})} + (q^{\pi_i(\{r\})} + q^{\pi_i(\emptyset)})T^{\pi_i(\{r\})}\right]} \tag{62}$$

And,

$$H_i^r(T^{\pi_i(\{r,r'\})}) = \frac{\sum_{\pi\in\Pi_i} d_i^\pi(T^{\pi_i(\{r,r'\})})}{d_i(T^{\pi_i(\{r,r'\})})}$$
$$= \frac{2\left[q^{\pi_i(\{r,r'\})}T^{\pi_i(\{r,r'\})} + q^{\pi_i(\{r\})}T^{\pi_i(\{r,r'\})}\right]}{2\left[q^{\pi_i(\{r,r'\})}T^{\pi_i(\{r,r'\})} + (q^{\pi_i(\{r'\})} + q^{\pi_i(\{r\})} + q^{\pi_i(\emptyset)})T^{\pi_i(\{r,r'\})}\right]} \tag{63}$$

Now, note that since $T^{\pi_i(\{r'\})} > T^{\pi_i(\{r,r'\})}$ and $\sum_{\pi\in\Pi_i} q^\pi = 1$, we have that $H_i^r(T^{\pi_i(\{r\})})$ is strictly greater than

$$\frac{q^{\pi_i(\{r,r'\})}T^{\pi_i(\{r,r'\})} + q^{\pi_i(\{r\})}T^{\pi_i(\{r\})}}{q^{\pi_i(\{r'\})}T^{\pi_i(\{r'\})} + (1 - q^{\pi_i(\{r'\})})T^{\pi_i(\{r\})}} \tag{64}$$

Denote the denominator in Equation (64) by W, and note that $T^{\pi_i(\{r,r'\})} < W < T^{\pi_i(\{r\})}$. Now, $H_i^r(T^{\pi_i(\{r\})}) > H_i^r(T^{\pi_i(\{r,r'\})})$ if the expression in Equation (64) is strictly greater than the expression in Equation (63). This inequality can be arranged to obtain

$$q^{\pi_i(\{r\})}(T^{\pi_i(\{r\})} - W) > q^{\pi_i(\{r,r'\})}(W - T^{\pi_i(\{r,r'\})}) \tag{65}$$

Equation (65) holds for $|\pi_i(\{r, r'\})|$ sufficiently small, which suffices to establish that $H_i^r(t)$ reaches a global maximum at $T^{\pi_i(\{r\})}$. □

Appendix II: Preferential attachment

Suppose that in each period $t \in \{1, 2, 3, ...\}$ every active agent i *interacts* with other agents as follows: Agent i selects the social group Γ_i^r with probability $q^r \geq 0$ for $r \in \mathcal{R}_+$ and for each $N_i(k) \subseteq \Gamma_i^r$, selects the group of agents of type k with probability $\frac{|N_i(k)|}{|\Gamma_i^r|}$. Agent i then sends a friend request to an agent j selected at random with *a probability that is proportional to j's degree* from among the active agents in $N_i(k)$ who are not yet i's friends.

We can re-derive Equations (12) and (13) as follows: For any agent i of type k, i interacts with agents of type k' with probability $\alpha^{kk'}$, and provided that $t \leq T^{kk'}$, agent

i sends a friend request to some active agent of type k' who is not yet friends with i. Therefore, i initiates $\alpha^{kk'}\mathbf{1}(t \leq T^{kk'})$ links with agents of type k' in period t (This is identical to Equation (12)). Similarly, in period t, an agent j of type k' interacts with agents of type k with probability $\alpha^{k'k}$. Provided that $t \leq T^{k'k} = T^{kk'}$ (see Remark 3), and conditional on interacting with agents of type k, agent j (of type k') sends an immediately accepted friend request to some remaining agent of type k (who is not idle and is not yet friends with j). Denote by $\mathcal{Z}_j^k(t)$ the *set* of active agents of type k who are not yet friends with j in period t. Since j selects agent i to send a friend request to with a probability that is proportional to i's degree, i is selected with probability $\frac{d_i(t)}{\sum_{h\in\mathcal{Z}_j^k(t)} d_h(t)}$. Finally, there are precisely $Z^{kk'}(t)$ agents j of type k' from which an agent i of type k could accept a friend request from in period t. From the above, it follows the expected number of friend requests that an agent i of type k receives (and immediately accepts) from agents of type k' in period t is given by

$$\alpha^{k'k}\frac{d_i(t)}{\sum_{h\in\mathcal{Z}_j^k(t)} d_h(t)}Z^{kk'}(t) \tag{66}$$

As long as they are not idle, any two agents that are of the same type are essentially indistinguishable. Therefore, for any agent j of type k', we have that $|\mathcal{Z}_j^k(t)| = Z^{k'k}(t)$, and for any agents i and j of type k, $d_i(t) = d_j(t)$.[31] From this it follows that Equation (66) can be re-written as

$$\alpha^{k'k}\frac{d_i(t)}{Z^{k'k}(t)d_i(t)}Z^{kk'}(t) \tag{67}$$

This is identical to Equation (13). So, none of our results change if the link formation process were governed by preferential attachment.

[31] The manner in which we simplify Equation (66) is usually not possible in a typical growing random network model because in such models the function expressing the degree of the i^{th} node in period t differs from that of the j^{th} node in period t. For example, see Jackson (2008, p. 131).

Appendix III: Data description

We use a dataset from Facebook that was anonymized by Adam D'Angelo (the then CTO of Facebook) and first made available to Traud et al. (2010). The dataset was also analyzed by Shaw et al. (2011), Traud et al. (2012), Tarbush and Teytelboym (2012).

The data represent a September 2005 cross-section of the complete structures of social connections on www.facebook.com *within* (but not across) the first ten American colleges and universities that joined Facebook. At the time, Facebook was available only to those registered as students or staff with .edu email addresses at selected American colleges and universities. Signing up was free. Users were able to search for and browse through profiles of other users of Facebook and send them "friend requests". If the friend request were accepted, the users became "friends" and could access further information on each others' profiles. The users' profiles contained their photo, a space for public comments (a "wall"), private information (such as their age or gender), contact information (such as their email address), courses, and a list of their friends.

For the first ten American colleges and universities that joined Facebook, the (anonymized) raw data contain over 130,000 nodes (users) and over 5.6 million links (friendships). We observe six social categories for each user: gender, year of graduation, major, minor, dorm, and high school. Since all personal data were provided voluntarily, some users did not submit all their information. We therefore cleaned the data as follows. We dropped any user (and their links), who had not provided all the personal characteristics other than high school.[32] In addition, some users were listed as faculty members and some students listed graduation years that were probably untruthful (e.g. 1926). We therefore dropped all faculty members and every user whose year of graduation is outside 2006-2009. There are 27,454 users and 492,236 links in our cleaned data, consisting only of students graduating between 2006 and 2009, who have supplied all the relevant personal characteristics (except high school).

A summary of the data can be found in Table 2.

[32] While high school is an interesting social category, the relative group sizes within colleges are too small to allow for a meaningful analysis.

College	Raw number of nodes	Raw number of edges	Nodes	Edges	Average degree	Women	Men	Avg. major size	Avg. minor size	Avg. dorm size	Avg. class size
Harvard U.	15126	824617	1325	18608	28.1	567	758	23.2	22.5	42.7	46.9
Columbia U.	11770	444333	2663	52697	39.6	1573	1090	29.6	29.9	54.3	65.7
Stanford U.	11621	568330	2254	55,124	48.9	1043	1211	30.9	30.1	25.6	55.0
Yale U.	8578	405450	1431	23847	33.3	639	792	19.6	19.1	68.1	38.2
Cornell U.	18660	790777	2509	26653	21.2	1078	1431	27.6	24.6	20.6	51.6
Dartmouth College	7694	304,076	1612	34030	42.2	780	832	29.9	29.3	23.0	45.0
U. of Penn.	14916	686501	3006	60516	40.3	1417	1589	28.4	27.1	50.9	77.0
M.I.T.	6440	251252	1563	32751	41.9	626	937	44.7	37.2	26.1	58.2
New York U.	21679	715715	5581	95968	34.4	3345	2236	53.7	52.2	105.5	99.7
Boston U.	19700	637528	5510	92042	33.4	3355	2155	37.5	34.7	91.8	90.8
Average	**13618**	**562858**	**2745**	**49224**	**36.3**	**1442**	**1303**	**32.5**	**30.7**	**50.5**	**62.8**

Table 2: Data summary

References

Bala, V., Goyal, S., 2000. A noncooperative model of network formation. Econometrica 68, 1181–1229.

Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O., 2012. The Diffusion of Microfinance. Working Paper 17743. NBER. http://www.nber.org/papers/w17743.pdf.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Botha, L., Kroon, S., 2010. A community-based model of online social networks, in: The 4th SNA-KDD Workshop on Social Network Mining and Analysis.

Boucher, V., 2012. Structural Homophily. Working Paper. Université de Montréal. http://www.vincentboucher economist.com/SH5juillet.pdf.

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

Breiger, R.L., 1974. The duality of persons and groups. Social Forces 53, 181–190.

Currarini, S., Jackson, M.O., Pin, P., 2009. An Economic Model of Friendship: Homophily, Minorities, and Segregation. Econometrica 77, 1003–1045.

Currarini, S., Jackson, M.O., Pin, P., 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.

de Marti, J., Zenou, Y., 2011. Identity and Social Distance in Friendship Formation. Working paper. Stockholm University. http://www.econ.upf.edu/~demarti/Articles/identity.pdf.

Easley, D., Kleinberg, J., 2010. Networks, Crowds, and Markets: Reasoning about a highly connected world. Cambridge University Press, Cambridge, UK.

Falk, A., Ichino, A., 2006. Clean evidence on peer effects. Journal of Labor Economics 24, 39–57.

Feld, S.L., 1981. The focused organization of social ties. American Journal of Sociology 86, 1015–1035.

Fosco, C., Vega-Redondo, F., Marsili, M., 2010. Peer effects and peer avoidance: The diffusion of behavior in coevolving networks. Journal of the European Economic Association 8, 169–202.

Golub, B., Jackson, M.O., 2012. How homophily affects diffusion and learning in networks. Quarterly Journal of Economics 127, 1287–1338.

Goyal, S., 2009. Connections: An Introduction to the Economics of Networks. Princeton University Press, Princeton, NJ.

Granovetter, M., 2005. The impact of social structure on economic outcomes. Journal of Economic Perspectives 19, 33–50.

Granovetter, M.S., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Iijima, R., Kamada, Y., 2014. Social distance and network structures. Working paper. Harvard University. http://www.ykamada.com/pdf/Clustering.pdf.

Jackson, M.O., 2008. Social and Economic Networks. Princeton University Press, Princeton, NJ.

Jackson, M.O., 2014. Networks in the understanding of economic behaviors. Journal of Economic Perspectives 28, 3–22.

Jackson, M.O., Rogers, B.W., 2007. Meeting strangers and friends of friends: How random are social networks? American Economic Review 70, 890–915.

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

Kandel, D.B., 1978. Homophily, selection, and socialization in adolescent friendships. American Journal of Sociology 84, 427–436.

Kremer, M., Levy, D., 2008. Peer effects and alcohol use among college students. Journal of Economic Perspectives 22, 189–206.

Kumar, R., Novak, J., Tomkins, A., 2010. Structure and evolution of online social networks, in: Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining, pp. 611–617.

Lattanzi, S., Sivakumar, D., 2009. Affiliation networks, in: Proceedings of the 41st annual ACM symposium on Theory of computing, pp. 427–434.

Leskovec, J., Kleinberg, J., Faloutsos, C., 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations, in: Proceedings of the 11th ACM SIGKDD international conference on Knowledge Discovery in Data Mining, pp. 177–187.

Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W., 2008. Statistical properties of community structure in large social and information networks, in: Proceedings of the 17th international conference on World Wide Web, pp. 695–704.

Mayer, A., Puller, S.L., 2008. The old boy (and girl) network: Social network formation on university campuses. Journal of Public Economics 92, 329–347.

McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444.

Moody, J., 2001. Race, school integration, and friendship segregation in america. American Journal of Sociology 107, 679–716.

Mouw, T., Entwisle, B., 2006. Residential segregation and interracial friendship in schools. American Journal of Sociology 112, 394–441.

Newman, M.E.J., 2010. Networks: An Introduction. Oxford University Press, Oxford, UK.

Newman, M.E.J., Watts, D.J., Strogatz, S.H., 2002. Random graph models of social networks. Proceedings of the National Academy of Sciences 99, 2566–2572.

Price, D.D.S., 1976. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science 27, 292–306.

Sacerdote, B., 2001. Peer Effects with Random Assignment: Results for Dartmouth Roommates. Quarterly Journal of Economics 116, 681–704.

Shaw, B., Huang, B., Jebara, T., 2011. Learning a distance metric from a network, in: Proceedings of the 25th annual conference on Neural Information Processing Systems, pp. 1899–1907.

Shrum, W., Cheek Jr., N.H., Hunter, S.M., 1988. Friendship in school: Gender and racial homophily. Sociology of Education 61, 227–239.

Tarbush, B., Teytelboym, A., 2012. Homophily in online social networks, in: Goldberg, P. (Ed.), Internet and Network Economics. Springer Berlin Heidelberg. volume 7695

of *Lecture Notes in Computer Science*, pp. 512–518. URL: http://dx.doi.org/10.1007/978-3-642-35311-6_40, doi:10.1007/978-3-642-35311-6_40.

Traud, A.L., Kelsic, E.D., Mucha, P.J., Porter, M.A., 2010. Comparing community structure to characteristics in online collegiate social networks. SIAM Review 53, 526–543.

Traud, A.L., Mucha, P.J., Porter, M.A., 2012. Social structure of Facebook networks. Physica A 391, 4165–4180.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. nature 393, 440–442.

Wimmer, A., Lewis, K., 2010. Beyond and below racial homophily: Erg models of a friendship network documented on facebook. American Journal of Sociology 116, 583–642.

Xiang, R., Neville, J., Rogati, M., 2010. Modeling relationship strength in online social networks, in: Proceedings of the 19th international conference on World Wide Web, pp. 981–990.

Zheleva, E., Sharara, H., Getoor, L., 2009. Co-evolution of social and affiliation networks, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pp. 1007–1016.