Social Science & Medicine xxx (2014) 1–13

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Choosing your network: Social preferences in an online health community

Damon Centola [a,*], Arnout van de Rijt [b]

[a] Annenberg School & School of Engineering, University of Pennsylvania, Rm. 306, 3620 Walnut St., Philadelphia, PA 19104, USA
[b] Department of Sociology & Institute for Advanced Computational Science, SUNY Stony Brook, Stony Brook, NY 11794, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Networks
Homophily
Selection
Social support
Fitness
Weight loss
Internet

ABSTRACT

A growing number of online health communities offer individuals the opportunity to receive information, advice, and support from peers. Recent studies have demonstrated that these new online contacts can be important informational resources, and can even exert significant influence on individuals' behavior in various contexts. However little is known about how people select their health contacts in these virtual domains. This is because selection preferences in peer networks are notoriously difficult to detect. In existing networks, unobserved pressures on tie formation – such as common organizational memberships, introductions by friends of friends, or limitations on accessibility – may mistakenly be interpreted as individual preferences for interacting/not interacting with others. We address these issues by adopting a social media approach to studying network formation. We study social selection using an *in vivo* study within an online exercise program, in which anonymous participants have equal opportunities for initiating relationships with other program members. This design allows us to identify individuals' preferences for health contacts, and to evaluate what these preferences imply for members' access to new kinds of health information, and for the kinds of social influences to which they are exposed. The study was conducted within a goal-oriented fitness competition, in which participation was greatest among a small core of active individuals. Our results show that the active participants displayed indifference to the fitness and exercise profiles of others, disregarding information about others' fitness levels, exercise preferences, and workout experiences, instead selecting partners almost entirely on the basis of similarities on gender, age, and BMI. Interestingly, the findings suggest that rather than expanding and diversifying their sources of health information, participants' choices limited the value of their online resources by selecting contacts based on characteristics that are common sources of homophily in offline relationships. In light of our findings, we discuss design principles that may be useful for organizations and policy makers trying to improve the value of participants' social capital within online health programs.

1. Introduction

Over the last decade, the Internet has become an increasingly important domain for health (Fogel et al., 2002; Thackeray et al., 2008; Chou et al., 2009; Hawn, 2009; McNab, 2009; Pampel et al., 2010; Salathe and Khandelwal, 2011). Recent surveys of Internet use for health estimate that 23% of US patients living with chronic illnesses, such as high blood pressure, diabetes, heart conditions, or cancer, use peer-to-peer online resources to help support their medical treatment and discovery processes (Fox, 2011). Even more striking, among populations with chronic diseases who are seeking "practical advice for coping with day-to-day health situations," patients were overall more likely to seek out informal sources of peer-to-peer assistance than consult with medical professionals (Fox, 2011). As this trend increases, social scientists interested in the social dimensions of health are increasingly concerned with characterizing the online social networks that people use. In particular, recent research has begun to explore the question of how online social networks influence the spread of health information and behavior change (White and Dorman, 2001; Japuntich et al., 2006; Hawn, 2009; Centola, 2010, 2011). Centola (2010, 2011) uses controlled online experiments to demonstrate the effects of both network structure and homophily in promoting the contagious spread of health behaviors. However, relatively little is known about how online health communities form, and what kinds of networks people "create" in these often anonymous environments (Wellman and Hampton, 1999; Wellman, 2001). Given the variety of online health contexts for information exchange and influence (Fox, 2011), we focus

* Corresponding Author.
E-mail address: dcentola@asc.upenn.edu (D. Centola).

http://dx.doi.org/10.1016/j.socscimed.2014.05.019
0277-9536/

our study on the increasingly popular domain of online fitness programs, which are designed to promote exposure to health information and increased fitness through peer to peer interaction (Centola, 2013).

The popular bromide that people select ties "homophilously" – i.e., based on preferences for others with similar characteristics – was formally introduced over a half a century ago by Lazarsfeld and Merton (1954). The goal of their study was to determine why strong correlations were regularly observed between people with specific demographic characteristics and those exhibiting certain beliefs, attitudes and behaviors. Their explanatory strategy was first to show that people with similar demographic traits selectively formed ties to one another, and then to show that people who were socially connected influenced each other's beliefs. However, while they found that friends influence friends, support for homophily in tie formation (henceforth "choice homophily") was variable, occurring in some situations, but not in others. As Lazarsfeld and Merton put it, "[T]he problem of selection [is] not adequately formulated by the familiar and egregiously misleading question: When it comes to close friendships, do birds of a feather actually flock together? Rather it is a more complex problem of determining the degree to which such selectivity varies for different kinds of social attributes, how it varies within different kinds of social structure, and how such selective patterns come about." (Lazarsfeld and Merton, 1954:18).

A large literature has since emerged on homophily in social relations. As the terminology has evolved, the term "homophily" has now come to refer to the observed population-level regularity that people within a community tend to be socially connected to others who are more similar to themselves than would be expected by random chance (Coleman, 1958). Researchers in this tradition have identified several, very different, mechanisms that can generate this regularity. The most obvious mechanism, initially identified by Lazarsfeld and Merton (1954), is "choice homophily": People preferentially make ties to others who are similar to themselves. However, inferring individual choice homophily from population level homophily risks running afoul of the ecological fallacy since choice homophily can be completely absent at the individual level even when populations exhibit high levels of observed homophily. This disjuncture between individual behavior and collective outcome is due to the variety of other mechanisms that can produce similar population-level patterns. For instance, a second mechanism, which has recently been widely discussed in the literature on networks and health is the process of social influence (McPherson and Smith-Lovin, 1987; Popielarz and McPherson, 1995; Christakis and Fowler, 2007). While homophily on some traits, like race and gender, cannot emerge through social influence, interpersonal correlations on other health characteristics, such as obesity, heart disease, or smoking, can be linked to social influences between contacts (Christakis and Fowler, 2007). Recent research has emphasized these mechanisms as competing explanations for patterns of observed homophily on obesity, giving rise to a dichotomization of the literature on homophily and health into the competing positions of "social influence" vs. "choice homophily". However, the scope of the problem of the origins of interpersonal correlations on health characteristics is actually much broader. Other explanatory mechanisms, which operate at the level of social structure rather than at the level of the individual or the dyad, are equally important factors in the emergence of correlations in social networks.

For instance, organizational and institutional sorting processes at schools and workplaces typically determine the set of potential social contacts that an individual is exposed to within a given context (Feld, 1982; McPherson et al., 2001; Moody, 2001; Ruef et al., 2003; Bertrand and Mullainathan, 2004). These structures often implicitly "preselect" individuals into homophilous groups (by race, class, gender, educational background, and so forth), thereby eliminating opportunities for heterophilous tie formation (Blau, 1977; Blau and Schwartz, 1984; McPherson and Smith-Lovin, 1987). These social processes can force homophilous tie formation even when the members of a population lack any particular preference for homophilous ties (McPherson and Smith-Lovin, 1987). Similarly, homophily can also emerge from the process of friends introducing friends to one another, or "triadic closure" in social networks (Kossinets and Watts, 2009). For instance, if a pair of friends, A and B, are homophilous, and B also has a friend C with whom she is similarly homophilous, then A may become friends with C by virtue of B's introduction. A homophilous tie between A and C can thus form by virtue of social structure, without A having any particular interest in "finding" someone similar to herself. More importantly, homophily can emerge in social networks even when individuals consciously prefer heterophily. In friendship networks, competitive preferences to form ties with the most healthy, most physically attractive or most successful individuals can create patterns of observed homophily via the endogenous exclusion of low-health or low-attractiveness members of the population, who are then forced to form ties with one another (Ali et al., 2012). Crosnoe et al. (2008) shows that this mechanism of social exclusion can generate explicit patterns of homophily on obesity. More generally, across a broad array of social characteristics in which actors have "aspirational" preferences to form ties to "desirable" alters, patterns of systematic exclusion of the less desirable individuals can lead to the false appearance of choice homophily in domains such as health (Ali et al., 2012), online dating (Hitsch et al., 2010), marriage markets (Mare, 1991; Kalmijn, 1994), scientific collaboration (Dahlander and McFarland, 2013), and residential segregation (Van de Rijt et al., 2009). Finally, selection on an unobserved trait may be mistaken for a selective preference for a correlated trait that is observed (Yamaguchi and Kandel, 1993; Kalmijn and Vermunt, 2007). For example, as fitness is related to age, a tendency for individuals to choose ties to others of a similar fitness observed in a study that measures subject fitness but not age may in actuality represent an unobserved tendency for subjects to select on the basis of age. Consequently, in evaluating the implications of social networks for health communications, observed patterns of homophily on health characteristics do not provide clear evidence for individuals' selective preferences for health contacts.

These issues become particularly salient in contexts where the selections that people make are typically sought after as informational or motivational resources. Within online fitness programs, the selection of health contacts explicitly serves the goal of providing a reference point for achievement within the program, and establishing a standard against which to evaluate success. Our goal is to determine how people select ties in these contexts, and thereby to understand how social selection both frames the scope of participants' exposure to novel and productive health information, and provides a motivational frame for future health. In particular, we are interested in whether participants select online health contacts who have levels of fitness and "status" on health characteristics that suggest aspirational goals in establishing ties, or whether ties are formed primarily to contacts with similar levels of fitness as themselves. This difference between "aspirational" tie formation, vs. "homophilous" tie formation is important for understanding the ultimate impact of online health networks on participants' health. One of the primary incentives for forming contacts within an online health program is because they provide a means for discovering new ways to lead a healthier lifestyle by providing exposure to new health information. Another reason that participants form ties is because they are seeking

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

connections with health "leaders," whom they may not have contact within their day to day routines, but who can provide peer-guidance on improving their fitness and lifestyle. However, these goals are primarily served only if the ties participants make actually connect them to people who extend their informational and motivational exposure.

Thus, we emphasize that our goal is not to explore the familiar tension between homophily and contagion as competing explanations for observed correlations between network ties and individual traits. Rather, we are interested in people's selection patterns on relatively stable health characteristics, which determine the kinds of informational and motivational exposure that the members of these online communities receive. The goal of the present study is to clearly identify individuals' preferences in forming fitness-specific online health contacts. There are important new methods (Steglich et al., 2010) that have been developed for identifying selection behavior in complex observational datasets (Mercken et al., 2009; Wimmer and Lewis, 2010; Lewis et al., 2012). Each of these methods is designed to solve problems of causality and identification that are caused by uncontrolled factors such as endogeneity, unobserved heterogeneity, and exogenous influence (Aral et al., 2009; Shalizi and Thomas, 2011). Our study was developed to eliminate these factors at the outset by using a controlled, randomized design, implemented within an existing online health program. Recently, many scholars have used randomization and experimental controls to eliminate the large number of factors that can prevent the identification of social influence in network contexts (Centola, 2010, 2011; Bond et al., 2012). By contrast, our design eliminates social influence, as well as the confounding factors of organizational grouping, hierarchical exclusion, friends introducing friends and exogenous influence, in order to isolate and identify individual preferences in tie selection.

Our approach to studying the process of network formation follows that of Lazarsfeld and Merton (1954), who argue that the problem of emergent patterns of association in social networks is not one of determining a general model of choice dynamics. Rather, in different contexts, different selection criteria guide individuals' preferences. Motivated by the growing importance of social media in peer-to-peer informational exchange and health related decision-making (Fox, 2011), this study shows how individual selection can shape the active communication channels in an online fitness program, and what this implies for participants' access to health information and social influence in this domain.[1]

2. Data

We partnered with an existing fitness-improvement program, which was designed to help motivate people to increase their daily exercise level through a series of weekly incentive offerings. We then created a peer-to-peer social network platform within the program that permitted participants in the fitness-improvement program to observe and learn from other members of the online community. Participants in the program were initially assigned a random peer-to-peer network of online health contacts, which they were permitted to change over the course of a five week period. This design allowed us to record the complete evolution of social network ties among the members of the fitness community.

Participants were recruited directly to our study from within the program registration process. All of the individuals who joined the program were given the opportunity to join our study, called the "The Health Improvement Network." 432 participants consented to participate in our study.[2] Participants registered by creating an anonymous on-line profile, which included their age, gender, ethnicity, BMI, fitness level, diet preferences, goals for the program, and favorite exercise, as well as a record of their average exercise minutes and intensity level. Subjects then provided informed consent for their participation in the study. They were then randomly assigned to a position in one of six, pre-existing, unpopulated network topologies. Each of these networks constituted its own, independent health community. Each network was designed with an identical network architecture. The number of "neighbors" or social "links", Z, for each node was identical for every person in every network ($Z = 6$). The level of "clustering," C, i.e., the fraction of a person's neighbors who were connect to each other, creating "triangles" in the network, was identical in every neighborhood of every network ($C = .4$). And, the size of the population, N, was identical for every network ($N = 72$). The subjects were randomly assigned across networks such that all six network populations were identically distributed, allowing for six independent community-level "observations" of the tie formation process. These independent trial-level observations permit a conservative statistical evaluation of choice dynamics, which overcomes traditional obstacles to statistical inference posed by interdependencies between observations in dyadic analysis of a single trial.

Participants' initial social contacts within the program were comprised of the randomly assigned members who occupied the nodes that were immediately adjacent to them in the network, i.e., their network "neighbors." All social ties in the study were symmetrical, so for every actor B who was a neighbor of A, A was also a neighbor of B. The initial randomization of subjects across network positions ensured that social ties were uncorrelated with subjects' identities. Thus, at the start of the study, traditional sources of unobserved heterogeneity in network composition, such as affect in social relations, historical familiarity, or shared friends in common, were controlled by our design, and could not have an effect on subsequent tie choice. Finally, by randomizing the subject pool into six independent and identically distributed populations, we could observe the dynamics of tie formation across multiple, independent networked populations, as discussed below.

Each participant was provided with a personalized on-line "health dashboard," which displayed all profile information and real-time health information for her and her health buddies. Every time a subject logged in to the health program, her health dashboard would display her complete profile with her exercise and health characteristics, along with those of each of her health contacts. Health contact avatars were listed in descending order according to the number of completed exercise minutes in the current week. This ranking was performed in real-time every time a subject accessed her health dashboard. This prevented any one health buddy from always being located at the top of the buddy list.

Once the participants completed the registration process and were assigned to a network position, the only people who could be directly observed by a participant were the individuals who were directly connected to her in the social network, i.e., her health contacts. To change contacts, a participant could select a "Change Your Health Contacts" link

[1] The theoretical implications of these selection dynamics for network topology and the dynamics of social influence are discussed in the Appendix.

[2] Approximately 20% of the program's 2000 members opted into the study. There were no significant differences along the observed characteristics between the subjects who enrolled in the study and those who did not participate.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

Table 1
Descriptive statistics for all six networks ($N = 432$).

Variable	Summary statistics							
Health-related								
Age	Min	Mean	Max					
	17	34.6	79					
Gender	Female	Male						
	276	156						
Ethnicity	Af-Am	Hisp	Asian	Euro	Other			
	23	20	63	254	72			
BMI	Min	Mean	Max					
	17.7	25.0	47.2					
Fitness	Poor	<Av.	Average	>Av.	High			
	9	51	165	175	32			
Diet preferences	Low Cal.	Veget.	Omniv.	Carniv.				
	31	45	265	91				
Exercise-specific								
Exercise intensity	Low	Medium	High					
	67	269	96					
Exercise minutes (per week)	Min	Mean	Max					
	0	183.9	1000					
Exercise goals	Lose weight	Look better	Feel healthy	Reduce risks	Reduce stress			
	73	53	232	37	37			
Favorite exercise	Swimming	Walking	Running	Bicylcing	Weights	Elliptical	Team sports	Other
	81	71	32	31	31	30	28	128

on the dashboard. This opened an Add/Drop page that listed all of the members of the participant's entire network, excluding themselves and their existing health contacts, with whom they could form a new tie. The Appendix shows this Add/Drop page (Fig. A1).

Inspecting their potential health contacts, participants could observe the general demographic traits (age, gender, and ethnicity), health-related characteristics (BMI, fitness level, and diet preferences), and exercise-specific attributes (exercise goals, typical exercise intensity, typical exercise minutes, and favorite exercise) of other community-members, but did not have any other information about their fellow participants, nor any knowledge of how they were connected to one another. By withholding information from participants on how buddies were connected to one another we prevented people from attaching to members simply because they were "popular" among other members, thereby ensuring that the ten visible traits were the only basis for tie formation. All of these traits were fixed for the duration of the study. Descriptive statistics for the entire subject pool are shown in Table 1. A majority of subjects were female, typically in their twenties, and of European ethnic descent. The median subject reported average fitness, average BMI was normal weight, and diet preferences were predominantly omnivore. The typical exercise intensity was medium, and average subjects reported they exercised about 3 h each week, mostly to feel healthy. Among the most popular exercises were swimming, walking, running, biking, working out on the elliptical, lifting weights and participating in team sports.

Once we initiated each of the six independent communities,[3] we observed participants' choices to add and drop ties to other members of the on-line community over a period of five weeks. Over the course of the study, subjects could add and drop ties as many health contacts as they wanted, and could reconsider past choices as many times as desired. However, to select a new health contact, a participant was required to drop an existing tie. Similarly, dropping an existing health buddy required that a participant make a new tie. This constraint provides some specific methodological advantages for our study.

First, it introduces a cost, if slight, into the decision process, and means that participants could not, for instance, simply add all of the members of their community to their contact list. Second, in order to see the activities and behaviors of a given community member, a participant was forced to drop an existing health contact. Thus, not only did the tie formation a decision have an explicit cost, but our design allows us to observe how preferences for tie formation also correspond with preferences for tie deletion — i.e., we are able to independently identify both the traits of the contacts that individuals preferred to attach to, as well as the traits of those that they preferred to remove. Third, this procedure permitted heterogeneity in the actual number of ties per person to change over the course of the network evolution, while also ensuring that the overall number of ties in the network remained constant.[4]

A final feature of the study design was that participants did not require consent from a new health contact to add a tie to her. This is an important feature of our study since observed homophily on health characteristics, obesity in particular, has been argued to

[3] Our motivation for running six independent "trials" of the same study was to permit two levels of statistical analysis. At the individual level, aggregated results across all communities allow us to identify individual tendencies in the tie formation dynamics. At the network level, comparing the outcomes across independent communities allow us to identify any significant trends that emerge across multiple identically and independently distributed observations of the network formation process. At the start of the study, there was no variation in individuals' initial structural positions either within or across the six communities, and no significant differences in population composition across communities, allowing us to treat individuals as identically situated decision-makers.

[4] This is an important element of our design since it allowed us to detect if individuals with certain desirable traits became "stars" (i.e., persons with many social contacts) in the emergent network, while preventing an abundance of ties from becoming a trivial feature of people who prefer to make many ties. That is, in order to become a "star" in our study's evolving network, an individual had to receive lots of ties due to having desirable traits, and could not simple be a "social" person who wanted lots of connections. Further, since the overall number of ties was held constant, the emergence of a "star" in the network signals a meaningful measure of members' preferences for that individual's traits since it also implies that other individuals, with less desirable traits, would have many fewer ties (due to overall tie conservation). Thus, by preventing network density from increasing (by keeping a constant number of ties in the overall network) our study maintained *i*) the individual level significance of tie formation, and *ii*) the network level significance of certain individuals accruing more ties than others.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

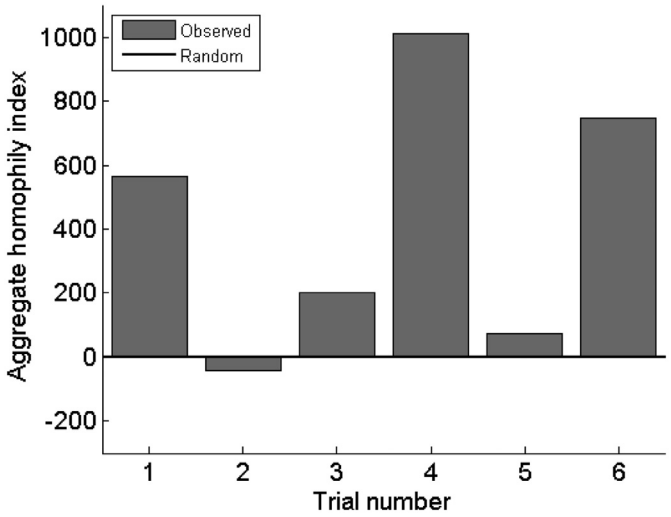

Fig. 1. Choice homophily in tie change for each network. Shown is the aggregate homophily index H, which is calculated as the difference between the observed number of traits in common and the number expected under random tie choice, summed across all new ties (see Appendix). Cumulative homophily is significantly present ($p = .031$ using a two-tailed signed rank test, $N = 6$).

emerge from the combination of preferential selection and social exclusion (Crosnoe et al., 2008). Our study explicitly eliminates this mechanism for homophily by allowing tie formation to be driven by individual attachment preferences. Notably, actors can always subsequently drop a tie. So, for example, if an obese individual A added a tie to a healthy individual B, the healthy individual B could then remove the newly formed tie from A, and replace this tie with a more desirable health contact. Our design allows us to observe this pattern of behavior as two explicit actions (an attachment preference by A, and a removal preference by B), which allows us to independently analyze both sides of health-based tie selection (i.e., addition and removal).

The controls created by our *in vivo* design necessarily also entailed limitations. Perhaps the major limitation of our design was that the level of observed activity in the study was directly tied to the level of engagement in the health program we partnered with. During our 5 week study, participation in the health program was extremely low, which translated directly into a limited number of observations. Among the 432 subjects enrolled in the study, only 18 engaged in active tie changes. Together, these 18 active participants ("tie initiators" hereafter) made a total of 51 tie changes.[5] A single tie change was made in the least active community while 19 were made by 6 distinct individuals in the most active community. Most of these tie changes (33) were made in the first week of the study, when overall subject participation on the site was generally the strongest. Each new tie was relatively independent, resulting in a permanent change in the network; i.e., there were no "cascading" effects of tie selection on others' tie selection.

Our primary concern was whether the low number of observations resulted in some form of sample bias within our data. To address this question, the Appendix provides a detailed analysis of activity levels among participants in the study. These analyses determine both if tie activity was correlated with any distinguishing features of particular individuals, and whether the lack of tie activity was the model behavior of active participants, or whether the people who failed to make tie changes were simply inactive members of the health program. We found no significant differences along health characteristics between participants who formed ties and those who were inactive, except that women were more active in changing ties than men. Overall, the most significant indicator of inactivity in tie formation was inactivity in the health program as whole, with nearly all the inactive subjects failing to click on the website at all during the observation period. The low levels of activity within the health program prevented us from detecting large scale topological patterns in network evolution. However, despite the limitations on statistical power created by small sample size, we found that active participants exhibited remarkably strong and significant patterns of choice behavior across each of the network communities. The analyses included in the Appendix demonstrate that these findings are robust even when the data are partitioned to exclude the most active members of the study. The results exhibit clear trends in behavior at the level of both the individual and the network, which provide insights into whether participants' selected contacts helped to support the program's goals. We conclude by discussing these implications and suggesting program strategies that may promote the selection of productive health networks, as well as increase program participation.

3. Results

3.1. Network-level patterns of tie choice

At the start of the study, conditions were equivalent in each of the six fitness communities. Every individual had a "balanced" neighborhood, in which their neighbors had a random distribution of each of the 10 measured health characteristics. As subjects began to add and drop ties, this created measurable, real-time changes in each individual's neighborhood composition. We used these changes in "average neighborhood composition" to evaluate the overall tendency in each of the six communities to evolve toward a distinct aggregate pattern.

There were no discernible aggregate tendencies toward preferential attachment, or emergent "stars," in any of the health communities. We measured the "popularity" of a participant as the number of fitness community members from whom he/she received new ties. We then compared the distribution of health buddy popularity in each population with the distribution of popularity expected under random tie choice. The results show that in none of the six trials were any health buddies chosen more than twice, and in only two trials was anyone chosen more than once. In each trial, the number of such duplicate choices (popularity of 2) was precisely equal to the expected number of duplicates under random tie choice. We also examined whether individuals had preferences to disproportionately connect to alters along any combination of the 10 traits (e.g., younger, fitter, better diet, etc.) within the empirical range of the population, and found no departures from random selection across all permutations.

We did, however, find a significant trend toward homophilous tie formation across the independent populations. Fig. 1 shows the aggregate homophily index observed in each of the six networks. We measured aggregate homophily using a network-level extension of Coleman's Individual Homophily Index (Coleman, 1958), which sums the degree of choice homophily on all ten observable attributes (age, BMI, favorite exercise, etc.) across all newly formed ties in the community (see Appendix). In five of the six trials aggregate homophily is greater than expected by random chance. The small negative index in trial 2 is based on a

[5] Excluded from these 51 tie changes are four instances in which a subject removed a tie shortly after adding it. Our findings do not change when these cases are included.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

Table 2
Revealed-preference model of tie addition. Effect sizes α are reported for popular trait effects, and effect sizes β for choice homophily effects, both with corresponding significance levels (p). The model was estimated using multivariate conditional logistic regression with cluster-robust standard errors ($N = 1170$). The coefficient α in the popular traits column represents the effect of a unit increase in a trait of a potential health buddy on the log odds that a subject will choose to form a tie to that person. The coefficient β in the choice homophily column represents the effect of increasing similarity of a potential health buddy on the log odds that a subject will choose to form a tie to that person.

Variable		Popular traits			Choice homophily		
		α	S.E.	p	β	S.E.	p
Health-related							
Age	(In years)	−.01	(.03)	.800	.16	(.04)	.000***
Gender	Male (vs. Female)	.71	(.60)	.237	2.27	(.62)	.000***
Ethnicity	Asian	.61	(.56)	.280	.67	(.46)	.143
	Hisp	1.04	(.90)	.247			
	Af-Am	.70	(.97)	.468			
	Other (vs. Euro)	1.11	(.60)	.065			
BMI		−.08	(.05)	.124	.18	(.07)	.009**
Fitness		−.11	(.26)	.690	.30	(.24)	.218
Diet preferences	Low Calorie Diet	−1.04	(.78)	.181	.01	(.47)	.986
	Vegan/Veget	−.77	(.46)	.094			
	Carnivorous (vs. Omnivorous)	.39	(.43)	.356			
Exercise-specific							
Exercise intensity		−.20	(.44)	.656	.05	(.33)	.870
Exercise minutes	(in hundreds)	.16	(.13)	.207	.15	(.08)	.058
Exercise goals	Reduce stress	.54	(.48)	.267	.15	(.33)	.658
	Reduce risk	.19	(.69)	.787			
	Look better	.57	(.55)	.299			
	Lose weight (vs. Feel healthy)	.81	(.45)	.074			
Favorite exercise	Walking	.01	(.71)	.989	.73	(.48)	.126
	Running	.26	(.88)	.767			
	Swimming	.11	(.46)	.817			
	Bicycling	.44	(1.67)	.793			
	Elliptical	.79	(.92)	.389			
	Weights	.23	(.68)	.741			
	Team sports (vs. Other)	.73	(.77)	.348			

single observation (it is the only fitness community in which only a single tie change occurred). The positive indices in the other trials are based on multiple tie changes in each community. The independence of the 6 trials permits a statistical evaluation of the null hypothesis that there was no independent trend toward homophilous tie selection across the six trials. Consistent with the homophily hypothesis, we found that there was a significant ($p < .05$ using a two-tailed signed rank test, $N = 6$) overall tendency for participants to initiate ties with homophilous health contacts. At this low level of resolution, with complete statistical independence, this finding shows that homophilous selection forms a dominant aggregate pattern across all active members of the population. Yet, while this indicates a clear trend toward homophilous behavior, it does not permit us to identify which traits participants preferred, and whether these trait preferences were consistent across the active participants.

3.2. Trait preferences in tie selection and removal

Our analysis now turns to the question of which health characteristics participants chose to select on. To begin with, we note that while preferred characteristics, or "desirable traits" did not emerge at the network level, they can yet be present at the individual level. For instance, subjects may have preferred to connect to health buddies who were different on some traits, while similar on others (e.g., same gender, same age, lower BMI). In order to provide a complete picture of individuals' selection preferences, we evaluated both models of selection for each of the 10 observable traits. We evaluated the likelihood that subjects selected on a desirable characteristic (e.g., high fitness, young, low BMI, etc.), and also the likelihood that subjects selected homophilously on each trait, in both cases controlling for all other traits.

We used a revealed preference model (McFadden, 1974) to estimate the independent weights of homophily vs. aspiration on each of the characteristics in the individual selection process.[6] This model simultaneously evaluates the homophilous and aspirational effects of all observable attributes, thereby identifying the specific tendency to make ties based on each characteristic net of all others (see Appendix). We estimated the model by performing conditional logistic regression with robust standard errors (see Appendix).

The effects reported in Table 2 show the log odds that a health buddy was chosen based on preferential attachment.[7] The coefficients in the α column show tie initiators' preferences to form health contacts based on specific "desirable" characteristics. For continuous values, the coefficient indicates incremental increases/decreases in the likelihood of attachment based on a potential contact having a given trait. For instance −.01 for age, means that a unit increase in a potential health buddy's age (e.g., 22 instead of 21) decreases the log odds of that person being chosen by .01. For nominal categories (gender, ethnicity, diet preferences, exercise

[6] See Appendix for a complete description of this model.

[7] Table 2 shows results from the full multivariate model with fixed effects for all homophily and aspirational preference variables, permitting a simultaneous evaluation of all hypotheses. However, the large number of variables and the limited number of positive values for the response variable (51) in this model poses issues of multicollinearity and small-sample bias in maximum likelihood estimation. Results from penalized-likelihood logistic regression with reduced numbers of predictors (not shown here; available from the authors upon request) are substantively the same, with the three homophily effects (age, gender and BMI) maintaining their statistical significance.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

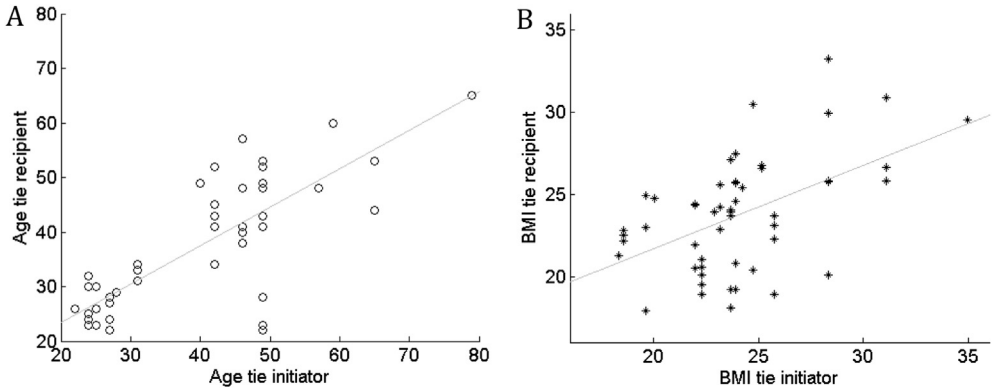

Fig. 2. Homophily on age and BMI. Age of tie initiator by age of tie recipient (panel A) and BMI of tie initiator by BMI of tie recipient (panel B). Panel A and B show a clear tendency for subjects of all ages and all BMI groups to initiate ties with health buddies of comparable age and BMI.

goals, and favorite exercise) we included a dummy variable for each trait category except the most common one, which was taken as the reference category. For instance, the value of .71 on gender indicates that the log odds of a male health buddy being chosen are .71 higher than the log odds of a female health buddy being chosen. This allowed any trait (male or female, high fitness or low fitness, etc.) to show up as more popular.

The results show no significant effects of any preferred traits on the likelihood of tie selection, providing no support for tie formation based on desirable traits, such as youth, low BMI, high fitness, or high exercise minutes. While limited statistical power prevents a hard conclusion about the absence of preferential selection, this finding is nonetheless noteworthy in light of the program goal of increasing participant fitness by providing members with incentives to aspire to more rigorous exercise routines than they would otherwise follow. While we did not observe participants form ties to the healthier members of the community, we did, however, find significant effects of health characteristics in the tie selection process (shown in the β column).

The coefficients in the β column in Table 2 show tendencies among tie initiators to form health contacts based on homophilous preferences. The coefficients in this column indicate bias toward choice homophily on each attribute, again controlling for any preferential or homophilous effects of the other attributes. We found significant choice homophily effects for three characteristics – age, gender, and BMI. For every additional year closer in age to a potential health buddy, subjects were 18% ($[e^{162}-1] * 100\%$) more likely to form a tie to that person.[8] For every BMI point closer in body mass, there was a 19% greater likelihood of forming a social tie. And, subjects were much more likely – 868% more likely – to connect with alters of the same gender than alters of the opposite gender.

These homophilous tendencies for health buddies of similar age, gender, and BMI are interesting not only given the striking absence of aspirational effects, but also because other traits that measure fitness more directly do not seem to have been relevant to subjects. Our conclusion that fitness homophily was absent from subjects' selection behavior is taken with caution, however we note that had the rationale behind subjects' tie choices been to seek a meaningful comparison group for their exercise goals, then we would have expected to see an overall tendency to match on fitness, exercise intensity, exercise minutes, and favorite exercise. Instead, what we observe is that subjects sought out ties to fellow members of categories that do not directly measure fitness or exercise routines, but provide a general, almost demographic reference for health.

Table 3
Revealed-preference model of tie removal. Effect sizes are reported for choice homophily effects with corresponding significance levels (p). The model was estimated using multivariate conditional logistic regression with cluster-robust standard errors ($N = 96$). A coefficient represents the effect of increasing difference on a trait with a health buddy on the log odds that a subject will choose to remove a tie to that person.

Variable	Choice homophily		
	β	(S.E.)	p
Health-Related			
Age (difference in years)	.10	(.05)	.036*
Gender	1.62	(.67)	.016*
Ethnicity	−.20	(.58)	.728
BMI	.14	(.09)	.128
Fitness	−.24	(.40)	.547
Diet preferences	−1.09	(.60)	.070
Exercise-specific			
Exercise intensity	.73	(.42)	.078
Exercise minutes (difference in hundreds)	.19	(.20)	.331
Exercise goals	− .50	(.94)	.590
Favorite exercise	.73	(1.09)	.506

This tendency was remarkably consistent across the full range of age and BMI values. Fig. 2 shows the scatterplot of age (Panel A, circles) for each tie initiator (x-axis) and tie recipient (y-axis); Panel B shows the corresponding scatterplot for BMI (plusses). Both the circles and the plusses follow a clear diagonal pattern from bottom left to top right. The best fitting line (using the method of "least squares") is drawn in both panels and has a steep positive slope in both cases, highlighting the homophilous pattern for tie initiators at all values of both traits. Both panels show a complete absence of violations of this tendency, with none of the 18 tie initiators adding even a single tie to anyone of a very different age or BMI.

Finally, we also observed similar patterns of homophilous bias in the ties that subjects removed. Table 3 shows results for the revealed preference model (same as used above) for tie removal, including both preferred trait effects and homophily effects for all ten attributes.[9] The power of our analysis for tie removal is weaker than for tie addition because the comparison set of possible ties to drop is only six ties, instead of the 65 that subjects could add; however we still found significant results for both age and gender. As was the case in Table 2, significantly positive coefficients in Table 3 indicate homophily and

[8] By "x% more likely" we mean that the odds of one tie being chosen over another tie is increased by x%.

[9] The limited number of cases in the tie removal regression prevents the full model with both homophily and preferred trait coefficients from converging. We experimented with subsets of controls and never found any "preferred trait" to be significant, and consistently found homophily on age, gender and BMI to be strongly significant.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

should be interpreted as increases in the chance of a subject removing a tie as a result of a larger difference along the respective trait. For each year difference in age between a subject and an existing contact, subjects were 10% more likely to remove the tie ($p < .05$). And, subjects were 405% more likely to remove ties to opposite gender partners ($p < .05$). The effect for BMI was in the correct direction – participants were more likely to remove people with larger absolute differences in BMI – however it was not significant. There were no significant effects of preferred traits on tie removal.

Since our study was implemented within a fitness-based social networking site, one particularity relevant question is what our findings imply for other kinds of on-line health environments. The dominant criteria for tie choice may be very different when people with a chronic disease seek emotional support, when people seek advice about the importance of screenings, or when adolescents seek information about safe methods of birth control. Each of these topics provide important directions for future research, which will hopefully offer a broad picture of how informational sources and targets of social comparison are chosen in specific health settings. Our results suggest at least one general implication that may apply across these different kinds of health networks. Namely, selection is biased toward homophilous traits even in contexts where heterophilous ties may be more beneficial.

Just as individuals who want to increase their fitness may select members from similar social categories as their best reference group, people may also elect to receive emotional support, diet information, and medical advice from people with recognizable characteristics. Familiar demographic and health traits may dominate selection choices even when a more medically appropriate fellow patient, or a more informed health resource is available. The consistency of our findings across network-level effects, preferences in tie addition, and preferences in tie removal, indicates strong behavioral trends in subjects' selection behavior, and suggests that as participants altered their health networks, they consciously aimed to surround themselves with health contacts that belonged to the same categories as themselves.

4. Conclusion

In Lazarsfeld and Merton's (1954) study of tie selection, they distinguish between two basic kinds of choice homophily: *value homophily*, based on similar attitudes, beliefs, and behaviors, and *status homophily*, based on nominal status characteristics, such as class, gender, or race (McPherson et al., 2001). Our results suggest that attitudinal factors, such as aspirational interests (i.e., "goals" for the fitness program) and health attitudes (i.e., "diet preferences"), were not primary considerations for tie selection among the subjects in our study. This is perhaps explained by the fact that the focus of the fitness program was particularly tailored to achieving weekly exercise goals. Yet, participants could also have selected on health-based ranking (i.e., "fitness level"), or a number of behavioral factors that were specifically relevant to the goals of the program, such as exercise minutes, exercise intensity, and favorite exercise. Value-based homophily on any of these factors might be motivated by participants' interest in finding relevant comparisons (Festinger, 1954) for evaluating their behavior against others with similar exercise routines and health habits. By connecting to people with similar minutes, intensity, fitness level, or exercises, participants could establish a benchmark with peers whose fitness profiles were similar to their own, and whose exercise goals and aspirations would also be similar. Yet, our results indicate that participants did not select ties based on any of these characteristics. Rather, they seem to have mostly ignored value homophily, and selected ties overwhelmingly based on status characteristics (McPherson et al., 2001).

In offline health networks, where participants have no pre-existing relationships, status characteristics, like age, gender, and BMI are intuitive selection factors because they are readily observable features, which can easily be used to infer a potential partner's relevance for one's own exercise behavior. They provide a simple and effective heuristic for selecting health contacts in the absence of easily identifiable traits, such as regular exercise intensity, minutes, or activities. However, to facilitate participants' ability to make the most relevant connections, our program explicitly revealed these fitness characteristics (i.e., intensity, minutes, and activities), which were specifically targeted to the task of increasing and maintaining participants' exercise levels. Participants' choices to instead select homophilously on familiar demographic and health traits suggests that not only did individuals not select "health leaders" or "desirable individuals," but they did not even select the individuals who might form the most apt comparison group for evaluating their weekly progress toward the program's goals. At both the high and low scales of health status, participants reproduced the basic forms of status homophily that might otherwise be created by social exclusion and institutional sorting (McPherson and Smith-Lovin, 1987; Crosnoe et al., 2008).

The endogenous, choice-based emergence of these status preferences raises the curious question of whether individuals chose these characteristics simply because they are already familiar with these attributes from their offline experiences, or whether they "intrinsically" prefer connecting to others with these characteristics. While individuals may indeed prefer homophily along observable characteristics, a longstanding sociological observation suggests that social structures frame individual expectations (Marx, 1977 [1867], Weber, 1978 [1922], Berger et al., 1977). Our results may thus suggest that the freedom of the online space may be bounded in distinct ways by the social traditions that precede it. Thus, while our fitness study was strategically constructed to eliminate all the constraints on tie choice that normally limit opportunities for interaction across social categories, participants nonetheless deliberately avoided selecting alters with significant differences on these categories. The conclusion from our findings is that in the online fitness context, people prefer to make ties to the "devil they know." By selecting ties based on familiar social characteristics, participants may unintentionally limit their available social capital, and restrict their opportunities for finding new health information from sources that they are not normally exposed to.

For organizations interested in using online health networks to promote informational access and greater social exposure, this suggests that they may need to provide participants with an interest in forming ties that cut across traditional boundaries. For instance, postings that encourage heterophilous or activity-specific ties may increase participants' likelihood of making these connections. Further, promoting tie formation across traditional status boundaries may require incentivizing higher health-status people to initiate tie formation, or perhaps providing program goals that are particularly tailored to encouraging participants to make social ties to health buddies whom they would not otherwise meet. Introducing these goals and incentives can both increase people's awareness of potential contacts, and highlight their salience for those outside their status community. Both of which may be necessary to achieve the goal of increasing participants' exposure to new sources of health information and influence. Finally, these incentives may also increase participation in health programs overall by highlighting the value of online social capital for discovering health resources that are less easily found in contexts with less transparent information, and higher barriers to introduction.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

Acknowledgments

We thank A. Wagner and T. Groves for website development; and K. Schive, M. Kirkbride, and M.I.T. Medical for assistance with site design and participant recruitment. DC is grateful for support from the James S. McDonnell Foundation and the MIT Research Support Committee. AR was supported by National Science Foundation Grant SES-1340122.

Appendix

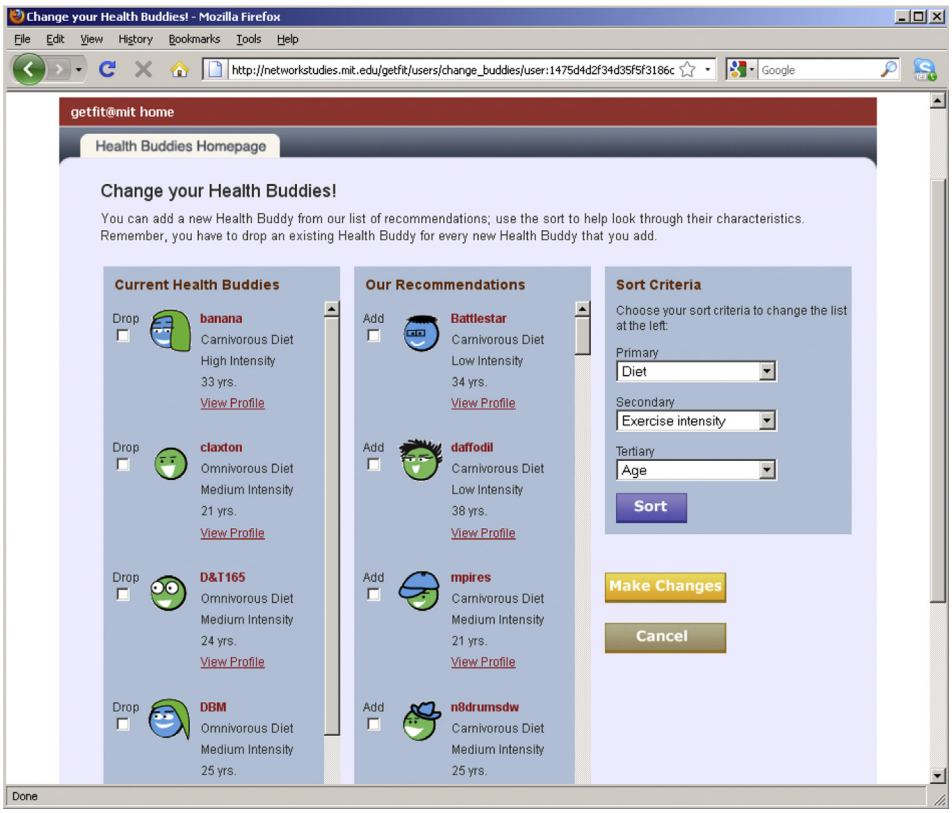

Fig. A1. The Add/Drop page for changing health buddies

Theoretical implications of tie selection for network dynamics

Our empirical approach places our investigation on a very specific theoretical footing. Once social influence is removed, and all of the social and structural constraints on tie formation are accounted for, what remains are two basic theoretical positions in the literature on individual tie selection. The first is that tie preferences are fundamentally homophilous. This position states that regardless of the abundance of other social mechanisms that can confound and obscure the effects of true choice homophily, individual preferences are indeed homophilous. In other words, once organizational constraints and influence processes are removed, network patterns of observed homophily will remain. Because of the widely observed patterns of homophily in social networks, this view is the default expectation, and it also frames our primary hypothesis for this study. However, an increasingly popular alternative view of selection, which is based on a growing interdisciplinary literature on social tie formation, is that individual preferences are fundamentally based on "preferential" interests in social contacts (Kalmijn, 1994; Crosnoe et al., 2008; Hitsch et al., 2010; Ali et al., 2012; Dahlander and McFarland, 2013). In this view, most members of a population will tend to disproportionately connect to a small number of select individuals, who have the most desirable characteristics. While patterns of everyday homophily may seem to belie this mechanism, as we observed above, preferential tie selection may itself generate patterns of observed homophily through the complex dynamics of competition and exclusion.

Once traditional forces constraining tie formation are removed, a signature difference between these two choice mechanisms is the network structures that will emerge. Because homophilous choice implies that people are similar to their friends, and their friends' friends are most likely also quite similar, homophilous preferences will tend to result in people's friends being connected to each other, creating lots of triangles, or "clustering," in the social network. As "neighborhoods" form, the distribution of ties over the population will typically be quite even, resulting in networks in which everyone has a similar number of social ties, and is connected in clustered, homophilous social cliques (Centola et al., 2007). By contrast, when people select ties preferentially, the emergent social network will have a skewed distribution in the number of ties per person (i.e., "degree"), resulting in most people having only a few ties, and a few people having a large number of ties. This is because most people attach to highly desirable social "stars," and not to one another, resulting in a network that has low levels of clustering, and high levels of heterophily. Highly skewed networks have been shown to occur in environments where tie formation is relatively

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

unconstrained, e.g., in human sexual contact networks (Schneeberger et al., 2004).

The dynamics of tie selection, and their resulting network structures have clear implications for the effective transmission of health information through social networks. Scale-free networks, with highly connected hubs and low clustering, can be extremely effective social structures for disease diffusion, as well as for the rapid transmission of new information. By contrast, clustered networks tend to reduce the novelty of the information that people are exposed to because their social contacts are primarily people who have the same resources and characteristics as themselves, and each other. While a large literature discusses the implications of network structures for information spreading (Granovetter, 1973, Watts, 1999; Centola and Macy, 2007), much less is known about which individual preferences govern tie choice, and what this portends for the process of network evolution within online communities. The present study thus helps to circumscribe how individual tie selection may aggregate into community-level pathways that shape members' access to new informational and behavioral influences.

Calculation of homophily measures

Homophily on continuous and ordinal attributes (age, BMI, fitness, exercise intensity & exercise minutes) was operationalized as the negative absolute difference between the tie initiator and tie recipient on each trait. For instance, if a tie initiator's age was '22' and a tie recipient's age was '30,' then the measured age similarity would be '-8.' The lower the absolute difference, the greater the similarity, with '0' being maximum similarity.[10] Similarity on nominal attributes (gender, ethnicity, diet preferences, exercise goals, favorite exercise) was defined as '1' in cases where ego and alter share a trait and '0' otherwise.

We measured aggregate homophily using a network-level extension of Coleman's Individual Homophily Index (Coleman, 1958), which sums the degree of choice homophily on all ten observable attributes (age, BMI, favorite exercise, etc.) across all newly formed ties in the community (see Appendix). In order for the ten attributes to all contribute equally to the aggregate index, the measure of homophily must be normalized across attributes (e.g., Age: 2 years apart, Gender: same gender, BMI: 3 BMI points different, etc.). We thus calculated the rank[11] of the level of homophily on each attribute for each of the 65 ties a tie initiator could have chosen. For example, a tie initiator who chose the health buddy that was the 3rd nearest in age received a rank of 3 on age. To obtain a measure of homophilous bias we used the baseline null hypothesis of random tie formation (Coleman, 1958; Fararo and Sunshine, 1964; Rapoport, 1979; Currarini et al., 2009). Accordingly we subtracted the *observed* rank from the *expected* rank under random tie choice, namely the mean rank of 33, resulting in a rank score between −32 (maximal heterophily) and +32 (maximal homophily). For the above example, in which an individual selects the 3rd closest person on age in the entire population, this results in a rank score of +30 (33−3), indicating strong homophily on age. We calculated the aggregate homophily index H for each community by summing these rank scores across all attributes and across all chosen ties:

$$H = \sum_{\text{new ties}} \sum_{\text{attributes}} \text{expected rank} - \text{observed rank}$$

[10] There were no general tendencies toward selection of partners who were either somewhat above or somewhat below the tie initiator's age, BMI, or fitness level.

[11] Alternatively, one could normalize by dividing demeaned scores by the standard deviation of the homophily variable. Analyses not shown here confirm that this alternative procedure yields an identical test result. We preferred sums of ranks, as for most attributes the homophily distribution does not approximate a normal.

Revealed preference model

To identify the individual decision-making process underlying the observed patterns of network formation, we used a revealed preference model (Thurstone, 1927; Mosteller, 1951; McFadden, 1974; Steglich et al., 2010) to estimate the independent weight of each of the characteristics in the individual selection process. This model evaluates whether individuals selected on particular traits, controlling for the selection effects of all other traits. The model also separates out aspirational tendencies to select individuals with certain popular traits vis-à-vis homophilous tendencies to connect to individuals with matching traits.

The model assumes that an individual ascribes a utility to being connected to a given health buddy x, denoted by $u(x)$. Utility depends on alter's traits as well as on how well alter's traits match ego's traits. Specifically, the individual's utility from connecting to alter x is conceived of as a linear combination of alter x's score m on each of the attributes, a, the focal individual's similarity to x, s, on each of the attributes, a, and a random utility term, ε:

$$u(x) = \sum_a \alpha_a m_a(x) + \beta_a s_a(x) + \varepsilon \tag{A1}$$

The addition of the random term ε to the utility function can be interpreted as rendering each individual's choice from the set of available ties as boundedly rational (Young, 1998). That is, individuals seek to optimize utility but do so only imperfectly. If ε is i.i.d. and Gumbel distributed, then the probability $p(x)$ that a tie to x is chosen among all candidate ties X is given by (Luce and Suppes, 1965):

$$p(x) = \frac{e^{\sum_a \alpha_a m_a(x) + \beta_a s_a(x)}}{\sum_{x \in X} e^{\sum_a \alpha_a m_a(x) + \beta_a s_a(x)}} \tag{A2}$$

Equation (A2) is a conditional logistic regression model (McFadden, 1974) with the coefficients α — shown in Table 2 — representing the relative popularity of the traits and β — also shown in Table 2 — representing the relative weight of similarity on each of the attributes in an actor's utility function. Maximum likelihood estimates of coefficients α and β can be estimated directly from the data. A positive (negative) coefficient α_a would indicate that — all else being equal — subjects in the study sought out (avoided) health buddies with trait a. A positive (negative) coefficient β_a would indicate that subjects sought out (avoided) ties to health buddies matching on trait a. We employed multivariate conditional logistic regression with cluster-robust standard errors to predict the log odds of tie choice on the basis of all ten attributes. For each of the individuals across the 6 networks who made at least one tie change (the "tie initiators"), we included in the choice set all 65 ties they could add, denoting all realized choices with a '1', and marking the possible tie with a '0' otherwise. In this conditional logistic regression only within-subject comparisons are made (i.e., fixed effects), ensuring that between-subject differences in passivity of behavior and in the availability of health buddies did not affect the results. In separate network-level fixed-effects regression, as well as unconditional logistic regression, we found the same attributes to affect choice. The conditional logit model further assumes independence of choice behavior across individuals. We believe this assumption is reasonable in the present context as almost all individuals in our study who made a tie change had passive neighbors.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

A final assumption the model makes is that of subject homogeneity in choice behavior whereby choices by all subjects are made on the basis of the same homophily considerations. This assumption risks running afoul of the ecological fallacy since individual subjects' preferences may well have varied significantly, even given the clarity of our finding of average population preferences for homophilous ties (Robinson, 1950). Such errors are common in the observational literature because the dominant theories of tie selection (preferential attachment and homophily) posit a general population tendency for tie formation, while foregoing exploration of possible variation in selection behavior across members. To investigate potential variability, we disaggregated the observed tie choices and evaluated them individually. We constructed a contingency table, cross-tabulating the degree of similarity of new ties on (4 rows) by the individuals who initiated those new ties (18 columns). The degree of similarity was measured as the sum of the traits on which ego and alter match among those traits that were found significant in the multivariate analysis: age, gender and BMI. For age a threshold difference of 9 years was used while for BMI a difference score of 4 points was used to differentiate between matches and non-matches. By dichotomizing homophily levels into a binary score (match vs. non-match) and summing these binary scores into a single homophily measure we prevented the contingency table from becoming too sparse for meaningful analysis. The cut-points of 9 for age and 4 for BMI split the population into approximate equal halves. Other cut-points yield similar results.

This contingency table is visualized in Fig. A2. Under homogeneity of choice behavior, individuals would all exhibit the same probability of initiating a tie of a given level of homophily. We found that only 1 participant (initiator 13) initiated ties to completely dissimilar others. Across all remaining initiators, the propensity of varying levels of homophily is approximately evenly distributed. We performed a test of this homogeneity assumption. Under the assumption of homogeneity in choice behavior, the rows and columns of the resulting table should be statistically independent. A Fisher Exact Test for independence of rows and columns confirms that tie choice was similarly homophilous across all individuals ($p = .137$). We repeated this test procedure for possible heterogeneity across subjects in different trials, and found no systematic variation across trials ($p = .667$).

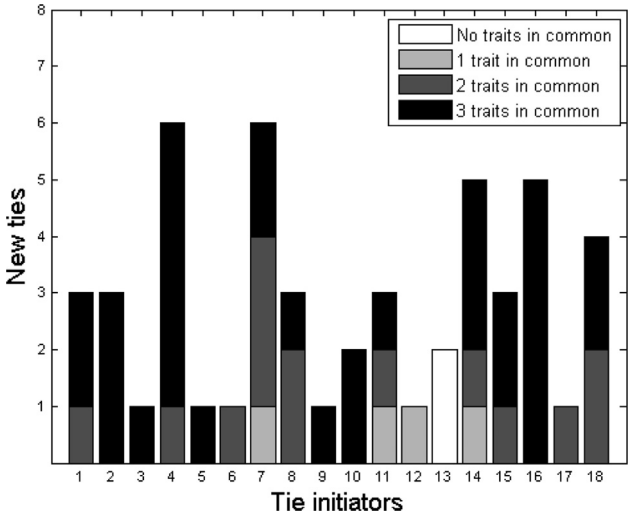

Fig. A2. Subject homogeneity of choice behavior. For each tie initiator each new tie to a "health buddy" ($N = 51$) is represented by a bar unit. Bar colors white, light-gray, dark-gray, and black correspond to having respectively 0, 1, 2, and 3 of the traits age (young/old), gender (female/male) and BMI (overweight/normal weight) in common with the new health buddy. Homophily does not significantly vary across subjects ($p = .137$ using a Fisher Exact Test).

Table A1
Revealed-preference model of tie addition, for all subjects ($N = 1170$; left column) and all but 2 most active subjects ($N = 1040$; right column). Effect sizes, β, are reported for choice homophily on each trait, with corresponding significance levels (p). Models were estimated using multivariate conditional logistic regression with cluster-robust standard errors. A coefficient β represents the effect of increasing similarity of a potential health buddy on the log odds that a subject will choose to form a tie to that person.

Variable	All subjects		All but 2 most active subjects	
	Choice homophily		Choice homophily	
	β	P	β	P
Health-related				
Age	.16	.000***	.21	.001**
Gender	2.27	.000***	2.09	.000***
Ethnicity	.67	.143	.60	.487
BMI	.18	.009**	.22	.012*
Fitness	.30	.218	.47	.039*
Diet preferences	.01	.986	.03	.951
Exercise-specific				
Exercise intensity	.05	.870	.30	.427
Exercise minutes	.15	.058	.10	.319
Exercise goals	.15	.658	.17	.684
Favorite exercise	.73	.126	.42	.452

Sensitivity of main results to exclusion of active subjects

As Fig. A2 shows, some subjects changed many more ties than others. Because of the potential for one or two individuals to influence the results, we evaluated the robustness of our findings across a series of subsamples of our data that exclude the most active subjects. We generated six reduced datasets, which eliminated each combination of two out of the four most active subjects. On these reduced datasets, we performed the aggregate homophily test reported in Fig. 1 which continued to show a significant ($p < .05$) effect of choice homophily in all ten cases. We also re-estimated the revealed preference model for all reduced datasets and found that the three main homophily effects (age, gender and BMI) continued to be statistically significant ($p < .05$). As an illustration, Table A1 compares the choice homophily estimates originally reported in Table 2 with corresponding estimates from a reduced dataset in which data from the two most active subjects (tie initiators 4 and 7 from Fig. A2) are excluded. In both cases, we find significant effects of choice homophily on age, gender, and BMI.

Differences between tie initiators and other subjects

Our primary goal with this study is to identify how participants in an online fitness program select their health contacts – whether they select them based on aspirational preferences for health leaders, or whether they select individuals who are primarily similar to themselves. As reported in the main text, there were unexpectedly low levels of tie selection behavior with 51 tie changes made by only 18 of the 432 subjects. This lack of activity can be interpreted as either indifference to social contacts, or lack of participation in the program as a whole. We consider these in turn.

First, lack of activity may suggest that the majority of participants were not discriminating about their health contacts. As a result, whatever contacts were provided for them initially would be

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

perceived as relevant social influences, and they would not voluntarily seek out new contacts. This would suggest that the findings on tie activity represent the behavior of a small fraction of the population whose tastes differ from the majority. Among active participants, we observed a remarkably strong signal regarding their selection preferences. Since the goal of our study is to identify how ties are selected, this would then suggest that networks are largely stable, even based on arbitrary initial assignments of ties, but to the degree that they do evolve, we are able to clearly identify basic preferences that drive tie formation among the active individuals.

Second, the lack of activity may suggest that these participants in the study were less engaged in the health program in which the study was embedded. In this case, our findings would accurately reflect the behavior of the active members of the health program; that is, among the subjects who were actually participating in the program, we observed significant trends in their tie preferences.

To investigate these possibilities, we compared the 18 tie initiators with other subjects to see if the former represented a particular demographic or rather a user base with greater user participation. Table A2 displays the health-related and exercise-specific traits of tie initiators and other subjects. For categorical variables Table A2 shows the most common category and the results of a Fisher exact test for differences between tie initiators and other subjects are reported. For continuous variables Table A2 shows the average value and the results of a rank-sum test for differences between tie initiators and other subjects are reported. There is a significant tendency for women, who constitute the majority of subjects in the study, to engage in more networking activity than men ($p = .005$). There are no other significant differences between tie initiators and other subjects on health-related and exercise-specific traits.

Table A2 also shows two measures of user participation in the study. The first measure, average # of clicks on buddies, captures the level of interest that participants had in comparing exercise activities and progress with their health buddies. Tie initiators were about 12 times as active in such as other subjects ($p = .000$). The second measure, average # of active weeks, measures the number of weeks in which at least some minimal online activity level was recorded for a subject. Tie initiators were active 2 of the 5 weeks while people who did not engage in tie activity were largely inactive (showing an average of .19 weeks of recorded activity, $p = .000$). Together, these results indicate that the dominant determinant of networking activity was overall participation in the fitness program.

Table A2
Differences between tie initiators and other subjects.

Variable	Tie initiators	Other subjects	Test for difference	Significance
Health-related				
Percent female	94%	63%	Exact	$p = .005^{**}$
Average age	41	35	Rank-sum	$z = 1.60$; $p = .109$
Most common ethnicity	White	White	Exact	$p = .459$
Average BMI	24	25	Rank-sum	$z = .669$; $p = .503$
Most common fitness	"Above average"	"Above average"	Exact	$p = .166$
Most common diet preference	Omnivorous	Omnivorous	Exact	$p = .374$
Exercise-specific				
	"Medium"	"Medium"	Exact	$p = .263$
Most common exercise intensity				
Average exercise minutes	211	183	Rank-sum	$z = 1.01$; $p = .314$
Most common exercise goal	"Feel healthy"	"Feel healthy"	Exact	$p = .375$
Most common favorite exercises	Swim & walk	Swim & walk	Exact	$p = .839$
User participation				
Average# clicks on buddies	12.6	1.04	Rank-sum	$z = 7.69$; $p = .000^{***}$
Average# active weeks	2.0	.19	Rank-sum	$z = 7.80$; $p = .000^{***}$
N	18	414		

References

Ali, M.M., Amialchuk, A., Rizzo, J.A., 2012. The influence of body weight on social network ties among adolescents. Econ. Hum. Biol. 10, 20–34.
Aral, S., Muchnik, L., Sundararajan, A., 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc. Natl. Acad. Sci. 106, 21544–21549.
Berger, J., Fisek, M.H., Norman, R.Z., 1977. Status Characteristics and Social Interaction: an Expectation-States Approach. Elsevier, New York.
Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am. Econ. Rev. 94, 991–1013.
Blau, P.M., 1977. Inequality and Heterogeneity: a Primitive Theory of Social Structure. Free Press, New York.
Blau, P.M., Schwartz, J.E., 1984. Crosscutting Social Circles: Testing a Macrostructural Theory of Intergroup Relations. Academic Press, New York.
Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., Fowler, J.H., 2012. A 61-million_person experiment in social influence and political mobilization. Nature 489, 295–298.
Centola, D., 2010. The spread of behavior in an online social network experiment. Science 329, 1194–1197.
Centola, D., 2011. An experimental study of homophily in the adoption of health behavior. Science 334, 1269–1272.
Centola, D., 2013. Social media and the science of health behavior. Circulation 127, 2135–2144.
Centola, D., Macy, M., 2007. Complex contagions and the weakness of long ties. Am. J. Sociol. 113, 702–734.
Centola, D., Gonzalez-Avella, J.C., Eguiluz, V., San Miguel, M., 2007. Homophily, cultural drift, and the co-evolution of cultural groups. J. Confl. Resolut. 51, 905–929.
Chou, W.S., Hunt, Y.M., Beckjord, E.B., Moser, R.P., Hesse, B.W., 2009. Social media use in the United States: implications for health communication. J. Med. Internet Res. 11, e48.
Christakis, N., Fowler, J., 2007. The spread of obesity in a large social network over 32 years. N. Engl. J. Med. 357, 370–379.
Coleman, J.S., 1958. Relational analysis: the study of social organizations with survey methods. Hum. Organ. 17, 28–36.
Crosnoe, R., Frank, K., Mueller, A.S., 2008. Gender, body size, and social relations in American high schools. Soc. Forces 86, 1189–1216.
Currarini, S., Jackson, M.O., Pin, P., 2009. An economic model of friendship: homophily, minorities, and segregation. Econometrica 77, 1003–1045.
Dahlander, L., McFarland, D.A., 2013. Ties that last. Tie formation and persistence in research collaborations over time. Adm. Sci. Q. 58, 69–110.
Fararo, T.J., Sunshine, M.H., 1964. A Study of a Biased Friendship Network. Syracuse University Press, Syracuse, NY.
Feld, S.L., 1982. Social structural determinants of similarity among adolescents. Am. Sociol. Rev. 47, 797–801.
Festinger, L., 1954. A theory of social comparison processes. Hum. Relat. 7, 117–140.
Fogel, J., Albert, S.M., Schnabel, F., Ditkoff, B.A., Neugut, A.I., 2002. Internet use and social support in women with breast cancer. Health Psychol. 21, 398–404.
Fox, S., 2011. The Social Life of Health Information. Pew Research Center Report.
Granovetter, M., 1973. The strength of weak ties. Am. J. Sociol. 78, 1360–1380.
Hawn, C., 2009. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. Health Aff. 28, 361–368.
Hitsch, G.J., Hortaçsu, A., Ariely, D., 2010. Matching and sorting in online dating. Am. Econ. Rev. 100, 130–163.
Japuntich, S.J., Zehner, M.E., Smith, S.S., Jorenby, D.E., Valdez, J.A., Fiore, M.C., Baker, T.B., Gustafson, D.H., 2006. Smoking cessation via the internet: a randomized clinical trial of an internet intervention as adjuvant treatment in a smoking cessation intervention. Nicot. Tob. Res. 8, S59–S67.
Kalmijn, M., 1994. Assortative mating by cultural and economic occupational status. Am. J. Sociol. 100, 422–452.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019

Kalmijn, M., Vermunt, J., 2007. Homogeneity of social networks by age and marital status: a multilevel analysis of ego-centered networks. Soc. Netw. 29, 25–43.
Kossinets, G., Watts, D.J., 2009. Origins of homophily in an evolving social network. Am. J. Sociol. 115, 405–450.
Lazarsfeld, P., Merton, R.K., 1954. Friendship as a social process: a substantive and methodological analysis. In: Berger, M., Abel, T., Page, C.H. (Eds.). Van Nostrand, New York, pp. 18–66.
Lewis, K., Gonzalez, M., Kaufman, J., 2012. Social selection and peer influence in an online social network. Proc. Natl. Acad. Sci. U S A 109, 68–72.
Luce, R.D., Suppes, P., 1965. Preference, utility, and subjective probability. In: Luce, R.D., Bush, R., Galanter, E.H. (Eds.), Handbook of Mathematical Psychology, vol. 3. Wiley, New York, pp. 249–410.
Mare, R.D., 1991. Five decades of educational assortative mating. Am. Sociol. Rev. 56, 15–32.
Marx, K., 1977 [1867]. Capital: a Critique of Political Economy, vol. I. Penguin Books, London.
McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.
McNab, C., 2009. What social media offers to health professionals and citizens. Bull. World Health Organ. 87, 566.
McPherson, M., Smith-Lovin, L., 1987. Homophily in voluntary organizations: status distance and the composition of face-to-face groups. Am. J. Sociol. 52, 370–379.
McPherson, M., Smith-Lovin, L., Cook, J., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27, 415–444.
Mercken, L., Snijders, T.A.B., Steglich, C., de Vries, H., 2009. Dynamics of adolescent friendship networks and smoking behavior: social network analyses in six European countries. Soc. Sci. Med. 69, 1506–1514.
Moody, J., 2001. Race, school integration, and friendship segregation in America. Am. J. Sociol. 107, 679–716.
Mosteller, F., 1951. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. Psychometrika 16, 3–9.
Pampel, F.C., Krueger, P.M., Denney, J.T., 2010. Socioeconomic disparities in health behaviors. Annu. Rev. Sociol. 36, 349–370.
Popielarz, P., McPherson, M., 1995. On the edge or in between: niche position, niche overlap, and the duration of voluntary association membership. Am. J. Sociol. 101, 698–720.
Rapoport, A., 1979. Some problems relating to randomly constructed biased networks. In: Holland, P., Leinhardt, S. (Eds.), Perspectives on Social Network Research. Academic Press, New York, pp. 119–164.
Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. Am. Sociol. Rev. 15, 351–357.
Ruef, M., Aldrich, H.E., Carter, N.M., 2003. The structure of founding teams: homophily, strong ties, and isolation among U.S. entrepreneurs. Am. Sociol. Rev. 68, 195–222.
Salathe, M., Khandelwal, S., 2011. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput. Biol. 7, e1002199.
Schneeberger, A., Mercer, C.H., Gregson, S.A., Ferguson, N.M., Nyamukapa, C.A., Anderson, R.M., Johnson, A.M., Garnett, G.P., 2004. Sex. Transm. Dis. 31, 380–387.
Shalizi, C.R., Thomas, A.C., 2011. Homophily and contagion are generically confounded in observational social network studies. Sociol. Methods Res. 40, 211–239.
Steglich, C., Snijders, T.A.B., Pearson, M., 2010. Dynamic networks and behavior: separating selection from influence. Sociol. Methodol. 40, 329–393.
Thackeray, R., Neiger, B.L., Hanson, C.L., McKenzie, J.F., 2008. Enhancing promotional strategies within social marketing programs: use of Web 2.0 social media. Health Promot. Pract. 9, 338–343.
Thurstone, L.L., 1927. The method of paired comparisons for social values. J. Abnorm. Soc. Psychol. 21, 384–400.
Van de Rijt, A., Siegel, D., Macy, M., 2009. Neighborhood chance and neighborhood change. Am. J. Sociol. 114, 1166–1180.
Watts, D., 1999. Small Worlds: the Dynamics of Networks Between Order and Randomness. Princeton University Press, Princeton, NJ.
Weber, M., 1978 [1922]. Economy and Society. University of California Press, Berkeley, CA.
Wellman, B., 2001. Physical place and cyber place: the rise of personalized networking. Int. J. Urban Reg. Res. 25, 227–252.
Wellman, B., Hampton, K., 1999. Living networked on and offline. Contemp. Sociol. 28, 648–654.
White, M., Dorman, S.M., 2001. Receiving social support online: implications for health education. Health Educ. Res. 16, 693–707.
Wimmer, A., Lewis, K., 2010. Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. Am. J. Sociol. 116, 583–642.
Yamaguchi, K., Kandel, D., 1993. Marital homophily on illicit drug use among young adults: assortative mating or marital influence? Soc. Forces 72, 505–528.
Young, H.P., 1998. Individual Strategy and Social Structure: an Evolutionary Theory of Institutions. Princeton University Press, Princeton, NJ.

Please cite this article in press as: Centola, D., van de Rijt, A., Choosing your network: Social preferences in an online health community, Social Science & Medicine (2014), http://dx.doi.org/10.1016/j.socscimed.2014.05.019