Supporting Information

Aral et al. 10.1073/pnas.0908800106

SI Text

Data. ***Yahoo! Go 2.0 mobile service application.*** Yahoo! Go is a mobile software application designed to deliver personalized content to users' devices and to enhance mobile search capabilities. Yahoo! Go provides personalized information across several domains, including news, sports, weather, and financial information, as well as access to e-mail, calendaring, location-based services, photo sharing, and web search services. Users are identified by their Yahoo! ID, which is used for their everyday interaction with the Yahoo! portal, instant messaging (IM), e-mail and other services from all platforms, including personal computer (PC) and mobile devices. Consequently, observations of activities and behavior across different platforms can be attributed to each unique user (Movies S1 and S2).

It should be noted that Yahoo! Go 2.0, the subject of this study, does not support IM services. In this respect Yahoo! Go is not networking software and does not exhibit direct network externalities, whereby the value of the product to a user is a function of the number of other users with whom one can connect or communicate by using the product. Therefore, its adoption is not likely to be driven by the desire to communicate with one's friends. It is likely that peer influence effects are quite different for products that exhibit network externalities. We therefore encourage the application of our methods to influence estimation in adoption processes of such products.

The collected dataset includes detailed demographic and daily usage behavior data for all 27.4 million users, including geographic location and demographic data, IM usage behavior, PC usage behavior, mobile usage behavior, Go usage behavior, and product adoption date for 5 months starting from June 1, 2007, to October 29, 2007 (see Table S1 for variable definitions). We focus on adoption dynamics after the official launch date (July 4, 2007) because prior users were "Beta testers" seeded by Yahoo!

We took exceptional care to preserve the privacy of the research subjects. Minors (users below the age of 18) where excluded from the sample and the data were anonymized by the company before it was released to us. We have obtained Institutional Review Board (IRB) approval for this study, including detailed procedures for security and privacy protection, and scrupulously fulfilled the IRB protocols.

Considering the size of the data (exceeding 150 Gb), the size of the network, and the complexity of the analysis, data processing was quite challenging. Few standard network or statistical analysis tools can directly handle datasets of this order of magnitude well, so the data were preprocessed to remove redundant information. Preprocessing was performed with optimized C++ code. Further statistical analysis was performed in MatLab by using its statistical toolbox, and network analysis was performed with the Complex Networks Toolbox for Matlab (1).

Descriptive Statistics. ***Individual behavioral data: online activity and browsing behavior (PC, mobile, and Go platforms).*** In Table S2 we list basic statistics of user surfing activities for different types of content across PC, Mobile, and Go platforms. All page view distributions are heavy tailed (Fig. S1). Most users surf infrequently, but some are remarkably active consumers of certain types of content such as news, finance, and sports. We find users' online behavior to be highly predictive of Yahoo! Go adoption and (because of homophily) the propensity of having adopter friends.

IM network and message traffic. Our sampling procedure (described in *Materials and Methods*) produced complete data in local network neighborhoods two steps away from any sampled seed node. The sampling strategy was designed to make sure we (*i*) captured all adopters of the product (so that we could comprehensively observe the adoption curve), and (*ii*) to have a representative sample of all Yahoo! users. These two seed samples (all adopters and the random sample) served this purpose. The two-step snowball sample was used to capture the local network neighborhoods of all of the initial seeds (both random seeds and adopters). Table S2 presents basic IM network and message traffic summary statistics and reports strength of tie measures based on the number and fraction of IM messages exchanged between users.

Combining the detailed demographic and behavioral data with the IM network, we measure each user's social environment by computing averages of their friends' characteristics. To assess homophily we compute average values across users' friends by using each of the online behavioral characteristics.

Demographic, device, and geographic location data. *Gender.* 25.0 million users (91%) report their gender. 10.3 million (41.2%) are female and 14.7 million (58.8%) are male. 99.8% of people who have accessed the Yahoo! website (active users) at least once report their gender.

Age. 22.3 of 27.4 million users (81.5%) and 19.9 of 22.5 active users (88.5%) report their age. However, the percentage of users reporting their age climbs to 97.6% for Go service adopters. In general, the propensity of a person to disclose his or her age increases with higher levels of activity. Fig. S2 *Left* shows the age distribution of Yahoo! users.

Device data. Mobile device type and the fraction of mobile page views from that device are reported daily for mobile users and Go adopters. A total of 2,030 distinct mobile devices were used to access the Yahoo! portal and 111 different devices were used to access Go.

Country. The dataset contains a self-reported record of each user's native country. In addition, there is a daily record of the main and secondary country locations from which users access Yahoo! services based on their internet protocol (IP) addresses (defined in Table S1), along with the number of page views originating from each of them. The Yahoo! portal is accessed from 239 distinct countries, areas, and administrative territories. However, the distribution is uneven. Ninety percent of Yahoo! users live in 19 countries. The top six countries are the United States (59.4%), India (7.6%), the United Kingdom (3.5%), the Philippines (2.7%), China (2.0%), and Taiwan (1.9%). Ninety percent of mobile users live in only 12 countries. The United States (52.4%), Taiwan (7.0%), India (6.4%), Romania (5.4%), and the Philippines (5.3%) are the top five most represented countries Fig. S2 *Right*). Adoption of Yahoo! Go is also not uniform across the globe. Ninety percent of all Go adopters operate from only 14 countries. The United States (62.2%), India (7.3%), the United Kingdom (5.3%), Indonesia (3.0%), the Philippines (2.5%), and Romania (2.0%) are the top six most represented countries.

Mechanisms of Assortative Mixing and Temporal Clustering. Assortative mixing and temporal clustering of node behaviors in networks have been documented in several recent studies across a wide range of social phenomena. A number of theoretical explanations could account for such clustering, and precise descriptions of the mechanisms that may produce such patterns are essential to validating causal claims of homophily or influence. Three major drivers of assortative mixing and temporal clustering have been proposed in the literature: (*i*) peer influence (individuals induce their friends to adopt similar behaviors), (*ii*) homophily (similar people tend to be friends), and (*iii*) confounding environmental factors (friends may

be exposed to the same exogenous environmental shocks and thus adopt similar behaviors) (2). The emergence of homophily itself is known to be the product of a complex process of selection and influence, whereby friends choose to be friends with those similar to themselves (selection), and friends tend to induce each other to become more similar (general influence) (3). Table S3 documents some potential mechanisms that may explain assortative mixing and temporal clustering in networks and gives examples of each in the context of a focal choice—the decision to exercise regularly.

Our analysis focuses on "direct influence," when a specific focal action induces the same focal action in one's friends (in our case product adoption), which we distinguish from a more general notion of influence that could include indirect influence effects such as the effect of my friends' proclivity to be "tech savvy" or early technology adopters inspiring me to be more tech savvy and therefore to adopt Go. Because it is highly unlikely that people become friends solely because of their adoption of Yahoo! Go, we do not consider the theoretical implications of selection.

Clustering of behaviors in networks is the result of complex processes involving selection, influence, homophily, and confounding environmental factors to which individuals are exposed. A distinct but related body of literature examines selection and influence processes in the co-evolution of behaviors and network structure in cases where tie formation is likely to be a function of the behavior in question (see foundational work by Tom Snijders and colleagues, e.g. ref. 3). In our context (and in many important contexts) we are interested in separating influence and homophily effects when link formation is not likely to be driven by the behavior in question (Go adoption is unlikely to drive friendship).

We see the IM network as a proxy for an underlying social network in which friends communicate with one another on a variety of dimensions through a variety of communication channels (IM, e-mail, face-to-face encounters, etc.) We simply assume that IM traffic between two users is a proxy for the increased likelihood that an adopter of Go will communicate his or her adoption and use of the application to friends through any number of channels.

Multivariate Survival Analysis. We employ Cox proportional hazards regression (4) to assess the effect of individual user characteristics on the rate of adoption. The regression,

$$h(t,X) = h_0(t)\exp[\alpha + \beta X + \varepsilon]$$

estimates the users' rate of Yahoo! Go adoption, where $h(t, X)$ represents the adoption rate, t is user time in the risk set, and $h_0(t)$ is the baseline adoption rate. The effects of independent variables are specified in the exponential. The coefficients in this model have a straightforward interpretation: β_i represents the percent increase or decrease in the adoption rate associated with a one-unit increase in any independent variable i. Coefficients greater than 0 represent an increase in the adoption rate; coefficients less than 0 represent a decrease. Survival analysis is applied to ≈3.5 million seed nodes, over half a million of which adopted the product during the sampled interval. Each user was characterized by the 35 covariates listed in Table S4.

The two main parameters we are interested in—the number and percentage of adopter friends in ego's local network—are functionally related. We therefore apply the proportional hazards rate model twice: first for the number of adopter friends (NAF) and then separately for the percentage of adopter friends (PAF). Countries were represented by dummy variables. Because the majority of users come from relatively few countries, the six most prevalent countries were used as dummies.

Tests of the proportional hazards assumption hold, but we do observe duration dependence (time-varying covariates) in our independent variables of interest—a complication of hazards modeling in this context that we take great care to address. The number or fraction of adopter friends in users' local networks changes over time. Therefore, assessment of the effect of these variables on the rate of Go adoption (the hazard rate) requires application of the time-varying specification of the proportional hazards model. Unfortunately, the time resolution, the number of samples in the dataset, and the number of covariates in our analysis make such estimation infeasible. We therefore estimate the Cox model on a subset of users who do not acquire adopter friends after the first 2 weeks of the observation period. For these users the number and percentage of adopter friends is time invariant.

Ten percent of adopters have friends who adopted before they adopted. This statistic is primarily driven by two factors: (*i*) the degree distribution is heavy tailed (most users have small numbers of friends) and (*ii*) adoption occurs very rarely in this network (0.4% of users adopt). The excessive number of links between adopters (which is much more than one would expect to occur randomly) is what makes some suspect the existence of influence. Fig. 2, the corresponding logistic and hazard rate models, and the random matching estimates all show that the observed increase in the likelihood of adoption in the presence of adopter friends is very large, meaning that even if adoption in the presence of adopter friends is rare, the likelihood of adoption is much higher once you have adopter friends. As shown in Fig. 2*C*) users are 2.2 times more likely to adopt if they have one adopter friend than if they have 0 adopter friends. It is important to note that this paper is not about whether or not this product is one that is typically adopted independently, but rather about whether the observed increase in the likelihood of adoption in the presence of adopter friends is evidence of peer influence. Even in the case where most adoption occurs independently, estimates of influence that track the increase in the likelihood of adoption in the presence of an adopter friend without controlling for homophily will still produce overestimates of influence as we demonstrate. If, compared with other products, this product happens to be one that is adopted independently more often than is typical, this would only serve to make our point more conservative because in that case it should be even harder for prior methods to overestimate peer influence.

Propensity Score Matching. The objective of propensity score matching is to assess the effect of a treatment by comparing observable outcomes among treated observations to a sample of untreated observations matched on the propensity of being treated (5). We define treatment as having one, two, three, four, or more adopter friends. We refine this treatment by also accounting for the recency of friends' adoption by defining treated users as those with friends who had adopted within certain time intervals (1 day, 2 days, 3 days, etc). We then perform matching for a range of treatments, varying the number of adopter friends and the recency of their adoption.

We estimate propensity scores by using logistic regression (6), which produces sufficient predictions when as many as 33 covariates (Table S5) are used. Other statistical methods could be used to estimate propensity scores. We rely on logistic regression because it is the standard in matched sample estimation. Country and device covariates (originally categorical variables) require special treatment because they are significant predictors of Go adoption. We find that instead of using these variables as categorical in multinomial regression analysis, replacing each of them with a population-wide likelihood of adoption for the corresponding country or device typically results in better matches. These quantities are computed for the time interval for which propensity matching is performed because they change significantly over time (for instance because of the release of Go software updates).

To account for systematic changes occurring over time, we perform propensity score matching in 14-day intervals (except for the period including and immediately following the product release date, which we consider as a distinct 5-day interval). Working with biweekly data allows us to observe changes of treatment effects over time. It also allows us to compare the relative homophily of early adopters to each other and to their nonadopter friends as well as to

the rest of the population. The propensity of treatment estimation improves significantly when performed on short time intervals, demonstrating the importance of matching dynamically. On the other hand, the intervals cannot be too short because there are not enough instances of treated individuals for the logistic regression (or any qualifying predictive model) to learn and produce sufficiently accurate estimates.

When pools of treated and untreated nodes used for matching are constructed, we exclude those users who cannot adopt Yahoo! Go (for instance, users without mobile devices, users having mobile devices that are incompatible with Go, and users living in countries where the service was not available). These properties are assessed at each time interval.

Once the model has been trained for a particular treatment and time interval, all treated nodes are matched with untreated nodes with the nearest propensity scores. Usually, close matches exist. We drop pairs for which the distance of propensity scores exceeds two standard deviations of the observed distribution of propensity score differences. For all treated nodes i,($\forall i$, $T_{it} = 1$) we choose an untreated match j such that $\|p_{it} - p_{jt}\|$ subject to $\min\|p_{it} - p_{jt}\| < 2\sigma_d$ where $d = p_{it} - p_{jt}$. Fig. S3 is a normal probability plot used to assess deviations of the propensity score differences distribution from the normal distribution to identify outliers. The data (blue) are plotted against a normal distribution with same mean and variance (red). Deviations from the normal distribution indicate the presence of fat tails in the observed distribution of propensity score differences. The horizontal (green) lines mark two standard deviations and identify outliers that are removed from the analysis.

Treatment outcome estimation is then defined as the ratio (n_+/n_-) of the number of treated adopters (n_+) over the number of nontreated adopters (n_-) in the propensity score matched sample. Adoption that can be explained by the homophily effect is the difference between the treatment outcome (Go service adoption) estimated by propensity score matching and the treatment outcome estimated by random matching, which ignores all user similarity on characteristics and behaviors.

Our matching process should account for homophily on all observed characteristics and those unobserved or latent characteristics that are correlated with the characteristics we do observe. The more strongly an unobserved characteristic is correlated with those we measure, the more our methods will account for it.

An important assumption of matching sampling techniques is the strong ignorability condition (also referred to as unconfoundedness, exogeneity, and conditional independence), which states that to identify causal treatment effects, the treatment and the reaction to the treatment should be independent conditional on the observable characteristics X (5). Although it is well known that that the strong ignorability condition is in principle untestable there could be unobserved factors in our setting that force us to relax this assumption. For this reason we state in the article that "although we measure individuals' dynamic characteristics, preferences, and behaviors in great detail, the data are not necessarily comprehensive. Although the matching process accounts for homophily on all observed characteristics and those unobserved or latent characteristics that are correlated with what we do observe, unobserved and uncorrelated latent homophily and unobserved confounding factors may also contribute to assortative mixing and temporal clustering. The methods therefore establish upper bounds of influence estimates that account for homophily, and these limitations in observability estimates of the homophily effect are necessarily conservative." We estimate upper bounds of influence because unobserved and uncorrelated latent homophily and unobserved confounding factors that contribute to assortative mixing and temporal clustering may still exist.

The comparison between the matched sample and the random sample is the best informed benchmark of the potential magnitude of the homophily effect. There are two types of observables that increase the likelihood of adoption—those that are homophilous and therefore confound peer influence estimates and those that are not homophilous and therefore do not confound peer influence estimates (but still affect the likelihood of adoption). The only observables that can confound peer influence estimates are those that both (*i*) make one more likely to adopt and (*ii*) are shared characteristics between peers who are more likely to adopt together. Both random matching and matched sample estimation ignore observables that are not homophilous.

Assume for example that one's age was the only variable that determined the likelihood of adoption. Consider further two scenarios. In scenario 1 there is homophily in the network on the age dimension, meaning people tend to be friends with others of the same age. In scenario 2 there is no homophily on age. In scenario 1 random matching would produce a certain (higher) level of estimated influence, whereas the matched sample would reduce those estimates by removing the homophily effect on age. In scenario 2, on the other hand, both methods (random matching and matched sample) would produce estimates of both no influence and no homophily (even though the age observable determines adoption). Our method and the comparison to random precisely picks out those observables that confound influence estimates because of homophily and is not biased by observables that affect the likelihood of adoption but are not homophilous.

Propensity score matching on the recency of friends' adoption (Δt). As we note in the main article, contemporaneous adoption among friends may be attributed to influence as well as to homophily. If influence exists, it is reasonable to assume that the effect of a friend's adoption on a user's adoption likelihood may change over time. To account for this possibility we refine treatment in the dynamic matching methodology. For a given "recency" (R) , we consider a user as treated if one of his friends had adopted Go within the specified time interval ($\Delta t \equiv t_i^a - t_j^a = R$), where t_i^a is the adoption time of the adopter i, and t_j^a is the adoption time of adopter j, a friend of i. We again use multinomial logistic regression to compute estimates of the propensity of a user to be treated, i.e., the likelihood to have had a friend who had adopted R days earlier. Once propensity scores are computed, we match treated users with untreated users having closest likelihood of being treated. Untreated users, as before, are those who have no adopter friends within the time window or in the observable future. We again drop pairs for which the distance of propensity scores exceeds two standard deviations of the observed distribution of propensity score differences. Influence estimates are thus bounded from above by the ratio of the number of treated adopters (n_+) to the number of untreated adopters (n_-).

We perform this procedure repeatedly for a range of time intervals from 0 to 6 days ($\Delta t \in [0, 6]$) (see Fig. 3*D*) where 0 corresponds to friends adopting Go on the same day.

Quantifying the aggregate effect of homophily on adoption in the network. We found that personal characteristics such as age and gender as well as the type of owned mobile device are crucially important for successful matching. Ignorance of these properties leads to a significantly higher difference between the number of matched treated and untreated adopters. To produce the best estimates, we excluded users who had either of these important characteristics missing. A small number (<1%) of matches are discarded because of an inability to find an untreated user with a sufficiently close propensity score.

Environmental Conditions. We estimate influence effects under various environmental conditions by holding out and varying one characteristic x_i while matching on all other characteristics X_{-i}. We find that the upper bounds of influence vary as conditions change. We compute the ratio of the number of treated (n_+) and untreated (n_-) adopters for various subsets of matched pairs corresponding to different values (or ranges) of the environmental conditions. As with the propensity score procedure described above, environmen-

tal conditions can be applied repeatedly for different time intervals to estimate their effect over time.

One can define several classes of environmental conditions (EC): (*i*) those representing individual characteristics of the treated node (the node that is subject to influence, e.g., age, gender, country, device, numbers of page views of treated node), (*ii*) those related to the individual characteristics of the treating, (*iii*) the characteristics of the treated node's local network neighbors, and (*iv*) relational characteristics of treated and treating nodes (e.g., age difference, gender difference, the effect of men on women, women on men, the old on the young, the young on the old, etc.).

Aggregative Effect of Peer Influence. Estimates of treatment effect computed for a subset of the population can be applied to a general population only if the following two conditions hold:

- First, the successfully matched subset of treated adopters must be representative of the entire population of treated adopters. We verify that this requirement holds by comparing distributions of various characteristics of the matched subset of treated adopters to the corresponding distributions of the entire treated population. This makes matching results applicable to the entire population of treated users and in particular to the subpopulation of treated adopters.
- Second, treated adopters might be fundamentally different from treated users who did not adopt (7). We therefore test whether the density of untreated adopters in the population matched to treated adopters is close to the density of untreated adopters in the population matched to treated nonadopters. We find these densities are very similar and stable over time, indicating that our matching results are not biased by the selection of the treated population being matched.

1. Muchnik L, Itschak R, Solomon S, Louzoun Y (2007) Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. *Phys Rev E* 76:016106.
2. Manski CF (1993) Identification of endogenous social effects: The reflection problem. *Rev Econ Stud* 60:531–542.
3. Snijders T, Steglich C, Schweinberger M (2006) Modeling the co-evolution of networks and behavior. *Longitudinal Models in the Behavioral and Related Sciences*, eds Montfort Kv, Oud H, Satorra A (Erlbaum, Philadelphia), pp 41–47.
4. Cox DR (1972) Regression models and life-tables. *J R Stat Soc B* 34:187–220.
5. Rosenbaum PR, Rubin DB (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70:41–55.
6. Hosmer DW, Lemeshow S (2002) *Applied Logistic Regression* (Wiley, New York).
7. Morgan SL, Winship C (2007) *Counterfactuals and Causal Inference* (Cambridge Univ Press, New York).

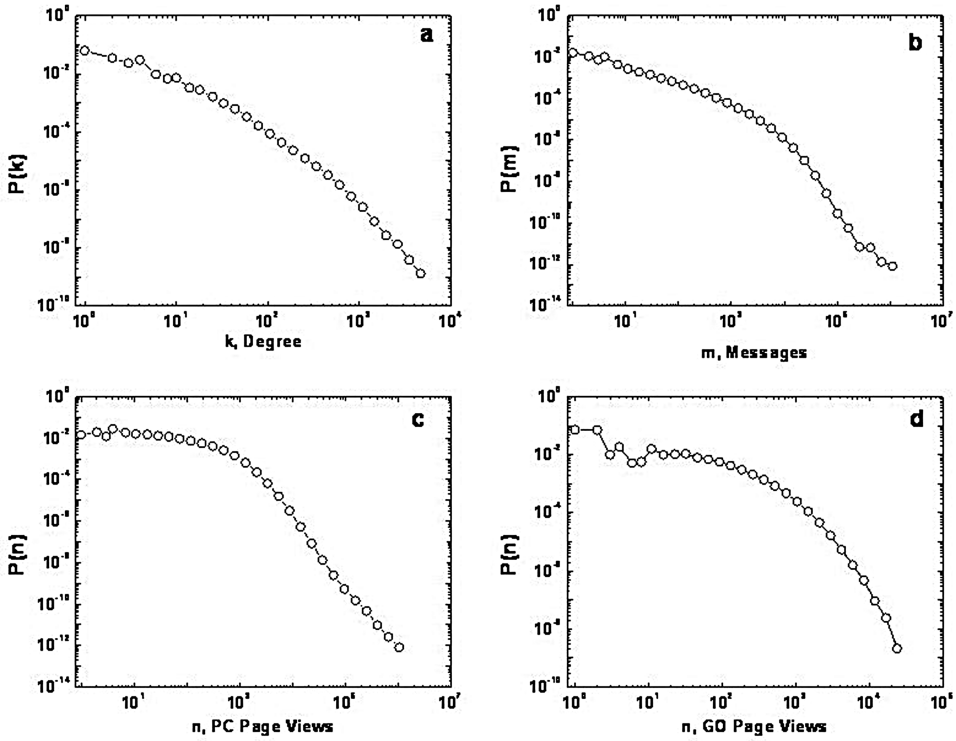


Fig. S1. Sample distributions of the number of IM buddies (*a*) number of IM messages sent and received (*b*), and the numbers of pages loaded during a sample month from PC (*c*) and Go platforms (*d*).

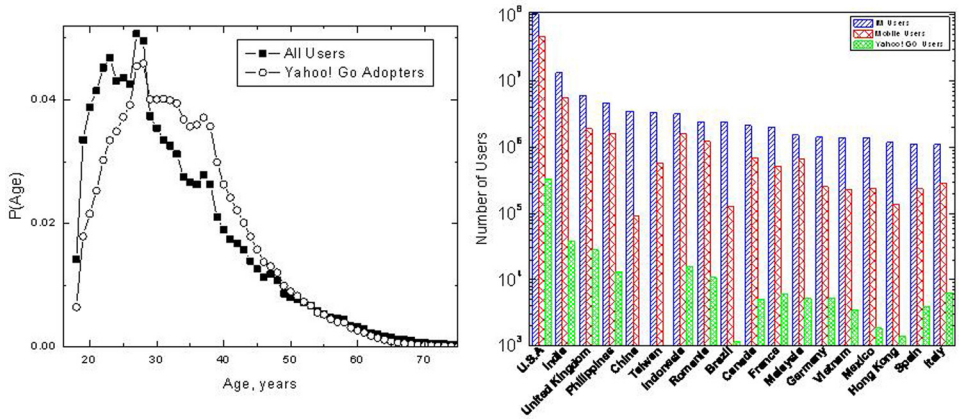


Fig. S2. (*Left*) Age distribution of all Yahoo! IM network users (filled squares) and Yahoo! Go Adopters (open circles). (*Right*) Histogram of the estimated number of Yahoo! IM users, Yahoo! mobile phone users, and Yahoo! Go adopters by country, based on a 2% random sample of the entire global Yahoo! IM Network and a complete sample of Yahoo! Go adopters.

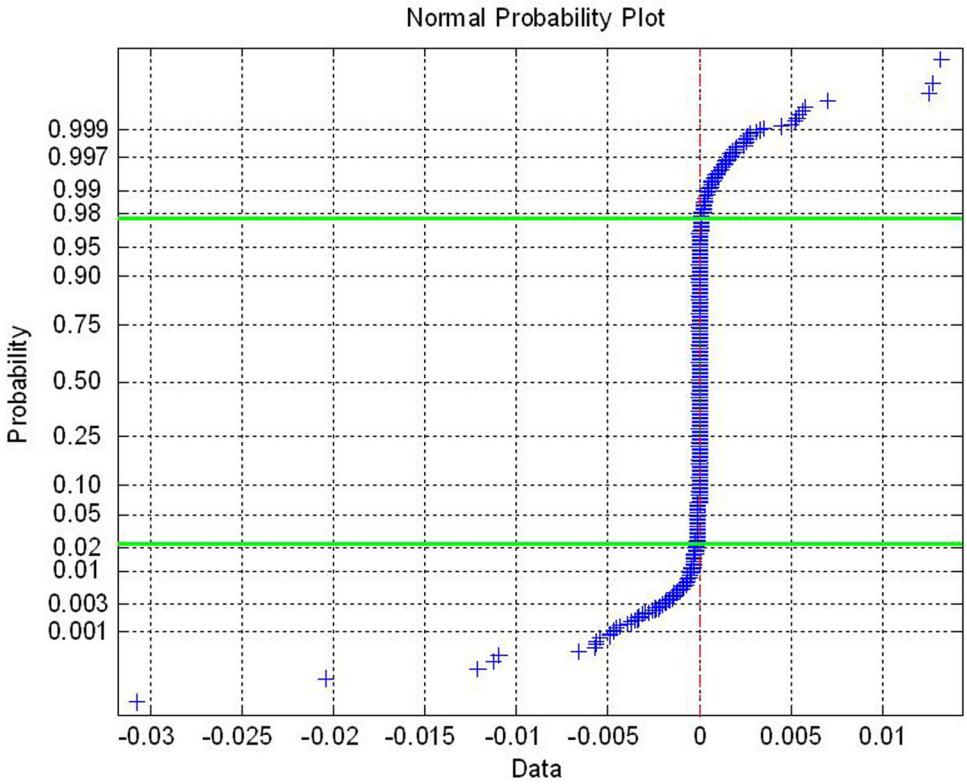


Fig. S3. Normal probability plot showing the observed distribution of differences in propensity scores of the matched treated and untreated nodes compared with a normal distribution with the same mean and variance. Light green horizontal lines correspond to the two standard deviation boundaries from the mean and indicate the cutoff we use to identify the outliers. These outliers are considered to be matches of insufficient quality and are excluded from the analysis.

Table S1. Demographic data details and user online activity (PC, mobile, and Yahoo! GO platforms)

Type	Details
Demographic data	
Gender	Self-reported gender of users.
Age	Self-reported age of users. Users below the age of 18 were excluded from the sample due to IRB requirements.
Primary country*	Observed daily. Refers to the country from which users accessed the portal most often.
Secondary country*	Observed daily. Refers to the country from which users accessed the portal second most often.
Mobile device†	Observed daily. The type of device most frequently used by the user to access Yahoo! services from a mobile platform. Includes 2,030 unique devices.
Go device‡	Observed daily. The type of device most frequently used by the user to operate Yahoo! Go software. Includes 111 unique devices.
IM network data	
Number of messages	Observed daily. Number of messages sent to and received from each Yahoo! Messenger contact.
Online activity and browsing behavior§	
Total page views (PV)	Total number of Web pages viewed on Yahoo! websites.
Front page PV	Total number of front page Web pages viewed on Yahoo! websites.
News PV	Total number of news-related Web pages viewed on Yahoo! websites.
Finance PV	Total number of finance-related Web pages viewed on Yahoo! websites.
Sports PV	Total number of sports-related Web pages viewed on Yahoo! websites.
Weather PV	Total number of weather-related Web pages viewed on Yahoo! websites.
Search PV	Total number of search-related Web pages viewed on Yahoo! websites.
Flickr (Photo-sharing) PV	Total number of Flickr (photo-sharing) Web pages viewed on Yahoo! websites.
e-mail PV	Total number of e-mail-related Web pages viewed on Yahoo! websites.

*Primary and secondary country data are recorded daily using the IP addresses of accessing devices.
†Designates data reported only for mobile users.
‡Designates data reported only for Yahoo! Go adopters.
§Behavioral data measure daily online activity and browsing behaviors across different micro content on the Yahoo! portal reported for every user. The records contain daily page views (PV) through October 2007 for browsing behaviors from stationary platforms (PC), mobile platforms, and from the Yahoo! Go application directly.

Table S2. Summary statistics of page views and IM network traffic

	PC			Mobile			Go		
	Mean	SD	Max	Mean	SD	Max	Mean	SD	Max
Page views by content category and platform									
Front page	0.3	4.4	9502.5	0.3	1.3	102.1	—	—	—
News	0.1	1.3	1742.8	0.0	0.8	156.3	0.6	1.6	120.1
Finance	0.1	9.1	22721.8	0.0	0.6	286.4	0.5	1.8	365.1
Sports	0.3	4.7	1361.7	0.0	0.3	100.25	0.5	1.3	122.3
Weather	0.0	0.1	250.9	0.0	0.0	17.3	0.5	1.3	94.3
Search	0.6	3.9	5189.6	0.1	1.9	512.0	1.2	4.4	634.3
Flickr	0.1	4.6	5654.1	0.0	0.5	324.5	0.9	4.6	709.0
E-mail	5.9	20.6	22153.2	1.3	4.8	598.9	2.8	9.4	696.7
IM network traffic (assessed using the random sample)									
In degree	2.9	20.8	4481						
Out degree	2.6	21.6	4674						
IM messages sent (average per day)	4.3	27.2	3342						
IM messages received (average per day)	4.4	41.9	3676						
Number of adopter-friends	0.02	0.2	36						
Fraction of adopter friends	0.01	0.07	1						
Clustering coefficient	0.005	0.04	1						

Table S3. Potential mechanisms of assortative mixing and temporal clustering in networks

Explanation	Implied relationships	Description	Example
Direct influence*	$a_i^* \leftarrow a_j^*$	Your focal action is influenced by the focal action of your friends.	When your friends exercise regularly, it influences you to exercise regularly.
Reflection effects†	$\left.\begin{matrix} a_i^* \leftarrow & c_{i.} & a_{i.} & s_i \\ a_j^* \leftarrow & c_{j.} & a_{j.} & s_j \end{matrix}\right\}$ + assortative mixing on *c*, *a*, or *s*.	Your own (and your friends' own) characteristics, actions and states influence your own (and your friends' own) focal action, and there is assortative mixing on the characteristics, actions, and states of you and your friends.	Eating healthy influences one to exercise regularly, and those who eat healthy tend to be friends.
Confounding environmental factors‡	$\left.\begin{matrix} a_i^* \leftarrow & e_i \\ a_j^* \leftarrow & e_j \end{matrix}\right\}^{\S}$ + *i* and *j* have identical *e*, or there is assortative mixing on *e*.	Unobserved factors that you experience influence you (and your friends) to take a focal action, and you and your friends experience the same unobserved factors or there is assortative mixing on these factors.	Employees of certain firms are financially rewarded when they exercise regularly. Friendships are formed at work.
Selection¶	$l_{i,j} \leftarrow \text{corr}(a_i^*, a_j^*)$	Links are formed due to shared or correlated actions or choices.	Those who exercise regularly tend to become friends.

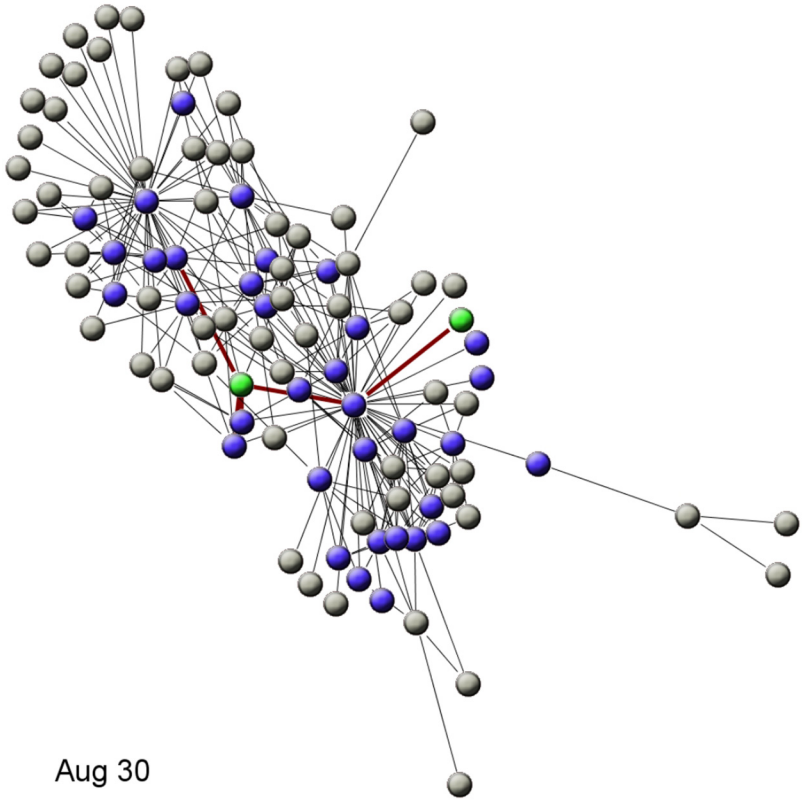
← denotes the direction of a causal relationship.
*a_i^*, a_j^*: Focal action or choice of interest taken by i or j (e.g., overeating, product adoption).
†a_i, a_j: Other actions or choices taken by *i* or *j*.
‡c_i, c_j: Characteristics of *i* or *j* (e.g. age, race).
§e_i, e_j: Unobserved "influencers" (e.g., neighborhood effects).
¶s_i, s_j: States of *i* or *j* (e.g. obesity, happiness). $l_{i,j}$ denotes a network link between *i* and *j* (e.g., friendship).

Table S4. List of individual characteristics used in proportional hazard rate model and cosine similarity estimates between users

Proportional hazard rate model (Cox)
Age
Gender
Fraction/Number of adopter friends
Number of IM messages
PC and mobile front page PV
PC and mobile weather PV
PC and mobile news PV
PC and mobile finance PV
PC and mobile sports PV
PC and mobile mail PV
PC and mobile search PV
Average friends' gender
Average friends' age
Average friends' degree
Average friends' IM messages
Average friends' finance PV
Average friends' search PV
Average friends' mail PV
Average friends' front page PV
Average friends' weather PV
Average friends' news PV
Average friends' sports PV
Location: USA
Location: India
Location: Romania
Location: UK
Location: Indonesia
Location: Philippines
Cosine similarity estimates
Gender
Age
Total page views
Weather PV
News PV
Finance PV
Sports PV
Mail PV
Search PV
Degree (number of IM buddies)
Average friends' gender
Average friends' age
Average friends' degree
Average friends' finance PV
Average friends' search PV
Average friends' mail PV
Average friends' front page PV
Average friends' weather PV
Average friends' news PV
Average friends' sports PV

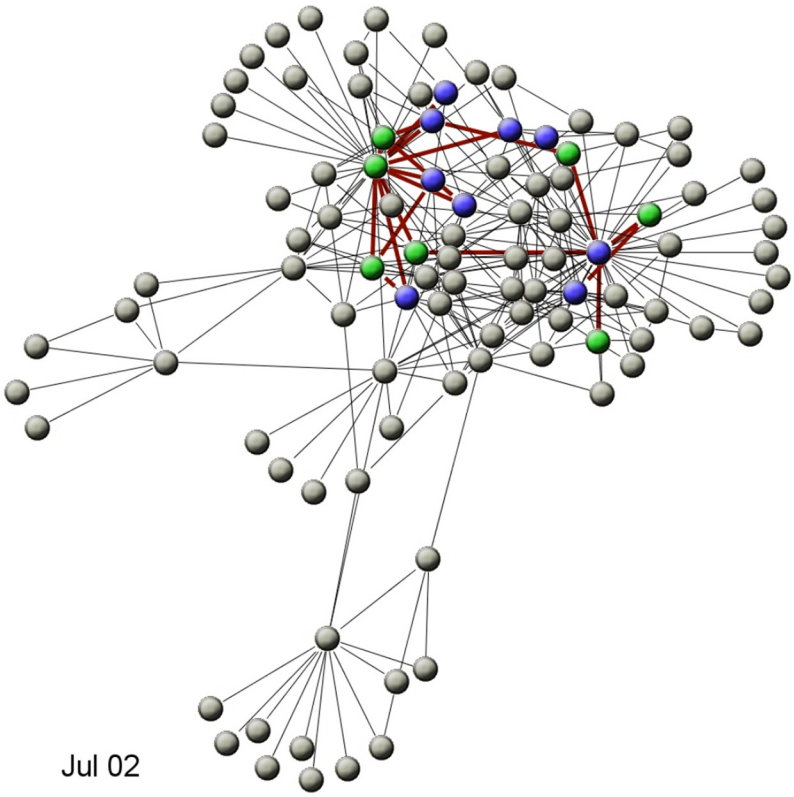
Table S5. Vector of individual characteristics (X_{it}): Covariates used to compute propensity of a node to be treated

Characteristic type	Characteristic
Personal	Gender
	Age
	Country
	Mobile device type
IM Network	Number of IM buddies
	Number of IM messages
	Mobile IM messages
Stationary platform behavior	Front page PV
	Weather PV
	News PV
	Finance PV
	Sports PV
	Mail PV
	Search PV
Mobile platform behavior	Mobile front page PV
	Mobile IM messages
	Mobile weather PV
	Mobile news PV
	Mobile finance PV
	Mobile sports PV
	Mobile mail PV
	Mobile search PV
Average friend's personal properties	Average friends' gender
	Average friends' age
Average friend's IM usage behavior	Average friends' degree
	Average friends' IM messages
Average friend's surfing behavior	Average friends' finance PV
	Average friends' search PV
	Average friends' mail PV
	Average friends' front page PV
	Average friends' weather PV
	Average friends' news PV
	Average friends' sports PV



Movie S1. A time-stamped animation of the diffusion of Yahoo! Go over time in a subsample of the instant messaging (IM) network. Gray nodes are yet to adopt. Blue nodes are old adopters from a different time period. Green nodes are recent adopters from the current day. Red lines connect recent adopters to old adopters or other recent adopters.

Movie S1 (AVI)



Movie S2. A time-stamped animation of the diffusion of Yahoo! Go over time in a subsample of the instant messaging (IM) network. Gray nodes are yet to adopt. Blue nodes are old adopters from a different time period. Green nodes are recent adopters from the current day. Red lines connect recent adopters to old adopters or other recent adopters.

Movie S2 (AVI)