JOURNAL OF COMPUTER-MEDIATED COMMUNICATION

Journal of Computer-Mediated Communication

The Diffusion of a Task Recommendation System to Facilitate Contributions to an Online Community

Y. Connie Yuan
Dan Cosley

Cornell University

Howard T. Welser

Ohio University

Ling Xia
Geri Gay

Cornell University

This paper studies the diffusion of SuggestBot, an intelligent task recommendation system that helps people find articles to edit in Wikipedia. We investigate factors that predict who adopts SuggestBot and its impact on adopters' future contributions to this online community. Analyzing records of participants' activities in Wikipedia, we found that both individual characteristics and social ties influence adoption. Specially, we found that highly involved contributors were more likely to adopt SuggestBot; interpersonal exposure to innovation, cohesion, and tie homophily all substantially increased the likelihood of adoption. However, connections to prominent, high-status contributors did not influence adoption. Finally, although the SuggestBot innovation saw limited distribution, adopters made significantly more contributions to Wikipedia after adoption than nonadopter counterparts in the comparison group.

Key words: diffusion of innovation, social network, tie cohesion, tie homophily, on-line community.

doi:10.1111/j.1083-6101.2009.01491.x

All communities, online and off, seek to motivate members to participate and continue contributing to the betterment of the group (Kanter, 1972; Olson, 1965). Whether posting messages, welcoming newcomers, building information databases, or helping to administrate the group's policy, online communities need member contributions to survive[1]. This is a serious problem for both new and existing communities because many face the challenge of undercontribution and/or inactivity over extended period of time (Cummings, Butler, & Kraut, 2002; Ling et al., 2005). Even in active communities, the levels of contribution among participants can be extremely uneven. For instance, in open-source development communities, Lakhani and von Hippel (2003) found that 4% of members contributed 50% of the answers on a user-to-user help site, while Mockus, Fielding, and Andersen (2002) found that 4% of developers contributed 88% of new code and 66% of code fixes. In our data we found that during a randomly selected 28-day data collection window for this study, 10% of the 6,570 randomly selected participants did not edit any Wikipedia content at all. In contrast, the most motivated contributor made 62,838 edits[2]; the top 5% of contributors made 44% of the total edits during this time. These contributors are obviously valuable, but uneven participation also has costs: It can lead to a few voices dominating the group and leave the group vulnerable if those few contributors depart. Thus, tools that encourage participation may help online communities thrive.

However, motivating contributions to these groups is difficult. As many scholars have observed, online communities and the resources they generate often take the form of a public good, in which all members of the community/public can enjoy the good regardless of their individual levels of contributions (e.g., Ling et al., 2005). Because community members can free ride on others' contributions, people will in general contribute less than would be optimal for the group. Although the critical mass model of collective action (Marwell & Oliver, 1993; Oliver & Marwell, 2001) predicts that a public good can be realized with the contribution of a small number of highly resourceful individuals so long as the provision level of the collective good reaches a level of self-sustainability, involving more contributors can make the group's participation patterns more democratic and robust.

One general strategy to increase participation is to reduce contribution costs, as suggested by the critical mass model of Marwell and Oliver (1993). The cost of contributing to online communities can take multiple forms, including financial cost, emotional cost, and cost in the time and effort in uploading/downloading information. Empirical studies on knowledge management in organizations show that employees were more likely to contribute their expertise to corporate knowledge repositories when contribution did not require too much time or effort (e.g., Yuan et al., 2005). Scholars of online communities have also found that reducing the cost of contribution by improving the design of technologies, e.g. by making it easier to find contributions a person would like to make, could motivate more contributions to a movie website's database (Cosley, Frankowski, Terveen, & Riedl, 2006) or to a discussion group (Ludford, Cosley, Frankowski, & Terveen, 2004). Following a similar logic, one of the authors created a recommendation tool, SuggestBot (Cosley, Frankowski, Terveen, & Riedl, 2007), and deployed it in Wikipedia to motivate more contributions to this online information commons. Wikipedia has hundreds of thousands of articles marked as needing improvement (usually lengthening), but no tools to help people find articles they are likely to be able to contribute to. Thus, there is a high cost to finding useful contributions to make. Building on the theory of collective action, SuggestBot uses a strategy called ''intelligent task routing'' to reduce a person's cost of finding articles to work on by recommending articles that both need attention and that are similar to articles that person has edited in the past[3]. Such articles are likely to be close to a person's interests, making it easier for them to contribute.

In this study, we examine two aspects of SuggestBot's use: First, how did it diffuse through the community, and second, how did it affect the contribution behavior of those who used it, compared to those who did not? Both questions are crucial when introducing technologies into online communities: If potential users do not adopt the tool, or it has minimal effects on their behavior, the technology will not benefit the community.

Diffusion of innovation has attracted decades of attention from scholars from diverse disciplines (Burt, 1987; Strang & Soule, 1998; Valente, 1996). However, the difficulties in tracking diffusion processes impose constraints on empirical research. Most studies use retrospective self-report data to examine the diffusion process, and the few studies that collect actual behavior data have sporadic information. For instance, in Coleman et al.'s (1966) study on the diffusion of tetracycline, doctors' prescriptions of the drug were sampled for only three consecutive days a month. Errors in recall or gaps in data sampling can add substantial noise to the data, influencing both statistical analysis and conceptual interpretation. The rise of the Internet has opened up new possibilities for observing diffusion processes. A plethora of digital traces of human online activities can be logged unobtrusively for academic research, giving scholars the opportunity to use objective measures of human behavior over time that are not contaminated with recall biases (Welser, Smith, Gleave, & Fisher, 2008). This may allow researchers to confirm and replicate findings from earlier studies on a much larger scale, as well as to investigate some of the issues that used to be too demanding to study empirically.

Wikipedia makes an excellent site for digital research because almost all activities on the site, including details of article edits and interpersonal communication, are archived and freely available for download. This data, plus our access to SuggestBot's internal logs, allowed us to obtain a complete, time-stamped record[4] of (a) who has adopted SuggestBot, (b) who has interacted with whom, and (c) who has edited articles suggested by SuggestBot. The data also allow us to find nonadopters, a much overlooked segment in existing diffusion research (Rogers, 2003) to compare and contrast with adopters. Finally, because all the data collected have a precise time stamp, the resulting empirical measurements can be arranged along a clear temporal order, which gives us more power to make causal inferences. Overall, we believe that our work can contribute to diffusion of innovation research from multiple dimensions.

In addition to furthering our understanding of diffusion, the project improves our understanding of how to motivate contributions to online communities. Motivating contribution to electronic commons is a challenging task because when community members are distributed globally, some conventional incentive strategies such as fostering strong local norms of cooperation (Coleman, 1988) become more difficult to implement. A promising alternative, suggested by Kraut (2003), is to use social science theories to inform the design of tools that motivate participation. SuggestBot, as briefly described above, dovetails with Kraut's call in that its design followed the basic premises of collective action theory (Olson, 1965, Marwell & Oliver, 1993) and theories of individual motivation to participate in groups (Karau & Williams, 1993), with a goal to involve more people in community development via cost reduction. Through examining SuggestBot's diffusion process, as well as the effect of its adoption, we can better understand how to motivate contributions to online communities.

Using a sample of 6,570 Wikipedia contributors, we explored possible answers to the following questions: (a) which factors influenced adoption of SuggestBot? And (b) has the adoption of SuggestBot made a difference in individuals' contributions to the community? In the following section of the paper, we will first review related literature about factors that may influence diffusion of innovation and online contributions. We then present an empirical test of the research questions/hypotheses raised. The paper ends with a discussion on substantive implications of our findings, practical implications, and directions for future research.

Diffusion of SuggestBot in Wikipedia

Existing studies on diffusion of innovation have identified a long list of factors that can influence the diffusion process. These factors can roughly be classified into two categories: attribute and relational factors (Scott, 1991/2004). Attribute factors focus on individual characteristics, such as innovativeness (i.e., willingness to try out new ideas/products (Rogers, 2003, p. 267–299)), exposure to mass media or metropolitan culture (e.g., Valente, 1996), and so on. Relational factors focus on structural properties of network relationships such as cohesion, tie homophily, and so on. Below, we will review both attribute and relational factors that we think may influence the adoption of SuggestBot in Wikipedia.

Individual Attributes

In Wikipedia, we anticipate that highly involved editors are more likely to adopt SuggestBot because these people are more committed to improve the quality of Wikipedia entries. Just as Rogers has found that innovativeness—valuing new technologies—leads to a greater propensity to adopt technologies in general (Rogers, 2003), we expect that involvement in Wikipedia—evidence of valuing the community—will lead to a greater propensity to adopt technologies related to Wikipedia. To operationalize level of involvement, we look at two behavioral factors that can be measured from activity logs: *admin status* and *preadoption contribution.*

"Admin" status—that is, being listed as an administrator on Wikipedia—is an important indicator of involvement. Only those who have made substantial contributions to Wikipedia over time and who are recognized as valuable, committed contributors by other admins can earn this status. Obtaining such a status can further motivate contribution for three reasons. First, it is a public acknowledgement of these people's sustained commitment to the community. Second, it gives these contributors additional privileges to help moderate Wikipedia entries and contributors, providing additional ways to be even more involved. Third, following Burke and Reitzes's (1991) identity theory of commitment, when commitment to a group is reinforced with a clearly assigned role, commitment to a role identity can further motivate engagement in activities that are associated with the role of a highly committed member (p.242). Given these contributors' intrinsic motivation to improve Wikipedia and the added incentive of having a more formally defined public role, we anticipate that they are more likely to adopt tools to cut their cost of contributing. Therefore, it is hypothesized:

> Hypothesis 1: Those Wikipedia contributors who have earned admin status before adoption are more likely to adopt SuggestBot than those who did not have such a status.

Admin status is a rather exclusive indicator of involvement in that it is awarded to a relatively small number of contributors. As of November 20, 2008[5], only 1,618 contributors had admin status out of over 8 million registered users, over 160,000 of whom were active within the last 30 days. Using admin status alone as an indicator of involvement therefore ignores many contributors who have made substantial contributions to Wikipedia, but who have not earned or sought admin status. Thus, we supplement admin status with another potential behavioral indicator of involvement: amount of preadoption contributions. Hechter (1987) defines involvement in terms of contribution, as the proportion of a person's resources that they dedicate to the goals of the group (p. 18). Kanter (1972) also emphasizes the importance of concrete practices that reinforce commitment and collectively held beliefs (p. 75). Both perspectives reinforce the notion that highly involved or committed members of a community should be more likely to adopt practices that either aid their contribution, or are consistent with the norms appropriate with high contributing members. Therefore, we believe that a positive relationship exists between levels of preadoption contribution and the likelihood of adoption because frequent contributors, regardless of their status, share a common motivation with the admin contributors, i.e. to improve the quality of Wikipedia entries. Based on these arguments, it is hypothesized:

> Hypothesis 2: High preadoption contribution predicts higher likelihood of adoption.

The Influence of Communication Networks

In addition to individual involvement, network connections among contributors can also influence adoption. The turbocharger effect happens when network variables explain additional variance in adoption beyond the direct effects of attribute variables (Rogers, 2003, p. 360). In Wikipedia, contributors communicate with each other through posting on each other's user-talk pages. Relationships formed over time through such social interactions can significantly influence the likelihood of adoption, causing differences in the time and probability of adoption even though SuggestBot is available to the entire community through Wikipedia's Community Portal page, a place where contributors can find tasks to do and tools to help do them.

Social influence through interpersonal exposure

Previous network studies of diffusion found that external influences in the form of subscription to medical journals, cosmopolitan connections, mass media coverage, and so on (DiMaggio & Powell, 1983; Rogers, 2003) can only inform potential adopters of an innovation. Interpersonal influence with friends and neighbors is often what leads to actual adoption (Valente, 1996, p. 80). It means that embedded in a network of adopters, a low-involvement contributor may still adopt an innovation following sufficient exposure to peers who have adopted it. Valente maintained that the impact of social influence through direct exposure needs to be accumulated over time. The larger the number of adopters that a focal node has in his/her ego network, the higher the chances of interpersonal exposure. When interpersonal exposures accumulate, potential adopters' familiarity with the innovation increases (Wejnert, 2002). Over time, the likelihood of adoption grows with reduction in people's fear and uncertainty about the innovation. When the level of interpersonal exposure exceeds a certain threshold, adoption happens. Because SuggestBot writes its suggestions on adopters' user-talk pages, people who interact with adopters are naturally exposed to SuggestBot. Thus, it is hypothesized that:

> Hypothesis 3: High interpersonal exposure predicts higher likelihood of adoption.

Other properties of communication networks can further enhance the persuasive power of interpersonal exposure. Such network properties include cohesion, tie homophily, and ties to opinion leaders (Rogers, 2003; Strang & Soule, 1998).

Interpersonal cohesion

In the diffusion of innovation research, cohesion refers to the strength of connection between ego and alter (Burt, 1999, p. 39). Cohesion implies a higher level of connectedness with other members in a community, as well as a stronger sense of belonging to the community. Cohesive ties can therefore contribute to higher likelihood of adoption because they enhance the power of social influence during interpersonal exposure (Strang & Soule, 1998). In addition, cohesion creates stronger incentives to comply with collective norms (Coleman, 1988). As Wejnert (2002) observed, in many cases adoption is a ''network-based decision ... as pressure toward conformity builds'' (p. 306). While people differ in their tendency to confirm to social norms, cohesively tied individuals are likely to mutually influence each other and jointly form a norm that they both are willing to buy into. Based on these arguments, it is hypothesized:

> Hypothesis 4: High interpersonal cohesion predicts higher likelihood of adoption.

Tie homophily

Rogers (2003) observed that "interpersonal diffusion networks are most homophilous" (p. 307) because commonalities between adopters and nonadopters increase the power of social influence. Extensive studies have found higher likelihood of establishing social ties among people having similar characteristics (e.g., Ibarra, 1992; Yuan & Gay, 2006). Monge and Contractor (2003) summarize two main lines of reasoning that support the theory of homophily, including Byrne's (1971) similarity-attraction hypothesis and Turner's (1987) theory of self-categorization. The similarity-attraction hypothesis predicts that people are more likely to interact with those who share similar traits. The theory of self-categorization proposes that people tend to categorize themselves and others in terms of race, gender, age, education, interests, and so on. Individuals classified into the same categories perceive themselves as more similar to each other. Because interpersonal similarity breeds connections (McPherson, Smith-Lovin, & Cook, 2001, p. 415), social interactions are more likely to happen among similar others. Moreover, because interpersonal similarity and effective communication breed each other (Rogers, 2003, p. 306), homophilous ties become effective means of social influence.

The finding of adoption clusters in the diffusion process provides strong evidence for homophilous influence in diffusion. When studying the diffusion of family planning methods in Korean villages, Rogers and Kincaid (1981) found "pill villages," "IUD villages," and "vasectomy villages," where women of the same village tended to adopt the same contraceptive method even though all the different family planning methods were introduced to each village at the same time. In the context of the current research, we anticipate that homophilous ties among contributors who had the same Wikipedian status (i.e. ties among admin contributors and among nonadmin contributors) would increase the influence of interpersonal exposure. Over time, peer-to-peer communication via homophilous ties can trigger bandwagon effects among contributors (Abrahamson & Rosenkopf, 1997), and consequently bring about widespread adoption. Summarizing these findings and reasoning, it is hypothesized:

Hypothesis 5: High levels of tie homophily predict higher likelihood of adoption.

Ties to opinion leaders

While homophilous ties can contribute to adoption, diffusion through homophilous ties tends to be horizontal and confined to people in the same social category (Rogers, 2003). Therefore, ties to people from different social categories are crucial for vertical diffusion throughout the whole system (p. 308). Among different cross-boundary ties, the most crucial ones are those with opinion leaders. Opinion leaders typically occupy higher status in a system, have greater exposure to external information, participate in more social circles, and so on (p.308). Connections with opinion leaders can therefore have a greater influence on adoption because opinions from high-status alters tend to carry more weight. Existing research on health intervention programs has found significant influence of opinion leaders, e.g. breast cancer survivors, on motivating people to adopt particular prevention practices (Earp et al., 2002).

In Wikipedia, we anticipate that people with admin status can function as opinion leaders. As discussed in Hypothesis 1, because only highly involved contributors can earn admin status, and because the status gives these contributors additional privileges and visibility in Wikipedia, we anticipate that these contributors are more likely than others to adopt SuggestBot. Building on this preassumption, we further hypothesize that ties to admin contributors would boost adoption across the whole community. Burt (1999) maintains that opinion leaders are actually information brokers who play a key role in bringing innovative information from one group to another. Given that opinion leaders are usually better connected (Rogers, 2003), we anticipate that admin contributors can spread SuggestBot beyond the circle of admin contributors and into the community at large. Based on these arguments and reasoning, it is hypothesized that:

> Hypothesis 6: More ties to opinion leaders predict higher likelihood of adoption.

The Impact of Adoption

As discussed earlier, SuggestBot was developed to facilitate contribution to Wikipedia. We were interested in evaluating the effect of adoption, i.e. whether the adoption has resulted in greater contribution to the community. Because the tool was designed to make it easier for contributors to locate which entries need developing or editing, we anticipate that adoption would increase contributions by reducing search cost. It is therefore hypothesized that

> Hypothesis 7: Post adoption, adopters of SuggestBot contribute more than nonadopters.

Method

Sample

Our data is based on a set of 6,570 Wikipedia editors. We first selected 2,190 editors who used SuggestBot at least once between March 8, 2006 and March 30, 2007. This is not quite all of the adopters, for technical reasons such as users changing names and occasional errors in the SuggestBot software, but it is the vast majority of adopters. The number of adopters was low compared to the number of contributors to Wikipedia, so we sampled nonadopters following King and Zeng's (2001) recommendation on sampling ratio (between 1:2 and 1:5) for rare event data. This type of case control sampling is a standard strategy in epidemiological studies of large, naturally occurring populations where positive cases are rare, and where cases are markedly heterogeneous on key control variables (see Stolley & Schlesselman, 1982). Given that our population of study shared the same characteristics, we used this sampling method in our study.

Specifically, for each adopter we sampled two nonadopters as control cases based on their activity level before adoption and the time they first started editing Wikipedia. These were important case control variables for the following reasons. First, matching on first edit time was important because even over the course of a year, the behavior and norms of Wikipedia change significantly. The number of registered users, total articles, and number of edits have all grown exponentially for several years. Likewise, the policies and guidelines governing behavior have evolved substantially over time. Social conditions have also changed—for instance, new contributors are much more likely to receive a personal message from an existing member now than they were 4 years ago. It is therefore important to ensure that on balance, conditions in Wikipedia were approximately the same for both the adopters and nonadopters in our sample. Second, matching on activity level before adoption was important to make sure that the adopter and nonadopter groups were comparable. The distribution of activity in online communities is often highly skewed (and often follows a power-law distribution). For instance, as of March 30, 2007, there were almost 1,190,000 registered editors who had a median of 3, a mean of 69, and a maximum of 162,138 edits[6]. Thus, a random sample of nonadopters that did not consider activity would primarily choose editors who had made very few contributions to Wikipedia. Further, because contributors of similar activity levels were more likely to have similar levels of social network ties, which are the key focus of our study, this sampling method gives more conservative estimates of the influence of these social network variables on adoption than simple random sampling.

To identify control cases, we looked at every adopter's first edit time and their adoption time, the date when they first used SuggestBot. For each adopter we chose two nonadopters whose first Wikipedia edit was within a week of the adopter's first edit and whose number of edits at the corresponding adopter's adoption time was within 10 percent of the adopter's number of edits at adoption time. Nonadopters were sampled without replacement (i.e., they would not be selected more than once in the dataset even though one nonadopter may match the profile of two adopters). For a few adopters, we could not find nonadopters who matched the corresponding adopter closely; in these cases, we incrementally widened the first edit date and activity level differences until we could find two matching nonadopters.

Because large samples make it easy to find significant results despite trivial effect sizes, we generated a subsample to evaluate and cross-validate the results. Using the random-generate function in SPSS 16.0, we randomly chose 15% (960 cases) of the original sample for results validation. All the hypotheses proposed were tested in both the whole sample and the subsample. Despite minor differences in the strengths of relationships, the patterns and directionalities of the tests were consistent across both samples.

Measurements

Because of the voluntary nature of participation in Wikipedia, contributors may be dormant for some time before they become active again. To reduce such noise in data, and exclude activities and social interactions prior to the launch of SuggestBot in Wikipedia, we decided to measure activity during the 28 days before and after a focal contributor adopted SuggestBot. The 28 day measurement window is intended to capture a snapshot of behavior that is representative of contributors' behavior during a period of recent potential influence.[7] Social network variables were also calculated using the time frame of 28 days before adoption to evaluate the extent of social tie formation through cross-postings on user talk pages. In the following section of the paper, unless specified otherwise, the variables calculated all refer to activities that took place within the 28 days measurement window. For nonadopters, the window was centered on the adoption time of their adopter counterparts to facilitate comparisons.

Individual attribute predictors include several variables that measure an editor's involvement and activity in Wikipedia. *Admin status* was a dummy variable with 1 representing "have obtained administrator status in Wikipedia by the time of adoption," and 0 representing "have not." *Pre-adoption contribution* measures the total number of edits that a contributor has made during 28 days before adoption. *Total preadoption contribution* measures the total number of edits that a contributor made before adopting. *Total postadoption contribution* measures the total number of edits that a contributor has made since adoption.

Network variables were generated by tracing Wikipedia contributors' posts on each other's user-talk pages. *Interpersonal exposure to innovation* was calculated by counting the number of adopters of SuggestBot that a contributor had direct contact with during 28 days prior to adoption. *Interpersonal cohesion* was measured by counting the number of reciprocal ties in a contributor's ego network. *Tie homophily* was calculated using the percentage of a contributor's same-status ties (admin vs. nonadmin). Finally, number of *ties to opinion leaders* was measured by counting the number of ties that an ego had with contributors who had admin status.

Control variable

In addition to the research variables, *number of active months* was calculated by counting the number of calendar months in which a contributor had made at least one edit before April, 2007. We believe this measurement of tenure with Wikipedia is a stronger control variable than a mere count of the number of months since a person has joined because contributors sometimes have dormant periods where they temporarily leave the community.

Analysis Method

To address the hypotheses around diffusion, we used logistic regression models because the dependent variable had only two response categories, with 1 = Adopted and 0 = Didn't Adopt. In reporting the results, comparison of odds ratios (o.r.), changes in significance, and model fit (as measured by pseudo R square) are used to illuminate how the effects of hypothesized variables change in the multivariate context. An odds ratio of greater than 1 indicates that an increased level of the variable is associated with an increased probability of adoption, while odds ratios between 1 and 0 indicate a diminished probability of adoption. A more concrete interpretation of the effects of predictor variables on the probability of adoption can be created by using the regression equation to calculate predicted probabilities for different levels of the predictor variables. The corresponding *p* value of the Wald test reveals whether a predictor variable can significantly influence the likelihood of adoption, controlling for all the other predictors in the equation. To emulate the multiple R^2 in ordinary least squares regression analysis, a number of pseudo R square measures, including Cox-and-Snell's, Nagelkerke's, and McFadden's R^2 have been proposed for logistic regression analysis. Among them Nagelkerke's pseudo R^2 is preferred (Garson, 2008) because it is a modified version of the Cox-and-Snell's R^2 value to ensure that its value falls between 0 and 1. To evaluate improvement in model fit, we used likelihood ratio (-2LL) tests between nested models. Since likelihood ratio approximates a χ^2 distribution, a significant drop in χ^2 implies significant improvement in fit of the regression equation with the data.

To address the question about the impact of adoption on contribution to Wikipedia, we conducted independent sample t-tests to compare the adopter and the nonadopter groups in their level of contribution to Wikipedia community both *before* and *after* adoption. Significant changes in contribution levels before and after adoption between adopters and nonadopters reveal the impact of adoption.

The descriptive statistics and zero-order correlations among the raw variables are reported in Table 1. We used point-biserial correlation coefficients to evaluate the relationship between admin status and other variables because admin status was a dichotomous variable; the rest were Pearson correlation coefficients.

Data Preparation

As described earlier, in most online communities, a small group of people tend to contribute a disproportionately high share of content, while most members are much less active, causing extremely high skewness in data distribution. Prior to conducting a series of tests of the hypotheses, we transformed a number of variables to reduce skewness. There is a major substantive reason to prefer transformations of key variables in studies that measure counts of behaviors in populations interacting across time. Utilizing untransformed variables makes the following assumption: A one unit increase at low levels of x has the same amount of effect on the causal mechanism as a one unit change at high levels of x. This assumption is almost certainly false. Instead, a much more likely situation is that a one unit change in x is decreasingly influential as x increases. Consider one measure of commitment, preadoption contribution. A shift from 10 edits to 20 edits is likely a substantively important difference in commitment, but for someone with 1000 edits, an extra 10 edits is not nearly as meaningful. There are also several general methodological reasons to perform transformations to reduce the skewness of predictors in general linear models, including logistic regression. While logistic regression does not require normal distribution of variables (Tabachnick & Fidell, 1996, p.575), highly skewed

Table 1 Descriptive Statistics and Zero-Order Correlations

	1	2	3	4	5	6	7	8	9	10	Mean	Std. Dev.
1. Active months		.37*	.06	.50*	.16*	.05	.10*	.07*	.14*	.08*	13.21	9.35
2. Admin status	.35*		.05	.42*	.09*	.25*	.21*	−.07*	.29*	.06	.03	.18
3. Preadoption contribution (28 days before adoption)	.14*	.10*		.33*	.46*	.28*	.35*	.28*	.40*	.10*	107.03	200.83
4. Total contribution before adoption	.47*	.43*	.46*		.26*	.10*	.15*	.07*	.20*	−.04	863.56	1657.38
5. Total contribution after adoption	.23*	.13*	.56*	.42*		.15*	.20*	.15*	.22*	.08*	582.29	1811.32
6. Interpersonal exposure	.06*	.12*	.19*	.11*	.10*		.79*	.16*	.78*	.24*	.66	2.10
7. Cohesion	.15*	.24*	.40*	.25*	.26*	.52*		.19*	.82*	.23*	1.74	5.32
8. Tie homophily	.13*	-.10*	.27*	.09*	.18*	.14*	.21*		.14*	.28*	.45	.44
9. Ties to opinion leaders	.16*	.30*	.33*	.24*	.23*	.53*	.80*	.12*		.21*	1.40	3.73
10. Adoption	.09*	.02*	.10*	−.03*	.06*	.13*	.22*	.27*	.18*		.33	.47
Mean	13.65	.04	112.75	1008.54	638.17	.69	1.82	.45	1.42	.33		
Standard Deviation	9.72	.19	239.22	2006.45	1726.20	2.87	4.91	.44	4.08	.47		

*p < .05

Note: The lower triangle reports the descriptive statistics and zero-order correlation using the whole sample (N = 6570); the upper triangle reports these using the random sample (N = 960).

variables often result in patterns of nonconstant error variance, excessive leverage, and inefficiency (Fox, 1991, 2002; Pregibon, 1981). Thus, though we acknowledge possible differences in results between transformed and untransformed variables, we believe that transforming the data increases the validity of our results.

Among the three commonly used data transformation methods to improve normality of distribution, i.e. square root, log, and inverse, Osborne (2002) maintained that inverse was most effective in transforming extremely skewed distributions. In addition, Tabachnick and Fidell (1996) have specifically recommended this data transformation method when the distribution of a variable resembles an "L" in shape (see specific pictures on p. 83 of their book), which was exactly how our independent variables were distributed. Following Osborne (2002), we first took the inverse of a variable, then multiplied the inversed value by −1 to preserve the rank order of the original variable, and finally added a constant to bring the minimum value above 1.0. These transformations significantly reduced the skewness levels.

Results

Hypotheses 1 to 6 examine factors that influence adoption. Since they shared the same dependent variable, the research variables were entered into the logistic regression model in steps; the models are shown in Table 2 for both the whole sample and the subsample. In general, results hold in both cases; we report the results from both samples below.

Model 1 contained only the control variable, number of active months. When number of active months was the only predictor variable in the analysis, it had significant influence on likelihood of adoption in both the whole sample and the subsample (B = .02, odds ratio = 1.02, $p < .05$ for both). The changes in the deviance scores from the null, baseline model showed that the improvement in model fit was significant ($\chi^2_{\text{whole sample}} = 51.70$, df = 1, $p < .05$; $\chi^2_{\text{subsample}} = 5.48$, df = 1, $p < .05$). However, Nagelkerke's pseudo R-square for Model 1 was .01, indicating that the overall fit was poor.

Model 2 added the attribute variables of admin status and total activity. Conceptually, these are both indicators of individual level involvement to Wikipedia. The control variable, number of active months, became nonsignificant when the attribute variables were added. Counter to Hypothesis 1, admin status was not a significant predictor of likelihood of adoption ($B_{\text{whole}} = .05$, o.r. $_{\text{whole}} = 1.05$, $p > .05$; $B_{\text{sub}} = .37$, o.r. $_{\text{sub}} = 1.45$, $p > .05$). That is, although the odds ratios showed that those with admin status had higher likelihood of adopting SuggestBot, the increased likelihood was not statistically significant. Consistent with Hypothesis 2, preadoption contribution was a significant predictor of adoption ($B_{\text{whole}} = 2.42$, o.r. $_{\text{whole}} = 11.23$, $p < .05$; $B_{\text{sub}} = 2.48$, o.r. $_{\text{sub}} = 11.89$, $p < .05$). That is, heavy contributors had much higher likelihood of adoption. Nagelkerke's pseudo R-square for Model 2 increased to .15, compared to .01 for Model 1; changes in likelihood ratios between the control-variable-only model and the current model showed that

Table 2 Results of Logistic Regression Analysis

	Model 1		Model 2		Model 3	
Variable	Whole Sample	Subsample	Whole Sample	Subsample	Whole Sample	Subsample
Number of active months	1.02**	1.02*	1.00	1.01	1.00	1.00
Admin status			1.05	1.45	.77	1.05
Preadoption contribution			11.23**	11.89**	5.14**	5.42**
Interpersonal exposure					3.08**	5.01*
Cohesion					1.93**	2.45*
Tie homophily					1.97**	2.34**
Ties to opinion leaders					.88	.46
R^2	0.01	0.01	0.15	0.15	0.22	0.24

**p < .01
*p < .05
Whole sample (n = 6,570)
Subsample (n = 960)

this improvement in model fit was significant ($\chi^2_{\text{whole}} = 709.64$, df = 2, $p < .05$; $\chi^2_{\text{sub}} = 105.56$, df = 2, $p < .05$). To sum up, although both administrator status and volume of contribution prior to adoption had clear theoretical reasons for being positively associated with adoption, the substantial increase in model fit was largely attributed to preadoption contribution, the only significant variable in Model 2.

Model 3 added four network variables. Consistent with Hypothesis 3, interpersonal exposure was a significant predictor of likelihood of adoption ($B_{\text{whole}} = 1.12$, o.r. $_{\text{whole}} = 3.08$, $p < .05$; $B_{\text{sub}} = 1.61$, o.r. $_{\text{sub}} = 5.01$, $p < .05$). Also consistent with Hypothesis 4, cohesion through reciprocal ties was a significant predictor of likelihood of adoption ($B_{\text{whole}} = .66$, o.r. $_{\text{whole}} = 1.93$, $p < .05$; $B_{\text{sub}} = .90$, o.r. $_{\text{sub}} = 2.45$, $p < .05$). Supporting Hypothesis 5, tie homophily was a significant predictor of likelihood of adoption ($B_{\text{whole}} = .68$, o.r. $_{\text{whole}} = 1.97$, $p < .05$; $B_{\text{sub}} = .85$, o.r. $_{\text{sub}} = 2.34$, $p < .05$). However, counter to Hypothesis 6, ties to opinion leaders did not increase likelihood of adoption ($B_{\text{whole}} = -.13$, o.r. $_{\text{whole}} = .88$, $p > .05$; $B_{\text{sub}} = -.79$, o.r. $_{\text{sub}} = .46$, $p > .05$). Nagelkerke's pseudo R-square for Model 3 increased to .22, compared to .15 for Model 2. Again, the change in likelihood ratios showed that the observed improvement in model fit from Model 2 was statistically significant ($\chi^2_{\text{whole}} = 359.46$, df = 4, $p < .05$; $\chi^2_{\text{sub}} = 71.32$, df = 4, $p < .05$).

Comparing odds ratios and model fit between Model 2 to Model 3 sheds light on how both individual commitment (preadoption contribution) and social influence (communication network variables) affect adoption. Adding social influence variables significantly increased model fit, and three out of four of the social influence variables had positive and clearly significant effects. However, the addition of these variables did not eclipse the role of involvement, as preadoption contribution remains strongly and positively associated with adoption. Thus, our results suggest that both individual involvement and social influence from network ties can have important and independent positive effects on the likelihood of adoption in communities of collaborators such as Wikipedia.

According to Models 2 and 3, comparison of odds ratios and model fit seems to suggest that preadoption contribution accounts for most of explanatory power of the final model. However, because variable effects are multiplicative and vary across the range of a predictor variable, odds ratios cannot be directly interpreted in terms of effect on probability. Further, because they are unstandardized, they cannot be easily compared across variables (see Menard 2001)[8]. To get a better sense for the relative importance of the model variables, we used Model 3 to compute the probability of adoption for several hypothetical users who differ on key theoretical variables that predicted adoption. Table 3 reports the results.

The first column of Table 3 shows the coefficients from Model 3 based on the whole sample. The second column presents a hypothetical average contributor who has the sample average for each of the variables, along with the probability that such a person would adopt SuggestBot. The next four columns explore how much impact each individual variable has by estimating the probability of adopting for a person who is average in every respect except that they have an unusually high value for one of the four key predictors. The last column presents the probability of adoption for a "supercontributor" who has high values for all four of the key predictors[9].

The model computes a baseline probability for the "average" contributor to adopt of .30[10]. Ceteris paribus, elevating tie homophily has only a modest effect on adoption (+.03, compared to the baseline). Increasing cohesion to a similarly high level almost doubles the increase in probability to +.07, setting contribution to a high level further increases the predicted probability by +.09 over the baseline, and interpersonal exposure has the greatest effect on the probability of adoption, raising it by +.15 over the baseline. Hypotheses 1–6 all referred to mechanisms that were predicted to increase the probability of adoption. While we found support for Hypotheses 2–5, these results suggest that the strongest effects arise from high levels of contribution (Hypothesis 2) and interpersonal ties (Hypothesis 3). Furthermore, when these mechanisms work in concert with tie homophily and cohesion, the model shows evidence of a strong increase in the probability of adoption by +.37 compared to the baseline.

The second research question explores how adoption affects contribution. Hypothesis 7 predicts that adopters would contribute more in the future than nonadopters. To test this, we conducted independent sample *t*-tests to compare the mean level of contributions between adopter and nonadopter groups both *before* and *after* adoption. Levene's test showed that equal variance between the two groups in the whole sample could not be assumed ($F(1, 6{,}568) = 6.37$, $p < .05$). The corresponding t value was statistically significant, $t(4{,}776) = -2.34$, $p < .05$, indicating

Table 3 Predicted Probabilities Computed From Model Three, Full Sample

Variable	Model coefficient	Sample average	High homophily	High cohesion	High contribution	High exposure	High composite
Intercept	−6.485						
Number of active months	−0.002	13.65					
Admin status	−0.258	0.037					
Preadoption contribution	1.635	1.735			1.995		1.995
Interpersonal exposure	1.122	1.165				1.745	1.745
Cohesion	0.654	1.248		1.745			1.745
Tie homophily	0.682	1.242	1.495				1.495
Ties to opinion leaders	−0.127	1.229					
Predicted probability		.30	.33	.37	.39	.45	0.67
Increase in predicted probability over baseline		—	+.03	+.07	+.09	+.15	+.37

Note: Table 3 compares the probability of six counterfactual people adopting SuggestBot. The baseline is the Sample Average, who is assigned sample means on all variables. Subsequent columns maintain the sample means for all variables except those indicated. The changing values represent the average of the third quartile and the maximum value for that variable. Models are displayed in ascending order of predicted probability.

that before adoption, adopters on average contributed less to the community than the nonadopters ($M_{adopters} = 926.52$, SD = 1875.71 versus $M_{nonadopters} = 1{,}049.55$, SD = 2,067.72). However, the contribution pattern was reversed after adopters adopted SuggestBot. An independent sample *t*-test on postadoption contribution found significant difference in means, $t(4{,}317) = 4.99$, $p < .01$ between adopters and nonadopters ($M_{adopters} = 788.86$, SD = 1,741.31 versus $M_{nonadopters} = 562.82$, SD = 1,713.83), indicating that adopters contributed significantly more than the nonadopters after they adopted SuggestBot.

The same independent sample *t*-tests were also conducted with the randomly generated subsample. The tests produced similar results. Before adoption, adopters and nonadopters ($M_{adopters} = 769.65$, SD = 1,613.94 versus $M_{nonadopters} = 910.51$, SD = 1,677.96) did not differ significantly in their levels of contribution to the community (equal variance of two groups can be assumed ($F(1{,}958) = 2.41$, $p > .05$), $t(958) = -1.24$, $p > .05$). However, after adoption, the adopters contributed significantly more to the community than nonadopters ($M_{adopters} = 787.36$ SD = 1805.67 versus $M_{nonadopters} = 479.75$, SD = 1,806.83), $t(638) = 2.49$, $p < .01$ (equal variance between the two groups could not be assumed ($F(1{,}958) = 4.30$, $p < .05$)). Taken together, the results based on both the whole sample and the subsample supported Hypothesis 7, suggesting that adoption of SuggestBot enabled adopters to make significantly more contributions to the community even though nonadopters contributed equally or more than adopters prior to adoption.

Discussion

Just as cities need a visitor's bureau and a sanitation department, online communities need to welcome newcomers, edit and modify existing content, and police problems. A key problem for these communities is motivating their members to perform this kind of maintenance activity. The critical mass model of public goods suggests that reducing the cost of contribution can motivate contributions (Marwell & Oliver, 1993). Building on this proposition, SuggestBot was developed to make it easier for Wikipedia editors to locate entries that match their abilities to contribute. In this study, we examined how SuggestBot diffused in Wikipedia and whether the adoption of SuggestBot has made any impact on people's contributions. The research aims at contributing to our understanding of both diffusion of innovation processes and the development of online communities. The results confirmed some of the insights from earlier studies and point out some areas of research worthy of further investigation.

Contributions

From hybrid corn to poison pills (Strang & Soule, 1998), diffusion of innovation is a research paradigm widely applicable across disciplines. The appeal and practical implication of the paradigm are strong because the success of an innovation would be incomplete without successful diffusion. However, some research issues remain inadequately explored even after decades of research on the topic. Rogers observed

(Rogers, 2003; 1971) a proinnovation bias in diffusion of innovation research in that researchers tend to focus more on adopters of successfully diffused innovations. As a result, we have very limited knowledge about nonadopters and unsuccessfully diffused innovations (Rogers, 2003). While adopters and nonadopters may share many commonalities, it is premature to accept this assumption without empirical evidence—but such evidence is hard to collect. Using retrospective self-reported data to study only adopters of successfully diffused innovation is unavoidable in many situations because diffusion is a process that has an unknown prospective duration of diffusion time. It is therefore hard for researchers to predict in advance the right time to track the actual behavior of every person in the population when working with limited resources. Thus, empirical difficulties in tracking diffusion processes can potentially impose major conceptual constraints on what questions can be asked and/or addressed.

Reaping the benefit of a growing body of digital traces of people's natural behavior on the Internet, we studied how individual and network factors influenced diffusion of an intelligent task recommendation tool, SuggestBot, in Wikipedia. Instead of focusing on adopters only, we first identified 2,190 adopters, obtained the public records of their contribution to Wikipedia, and then matched each adopter in our sample with two nonadopters by their starting time and their overall activity level prior to adoption. The inclusion of these nonadopters helped us gain a better understanding about the key factors that influence adoption in the whole population. Moreover, matching adopters and nonadopters on their preadoption activity levels and time of joining the community provides an even more conservative test of the research hypotheses because the sampling procedure made nonadopters in our comparison group resemble adopters more closely than a randomly selected nonadopter in multiple dimensions, including likelihood of adoption, network variables such as tie cohesion and homophily, and individual attribute variables such as involvement with Wikipedia.

About individual attribute variables, we anticipated that those contributors with admin status would show higher likelihood of adoption, given they are more likely to have sustained interest in improving Wikipedia. This was not the case; though admin contributors had a higher likelihood of adoption, the difference was not statistically significant. On the other hand, the hypothesis that high levels of contribution, as a second indicator of level of involvement with Wikipedia, would predict adoption was strongly supported. Because adopters and nonadopters were matched on overall preadoption activity levels[11], and the variable measuring contribution in the analysis focused on contributions within 28 days prior to adoption, the result suggests that the intensity of involvement right before adoption is a more important predictor of adoption than overall activity.

Consistent with our hypotheses, the network factors of interpersonal exposure, cohesion through interpersonal ties and tie homophily were all found to be significant predicators of higher likelihood of adoption in both the whole sample and the subsample. In combination, the results suggest that an important tool for increasing the likelihood of innovations spreading through a community is building stronger ties between members, giving members more opportunities to receive direct exposures to an innovation (Hypothesis 3) via cohesive ties (Hypothesis 4) among contributors of similar characteristics (Hypothesis 5). The analysis of predicted probabilities suggested that the two strongest factors affecting adoption were involvement (as measured by contribution) and interpersonal exposure. Both of these findings present important insights for the research literature. Highly involved community members were substantially more likely to adopt the innovation, suggesting the need for greater attention to measuring and theorizing what it is about involvement that makes people more likely to adopt innovations. Our strong evidence of the importance of interpersonal exposure is consistent with much prior research (Coleman, Katz, & Menzel, 1957, Valente, 1996, Wejnert 2002) and indicates a heightened need to specify exactly what it is about the additional personal ties that increases likelihood of adoption.

Overall, our hypotheses focused on the importance of cohesion for diffusion. Burt (1987, 1999), on the other hand, provided some strong arguments about the importance of structural equivalence for diffusion in a variety of offline contexts. Because both online ego networks and complete social networks are likely to have fuzzy, ephemeral boundaries, evaluation of structural equivalence among a random sample of 6,570 contributors among 8 million may be difficult. As Scott pointed out (1991/2004), structural equivalence mainly focuses on comparisons among different nodes within the same finite network. When the network is huge and its boundary is unclear, the influence of structural equivalence on diffusion is murkier. We do not want to rule out completely the importance of structural equivalence for diffusion in an online environment. On the other hand, it may be that people find it harder to identify their structurally equivalent counterparts online unless they are connected to their counterparts via cohesive ties at the same time. As a result, it may be hard to show that the drive for competition would motivate adoption among structurally equivalent actors online, as Burt has described. Even if Burt's underlying idea is sound when studying networks of finite boundaries, a more approximate measure of positional and role similarity that is not as overprecise as the current structural equivalence measure (Welser, Gleave, Fisher, & Smith, 2007) may be needed to examine how structural similarity or holders of similar structural roles shape diffusion in these online communities.

Finally, we hypothesized that admin contributors would function as both opinion leaders and information brokers (Burt, 1999) to help spread the innovation through the community. Counter to our prediction, ties to admin contributors did not increase likelihood of adoption. Since admin status did not predict adoption (Hypothesis 1), this finding became less surprising. The lack of wide support from across the high-status members of the community may explain the relatively low level of adoption of SuggestBot, despite the observed significant increase in postadoption contribution by adopters (Hypothesis 7). That is, without opinion leaders leading the way, SuggestBot did not diffuse widely throughout the community, which adds support to earlier studies' findings about the importance of opinion leaders in diffusion processes (Earp et al., 2002; Rogers, 2003).

Empirically, our study contributes to a growing trend in social science research to use large records of people's activities online to study human behavior (Crandall, Cosley, Huttenlocher, Kleinberg, & Suri, 2008; T. Turner, Smith, Fisher, & Welser, 2005). With the use of these large behavioral datasets, however, come methodological challenges. Behavioral data is often nonnormally distributed, affecting both activity measurement and sampling strategies (Welser et al., 2008). Naïve sampling methods may over- or underrepresent the null case in logistic regression models, as King & Zeng (2001) point out. We point out here that naïve random sampling may also lead to inappropriate comparison groups, without careful consideration of the variables on which the comparison should be matched.

Wikipedia has grown exponentially for years, and activity is roughly exponentially distributed; our sampling strategies needed to take this into account. Our choice of 28-day windows was driven by the need to walk a fine line between failing to detect individual-level change (longer periods) and being driven by the natural, intermittent nature of activity in online contexts, where external factors such as a deadline might curtail a person's activity in the community. Our windows are also participant-specific, centered around events of particular interest, which is potentially more informative than standard strategies that use calendar time to divide time series data into periods. Working with computer programmers and large datasets allowed us to do complex, nuanced sampling, and this is likely to become more common and more important in social science research going forward.

Limitations

One major limitation of the current research is that we used only archival data to study Wikipedia contributors' online activities. While the archival data provided very precise, objective measures of a number of key variables related to diffusion, network relationships, social interactions, etc., using archival data alone has some limitations. For instance, archival data provides behavioral indicators around contribution, but no psychological measures of what motivates contribution to the Wikipedia community. In addition, because we relied on Wikipedia archival data exclusively, we did not have information about the frequency or reciprocity of communication outside of Wikipedia, either in other online channels or in offline contexts. Wikipedians, especially admins and other committed members, do have offline meetings and dedicated communication channels outside of Wikipedia; how much these affect their behavior is impossible to tell from the available archives[12]. Richer data, including survey and interview responses along with information from other communication media, can reveal much more information about the differences between active and not-so-active contributors in their interests, motivations, experiences with the Wikipedia interface, emotional involvement with the community, and so on. Using multiple methods can both give a richer picture of what is happening and perhaps, as in Crandall et al. (2008), directly help with quantitative modeling.

It is also fair to ask about the generality of our results, as they are based on observations of a single community. Most prior studies of diffusion of innovation examine entirely new behaviors and technologies, such as adopting contraceptive methods, or technologies that compete with existing ones, as in the studies around hybrid corn adoption. SuggestBot was carefully designed to tie closely to the existing goals and practices of the Wikipedia community in order to be acceptable to the community and reduce the cost of adoption. Further, SuggestBot diffused naturally through existing social ties in the community by posting its suggestions on user talk pages. These choices might have reduced the role opinion leaders played in adoption compared to most other settings, and increased the effect of simply having interpersonal ties. We believe that these are general results, and that innovations that cost less to adopt and flow through social networks in natural ways will find that network ties matter more and opinion leaders matter less.

Direction for Future Research

This paper has only begun to reveal the potential of studying diffusion processes using log files of people's behaviors online. We see several directions for further research that advances our understanding of diffusion. First, as Rogers (2003) pointed out, while it has been widely studied what factors influence one-time adoption, little work has examined what factors will influence continued usage—or discontinuance—of an innovation (p. 110). Rogers and Shoemaker (1971) first observed this proinnovation bias in diffusion of innovation research around 4 decades ago. Still not much has been done to tackle the issue (Rogers, 2003). As a result, we know much more about adoption and use than about continued use and discontinuance (Rogers, 2003, p. 111). Yet continued usage of innovation is important because it is a more powerful indicator of success. We believe that the difficulties in collecting empirical data contribute to the focus on one-time adoption. In the current research, we were able to collect data on whether adopters of SuggestBot used the tool repeatedly to locate entries that need their work. While adopters tended to contribute more to Wikipedia postadoption (as shown in the independent sample t-test reported earlier), the diffusion of SuggestBot was not as successful as we have anticipated in either the scope or the depth of adoption. SuggestBot adopters account for well under 1% of the whole population of registered Wikipedia users. Further, relatively few adopters, about 15%, use SuggestBot repeatedly. This number has grown since SuggestBot offered a subscription option that makes it easier to receive suggestions on a repeated basis, but it is still small. It would be interesting to conduct a follow up study to find out why nonadopters fail to use SuggestBot. It would also be interesting to survey or interview those adopters who stopped using the tool. Insights from these studies might lead to a better design for SuggestBot; more generally, they may point to factors that could inform future diffusion research. For example, our prior experience developing innovations leads us to believe continued use heavily depends on initial experiences with an innovation, but empirically showing how important this is compared to attribute and relational factors could inform both design and research around innovation.

Second, in posthoc exploratory analysis, we found that structural properties of social network, including interpersonal cohesion and ties to opinion leaders did not make a huge difference in the level of postadoption contributions to Wikipedia. One primary reason is that the primary goal of SuggestBot is to motivate contributions from each individual editor. While new network ties may develop among editors who receive recommendations to edit similar articles, the tool did not have a component designated for fostering the development of social ties among contributors. In future research, it would be interesting to explore whether the addition of such a component would help foster a greater sense of belonging to a community, and thereafter motivate more contributions to Wikipedia. It will also be interesting to compare whether a tool with a networking component diffuses differently from a tool that focuses exclusively on facilitating individual contributions. Again, using sites with an explicit social networking component to explore this question would be interesting.

Practical Implications

While the adoption of SuggestBot in Wikipedia is not as widespread as we would have liked, the results showed that effective implementation of intelligent task routing systems can significantly increase contribution to online communities. The research shows the exciting promise of using technology as an intervention to boost online community involvement. Articles SuggestBot suggests are edited about four times as often as randomly chosen articles; it has received dozens of positive comments and several awards; and other wikis are interested in SuggestBot. Intelligent task routing also works in other communities such as MovieLens (http://www.movielens.org/) for public goods such as databases of movie information. The combination of its success in multiple communities, the theories of motivation and collective goods that underlie the idea, and the fact that simple recommendation strategies are effective all suggest that intelligent task routing is a valuable, general idea that could help make many online communities better.

This work also points to important design considerations for increasing the likelihood that an idea or innovation will propagate through a community. The fact that people were exposed to SuggestBot by seeing it on other users' talk pages was a fortunate side effect of the way it presented its suggestions, rather than a planned strategy for increasing its diffusion. Designing the interface and appearance of an innovation so that it naturally taps into the power of network variables for supporting diffusion is an important strategy, especially in the rapidly growing world of online social systems, from Wikipedia to blogs to Facebook. Badly done, this can hurt innovation—for instance, Facebook applications sometimes encourage people to send unsolicited invitations to everyone in their social network. Most of these ties are low cohesion and have little homophily of interests, so in light of our findings it is not surprising that this practice often fails to spread the application (and in fact often creates a backlash against it). Our work suggests that using only strong links and links between people who are fairly similar may be a more effective strategy for effectively diffusing an innovation.

For new or infrequent contributors, other strategies might support diffusion. One idea is to explicitly create ties. In Wikipedia, a group called the Welcome Committee seeks out new contributors and makes an initial post to their user talk pages in an effort to encourage them to contribute more; this idea is supported by studies showing that newcomers to discussion groups are more likely to return if their first post receives a reply (Joyce & Kraut, 2006). In principle, SuggestBot could do something similar, helping new members find both articles to work on and people to talk to, in an effort to both build ties and directly spread itself.

Conclusion

Overall, this work confirms, extends, and informs both intuitions about influence in social networks and studies of actual diffusion in networks. It demonstrates important factors in the diffusion of SuggestBot through the Wikipedia community and suggests that such tools will work to increase contributions to these communities. It also points to important factors in diffusion that may signal potential design opportunities for future innovators. And most generally, it shows the promise of carrying out diffusion studies in large-scale online interaction log data while pointing out important methodological issues. Effective use of this kind of data will open new ways for understanding diffusion processes.

Acknowledgement

The paper is supported by two NSF grants studying networks and social dynamics (SES-0433086) and cyber infrastructure (BCS-0537606).

Notes

1 Scholars disagree in their definitions about what constitute online communities. Following Preece and Maloney-Krichmar (2005), we think it is more important to ‘‘concentrate on more substantive issues such as how communities are created, evolve or cease to exist online’’ (p. 2) than to clean the ‘‘fuzzy boundaries’’ of a concept ‘‘that is more appropriately defined by membership’’ (p. 2).

2 It is likely this editor used a tool, in Wikipedia parlance called a ‘‘bot,’’ that helps editors make a large batch of related edits (such as correcting a particular misspelling in a number of articles).

3 SuggestBot uses several technologies for recommending articles, similar to technologies that help companies like Amazon recommend books and music. The related algorithms and other technical details of the tool are described in detail in Cosley et al. (2007).

4 A data file was generated partially from the public record of who has talked to whom, who has edited which articles, etc. that all Wikipedia editors can see, and partially from the log file of SuggestBot usage activities.

5 Statistics are based on http://en.wikipedia.org/wiki/Special:Statistics.

6 Programs such as SuggestBot can also edit Wikipedia, and in fact a program called "SmackBot" made almost 500,000 edits; the 162,138 edits came from an editor named Rich Farmbrough.

7 All such windows entail trade-offs between windows that are too long (thus masking long-term individual-level change) and windows that are too short (and thus are more strongly affected by the intermittent nature of participation).

8 While a greater magnitude in odds ratio for the same variable but across different models indicates a stronger effect on the probability of the outcome, the amount of that change cannot be directly inferred (See Menard, 2002, pp. 48–57).

9 We defined "high" as the average of the third quartile and the maximum value. The specific value is not critical for the "high" condition. Choosing other values between the 3rd quartile and the maximum does not substantially change the interpretation of the reported probabilities.

10 Note that these probabilities are not inherently meaningful, because the intercept in case control studies is determined by the sample structure (Stolley & Schlesselman, 1982). Instead our focus is on comparing the relative effects of the key variables on the likelihood of adoption. Therefore we use the Sample Average model as a baseline to evaluate the relative magnitude of change in the probability as we set each variable to a high level.

11 The independent sample *t*-test conducted to test Hypothesis 7 confirmed the success of matching because adopters and nonadopters were not statistically different from each other in their overall levels of contribution prior to adoption.

12 In some ways, this parallels the problem of incomplete sampling mentioned earlier. We do not claim that using online traces of data is a panacea; just that it provides new opportunities (and challenges).

References

Abrahamson, E., & Rosenkopf, L. (1997). Social network effects on the extent of innovation diffusion: A computer simulation. *Organization Science*, **8**(3), 289–309.

Burke, P., & Reitzes, D. (1991). An identify theory approach to commitment. *Social Psychology Quarterly*, **54**(3), 239–251.

Burt, R. S. (1987). Social contagion and innovation: Cohesion versus structural equivalence. *American Journal of Sociology*, **92**(6), 1287–1335.

Burt, R. S. (1999). The social capital of opinion leaders. *The Annals of the American Academy of Political and Social Science*, **566**(1), 37–54.

Byrne, D. E. (1971). *The attraction paradigm*. New York, NY: Academic Press.

Coleman, J. (1988). Social capital in the creation of human-capital. *American Journal of Sociology*, **94**, S95–S120.

Coleman, J., Katz, E., & Menzel, H. (1966). *Medical innovation: A diffusion study*. New York, NY: Bobbs Merrill.

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2006). Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries & G. Olson (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1037–1046). Montréal, Québec, Canada: ACM New York, NY, USA.

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2007). *SuggestBot: using intelligent task routing to help people find work in Wikipedia*. In *Proceedings of International Conference on Intelligent User Interfaces* (pp. 32–41). New York, NY: ACM.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceeding of the 14th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (pp. 160–168). New York, NY: ACM.

Cummings, J. N., Butler, B., & Kraut, R. (2002). The quality of online social relationships. *Communications of the ACM*, **45**(7), 103–108.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, **48**(2), 147–160.

Earp, J. A., Eng, E., O'Malley, M. S., Altpeter, M., Rauscher, G., Mayne, L., et al. (2002). Increasing use of mammography among older, rural African American women: Results from a community trial. *American Journal of Public Health*, **92**(4), 646–654.

Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage Publications.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Newbury Park, CA: Sage publications.

Garson, G. D. (2008). Logistic regression, from S*tatnotes: Topics in multivariate analysis.* Retrieved on 03/31/09 from http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm

Hechter, M. (1987). *Principles of group solidarity*. Berkely, CA: University of California Press.

Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative Science Quarterly*, **37**(3), 422–447.

Joyce, E., & Kraut, R. E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, **11**(3), 723–747.

Kanter, R. M. (1972). *Commitment and community: Communes and utopias in sociological perspective*. Cambridge, MA: Harvard University Press.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, **65**(4), 681–706.

King, G., & Zeng, L. (2001). Logistic Regression in rare event data. *Political Analysis*, **9**(2), 137–163.

Kraut, R. E. (2003). Applying social psychological theory to the problems of group work. In J. Carroll (Ed.), *HCI models, theories and frameworks: Toward a multidisciplinary science* (pp. 325–356). New York: Morgan Kaufman.

Lakhani, K. R., & Hippel, E. V. (2003). How open source software works: "Free" user to user assistance. *Research Policy (Special Issue)*, **32**(6), 923–943.

Ling, K., Beenen, G., Ludford, P., Wang, X. Q., Chang, K., Li, X., et al. (2005). Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, **10**(4). Retrieved on 11/30/2008 from http://jcmc.indiana.edu/vol10/issue4/ling.html.

Ludford, P. J., Cosley, D., Frankowski, D., & Terveen, L. (2004). Think different: Increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 631–638). New York, NY: ACM.

Marwell, G., & Oliver, P. (1993). *The critical mass in collective action: A micro-social theory*. New York, NY: Cambridge University Press.

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

Menard, S. (2002). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Mockus, A., Fielding, R. T., & Andersen, H. (2002). Two case studies of open source software development: Apache and Mozilla. *ACM Transactions on Software Engineering and Methodology*, **11**(3), 309–346.

Monge, P. R., & Contractor, N. (2003). *Theories of communication networks*. New York, NY: Oxford University Press.

Oliver, P., & Marwell, G. (2001). Whatever happened to critical mass theory? A retrospective and assessment. *Sociological Theory*, **19**(3), 292–311.

Olson, M. J. (1965). *The logic of collective action*. Cambridge, MA: Harvard University Press.

Osborne, J. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, **8**(6). Retrieved on 11/30/2008 from http://PAREonline.net/getvn.asp?v=8&n=6

Preece, J., & Maloney-Krichmar. (2005). Online communities: Design, theory and practice. *Journal of Computer-Mediated Communication*, **10**(4). Retrieved on 11/30/2008 from http://www3.interscience. wiley.com/cgi-bin/fulltext/120837966/HTMLSTART.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, **9**(4), 705–724.

Rogers, E. M. (2003). *Diffusion of innovation* (5th ed.). New York, NY: Free Press.

Rogers, E. M., & Kincaid, D. L. (1981). *Communication networks: Toward a new paradigm for research*. New York, NY: Free Press.

Rogers, E. M., & Shoemaker, F. E. (1971). *Communication of innovations: A cross-cultural approach*. New York: Free Press.

Scott, J. (1991/2004). *Social network analysis*. Thousand Oaks, CA: Sage.

Stolley, P. D., & Schlesselman, J. J. (1982). *Case-control studies: Design, conduct, analysis.* Oxford: Oxford University Press.

Strang, D., & Soule, S. A. (1998). Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, **24**, 265–290.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York, NY: HarperCollins Publisher

Turner, J. C. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford, UK: Basil Blackwell.

Turner, T., Smith, M. A., Fisher, D., & Welser, H. T. (2005). Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, **10**(4). Retrieved on 11/30/2008 from http://www3.interscience.wiley.com/cgi-bin/fulltext/120837981/HTMLSTART.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. *Social Networks*, **18**(1), 69–89.

Wejnert, B. (2002). Integrating models of diffusion of innovations: A conceptual framework. *Annual review of sociology*, **28**, 297–326.

Welser, H. T., Gleave, E., Fisher, D., & Smith, M. A. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, **8**(2). Retrieved on 11/30/2008 from http://www.cmu.edu/joss/content/articles/volume8/Welser/.

Welser, H. T., Smith, M. A., Gleave, E., & Fisher, D. (2008). Distilling digital traces: Computational social science approaches to studying the Internet. In N. Fielding, R. L. Lee & G. Grant (Eds.), *Handbook of online research methods* (pp. 116–140). London: Sage Publications.

Yuan, Y. C., Fulk, J., Shumate, M., Monge, P., Bryant, J. A., & Matsaganis, M. (2005). Individual participation in organizational information commons: The impact of team level social influence and technology-specific competence. *Human Communication Research*, **31**(2), 212–240.

Yuan, Y. C., & Geri, G. (2006). Homophily of network ties, and bonding and bridging social capital in distributed teams. *Journal of Computer-Mediated Communication*, **11**(4), 1062–1084.

About the Authors

Y. Connie Yuan is an Assistant Professor in the Departments of Communication and Information Science at Cornell University. She received her Ph.D. from the University of Southern California. Dr. Yuan's research focuses on social networks, communication technology, online communities, knowledge management and computer-supported distributed work in organizations. She has recently received funding from the National Science Foundation (2008–2011) to study how the development of network relations and the usage of communication technology influence the transfer and retention of organizational knowledge via the development of transactive memory systems. She has received best papers or distinguished article awards from the annual conferences of the Academy of Management and the National Communication Association. Her work has been published in *Communication Research, Human Communication Research, the Information Society*, and *Journal of Computer-Mediated Communication*, among others. She is on the editorial board of *Journal of Applied Communication Research* and *Journal of Computer-Mediated Communication*.

Address: 308 Kennedy Hall, Department of Communication, Cornell University, Ithaca, NY 14853.

Dan Cosley is an assistant professor of information science at Cornell University. His primary interest is helping groups make sense, use, and reuse of information; he studies this by using social science theory, HCI design principles, and models of behavior to build and evaluate systems in real contexts. His work spans a broad range of problems, most recently using people's behavior around community goods such as Wikipedia to motivate them to contribute more, and re-using content people create in social media systems to support individual and social reminiscence. He is also interested in the more general problem of how to move from theory, principles, and models to actual design, and how those designs can then inform the principles that inspired the designs. He has a Ph.D. in computer science from the University of Minnesota and is the recent recipient of an NSF CAREER award.

Address: 301 College Ave., Department of Information Science, Cornell University, Ithaca, NY 14850.

Howard T. Welser is an Assistant Professor of Sociology at Ohio University. He received his Ph.D. in 2006 from the University of Washington. Dr. Welser's research investigates how microlevel processes generate collective outcomes, with application to status achievement in avocations, development of institutions and social roles, the emergence of cooperation, and network structure in computer mediated interaction. He recently received a grant from Microsoft Research to study emergent social roles in online question and answer systems. His work has been published in *Rationality and Society*, *Journal of Social Structure*, *Journal of Computer-Mediated Communication*, the proceedings of*International AAAI Conference on Weblogs and Social Media, Hawaii International Conference on System Science*, *World Wide Web Workshop*, as well as chapters in *e-Research: Transformation in Scholarly Practice*, and the *Sage Handbook of Online Research Methods*.

Address: Bentley Annex 109, Department of Sociology & Anthropology, Ohio University, Athens, Ohio 45701.

Ling Xia is a graduate student in the Department of Communication at Cornell University. She received her M.S. degree in Communication from Cornell University. She is interested in knowledge management and social network analysis, especially in the context of the adversarial network. She has received top paper award from the annual conferences of National Communication Association. Her work has been published in *Journal of the American Society for Information Science and Technology* and *Management Communication Quarterly*.

Address: 209 Kennedy Hall, Department of Communication, Cornell University, Ithaca, NY 14853.

Dr. Geri Gay is the Kenneth J. Bissett Professor and Chair of Communication at Cornell University and a Stephen H. Weiss Presidential Fellow. She is also a member of the Faculty of Computer and Information Science and the director of the Human Computer Interaction Lab at Cornell University. Her research focuses on social and technical issues in the design of interactive communication technologies. Specifically, she is interested in social navigation, affective computing, social networking, mobile computing, and design theory. Professor Gay has received funding for her research and design projects from NSF, NASA, the Mellon Foundation, Intel, Google, Microsoft, NIH, the Robert Wood Johnson Foundation, AT&T Foundation, and several private donors. She teaches courses in interactive multimedia design and research, computer-mediated communication, human-computer interaction, and the social design of communication systems. Recently, she has published in *IEEE*, *International Journal of Human-Computer Interaction*, *Journal of Computer-Mediated Communication*, *Journal of Communication*, *CHI*, *HICCS*, *ACM Digital Libraries*, *SIGIR*, *JASIST*, and *CSCW*.

Address: 325 Kennedy Hall, Department of Communication, Cornell University, Ithaca, NY 14853.