"Friending" in Online Fitness Communities: Exploring Activity-Based Online Network Structure

Li Zeng
University of Washington
lizeng@uw.edu

Zack W. Almquist
University of Minnesota
almquist@umn.edu

Emma S. Spiro
University of Washington
espiro@uw.edu

Abstract

Individuals are influenced by both direct and indirect interaction with their social contacts. While peer influence is known to affect health-related outcomes such as exercise, limited work has fully explored how social networks are structured to support (or inhibit) interaction that could lead to positive health behaviors. With the development of pervasive technology and rise of personal health and wellness tracking, increasing attention has been paid to promoting positive fitness behaviors through social interaction mechanisms in online fitness communities. This trend offers a unique opportunity to understand the opportunity structures for personal health and wellness support. Utilizing a large-scale behavioral trace dataset from the online fitness community Strava, we examine how the size of people's personal network is structured by demographics (e.g. gender and age) and an economic indicator (i.e. if they pay for a premium account). We employ stochastic process models to characterize the empirical network degree distributions in this population of fitness community members. We find that gender, age and account status are associated with distinct network structure. Results have implications in the analysis and the design of health interventions that make use of network relationships in online settings.

1. Introduction

People are influenced by the social networks in which they are embedded. It is known that direct and indirect social interaction affect not only people's opinions and attitudes, but also the actions they take and decisions they make [1, 2]. Social interaction among individuals in a social network leads to social integration, social cohesion and social support [3, 4]. A number of prior studies have demonstrated that social support and peer influence have positive effects on health outcomes, for example motivating people to exercise more, lose weight and quit smoking [5, 6].

However, research into the social dynamics of health behaviors is still in its infancy. Indeed, while much previous research has studied the function of social support [4, 5], limited work has focused on the structural mechanisms that encourage or suppress certain social interactions, and therefore pathways for peer influence and social support.

Recent years have seen a rapidly growing body of research on promoting health and fitness through social interaction in online fitness communities [7, 8, 9]. Online fitness communities not only track and archive activities, but also adopt social networking features meant for increasing user engagement and retention [9, 10]. One example of such a platform is Strava. Strava allows users to post activity updates, interact with others by commenting and liking others' activities, and designate subscription relationships to keep up to date on the activities posted by friends and other users of interest. Prior research demonstrates that people think carefully when deciding with whom to interact in online health communities [11]. We build on this work and ask the question: what do people's personal networks look like in online fitness communities, and can we understand underlying "friending" processes from the aggregate characteristics of networks across the population? Understanding of the underlying mechanisms of relationship construction in this context is an essential step in analysis and design of health interventions that aim to utilize network structure and processes to promote specific behaviors.

Online activity tracking platforms record digital traces of human behavior as people use the service, offering a new data source and an unprecedented opportunity to study social networks at scale and with finer temporal granularity. More importantly, behavioral trace data allows researchers to observe how individuals really engage in physical activity and social interaction, rather than their reported participation as in much prior work in this area. Researchers can test existing theories on the large-scale data, as well as study new forms of activity-related behaviors in online communities.

HİCSS

In this work we examine network structure, exploring potential underlying mechanisms for the formation of personal networks. We utilize a large-scale collection of behavioral traces of social relationships in the online fitness community Strava over eight years. Specifically, we aim to address the following research questions: (1) *How are activity-based egocentric networks patterned (in terms of in-coming and out-going ties), and are these patterns structured along demographic and economic lines?*, and (2) *Does observed network structure match specific underlying mechanisms for the formation of social relationships? If so, what differences are observed between social and demographic groups?*

2. Related Work

2.1. Personal Networks

Demographic and socioeconomic factors are known to structure real-world networks; for example, [12] analyzed a network of Christmas card giving and found that age, household type, and the relationship to the card giver are associated with the resulting network structure. Homophily is observed time and time again in real-world networks. Individuals associate with others who have similar characteristics, interests, opinions and behaviors. Homophilous relations can be formed based on shared demographics including race, ethnicity, age, religion, education [13, 14]. Geographic propinquity has also been found to structure human interaction and hence social relationships [14]. [11] found that users in online health communities prefer to interact with those who are in similar situations whether it's losing weight, managing diabetes or having similar health goals.

Prior work has studied not only the structure of social networks, but also the processes that could construct them. The implications of understanding the underlying social processes vary from case to case. For example, [15, 16] proposed a set of stochastic/generative models for personal networks. They demonstrate how understanding the potential underlying mechanism of relationship construction in this type of network is vital for public health interventions – attempts to control and eventually eradicate sexually-transmitted diseases. Another work [17] analyzed the structure of multiple online social networks (e.g. Flickr, YouTube, LiveJournal, and Orkut) and found a much larger proportion of symmetric ties and much higher level of local clustering in these social networking sites than other networks. In this case, the authors argue that understanding the structure of online social networks helps explain the robustness and security of such networks. In each of these cases, researchers aim to understand how opportunities for interaction lead to a specific pattern of social relationships, which in turn affects processes that occur via the network ties.

2.2. Online Fitness Communities

Online fitness communities are web and mobile application-based environments allowing users to interact in order to support and encourage others to achieve fitness goals. Recently, we have seen the release of many online fitness communities, paralleling the explosion of personal health and wellness tracking wearable devices. Strava, RunKeeper, Nike+, and MapMyRun are just some of the popular new platforms. Online fitness communities usually combine both activity tracking features and social networking features [18, 10]. Strava illustrates this trend. It has marketed itself as an online fitness community for runners and cyclists, allowing users to post activities and "prove it" to their peers and friends. The platform also allows users to follow others – both recreational and professional athletes, view their activities and interact with others by making comments, giving kudos, etc. on specific activities posted.

Online fitness communities record large amounts of behavioral trace data generated by users – both via their exercise behaviors as well as through their online social interaction. Compared to self-reported data, which measures individuals' perceptions of their workout routine and social engagement, behavioral trace data has many advantages including the ability to record people's exact activity performance and precise social interaction (e.g., running in groups, comments and likes given to someone's activity). Behavioral trace data is not subject to the same issues of accuracy and recall error as more traditional data collection methods. This data source also has advantages in terms of scale, granularity and observation duration against data collected via expensive sensors [19]. Thus it offers a new window into how social interaction unfolds in large-scale online social networks.

Most research on online fitness communities focuses on design of incentives and interventions [5, 20, 21], or on the technical potential of wearable sensors and other tracking technologies [22, 23, 24]. A much less explored aspect of this phenomena is the dynamics of social interaction and how the resulting networks offer opportunities and constraints on peer influence and health outcomes. [25], for example, found that exercising with peers physically co-present is associated with greater exercise performance. Clearly more work is needed to characterize and explore the consequences

of social network structure in these settings.

3. Data

This work utilizes behavioral trace data from an online activity-tracking platform, Strava. Since its launch in 2009, Strava has been a rapidly growing social network for athletes around the world. Strava incorporates features from both activity tracking applications and social media. Users can track, record and share physical activities (e.g. running, cycling) via the Strava website and/or mobile application. The social media side of Strava is similar to social networking sites such as Facebook or Twitter. The platform allows users to explore and search for other users. Before connecting to peers on the platform, users can view their profiles (e.g. name, photo, location), activity posts, achievements, and their personal networks. Users can create out-going connections to their peers by simply clicking a "Follow" button (as on Twitter) unless follow requests are required by the target user. Users will be notified if someone has followed them and can create mutual connections if interested. As discussed subsequently, many relationships remain asymmetric in this setting.

This work studies personal networks of 69,658 anonymized Strava users in the community. The dataset used was obtained directly from the research team at Strava through a data sharing agreement. Data focuses on all active users located in two major metropolitan areas San Francisco, CA and Boston, MA. A geographic-based sampling strategy was chosen to allow for geographic comparison, in this case East versus West coast of the United States. San Francisco and Boston are very similar metro areas in terms of their mix of jobs, though separated by over 2500 miles [26]. They also have similar land area and population. Boston is slightly less dense and younger in terms of media age. San Francisco, CA is where Strava started its business; then the company has gained increasing popularity in the U.S. East coast and globally. Both locations have a large set of active Strava users; in fact, coastal regions and metropolitan areas have much higher rates of usage than elsewhere[1]. While sampling may limit the generalizability of any geographic-specific findings, this dataset enables us to take the first step to study potential patterns of personal network structure and how these patterns are consistent or different between these two metro areas. In the remainder of this paper, we refer to these datasets as the CA and MA samples.

Compared to public Strava data available on the

Strava Application Programming Interface (API), this collection includes more details for anonymized Strava users (e.g. age and number of years on the service). Importantly, this data records actions of personal networks' tie formation for each user and historical logs of their posted activities. The data consists of but is not limited to the following three major components: (1) user profile information, such as gender, age, location, sign-up date, enrolled plan (i.e. free or paid plan); (2) network subscription records: who followed whom and the time-stamp at which the connection was made; (3) detailed representation of activities posted by users over time, e.g. activity description, activity performance, etc.

Table 1 presents a descriptive summary of the dataset used in this work. Since Strava started and continues to be headquartered in CA, it is unsurprising that athletes in CA adopted this new fitness application earlier; CA enjoys a larger body of users. Strava has a much higher proportion of male users, reflected in both location-specific datasets. Around 40% of users reported their age information. Among them the largest age group is the 30-39 group for both locations.

Table 1. Descriptive Statistics of Strava Dataset

	California	Massachusetts
# of Users	46,181	23,477
Gender Proportion		
Male	0.757	0.744
Female	0.243	0.256
Age Proportion		
18-29	0.215	0.235
30-39	0.449	0.369
40-49	0.236	0.243
50-60	0.100	0.153
Enrolled Plan Proportion		
Free	0.923	0.927
Paid	0.077	0.073

4. Methods

4.1. Degree Distribution Models

To characterize and understand the process by which users intentionally build their personal networks in online fitness communities, we focus on the distribution of network size across the population of users in the study. The degree of a node in a social network is the number of ties attached to it. The degree distribution then characterizes the frequency of degree among network participants and is one of the fundamental of network properties [27]. While degree distributions do not give complete information about network structure,

---

[1] https://www.strava.com/heatmap#2.29/-45.89087/0.00000/hot/all

their form and features communicate very important information, including possible mechanisms by which ties could have been constructed. Modeling degree distributions is an area of ongoing work across many domains in the field of social network analysis. Most networks in the real world have highly skewed degree distributions [27]. In particular, a wide range of networks approximately follow a power-law degree distribution including but not limited to, the world wide web, protein-protein interaction and some social networks [28, 29, 30]. Such networks are referred to as scale-free networks [29]. This "scale-free" property implicitly implies that there exists underlying stochastic mechanisms for the formation of network structure.

Power-law degree distributions have been observed in many cases, however, they are not the only form observed in real-world social networks [31]. There are many theoretical models for degree distributions and each one implies a type of mechanism of the formation of networks. Thus, these models are important statistical tools for analyzing and understanding the structure and formation of social relationships. In this work, we employ the methods developed by Handcock and Jones [15]. Their work discusses three general classes of stochastic process models for network formation: (1) non-homogeneous Poisson; (2) preferential attachment; and (3) "vetting" models. While originally developed to capture potential mechanisms for choice of sexual partners, these models are general partner choice models. In particular, they are likely to apply in cases where individuals have to make decisions about which relationships to invest in, given a limited capacity for investment. In the case of online fitness communities, individuals may choose partners on the basis of social support, informational resources, potential for co-presence and social exercise, etc. In each case, users are forming social ties in a deliberate and intentional way. Next, we briefly introduce these three types of degree distribution models. An expanded discussion of the models can be found in [15]. All models used here have been implemented in the *R* package *degreenet* [32].

Non-homogeneous Poisson: Consider the population K of individuals with at least one tie in a given time period. Suppose that the number of additional ties $K - 1$ that the person has in the time period follows a Poisson distribution with expected value λ. This process (also referred to as homogeneous Poisson process) often fails to describe data realistically because it assumes all individuals have identical propensities to form ties in networks. Thus, more general non-homogeneous models, or extensions thereof, which account for within-population heterogeneity are more commonly used to study

real-world problems (e.g. [33, 34]). For example, if the structural mechanism differs by user demographics or socioeconomic status, non-homogeneous Poisson process is a better choice than homogeneous Poisson process.

As with [15], we model within population heterogeneity by representing individual expected values of λ as independent draws from a Gamma distribution with mean μ and standard deviation σ. The resulting distribution can have an extreme long tail but finite variance. It is called *shifted negative binomial distribution* in [15].

Preferential Attachment: Preferential attachment indicates a process in which new nodes prefer to attach to well-connected nodes over less-well connected nodes. This process is also referred to under the names "cumulative advantages", the "Matthew effect" and more colloquially, "the rich get richer." [15] proposes two distribution models that result from preferential attachment processes; they are the Yule distribution [35] and the Waring distribution [36].

Beginning with a network of r connections, the Yule distribution is motivated by assuming that (1) there is a constant probability $\frac{\rho-2}{\rho-1}$ that the $r + 1$st relationship in the population will be initiated from a randomly chosen person to a previously inactive person, and (2) otherwise the probability that $r + 1$st relationship will be to a person with exactly k relations is proportional to $k \cdot f(k|r)$. Here, $f(k|r)$ is the frequency of persons with exactly k connections out of the r total connections in the population [35]. The Yule distribution will exhibit power-law behavior.

The Waring distribution is motivated similarly as the Yule distribution, but in this case the constant probability that the $r + 1$st relationship in the population will be initiated from a randomly chosen person to a previously inactive person is $\frac{\rho-2}{\rho+\alpha-1}$. The Waring distribution is a natural generalization of the Yule distribution [36].

Vetting Models: This class of models involve a two-stage process. In general, the process first generates a list of potential contacts and then forms relationship based on this list. The vetting models are very flexible, allowing to specify any distribution for both of the two processes. In this work, we consider the Yule-vetting models (see [15] for more detail).

For the Yule-vetting models, the process first generates a number of A potential contacts from a distribution $P(A = a)$ which may exhibit power-law behavior. Next, the process assumes the potential number of friends L that a person has at the time follows a distribution that is short-tailed in most cases (e.g. geometric or negative binomial distribution). However,

the *actual* number of friends that a person has K is bounded by the number of potential contacts A that a person has - that is $K = \min(A, L)$. One example of Yule-vetting models is the negative binomial Yule. The number of potential contacts are formed from a Yule distribution and the potential friendship distribution is negative binomial.

4.2. Estimating Model Parameters

We adopt a likelihood framework to estimate model parameters and to compare between candidate models. There are substantial differences when the process fits low and high network degree. Therefore, the proposed likelihood model includes extra parameters to fit low-degree observations. Here we apply non-parametric fitting and standard model selection techniques.

Denote y as the data - a vector of degree counts and define k as the threshold above which the parametric model is fit, π as a vector of lower-tail probabilities for the non-parametric component, and p as a vector of upper tail probabilities for the parametric component which is $p_i = Pr(d = i | d > k_{min})$.

The log-likelihood of the data y is:

$$\log Pr(y|k_{min}, \pi, p) = \sum_{i=0}^{k_{min}} y_i \log \pi_i$$

$$+ \Big(N - \sum_{i=0}^{k_{min}} y_i\Big) \log \Big(1 - \sum_{i=0}^{k_{min}} \pi_i\Big) + \sum_{i=k_{min}+1}^{\infty} y_i \log p_i$$

4.3. Model Selection

We select the best fitted model among a set of candidate models implying certain types of social mechanisms of friend formation as discussed above. A reasonable model selection technique will balance goodness of fit with model complexity. Here we employ the model selection process stated in the work [15, 37]. Due to the finite nature of social networks [38], we use the Akaike information criterion with a correction for finite sample sizes (AICc) as the indicator of the best performing models [39]. For convenience, BIC [40] is also provided in later results. A model with smaller AICc or BIC is judged better than another model. Under the assumptions of MLE, the AICc is given by: AICc = AIC $+ \frac{2p(p+1)}{n-p-1}$, where n is the number of observations and p is the number of model parameters.

5. Results

5.1. Empirical Egocentric Network Characteristics

Table 2 shows descriptive statistics for the personal networks of users. Personal network size is closely related to the length of time since an individual joined the Strava platform. As seen in the Table 2, men consistently have longer tenure on the site than women. Therefore, rather than comparing the raw counts of incoming and outgoing ties, we divide an individual count of both types of ties by a user's tenure and compare the resulting normalized numbers by gender, age, location and plan enrolled groups.

First, we compare network characteristics between free-plan and paid-plan users. It is not surprising to see that paid-plan users, on average, have a much greater number of network connections than free users do. However, it is interesting to observe that both average incoming and outgoing ties for women who enrolled in paid plans are greater than paid-plan men – the pattern is consistent for both locations. Moreover, in the MA data women with free plan also tend to have larger in- and out-degree than counterpart men.

Recall that only about 40% of users report age information in their profiles. Therefore, the network statistics by age group in the Table 2 are obtained based on the personal networks whose age information is available. We find that a large gender difference exists in the group 50-60 years of age for both CA and MA; men tend to have larger counts of both in-coming and out-going ties than women do. For the mid-aged groups (e.g. 30-39 and 40-49), we see less substantial gender differences in averages of both in- and out-degree.

These results indicate gender and age differences in personal network structure. This may imply that underlying mechanisms of relationship formation differ by plan enrolled and demographics.

Further, Figure 1 shows the degree distributions for egocentric networks. Specifically, we compare distributions of network size (i.e. total degree) along gender, age and economic lines; we also compare how in- and out-degree are distributed by gender. Three sub-figures in the first row present degree distributions of male personal networks, whereas the second row represents that of female personal networks.

Figure 1(a) compares the distributions of personal network size for free-plan and paid-plan users by gender. Consistent with the results in the Table 2, paid-plan users on average have larger networks than free-plan users. Interestingly, we see a salient gender difference in network size among free-plan users;

Table 2. Descriptive Statistics of Personal Networks on Strava

California

Men	Tenure	Avg. $\frac{\text{In-Degree}}{\text{Tenure}}$	Avg. $\frac{\text{Out-Degree}}{\text{Tenure}}$	Women	Tenure	Avg. $\frac{\text{In-Degree}}{\text{Tenure}}$	Avg. $\frac{\text{Out-Degree}}{\text{Tenure}}$
Plan				Plan			
Free	27.818	0.076	**0.358**	Free	26.219	**0.098**	0.323
Paid	50.821	0.404	1.187	Paid	45.331	**0.590**	**1.369**
Age				Age			
18-29	29.049	0.174	**0.702**	18-29	26.512	**0.242**	0.614
30-39	39.790	0.194	**0.615**	30-39	36.646	**0.217**	0.559
40-49	44.979	**0.155**	0.575	40-49	40.168	0.137	0.498
50-60	39.100	**0.096**	0.418	50-60	36.388	0.060	0.242

Massachusetts

Men	Tenure	Avg. $\frac{\text{In-Degree}}{\text{Tenure}}$	Avg. $\frac{\text{Out-Degree}}{\text{Tenure}}$	Women	Tenure	Avg. $\frac{\text{In-Degree}}{\text{Tenure}}$	Avg. $\frac{\text{Out-Degree}}{\text{Tenure}}$
Plan				Plan			
Free	27.827	0.100	0.309	Free	22.468	**0.107**	**0.324**
Paid	43.661	0.376	1.021	Paid	33.790	**0.485**	**1.382**
Age				Age			
18-29	29.414	0.176	0.554	18-29	24.313	**0.216**	**0.590**
30-39	37.698	**0.173**	**0.527**	30-39	31.447	0.172	0.498
40-49	40.640	0.159	**0.465**	40-49	31.597	**0.165**	0.460
50-60	39.536	**0.167**	**0.412**	50-60	31.332	0.090	0.306

(a) Free vs Paid   (b) Age Group   (c) In- vs Out-ties
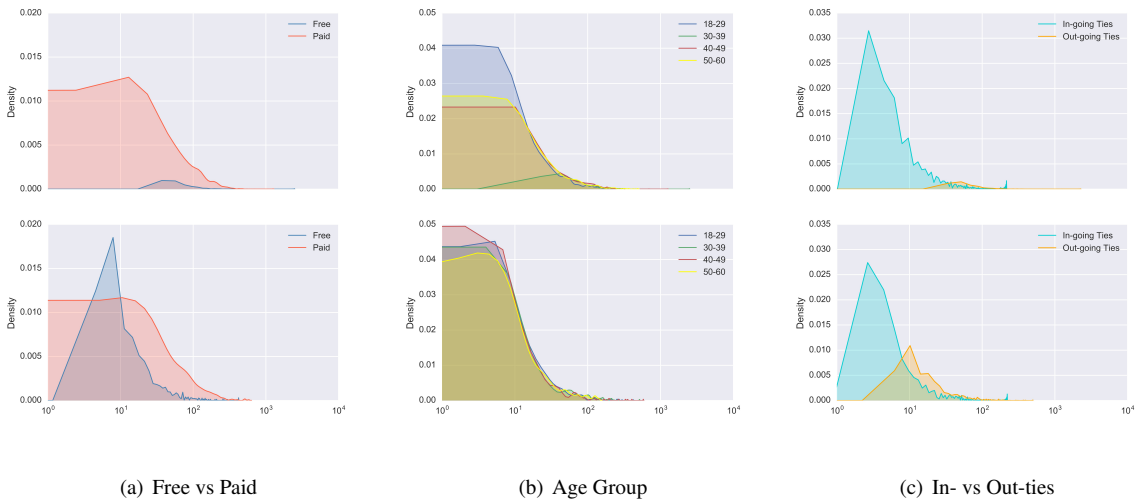
Figure 1. Network Degree Distributions: subfigures in the first and the second row show degree distributions of male and female personal networks, respectively.

overall, female free-plan users have smaller network size than counterpart males. Figure 1(b) shows the distributions of network size for users in different age groups. We observe that the degree distribution varies most by gender for the age group 30-39. Figure 1(c) presents the in- and out-degree distributions for male and female users. The average in-coming degrees for both men and women are lower than the average out-going degrees. Overall, in-degrees of female users' networks are slightly greater than male users'. But this gender difference is not quite noticeable as in the out-degree distributions. The out-degree distributions present a gender difference that women on average have fewer out-going ties than men do. Note that here we only show the figures based on the MA dataset due to space constraints, but the patterns remain consistent across two areas.

5.2. Modeling Degree Distributions

Table 3 and Table 4 present model selection results for the degree distribution modeling. We consider all stochastic models discussed in the Section 5. Here we show only the best fit model for each user group in both tables, but our model selection process went up to 10 for the non-parametric component.

Table 3 shows the most "likely" degree distribution model for users in free and paid plan; we also model men and women separately. For in-degree distributions which can be viewed as number of followers, we find that Strava users with free plans have an in-degree distribution most consistent with a Negative Binomial Yule distribution, while users with paid plans are best fit by a Negative Binomial distribution. This suggests that free-plan users are employing a vetting process in which first there is a list of potential users who may become ego's followers and then people in the list have equal chance to follow ego. On the other hand, paid-plan users are employing a non-homogeneous Poisson process. One interpretation in this context is that each user in the paid-plan group has a constant probability of being followed by other users, but this probability varies by users. This process would be consistent with an environment in which athletes have different levels of popularity or desirability to be targets of attention. The results of out-degree distribution modeling indicate that users with both free and paid plans are employing a non-homogeneous Poisson process when following other users. These results do not show substantial gender differences in each plan group.

Table 4 shows the best fit for personal networks of Strava users categorized into four age groups. Except for the 18-29 years of age group, we find that

the formation of in-coming relations for men most consistent with a Negative Binomial distribution and that female's in-degree formation most likely follows a preferential attachment process (e.g. the Yule and Waring distribution). We observe some age differences in out-degree distribution modeling. While men across all age groups employ a Negative Binomial process, women in the age group of 18-29 and 40-49 employ a preferential attachment process. All best-fitted degree distribution models have very low cut-offs (i.e. k = 1 and 2 as seen in the Table 3 and 4). Therefore, these models exhibit distinct processes at low degree; users with just one or two followers may be new or only recently active. They might also be interested in self tracking more than social networking features of the platform.

6. Discussion

Previous literature points out that women and men often have distinct social networks [41, 42]. These documented differences have important implications for information access, social support and health. Utilizing large-scale behavioral trace data, our work presents empirical evidence for variability in network size across users of the online fitness community Strava; importantly these differences are associated with demographic and economic categories. Men, users who pay for the service, and 30-49 year old users tend to have larger networks, i.e. more peers with whom they engage on the platform. Women, along with the youngest and oldest users have smaller personal networks – fewer individuals to turn to for social support and motivation in reaching fitness goals. These results are not just for a small sample of participants, but instead represent large-scale analyses of users in two metro areas on both coasts of the United States. Beyond characterizing empirical patterns in network size, this study also explores the mechanisms of the observed empirical distribution of network size in the population.

This investigation reinforces longstanding empirical patterns of inequality in network positions. Strava users have different opportunities for social interaction. They make different choices about which relationships to initiate, invest in, and pursue. These choices, in turn, have important consequences for receiving social support and resources, looking to others for motivation or reinforcement in achieving goals, and long term engagement versus drop out. Our findings suggest that men and women, for example, may approach choices about social interaction and investment in social relationships in a fitness setting in different ways. Initiating relationships based on popularity (e.g. a preferential attachment process) versus a two-stage

Table 3. The best fit models for in- and out-degree distributions of male and female users who enrolled in free plan versus paid plan in Strava.

In-Degree Distribution Models							
Plan	Gender	Model	k	np	log-lik	AICc	BIC
Free	M	NegBinYule	1	5	-19712.35	39434.7	39471.66
	F	NegBinYule	2	5	-6865.921	13741.86	13773.78
Paid	M	NegBinom	1	3	-5318.866	10643.75	10659.56
	F	NegBinom	1	3	-860.3571	1726.815	1737.193

Out-Degree Distribution Models							
Plan	Gender	Model	k	np	log-lik	AICc	BIC
Free	M	NegBinom	1	3	-27185.63	54377.27	54399.45
	F	NegBinom	1	3	-9205.993	18417.99	18437.15
Paid	M	NegBinom	1	3	-6706.873	13419.76	13435.57
	F	NegBinom	1	3	-1091.877	2189.855	2200.234

Table 4. The best fit models for in- and out-degree distributions of male and female users in different age groups in Strava.

In-Degree Distribution Models							
Age	Gender	Model	k	np	log-lik	AICc	BIC
18-29	M	Waring	1	3	-3157.574	6321.164	6336.88
	F	Waring	1	3	-1358.552	2723.145	2736.28
30-39	M	NegBinom	1	3	-6127.021	12260.05	12277.39
	F	Waring	1	3	-1793.552	3593.136	3606.98
40-49	M	NegBinom	1	3	-4400.432	8806.879	8823.196
	F	Waring	1	3	-770.2596	1546.587	1558.177
50-60	M	NegBinom	1	3	-2719.809	5445.64	5460.594
	F	Yule	2	3	-385.2882	776.689	786.716

Out-Degree Distribution Models							
Age	Gender	Model	k	np	log-lik	AICc	BIC
18-29	M	NegBinom	1	3	-4330.269	8666.556	8682.272
	F	Waring	1	3	-1759.673	3525.386	3538.521
30-39	M	NegBinom	1	3	-8147.217	16300.44	16317.78
	F	NegBinom	1	3	-2294.011	4594.053	4607.898
40-49	M	NegBinom	1	3	-5882.915	11771.84	11788.16
	F	Waring	1	3	-1049.921	2105.91	2117.501
50-60	M	NegBinom	1	3	-3426.347	6858.716	6873.67
	F	NegBinom	1	3	-528.0239	1062.16	1072.188

vetting process implies different underlying process considerations. Not only does this imply inequality in terms of network position, but it also suggests behavioral differences in the ways in which individuals make use of the platform itself. Further explorations, perhaps participant interview for example, might offer insight into the approaches that Strava users consider when constructing social networks to support fitness goals.

Previous literature on homophily and social support suggests that the individuals with certain health conditions (e.g. obesity) or with lower social status may be more dependent on the composition of their social networks than healthier or high-status individuals [43, 44]. Our work suggests that network-based interventions for health should take structural differences, and their associated implications about mechanisms for tie formation, into account. In fact, accounting for differences might be especially vital for people who are most in need of support to achieve healthier behavior. Moreover, this study points to the need for a better understanding of the dynamics of relationship construction and maintenance. For new participants to an online fitness community, with whom are their first relationships forged? Are they with friends from offline settings? Are they with high status and high visibility professional athletes? Do they look for others like themselves, in terms of social and demographic characteristics? Answering these questions could not only improve the design of these systems and their affordances, but they could also have important implications for health behavior change.

Previous work suggests that social motives are vital for persistent use of activity-tracking platforms [18, 45]. For current users of online fitness communities, stronger social motives and peer support may motivate them to stay with the fitness service for a longer time and be more likely to develop long-term healthy habits. Being embedded in an active and supportive network of peers within online fitness communities may be related to persistence of use. Understanding the processes individual platform users behave in accordance with may inform the design of tools and/or systems to support healthy personal networks. In fact, users may not behave in the ways that would lead to an optimally supportive mix of personal contacts. In such cases, platforms themselves should be able to suggest missing connections or social contacts that can support certain goals. Many of these important questions could be addressed in a follow-up study, perhaps one that also compiles participant interview or survey responses to explore motivations behind specific relationship construction choices.

7. Conclusion

Modeling the distribution of personal networks in this online community demonstrates that individuals are embedded such that they experience systematic differences in the social environment based on demographic and economic factors such as age groups, genders, and economic brackets on Strava and likely other activity based platforms. These differences have numerous implications for network-based interventions that leverage peer influence, social support, and other applications in efforts to change behavior in this setting.

The curve fits of network degree distributions presented in this study are just the first step in exploring and understanding the social dynamics at play in these new social settings. Future research should directly explore the stochastic mechanisms which generate empirical networks in these settings through complex dynamic models or careful observational studies and/or follow up surveys.

References

[1] P. V. Marsden and N. E. Friedkin, "Network studies of social influence," *Sociological Methods & Research*, vol. 22, no. 1, pp. 127–151, 1993.

[2] S. Aral and C. Nicolaides, "Exercise contagion in a global social network," *Nature communications*, vol. 8, p. 14753, 2017.

[3] N. Lin, "Building a network theory of social capital," in *Social capital*, pp. 3–28, Routledge, 2017.

[4] K. Miyata, "Social support for japanese mothers online and offline," *The internet in everyday life*, pp. 520–548, 2002.

[5] Y. Chen and P. Pu, "Healthytogether: exploring social incentives for mobile fitness applications," in *Proc. of the Second International Symposium of Chinese CHI*, pp. 25–34, ACM, 2014.

[6] C. Pechmann, L. Pan, K. Delucchi, C. M. Lakon, and J. J. Prochaska, "Development of a twitter-based intervention for smoking cessation that encourages high-quality social media interactions via automessages," *Journal of medical Internet research*, vol. 17, no. 2, 2015.

[7] D. Centola, "Social media and the science of health behavior," *Circulation*, vol. 127, no. 21, pp. 2135–2144, 2013.

[8] X. Chen, Z. Zhu, M. Chen, and Y. Li, "Large-scale mobile fitness app usage analysis for smart health," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 46–52, 2018.

[9] L. Zeng, Z. W. Almquist, and E. S. Spiro, "Stay connected and keep motivated: Modeling activity level of exercise in an online fitness community," in *International Conference on Social Computing and Social Media*, pp. 137–147, Springer, 2018.

[10] T. Althoff, P. Jindal, and J. Leskovec, "Online actions with offline impact: How online social networks influence online and offline user behavior," in *Proc. of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 537–546, ACM, 2017.

[11] M. W. Newman, D. Lauterbach, S. A. Munson, P. Resnick, and M. E. Morris, "It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health," in *Proc. of the ACM 2011 conference on Computer supported cooperative work*, pp. 341–350, ACM, 2011.

[12] R. A. Hill and R. I. Dunbar, "Social network size in humans," *Human nature*, vol. 14, no. 1, pp. 53–72, 2003.

[13] D. B. Kandel, "Homophily, selection, and socialization in adolescent friendships," *American journal of Sociology*, vol. 84, no. 2, pp. 427–436, 1978.

[14] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[15] M. S. Handcock and J. H. Jones, "Likelihood-based inference for stochastic models of sexual network formation," *Theoretical population biology*, vol. 65, no. 4, pp. 413–422, 2004.

[16] M. S. Handcock and J. H. Jones, "Interval estimates for epidemic thresholds in two-sex network models," *Theoretical Population Biology*, vol. 70, no. 2, pp. 125–134, 2006.

[17] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42, ACM, 2007.

[18] J. Stragier, M. V. Abeele, P. Mechant, and L. De Marez, "Understanding persistence in the use of online fitness communities: comparing novice and experienced users," *Computers in Human Behavior*, vol. 64, pp. 34–42, 2016.

[19] T. Althoff, "Population-scale pervasive health," *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 75–79, 2017.

[20] D. Foster, C. Linehan, B. Kirman, S. Lawson, and G. James, "Motivating physical activity at work: using persuasive social media for competitive step counting," in *Proc. of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments*, pp. 111–116, ACM, 2010.

[21] D. N. Cavallo, D. F. Tate, A. V. Ries, J. D. Brown, R. F. DeVellis, and A. S. Ammerman, "A social media–based physical activity intervention: a randomized controlled trial," *American journal of preventive medicine*, vol. 43, no. 5, pp. 527–532, 2012.

[22] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, *et al.*, "Activity sensing in the wild: a field trial of ubifit garden," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1797–1806, ACM, 2008.

[23] C. Lister, J. H. West, B. Cannon, T. Sax, and D. Brodegard, "Just a fad? gamification in health and fitness apps," *JMIR serious games*, vol. 2, no. 2, 2014.

[24] J. H. West, P. C. Hall, C. L. Hanson, M. D. Barnes, C. Giraud-Carrier, and J. Barrett, "Theres an app for that: content analysis of paid health and fitness apps," *Journal of medical Internet research*, vol. 14, no. 3, 2012.

[25] L. Zeng, Z. W. Almquist, and E. S. Spiro, "Let's workout! exploring social exercise in an online fitness community," *iConference 2017 Proc. Vol. 2*, 2017.

[26] T. N. Y. Times, "What is your citys twin?." https://www.nytimes.com/interactive/2018/04/03/upshot/what-is-your-citys-twin.html, 2018.

[27] M. Newman, *Networks*. Oxford university press, 2018.

[28] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[29] A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific american*, vol. 288, no. 5, pp. 60–69, 2003.

[30] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. of the national academy of sciences*, vol. 106, no. 36, pp. 15274–15278, 2009.

[31] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[32] M. S. Handcock, "Degreenet: Models for skewed count distributions relevant to networks," *Statnet Project, Seattle, WA. Version*, vol. 1, 2003.

[33] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proc. of the 19th international conference on World wide web*, pp. 851–860, ACM, 2010.

[34] J. Mahmud, J. Chen, and J. Nichols, "When will you answer this? estimating response time in twitter.," in *ICWSM*, 2013.

[35] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, pp. 425–440, 1955.

[36] J. O. Irwin *et al.*, "The place of mathematics in medical and biological statistics.," *Journal of the Royal Statistical Society*, vol. 126, no. Pt. 1, pp. 1–41, 1963.

[37] Z. W. Almquist and C. T. Butts, "Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics," *Sociological methodology*, vol. 44, no. 1, pp. 273–321, 2014.

[38] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

[39] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[40] R. E. Kass and L. Wasserman, "A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion," *Journal of the american statistical association*, vol. 90, no. 431, pp. 928–934, 1995.

[41] M. S. Granovetter, "The strength of weak ties," in *Social networks*, pp. 347–367, Elsevier, 1977.

[42] D. Centola and M. Macy, "Complex contagions and the weakness of long ties," *American journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007.

[43] D. Centola, "An experimental study of homophily in the adoption of health behavior," *Science*, vol. 334, no. 6060, pp. 1269–1272, 2011.

[44] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[45] S. Consolvo, K. Everitt, I. Smith, and J. A. Landay, "Design requirements for technologies that encourage physical activity," in *Proc. of the SIGCHI conference on Human Factors in computing systems*, pp. 457–466, ACM, 2006.