Feedback Effects between Similarity and Social Influence in Online Communities

David Crandall
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
crandall@cs.cornell.edu

Dan Cosley
Dept. of Communication
Cornell University
Ithaca, NY 14853
drc44@cornell.edu

Daniel Huttenlocher
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
dph@cs.cornell.edu

Jon Kleinberg
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
kleinber@cs.cornell.edu

Siddharth Suri
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
ssuri@cs.cornell.edu

ABSTRACT

A fundamental open question in the analysis of social networks is to understand the interplay between similarity and social ties. People are similar to their neighbors in a social network for two distinct reasons: first, they grow to resemble their current friends due to social influence; and second, they tend to form new links to others who are already like them, a process often termed *selection* by sociologists. While both factors are present in everyday social processes, they are in tension: social influence can push systems toward uniformity of behavior, while selection can lead to fragmentation. As such, it is important to understand the relative effects of these forces, and this has been a challenge due to the difficulty of isolating and quantifying them in real settings.

We develop techniques for identifying and modeling the interactions between social influence and selection, using data from online communities where both social interaction and changes in behavior over time can be measured. We find clear feedback effects between the two factors, with rising similarity between two individuals serving, in aggregate, as an indicator of future interaction — but with similarity then continuing to increase steadily, although at a slower rate, for long periods after initial interactions. We also consider the relative value of similarity and social influence in modeling future behavior. For instance, to predict the activities that an individual is likely to do next, is it more useful to know the current activities of their friends, or of the people most similar to them?

Categories and Subject Descriptors: H.2.8 Database Management: Database Applications – Data Mining

General Terms: Measurement, Theory

Keywords: social networks, online communities, social influence

1. INTRODUCTION

Social influence and selection. A fundamental property of social networks is that people tend to have attributes similar to those of their friends. There are two underlying reasons for this. First, the process of *social influence* [7] leads people to adopt behaviors exhibited by those they interact with; this effect is at work in many settings where new ideas diffuse by word-of-mouth or imitation through a network of people [19, 22]. A second, distinct reason is that people tend to form relationships with others who are already similar to them. This phenomenon, which is often termed *selection*, has a long history of study in sociology [13, 16].[1]

The two forces of social influence and selection are both seen in a wide range of social settings: people decide to adopt activities based on the activities of the people they are currently interacting with; and people simultaneously form new interactions as a result of their existing activities. In most settings, studying these forces and their interplay has been very difficult because collecting data about an individual's social network and activities over time is both expensive and error-prone. Online communities, however, provide an excellent opportunity to study large-scale social phenomena of this type. In our case, many of the online systems created in recent years offer both a rich set of activities and the opportunity for extensive interactions. These online systems

Supported in part by NSF grants CCF-0325453, CNS-0403340, BCS-0537606, and IIS-0705774, and by funding from Google, Yahoo!, and the John D. and Catherine T. MacArthur Foundation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.
Copyright 2008 ACM 978-1-60558-193-4/08/08 ...$5.00.

[1]The term *homophily* — the idea that "birds of a feather flock together" — is also used in the sociology literature to denote this process. However, since homophily is also used increasingly to denote the simple cumulative fact that neighbors in a social network are similar — regardless of the underlying process — we have adopted the term "selection" for the process itself.

often record both activities and interactions, thereby enabling large-scale studies of social influence, selection, and the interplay between the two.

Both social influence and selection lead to a common aggregate effect, namely that neighboring nodes in a social network tend to look similar to each other. But it is important to isolate the respective effects of these forces for several reasons. Social influence and selection produce homogeneity in ways that have very different structural consequences for the network: social influence can produce network-wide uniformity, as a new behavior spreads across the links, while selection tends to drive the network toward smaller clusters of like-minded individuals [10], a process sometimes called *balkanization* [24]. Moreover, because the two forces are based on different effects — interaction and similarity, respectively — the distinctions between them reflect analogous contrasts that arise in current computing applications that mine social network data. In particular, applications such as *viral marketing* [6] are rooted in the premise that a person's social contacts provide a valuable predictor of their future behavior, while *recommender systems* (e.g., [9, 18, 21]) build predictions based on the behavior and opinions of others who share similar behaviors and opinions. These relationships between viral marketing and recommender systems thus parallel, in important respects, the relationships between social influence and selection. Finally, from a social-science perspective, understanding how these phenomena operate is important in its own right [15].

Despite the basic nature of the questions here, it has been difficult to gain insight into effects of this kind in real social networks. Understanding the relationship between social interaction and similarity requires not just static analysis of social network structures [27], or even models of the evolution of these structures over time (e.g., [1, 11, 12]), but analysis of the detailed dynamics of *social processes*, as people's behavior and interaction patterns both shift over time.

The present work. In this paper we describe a framework for using data from large online communities to analyze the interactions between social influence and selection. We develop the framework using data from Wikipedia, which combines crucial ingredients needed for a study of this type: it contains the records of editors' activities and interactions over time in a setting where activities arise from a mix of intrinsic interest and social interaction. Within this setting, we focus on two central questions:

(i) Can we quantify the ways in which social influence and selection work together to affect people's interests and interactions? That is, can we characterize how social interactions affect interests and vice versa?

(ii) To what extent do similarities and social interactions between people serve as predictors of future behavior? That is, can we characterize the relative degree to which interests and social interactions affect what people do?

We now discuss our work on each of these questions in turn.

Interplay between social influence and selection. For the first of the above questions we both analyze online activities and interactions, and also build theoretical models of the underlying phenomena.

In Wikipedia, we focus on two types of recorded actions: editing articles, and editing the discussion page associated with a particular user. Edits to articles are indications of interests as each article is on a particular topic. We take the history of article edits for a user at a given time as a vector encoding that user's activities, and consider the time series of these vectors as representing the evolution of users' interests and activities as they edit different articles. Edits to discussion pages are indications of social ties between the users who are communicating.

In examining the relationship between selection and social influence, we consider pairs of Wikipedia editors who have communicated with one another and analyze how the similarity between their activity vectors evolves both before and after their first interaction. Aggregating over all such pairs, we find that there is a sharp increase in the similarity between two editors just before they first interact, with a continuing but slower increase that persists long after this first interaction (see Figure 1). This suggests that people encounter each other due to overlap in their interests as measured by article editing, but that the consequences of these encounters can lead to further effects that are visible many months later.

We postulate that the observed results are due to feedback between social influence and selection, where similarity leads to interaction but then interaction leads to further similarity. We investigate the degree to which such a feedback loop can be characterized by a simple probabilistic model. We are motivated by an interesting random-graph model recently proposed by Holme and Newman [10] in which each node has a single categorical attribute (an "opinion" in their terminology), and in each time-step a node either changes its opinion to match a neighbor's, or re-wires one of its links to connect to someone of the same opinion. In a very simple way, then, their model captures the dual processes of social influence (via opinion changes) and selection (via re-wiring of connections).

We find, however, that the Holme-Newman model is too simple to produce the effects we see in Wikipedia. Consequently, we propose a more expressive model in which there is a large space of possible *activities*. Nodes alternately either engage in an activity by sampling from this space, or interact with another node with whom they share an activity. Thus, a node's behavior is not just represented by the value of a single opinion, but by a vector of activities that have been performed, which allows for more subtle notions of similarity that drive interaction. Our model can be viewed as extending the well-studied concept of an *urn process* from discrete probability [17] to a network setting, and suggests a number of interesting open questions in the analysis of random graphs.

Predictive value of social interactions and similarity. To address the second central question of this paper, regarding the use of social interactions and personal interests to predict future behavior, we use data both from Wikipedia and LiveJournal, a site that combines blogging and social-networking features. We extend a methodology explored recently in several papers on social influence in online datasets [1, 11, 14]; these papers investigate how social ties affect behavior by studying the extent to which various properties of a social network predict future activities or behaviors. For instance, they study the probability of engaging in a new

behavior as a function of the number of neighbors in the social network who have already adopted the behavior.

We argue here that this approach can be extended to a framework in which the probabilistic effects of social interaction and similarity can be compared on a common footing. In particular, we compare the predictive power of networks derived from social interaction to the predictive power of *similarity networks* that are constructed by connecting people to other individuals with the most similar interests, regardless of whether they have interacted socially or not. For LiveJournal we predict the behavior of joining a group using a social network formed based on previously existing friendship links and a similarity network based on which groups one belongs to. In Wikipedia, we predict the behavior of editing an article for the first time using a social network formed based on previous interactions with other editors and a similarity network based on the articles that one has previously edited.

We find that in Wikipedia, properties of the social network are better predictors of future behavior than are properties of the similarity network, whereas the opposite is the case in LiveJournal. At first this result seems counter-intuitive, as Wikipedia is focused on content creation rather than social interaction, whereas LiveJournal strongly emphasizes social-networking features. However a more detailed consideration of the underlying communities helps explain the difference: the social interactions on Wikipedia are an integral part of the article creation process, whereas on LiveJournal friendship links are more a statement of who one knows, and often less reliably a source of direct interaction. We also find that combining features based on social ties and similarity is more predictive of future behavior than either social influence or similarity features alone, showing that both social influence and one's own interests are drivers of future behavior and that they operate in relatively independent ways.

Outline of the paper. We next consider each of our two main questions in detail, following the general outline above. We conclude by discussing how our contributions might influence research and practice in recommender systems, online communities, and data mining using datasets derived from these communities.

2. SOCIAL INFLUENCE AND SELECTION OVER TIME

To measure how selection and social influence operate, we track a time-evolving vector representing each person's activities, and we study how the vectors of two people are changing directly around the moment when they first interact. Thus, we consider systems in which there is a set of m possible *activities*, and each person v at time t has an m-dimensional vector $\vec{v}(t)$, where the i^{th} coordinate $\vec{v}(t)_i$ represents the extent to which person v is engaging in activity i. The vectors may be binary, with $\vec{v}(t)_i$ equal to 0 or 1 depending on whether v has ever engaged in activity i; or they may be weighted vectors, with $\vec{v}(t)$ derived in some way from the number of times v has engaged in i.

We will use standard vector-similarity definitions to track the changing similarities between people's activity vectors. The information retrieval and data mining literatures provide a number of such standard measures (e.g. [20]). We use one of the more common measures, the cosine metric,

$$\text{Cosine}(\vec{u},\vec{v}) = \cos \vec{u} - \vec{v} = \frac{\vec{u}\cdot\vec{v}}{||\vec{u}||_2||\vec{v}||_2}, \quad (1)$$

where $||\vec{v}||_2$ denotes the Euclidean norm of v.

While a comparison of similarity measures is not the focus of our current work, we have evaluated a wide range of measures for our purpose. We use the cosine metric here because it is independent of the rate at which people are editing, which increases over time. Another common metric, the weighted Jaccard coefficient, performs better as a predictor of future behavior and we use it in Section 3. Here, however, it is difficult to use because the baseline Jaccard similarity between pairs of users rises over time, making it harder to interpret the behavior of the other similarity curves.

2.1 Analyzing Similarity over Time

The Wikipedia dataset. Wikipedia is a large task-focused community whose goal is to produce a free online encyclopedia. The entire edit history of Wikipedia is freely available for analysis, making it attractive for research (see e.g. [5, 25, 23, 26]). Wikipedia has a rich social structure in which a large number of users interact during the course of producing articles. To facilitate social interaction, Wikipedia allows free-form discussion pages to be created for each article and Wikipedia user. As with the articles themselves, anyone can edit these discussion pages. In our study we consider users of English Wikipedia who have created an account and have a user discussion page; there were approximately 510,000 such users as of April 2, 2007. These users were responsible for 61% of edits to the roughly 3.4 million articles. We ignore anonymous edits, as these are recorded only by IP address and may combine the activities of many people, and we ignore actions by users without discussion pages, who tend to have very few social connections.

Following the approach outlined above, we define a user's activity vector $\vec{v}(t)$ to specify the number of times that he or she has edited each article up to that point in time. We define the time of first meeting for two users u and v as the time at which one of them first makes a post on the user discussion page of the other. In principle, we could also try to infer social interactions based on posting to the same article's discussion page. However, it is often unclear who is interacting with whom on these pages without analyzing the content of the discussion, so we restrict our focus to direct user-to-user connections through user discussion pages. Moreover, we found that using simple heuristics to infer interaction based on posts to article discussion pages produced closely analogous results to what we obtain from analyzing user discussion pages.

Similarity around the time of first interaction. With this framework, we are ready to ask a basic question that directly addresses social influence and selection effects: how does the similarity between two people vary in the time window around their first interaction with each other? An elevated level of similarity just before meeting indicates a type of selection at work, while increasing similarity following this meeting provides evidence for social influence.

Figure 1 shows results for this analysis over the entire edit history of Wikipedia through April 1, 2007. For all pairs of people who have ever interacted during this period, the plot

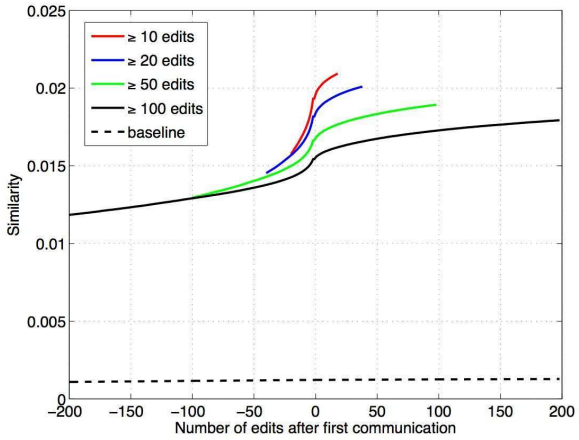


Figure 1: Average cosine similarity of user pairs as a function of the number of edits from time of first interaction, for Wikipedia.

averages the the cosine similarity of their vectors as a function of time, where time t on the x-axis is relative to the moment of first interaction for each pair. (That is, all time-series are shifted so that 0 is the time of first interaction for each pair.) Separate plots are shown for pairs of users with different activity levels; in particular we consider pairs of users who have each performed at least k edits before and k edits after the first interaction, for several values of the threshold k. To prevent mixing of populations with different activity volumes, we only show $2k$ edits on each side of 0 for the group associated with threshold k. The figure also shows a baseline representing the (much lower and essentially constant) average similarity for pairs of users who have not interacted, aligned at an arbitrary moment (in this case, midnight on January 1, 2006) instead of the time of first interaction.

The most prominent feature of the plots is the sharp increase in similarity immediately before the first interaction. There is also a continuing but slower increase that persists long after this first interaction. Thus, there appear to be strong selection effects, in which the steepest increase in similarity is taking place just before two people interact. However, there is also an ongoing effect of social influence, with similarity continuing to increase faster than the baseline for a long period of time.

It seems likely, then, that explicit coordination between people is not the main force driving increasing similarity, even in a setting such as Wikipedia where extensive coordination takes place: similarity is increasing most sharply *before* people first meet, and there are also long-term effects (rather than just a short term "bump") after people first meet. Differing activity levels do not change the results. Although less-active users become similar more quickly, the qualitative picture is the same whether users are less active ($k \geq 10$) or more active ($k \geq 100$). We have repeated this analysis while excluding Wikipedia administrators, whose behavior is often driven by administrative tasks like fixing vandalism. There, too, the results were all qualitatively similar to the plots in Figure 1. In short, the effects of selection and social influence are fairly robust in the Wikipedia data.

2.2 Modeling the Effect of Social Interaction

The phenomenon illustrated in Figure 1 has potential analogues in many domains. In essentially any type of social-media setting, people pursue activities and encounter others through these activities; based on these encounters, their pattern of activities may shift further due to the resulting social interaction. This captures examples such as editing an article on Wikipedia, contributing to a group discussion on a site such as Facebook, commenting on a story in a news-sharing site, and in a range of other settings. It also plays out clearly in the off-line world, through the activities that people engage in and the people they meet.

A natural question is whether the effect we see in the figure — rapidly rising similarity before first meeting, and slower but steady increase long afterward — can be explained by a model of individual behavior and interaction in a social network. Here we will develop such a model, which seeks in a minimal way to incorporate the forces that drive activities and interaction in an online community. When the parameters of the model are derived from Wikipedia data, we find that the model produces simulated behaviors for which the resulting curves are in striking agreement with the plot of Figure 1.

Activities and interactions. It is useful to fix a bit of terminology first. An *activity* is something that can be performed multiple times; each time it is performed is called an *instance* of the activity. A user's *activity history* $E(u)$ is simply the sequence of all its instances of past activities. For example, consider a user u who has performed five distinct edits to Wikipedia articles: three of these edits were to article A and two were to article B. Then u's activity history $E(u)$ has five entries, consisting of three instances of activity A and two instances of activity B.

Communication with others can be described in an analogous way. Each time a user u communicates with a user v, this can be viewed as an instance of communication, which we will call an *interaction* with v. (Interactions can be done many times with the same person, just as an activity can be performed many times.) User u's *interaction history* $N(u)$ is the sequence of all their past interactions.

Modeling the choice of activities. Now, we model a user's choice of activities through the following ingredients. First, people's future behavior in any of these settings is strongly correlated with their past behavior, and so one possibility is that a user chooses what to do next by sampling from his or her own past activity history. Second, the effects of social influence can be captured by assuming that a user may also choose a next activity by sampling from the past history of his or her friends in the social network.

We represent such sampling from the past using what is known in discrete probability as an *urn process* [17]. We model a user as selecting one instance uniformly at random from their activity history, and then performing another instance of the same activity. In this way, the user is more likely to engage in an activity that he or she has done more often in the past.[2] There is thus a close connection between

[2]The term "urn process" comes from the standard picture of this process in probability, due to Pólya. We view the activity history as an urn with differently colored balls; in each step, we draw a ball from the urn and increase the

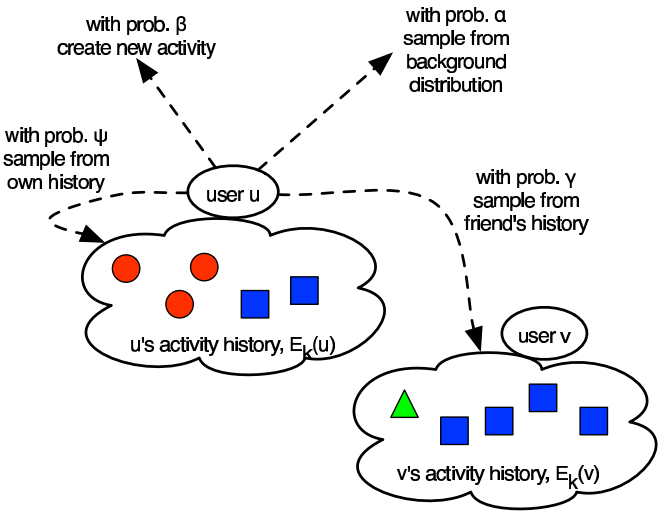


Figure 2: A schematic of the model for selecting activities.

urn processes and "rich-get-richer" or *preferential attachment* processes that produce the kinds of heavy-tailed activity distributions one sees in online behavior [2], and this has been an active area of study. In our case, when a user can sample from a neighbor's activities as well as its own, we are dealing with a novel kind of *networked urn process*, where the sampling is correlated across edges of a graph. Our results suggest that networked urn processes can be a powerful way to capture the evolution of behavior in a social network.

However, these two ingredients — sampling from one's own history or sampling from a neighbor's history — can't by themselves model a user doing an activity for the first time. As a result, we also allow users to choose next activities by sampling from the concatenation of all users' activity histories (essentially, a changing background frequency distribution on activities); or by generating a new activity that has never been done by anyone, and doing this.

Finally, there is one further aspect that is useful to capture: the idea that users are more likely to repeat recent activities than long-ago ones. Thus, for a constant k, we define $E_k(u)$ to be the activity history of u truncated to contain only the k most recent instances. We then model the sampling as taking place from $E_k(u)$, for a parameter k, rather than the full history $E(u)$.

Thus there are four possible ways in which a user selects an activity to add to their history $E(u)$, which is depicted schematically in Figure 2:

- *Sample from one's own history.* With probability ψ, choose a random instance from $E_k(u)$ and perform a new instance of this activity.
- *Sample from a neighbor's history.* With probability γ, choose a random interaction from $N(u)$; suppose this interaction was with user v. Choose a random instance from $E_k(v)$ and perform a new instance of this activity.

contents of the urn by adding another ball of the same color.

- *Sample from the world's history.* Let $\mathcal{E}$ be the concatenation of all users' entire activity histories. With probability α, choose a random instance from $\mathcal{E}$, and perform a new instance of this activity.
- *Start a new activity.* With probability β, create a new activity that has never been done before, and perform one instance of this activity.

There are four parameters to this part of the model: three of the four probabilities ψ, γ, α, and β (which sum to 1), plus history length k.

Modeling the choice of interactions. We also need to model how the interactions happen, and urn processes can be naturally used here as well. Specifically, we choose a new interaction for u as follows, based on an additional parameter δ.

- *Talk to a random person.* Let $\mathcal{U}$ be the sequence of all user activity instances thus far (i.e., containing each user in proportion to the number of actions they've taken). With probability δ, choose a user v at random from $\mathcal{U}$ and perform an interaction with v. (There clearly needs to be an initialization for $\mathcal{U}$ which we describe below.)
- *Talk to someone based on a common activity.* With probability $1-\delta$, choose a random instance of an activity a from $E_k(u)$ which we denote a. Randomly choose a person from the most recent k people to have performed an instance of this activity and perform a new interaction with them.

The full model. The full probabilistic process proceeds in discrete steps. At any time-step, a user u is chosen uniformly from $\mathcal{U}$. With probability ϕ, user u selects an activity using the procedure above; with probability $1-\phi$, u performs an interaction using the procedure above.

As noted in the introduction, our approach is motivated by a simpler model proposed by Holme and Newman [10], in which each user holds a single mutable opinion. In each time step of their model, a user u is chosen. With probability ϕ, u changes its opinion to match a random neighbor's, and with probability $1-\phi$, u re-wires one of its links to connect to a random user who holds the same opinion. We have tried a number of direct adaptations of the Holme-Newman model to settings in which users have access to a range of activities, rather than just holding a single opinion; we find, however, that none of these are capable of matching the qualitative shapes of the effects that we see in Figure 1.

As a result, we were led to the present model, which explicitly maintains each user's distribution over past activities. We will see that sampling based on urn processes, which is known to produce the kinds of heavy-tailed distributions one sees in practice [2], is also effective at producing the kinds of selection and social influence effects that arise in real data. It also provides a rich opportunity for extending classical urn models to the setting of an arbitrary network.

We now turn to a procedure by which we can learn parameters for our model from Wikipedia data. Following this, we will see that a version of the curve from Figure 1, simulated using our model, matches the qualitative features of the real curve with surprising fidelity.

Estimating model parameters from data. Of the model parameters $\theta = (\phi, \psi, \beta, \gamma, \alpha, \delta, k)$, two can be estimated by direct observation of the Wikipedia data: ϕ is the ratio of user discussion page edits to total edits, and β is the proportion of edits that create new articles. We use maximum-likelihood estimation to set the values of the unobservable parameters ψ, α, γ, and δ (of which only three are independent).

Given the model parameters θ it is straightforward to estimate the likelihood of the actual Wikipedia data W given the model, $P(W|\theta)$. This probability is the product of the probability of each observed event in W given the state of the model $S = (\mathcal{E}, \mathcal{U}, E_k(u), \bar{E}_k(a), N(u))$, where recall $\mathcal{E}$ is the concatenation of all users' activity instances, $\mathcal{U}$ is all users in proportion to the number of activities they've undertaken, $E_k(u)$ is the k most recent instances of activities for each user u, $N(u)$ is u's past interactions, and $\bar{E}_k(a)$ are the k most recent users to have performed activity a. We will let $P(U = u)$ denote the probability of sampling a given object u from the sequence U.

There are two types of observed events in W. First, we consider the probability of an event in which user u chooses to interact with user v, which we denote $C_{u,v}$, is given by,

$$P(C_{u,v}|\theta, S) = \phi\left[\delta P(\mathcal{U} = v) + (1 - \delta) f(u, v|S)\right],$$

where $f(u, v|S)$ is the probability that u chooses v by first sampling an activity from u's history and then sampling v from the users engaged in that activity, which we obtain by marginalizing over activities a,

$$f(u, v|S) = \sum_{a \in \mathcal{E}} P(E_k(u) = a) P(\bar{E}_k(a) = v).$$

Second, we consider the probability of u doing an activity a, denoted $D_{u,a}$, given the model parameters and current state is,

$$P(D_{u,a}|\theta, S) = (1 - \phi) \quad [\psi P(E_k(u) = a) + \gamma g(u, a|S) + \alpha P(\mathcal{E} = a) + \beta h(a|S)]$$

where $g(u, a|S)$ denotes the probability of u picking some user v from $N(u)$ and then picking activity a from $E_k(v)$, marginalized over users v,

$$g(u, a|S) = \sum_{v \in \mathcal{U}} P(N(u) = v) P(E_k(v) = a),$$

and $h(a|S)$ is an indicator function that is 1 if and only if a is a new activity.

To compute the likelihood of the Wikipedia data given specific model parameters, $P(W|\theta)$, we first initialize the set of sequences using the first t article and user discussion page edits. We then process each remaining edit event in order of increasing timestamp. For a user discussion page edit in which u interacts with v, we compute $P(C_{u,v})$ using the equation above, and then update the the interaction histories by adding a u interaction to $N(v)$ and a v interaction to $N(u)$. For an article edit event in which u edits article a, we compute $P(D_{u,a})$ and then update the state of the model by adding an instance of a to $\mathcal{E}$, an instance of u to $\mathcal{U}$, an instance of a to $E(u)$, and an instance of u to $\bar{E}(a)$.

We seek parameter values that maximize the probability of the Wikipedia data given our model. We estimate these parameters by brute force search over a grid of quantized values. To make this tractable, we first search over a relatively coarse quantization grid: for each parameter we search over the range $(0, 1)$ at a spacing of 0.05. We then conduct a finer-scale search around the maximizing parameter values using a quantization of 0.005 for each parameter.

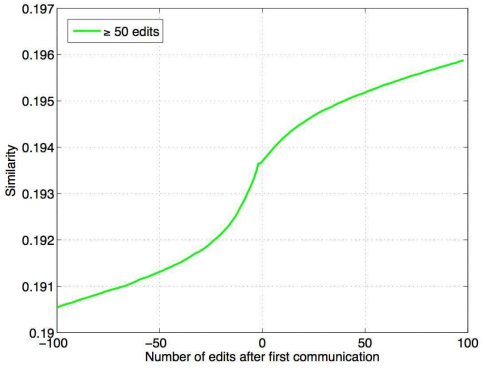


Figure 3: Result of simulation using our model, showing average similarity of user pairs as a function of the number of edits from time of first interaction.

We performed this parameter learning procedure on the entire history of Wikipedia through April 1, 2007, which consists of roughly 49 million article edit events and 3 million user interaction events. We set the history length parameter k to 10 for both the learning and the simulations presented in the next section, although we find that the simulation results are insensitive to k over the range $[5, 100]$. The result of the parameter estimation process was as follows. The parameter ϕ, the probability of communicating versus editing, was 0.058. Given that an edit is being performed, the article is chosen from one's own interests with probability $\psi = 0.35$, from a neighbor's interests with probability $\gamma = 0.081$, from the overall interests of Wikipedia editors with probability $\alpha = 0.5$, and by creating a totally new article with probability $\beta = 0.069$. In the interaction case, the user to communicate with is chosen randomly from the overall set of users with probability $\delta = 0.71$.

Simulating Wikipedia using the model. For the simulation, we initialized the model with 100,000 users as follows. For each user u, we sampled an edit count r_u from the empirical distribution of Wikipedia editor activity. We then chose an initial distribution of edits for u by sampling r_u articles from the empirical distribution of article edit frequency. This process gave initial values for the sequences $E(\cdot)$, $\bar{E}(\cdot)$, $\mathcal{U}$, and $\mathcal{E}$. To initialize the interaction histories, we set $N(\cdot)$ to be empty and then ran the model for 1,000,000 time steps. After initialization, we ran the model for an additional 3,000,000 time steps. Using the activities and interactions that occurred during this simulation, the average similarities of pairs of users were plotted as a function of time, aligned as before by the time of first interaction. Figure 3 shows the result of this simulation. This plot has a remarkably similar overall shape to the empirical data in Figure 1, and in particular exhibits the two key features of a rapid rise in similarity *prior* to the first interaction and a continued slow rise in similarity afterwards.

The model and the parameter settings derived from Wikipedia help us to better understand the effects of selection and social influence in this dataset. First, the vast majority

of activities are edits (94%) rather than interactions (6%). In choosing what to edit users are most influenced by the overall distribution of editing in Wikipedia (50%), then by their own edit history (35%), and finally by the the edit history of their neighbors (8%). In addition, 7% of the time they choose to start a new article rather than editing anything that already exists. We see that neighbors' past activities influence people's future behavior, but that this effect is less strong than people's intrinsic interests, and also less strong than effects that are not explicitly captured by the model and thus are captured by the background activity level across all articles.

We tried several simplifications to the model, but found that all of its ingredients are necessary to produce the sigmoid-shaped plots observed in the real data. For example, if the space of articles is kept constant by setting $\beta = 0$, then the baseline similarity between pairs of users increases at an excessive rate. Without the recency property of the sequences, the plot is approximately piecewise-linear instead of exhibiting the distinctive sigmoid shape. In particular, if the recency bias is removed from the user sequences (by setting $k = \infty$), then the convexity before first interaction disappears; if recency is removed from the article sequences, then the the concavity of the plot after the interaction disappears.

2.3 Insight from Selected Interactions

In order to better understand the aggregate behaviors we have been measuring and modeling, it is useful to take a more detailed look at the kinds of interactions occurring between people when they first meet on Wikipedia. There are a number of reasons why Wikipedia users might meet. For instance, a group of self-selected Wikipedians known as the Welcoming Committee writes greetings to a large number of new users; administrators (and others) post instructions and warnings to users who violate the community's norms; people notice others through their contributions to articles they are reading; and so on. Finally, as we suggested earlier, people might meet through shared activity around the construction of articles.

To explore the reasons people meet, we randomly selected 30 instances of two users meeting for the first time. We examined the content of the initial communication and any reply, looking for references to specific articles or other artifacts in Wikipedia. We also compared the edit history of the two users. Of the 30 messages, 26 referenced a specific article, image, or topic. In 21 cases, the users had both recently worked on the artifact that was the subject of conversation. The gap between co-activity and communication was usually short, often less than a day, though it stretched back three months in one case. Informally, communications tended to fall into a few broad categories: offering thanks and praise, making requests for help, or trying to understand the editing behavior of the other person.

This sample of interactions suggests that people most often come to talk to each other in Wikipedia when they become aware of the other person through recent shared activity around an artifact. Awareness then leads to communication, and often coordination. However, despite the fact that interaction often happens because of recent shared activity, we have seen this short-term effect has surprisingly long-term consequences in aggregate, as similarity continues to increase long after the first interaction. Taken together, then, these results emphasize the importance of building awareness of others in systems that seek to facilitate or exploit social interactions. Such awareness is fundamental to supporting distributed collaboration [8] and the efficient operation of teams [4]. Since awareness often leads to communication, systems that can identify needed social connections might deliberately present artifacts that lead people to become aware of each other and start communicating.

3. PREDICTIVE VALUE OF SOCIAL INTERACTIONS AND SIMILARITY

We now turn to our second main question, which addresses the relative power of similarity and social network links for predicting future behavior. We compare these two sources of information by putting them on a common footing as follows. First, we use a technique from [1] to compute the probability that a node adopts a given behavior given that k of its neighbors have done so; this measures the effect of social-network links. We then compare this to the probability that a node adopts a behavior given that the k most similar nodes have already done so.

In particular, we create two networks, one based on social interaction and one based on similarity of past behavior. For the interaction network, we create directed edges (v, w) between any users v and w such that v has edited w's user discussion page at some point before a given reference time t_1. We then create a *similarity network* with the same ordered degree sequence: if node v has d_v out-neighbors in the interaction network, we connect it to the d_v nodes whose activity vectors are the most similar to v's activity vector $\vec{v}(t_1)$. These activity vectors are constructed as in the previous section. We have considered a range of possible vector similarity measures, and in this section we focus on the *weighted Jaccard coefficient*:

$$\mathrm{Jac}(\vec{x}, \vec{y}) = \frac{\sum_{j=1}^{m} \zeta_j \min(x_j, y_j)}{\sum_{j=1}^{m} \zeta_j \max(x_j, y_j)} \qquad (2)$$

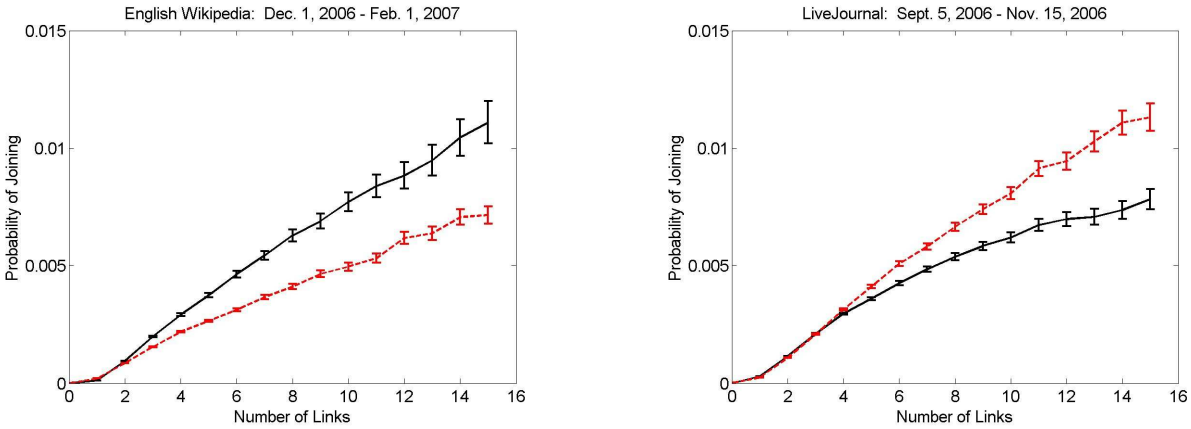
where the weights ζ_j are inversely proportional to the number of people who have engaged in each activity j. (We discuss our choice of similarity measure below.)

Once we create the networks, we use them to estimate the probability of future activities as follows. For either network, we say that an individual u is *k-exposed* to an activity a at time t_1 if that individual is a non-adopter of a at t_1 and has exactly k neighbors who are adopters at t_1. We then see whether u adopts a sometime between t_1 and a later time t_2. Let $p(k)$ be the fraction of cases in which a user was k-exposed to activity a at t_1 and then adopted that behavior between t_1 and t_2.

We have constructed the social influence and similarity networks for two different datasets. The first is Wikipedia, with the definitions of interactions and activities as defined previously; the second is LiveJournal, where we define interactions based on declared friendship links, and each LiveJournal community defines an activity that a user can engage in simply by joining the community.

Figures 4(a) and 4(b) show $p(k)$ as a function of k.[3] The

[3]For these plots, t_1 is 2006/12/01 and t_2 is 2007/02/01 for Wikipedia; for LiveJournal, t_1 is 2006/09/05 and t_2 is 2006/11/15. These times are arbitrary; moving forward or backward in time or changing the interval between the two times does not qualitatively affect the results.



(a) *Wikipedia* (b) *LiveJournal*

Figure 4: (a) Probability of joining a community based on k exposure via social ties versus similarity ties for (a) Wikipedia and (b) LiveJournal. The solid black curves corresponds to social ties and the dashed red curves to similarity ties. The error bars represent ± 2 standard errors.

solid black curves are drawn using neighbors in the social influence graph for each community, while the dashed red curves are drawn using neighbors in the corresponding similarity graphs. For Wikipedia, the interaction networks produces higher probabilities of future activities, whereas for LiveJournal the opposite is true. For a wide range of other vector-similarity measures, the shapes of the curves are qualitatively closely similar. The weighted Jaccard coefficient produces the curves that rise the most steeply among all the similarity measures we considered, hence providing the largest probability estimates.

The fact that the relative predictive values of similarity and social influence are reversed between Wikipedia and LiveJournal seems initially counter-intuitive: Wikipedia is focused on the creation of topic-specific articles of interest, while LiveJournal is more focused on social interaction, which might lead one to expect the opposite results. As with our observation of specific interactions in section 2.3, however, a detailed examination of behaviors provides further insight. In Wikipedia, much (perhaps most) of the work in editing articles involves activities such as fixing minor errors, editing for grammar, and fighting vandalism. These activities are important, but require little knowledge of or interest in the topic of the article. That is, editing an article is not a perfect indicator of interest in the topic, and having these "noisy" edits in one's profile reduces the performance of similarity-based predictors. The dynamics in LiveJournal are different: the creation of friendship links often has a primarily gestural significance, and does not necessarily indicate communication. However community membership in LiveJournal is arguably a good implicit indicator of interest in the topic of the community, as it does not serve the more mechanical sorts of activities (fixing errors, fighting vandalism) that one sees in Wikipedia.

This kind of analysis also sheds further light on the role of short-term coordination in Wikipedia. Recall that the curve in Figure 4(a) is produced by using article edits and social ties as of time t_1 to predict first edits between then and a later time t_2, where t_1 and t_2 were set to be two months apart. Many first edits close to t_1 would suggest effects based on short-term processes, such as immediate coordination around a given activity, whereas edits later in the time interval would suggest longer-term influence. We repeated the analysis with one modification: we considered only behavior that occurred in the latter half of the time interval between t_1 and t_2, still predicting based on the activities and social ties as of t_1. That is, t_1 and t_2 are still two months apart, but we only consider adoption of new activities during the second of the two months. The result is similar to that in Figure 4(a), further supporting the view that social influence effects in our data are not simply the result of short-term coordination.

4. DISCUSSION AND CONCLUSION

Our work examines two main questions, the first of which is the interplay between social interaction and similarity. We provide two main contributions around this question through empirical analysis of Wikipedia data. First, we show that in Wikipedia, people rapidly become more similar shortly before their first communication and continue to become more similar for a long time afterward. In other words, social interaction is both an effect and a cause of selection, and theories and models that relate them will need to consider their interaction. Second, we find strong evidence that people become aware of others through shared, recent activity around artifacts. This parallels the relationship between social interaction and selection in the physical world: people are more likely to talk to others they encounter in the same church, school, or workplace. Opportunities for these encounters are in turn driven by factors associated with selection such as income, race, location of residence, and education level, all of which are relatively immutable. Designers of online spaces have much more flexibility in structuring their systems to control how people come to interact with artifacts, and thus also to exert control over the nature and frequency of social interactions that arise.

Our model of the relationship between social influence and similarity provides a number of interesting questions for fur-

ther research. One main contribution of the model is that it provides a richer framework for exploring the dynamics of opinion change and behavior adoption in social networks than earlier models do. Another contribution of this model is more theoretical in nature. There is a rich mathematical literature that focuses on urn processes [17], and this model motivates new theoretical questions in this field — in particular, the study of urn processes that are coupled across the edges of a graph.

The second main question we have addressed is the extent to which similarities and social interactions serve as predictors of future behavior. We find that in Wikipedia social interaction is a better predictor of future behavior than similarity of interests, while the opposite is true in LiveJournal. We find that the dynamics of the community are important in understanding the relative power of social interaction and similarity-based predictors. Using both as sources of information might be a useful approach for recommender systems. For instance in Wikipedia, the social interaction and similarity graphs have little overlap, sharing fewer than 15% of their edges in common. This suggests future work investigating combining the two in a hybrid recommender system [3] may be a promising approach, especially in domains such as Facebook where social information and easily-computable similarity information coexist.

Finally, we find that examining specific instances of behavior can be a valuable complement to studying the data in aggregate. Looking at the content and context of people's first meetings informed our understanding of the relationship between interaction and similarity as well as creating a model of behavior adoption. Likewise, our observation that many Wikipedia edits are minor or systematic helped us understand why interest similarity is not as predictive as we might have expected.

Disciplines from sociology to economics are increasingly interested in exploring the large, rich datasets of behavioral and interaction data that can be captured when people interact through computers. We expect that mixed-method approaches that explicitly blend large-scale data mining techniques with deep understanding of the processes that generate the data will become an interesting area to explore.

5. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. *KDD*, 2006.

[2] N. Berger, C. Borgs, J. T. Chayes, and A. Saberi. On the spread of viruses on the Internet. *ACM Symposium on Discrete Algorithms*, 2005.

[3] R. Burke. Hybrid recommender systems: Surveys and experiments. *User Modeling and User-Adapted Interaction*, 12(4), 2002.

[4] M. Cataldo, P. Wagstrom, J. Herbsleb, and K. Carley. Identification of coordination requirements: Implications for the design of collaboration and awareness tools. In *CSCW '06*, 2006.

[5] D. Cosley, D. Frankowski, L. Terveen, and J. Riedl. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. *IUI*, 2007.

[6] P. Domingos. Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1), 2005.

[7] N. E. Friedkin. *A Structural Theory of Social Influence*. Cambridge University Press, 1998.

[8] C. Gutwin and S. Greenberg. The importance of awareness for team cognition in distributed collaboration. In E. Salas and S. M. Fiore, editors, *Team cognition*, APA Press, 2004.

[9] W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *CHI '95*, pages 194–201, 1995.

[10] P. Holme and M. E. J. Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74:056108, 2006.

[11] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(2006).

[12] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. *KDD*, 2006.

[13] P. Lazarsfeld and R. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Bergen, T. Abel, and C. Page, editors, *Freedom and Control in Modern Society*. Van Nostrand, 1954.

[14] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM EC*, 2006.

[15] M. Macy, J. Kitts, A. Flache, and S. Benard. Polarization in dynamic networks. In R. Breiger, K. Carley, P. Pattison (eds.), *Dynamic Social Network Modeling and Analysis*, Natl. Acad. Press, 2003.

[16] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 2001.

[17] R. Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007.

[18] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. *CSCW*, 1994.

[19] E. Rogers. *Diffusion of Innovations*, Free Press 1995.

[20] G. Salton. *Introduction to Modern Information Retrieval (McGraw-Hill Computer Science Series)*. McGraw-Hill Companies, September 1983.

[21] U. Shardanand, P. Maes. Social information filtering: Algorithms for automating "word of mouth". *CHI'95*.

[22] D. Strang, S. Soule. Diffusion in organizations and social movements. *Ann. Rev. Soc.* 1998.

[23] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In F. Naumann, M. Gertz, and S. E. Madnick, editors, *IQ*. MIT, 2005.

[24] M. Van Alstyne, E. Brynjolfsson. Global Village or CyberBalkans: Modeling and Measuring Integration of Electronic Communities. *Mgmt. Sci.*, in press.

[25] F. B. Viegas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in Wikipedia. In *HICSS 2007*, pages 78–87, 2007.

[26] J. Voss. Measuring Wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics*, 2005.

[27] S. Wasserman, K. Faust. *Social Network Analysis*. Cambridge Univ. Press, 1994.