Sequential Influence Models in Social Networks

Dan Cosley
Information Science
Cornell University
Ithaca, NY 14850
danco@cs.cornell.edu

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

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

Xiangyang Lan
Department of Computer Science
Cornell University
Ithaca, NY 14853
xylan@cs.cornell.edu

Siddharth Suri
Yahoo! Research
111 W. 40th St., 17th Fl.
New York, NY 10018
suri@yahoo-inc.com

Abstract

The spread of influence among individuals in a social network can be naturally modeled in a probabilistic framework, but it is challenging to reason about differences between various models as well as to relate these models to actual social network data. Here we consider two of the most fundamental definitions of influence, one based on a small set of "snapshot" observations of a social network and the other based on detailed temporal dynamics. The former is particularly useful because large-scale social network data sets are often available only in snapshots or crawls. The latter however provides a more detailed process model of how influence spreads. We study the relationship between these two ways of measuring influence, in particular establishing how to infer the more detailed temporal measure from the more readily observable snapshot measure. We validate our analysis using the history of social interactions on Wikipedia; the result is the first large-scale study to exhibit a direct relationship between snapshot and temporal models of social influence.

Introduction

The ways in which people influence each other through their interactions is a powerful but subtle process that is pervasive in social networks. There is a long history of empirical work on this topic in sociology, through studies of effects such as opinion formation and the diffusion of innovations (Rogers 1995; Strang and Soule 1998); in economics, theoretical models have been developed to cast social influence as a process by which individuals in a social network tend to coordinate (or anti-coordinate) their decisions (e.g. (Blume 1993; Young 1998)). More recently, computer scientists have begun developing models for influence in social networks, motivated by applications such as viral marketing (Domingos and Richardson 2001; Kempe, Kleinberg, and Tardos 2003; Leskovec, Adamic, and Huberman 2006), the spread of online news (Gruhl et al. 2004; Leskovec et al. 2007), and the growth of on-line communities (Backstrom et al. 2006).

It is natural to model influence in a social network using a probabilistic framework: as a behavior spreads through a population, we can examine the probability that a particular individual adopts the new behavior, given that k of his or her neighbors in the social network have done so. By "neighbors" here, we mean the people to whom the individual has direct social network links; we will refer to people as "adopters" or "non-adopters" at any point in time, depending on whether they have exhibited the new behavior by that time. Within the past few years, data from online settings has enabled the estimation of such probabilities for behaviors spreading on very large populations with detailed information about network structure, including the probability of a product purchase as a function of the number k of recommendations by e-mail (Leskovec, Adamic, and Huberman 2006), and the probability of joining an online community as a function of the number k of neighbors belonging to the community (Backstrom et al. 2006; Shi et al. 2009).

Definitions of Social Influence. It has become clear through these initial studies that there is not yet a standard model for representing influence, and this has become a source of difficulty in understanding large-scale network data on social influence. In fact, the relationships among the possible models here are subtle, and understanding these relationships forms the basic motivation for the present work. To begin with, here are two natural but distinct notions that one might mean by the probability of adoption, expressed as a function of the number k of neighbors who have already adopted. To make these notions easier to express, we say that an individual is *k-exposed* to the behavior at a particular point in time t if they are a non-adopter at time t, but they have exactly k neighbors in the network who are adopters at time t.

Two natural definitions follow from the notion of being k-exposed:

- **Ordinal-time definition.** Consider a complete time sequence of an evolving social network that includes each time a new network link is formed and each time an individual adopts a new behavior. For each k, consider the set of all individuals who were ever k-exposed at any time, and define $p_o(k)$ to be the fraction of this set that became adopters before acquiring a $(k+1)^{\text{st}}$ neighbor who is an adopter.

- **Snapshot definition.** Consider two snapshots of the network at different points in time. For each k, consider the set of all individuals who are k-exposed in the first snapshot. Let $p_s(k)$ be the fraction of individuals in this set who have become adopters by the time of the second snapshot.

In other words, for the ordinal-time definition, we imagine that at the moment a non-adopter acquires their k^{th} neighbor who is an adopter, he or she flips a coin of fixed bias $p_o(k)$ to decide whether to adopt. For the snapshot definition we imagine that everyone who is k-exposed in the first snapshot flips a coin of fixed bias $p_s(k)$ to decide whether to adopt. In the two cases, we determine the maximum-likelihood values of these fixed probabilities, $p_o(k)$, $p_s(k)$, respectively. The measurements of influence probabilities by Kossinets and Watts (2006), Backstrom et al. (2006), and Shi et al. (2009) used the snapshot definition—though with widely varying numbers of snapshots—while the work of Leskovec et al. (2006) used something closer to, though different than, ordinal time. This difference in models, without an understanding of the relationship between them, makes it difficult to compare results. Moreover, no direct implementation of the ordinal-time definition on large-scale data has appeared in the literature, thus limiting our ability to draw conclusions about the temporal evolution of adoption behavior.

While the ordinal-time definition is appealing in positing an operational procedure by which influence is manifested, the snapshot definition is more widely applicable; it is amenable to settings, such as Web crawls or some types of on-line communities, in which one can only take periodic mass observations of the network, without the ability to perform moment-by-moment measurement. Here we address the following questions. What is the relationship between these definitions? How do they differ when both applied to a single dataset? And how can one infer an ordinal-time measurement with reasonable accuracy given only snapshots?

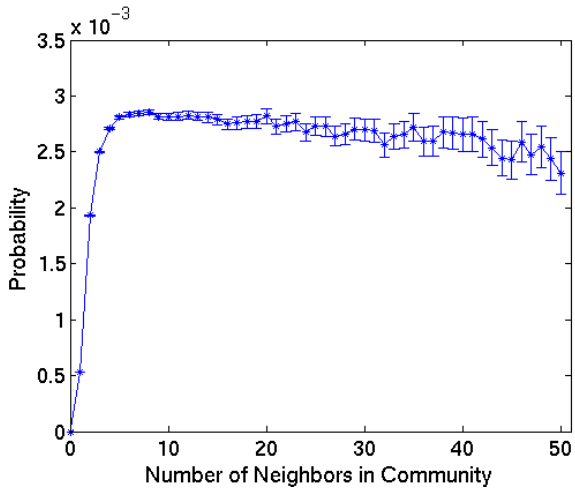
Wikipedia as a Dataset for Analyzing Social Influence. We address these questions both through analysis and through the study of a large dataset, based on Wikipedia, in which social influence can be measured using both models. Wikipedia has been the subject of research both because its entire history is freely available for analysis, and because, beyond serving as an encyclopedia, it is also a community (see e.g. (Viegas et al. 2007; Stvilia et al. 2005; Voss 2005; Wilkinson and Huberman 2007; Crandall et al. 2008)). Indeed, while most people may experience Wikipedia simply by reading its articles, it also has a rich social structure in which several hundred thousand contributors interact with one another in the process of creating those articles. In particular, many contributers maintain *user-talk pages* and talk to one another through postings on these pages.

For our purposes, this makes Wikipedia an ideal system in which to study the spread of behaviors in the presence of social influence. We can define an (undirected) social network link between two Wikipedia editors u and v if one has written on the user-talk page of the other, and we can define a "behavior" associated with the editing of each individual article. Since user-talk page interactions are often concerned with the content being created, there is a strong connection between these interactions and the kinds of influence that lead a user to edit a particular article. Thus it is not surprising that a user's interaction with editors of a particular article indeed increases that user's probability of subsequently editing the article. How this probability depends on the social interaction is precisely the effect we wish to measure, using both the snapshot and ordinal definitions: in other words, we ask how the probability of adopting a behavior, which in this setting corresponds to editing an article, depends on the number k of links one has had to previous editors of the article. Wikipedia's edit history provides us with ordinal time data, which we can then use to generate snapshot data at any point in time.

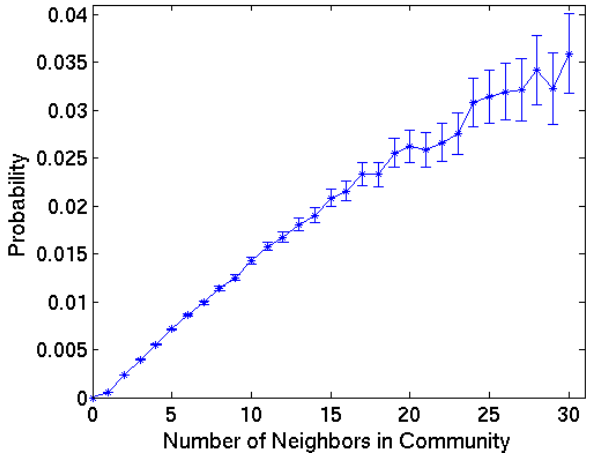
A final interesting feature of Wikipedia is the fact that there is actually more than one Wikipedia: over a dozen languages have their own Wikipedias with at least 100,000 articles; another 50 or so have at least 10,000. Although some contributors participate in multiple languages, each Wikipedia evolves independently and provides its own historical data. We can therefore assess the generality of our models and results by evaluating them on multiple large Wikipedias. In this paper, we will discuss results primarily in terms of the English Wikipedia, but the findings carry over remarkably closely to German and French. Although all are from the same domain of creating encyclopedias, the English, French and German Wikipedias are all large, independent data sets, with largely disjoint sets of editors. Each one also has their own set of cultures, norms, and administrative processes, the differences of which have been studied elsewhere (such as (Pfeil, Zaphiris, and Ang 2006); see also the article at http://en.wikipedia.org/wiki/User:Elian/comparison by a longtime Wikipedia contributor who participates in both the German and English versions). Thus, it is important to note that the empirical evaluation has been done on three quite different datasets.

Relationships Among the Definitions. Figure 1 shows the basic contrast between the shape of influence curves for the snapshot and ordinal-time definitions using the Wikipedia data. The snapshot-based curve is qualitatively consistent with what one finds in other datasets, such as community membership in LiveJournal (Backstrom et al. 2006) and engaging in email correspondence (Kossinets and Watts 2006): the extent of influence steadily increases with more links, but the marginal influence of each additional link slowly decreases. In apparent contrast, the ordinal-time curve shows that the first five links have increasing effect, after which subsequent links have relatively constant effect (with a gradual decline).

As one of our main results, we will provide a method for converting from influence probabilities based on snapshots to influence probabilities based on the more fine-grained notion of ordinal time. On the Wikipedia data, we find that in fact the shape of the ordinal-time plot—including the in-



(a) *Wikipedia: Ordinal-Time Definition*



(b) *Wikipedia: Snapshot Definition*

Figure 1: The probability of editing an article on Wikipedia, as a function of the number k of previous editors of that article with whom one has had user-talk-page interactions. Results are shown for (a) the ordinal-time definition and (b) the snapshot definition. The error bars represent ± 2 standard errors.

crease over the first five values of k followed by leveling off—can be approximated from the data inherent in just a single snapshot. Thus, the specifics of the ordinal-time process can be approximately inferred even in cases when one is given coarse snapshot views of the network, rather than moment-by-moment temporal data. This offers the promise of applying more detailed social influence models to domains in which one has much less finely resolved views of the underlying dynamics.

Analysis of Adoption Behavior

We restrict our attention to users who have both Wikipedia user ID's and user-talk pages. Anonymous edits are recorded by IP address, which might combine the activity of many people, while users without user-talk pages have very few social connections. As of April 2, 2007, for English Wikipedia there are approximately 510,000 users with user ID's and user-talk pages. These users were responsible for 61% of all edits to articles on the English Wikipedia.

We now provide a bit of additional terminology that will be useful in what follows. First, if nodes u and v are connected by an edge in a social network, we refer to them as "neighbors". In the context of Wikipedia, we create an edge between u and v at the first time either one edits the other's talk page; they are neighbors from that time onward. Here we will consider only undirected relations, although there are also interesting questions involving directed edges, emphasizing the distinction between u writing on v's user-talk page and v writing on u's.

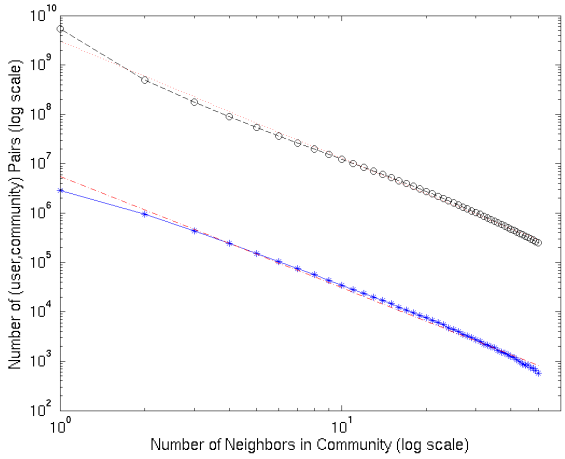
A second piece of terminology is the following: each behavior defines a *community* of nodes, simply consisting of all those who have engaged in the behavior. (Thus the term "community," like "neighbor," is meant in this paper in a specific technical sense.) The membership of a community grows over time as more people adopt the behavior; the spread of a behavior through the network can therefore be equivalently viewed as the growth of the associated community. In the case of Wikipedia, there is a community associated with each article; it is the set of all users who have ever edited the article. The crucial question is how a node's probability of joining a community depends on the set of neighbors it has within the community.

Ordinal time. We begin by recalling the definition of $p_o(k)$ from the introduction, broadened here to reflect the fact that there are multiple communities being studied on the same network. To begin with, we say that a node u is *k-exposed* to a community C at a time t if it has k neighbors in C at time t, but does not belong to C at time t.

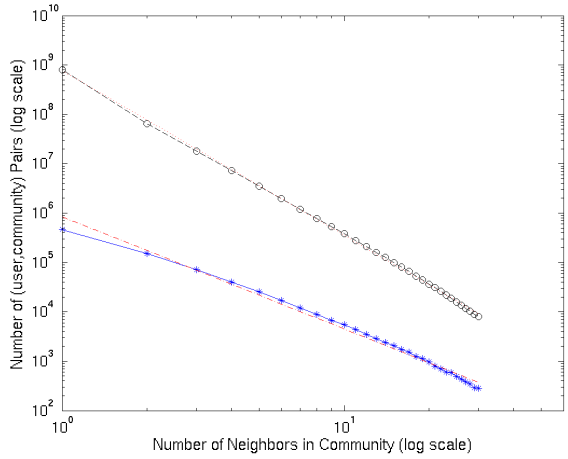
Informally, $p_o(k)$ is the fraction of cases in which a node that is k-exposed to a community C proceeds to join C before acquiring a $(k+1)^{\text{st}}$ neighbor in C. That is, we define $p_o(k)$ as the ratio of two quantities, $p_o(k) = n_o(k)/d_o(k)$, where $d_o(k)$ is the number of triples (u, C, k) for which u was ever k-exposed to C, and $n_o(k)$ is the number of triples (u, C, k) for which u was k-exposed to C, and then joined C before acquiring a $(k+1)^{\text{st}}$ neighbor in C.

We can now investigate these quantities in the context of Wikipedia. Because Wikipedia maintains such a finely time-resolved history of edits to articles and user-talk pages, we know the precise interleaving of social network link formation and community joining behavior, allowing us to compute the quantities $d_o(k)$ and $n_o(k)$. We do this analysis for all of the edits that have occurred since the beginning of Wikipedia on January 15, 2001 until April 2, 2007.

For the English Wikipedia data, Figure 2(a) shows the plots of $n_o(k)$ and $d_o(k)$ on a log-log scale, along with the best linear fit. A linear model accounts relatively well for the data over a large range, suggesting a power law is a reasonable approximation for each of these quantities. This is not



(a) *Ordinal Time: Numerator and Denominator*



(b) *Snapshot: Numerator and Denominator*

Figure 2: The numerator and denominator for ordinal-time and snapshot on a log-log scale along with their linear fits. The numerators are solid and their fit is dotted-dashed. The denominators are dashed and their fit is dotted. In both plots the denominators appear above the numerators.

surprising, given that $d_o(k)$ can be viewed as a variation on the standard degree distribution: it measures the distribution of the "degree" (number of edges) of each node into each community. The corresponding plot of $p_o(k)$ was shown in the introduction as Figure 1(a), where it was observed that the probability increases for the first five neighbors before becoming roughly constant (with a gradual decline). In the next section, we will look more closely into how the shape of the $p_o(k)$ curve relates to power laws fitted to $n_o(k)$ and $d_o(k)$.

Snapshots. Analogously, we can extend the snapshot definition of influence $p_s(k)$, given in the introduction, to multiple communities. Suppose that we take two snapshots of all community memberships in the network: the first at time t_1 and the second at time t_2. As with ordinal time, it is now useful to define $p_s(k)$ as the ratio of two quantities, $p_s(k) = n_s(k)/d_s(k)$. Here $d_s(k)$ is the number of triples (u, C, k) for which u was k-exposed to C at time t_1, and $n_s(k)$ is the the number of triples (u, C, k) for which u was k-exposed to C at time t_1, and such that u joined C between times t_1 and t_2. For the English Wikipedia data, Figure 2(b) shows the plots of $n_s(k)$ and $d_s(k)$ on a log-log scale, along with the best linear fit. Here again, the linear model accounts relatively well for these two quantities indicating that a power law is a good approximation for each of them.

Recall that there are many settings with on-line social interaction data in which it is feasible to produce snapshots of the system, but not to obtain the kind of time-resolved data necessary to compute the ordinal-time measure $p_o(k)$. Using the Wikipedia ordinal time data we generated two snapshots, choosing November 1, 2005 and November 6, 2006 as two relatively arbitrary moments at which which we measure the full set of community memberships. We then computed $p_s(k)$ using the snapshot method, shown in Figure 1(b).

Clearly the snapshot curve looks qualitatively very different from the ordinal-time curve, and understanding the relationship between the two will be the focus of the next section. It is interesting that it is quite similar to other snapshot-based influence curves from earlier work, such as for LiveJournal community memberships (Backstrom et al. 2006), considering that these other curves come from very different domains with different numbers of users and communities and different reasons for joining communities. In each case, the snapshot curves show a general sublinear increase with a marked dip at $k = 1$ and rise at $k = 2$. However, the Wikipedia curve is much closer to linear than the LiveJournal curve up to moderate k.

Relating Snapshot Measurements with Ordinal Time Measurements

In this section we investigate how the snapshot definition and ordinal-time definition are related to each other. To describe this relationship we first define some notation for various sets of tuples in terms of the two snapshots. Informally, B will denote the set of instances in which a user joined a community before the first snapshot, J will denote the set of instances in which a user joined a community between the two snapshots, and N will denote the set of instances in which a user did not join a community by the second snapshot. More formally,

- $B(t_1) = \{(u, C, k_1) \mid u$ joined C before t_1 and u had k_1 neighbors in C at $t_1\}$
- $J(t_1, t_2) = \{(u, C, k_1, k_2) \mid u$ had k_1 neighbors in C at t_1, u joined C between t_1 and t_2, u had k_2 neighbors in C at $t_2\}$
- $N(t_2) = \{(u, C, k_2) \mid u$ did not join C before t_2 and u had k_2 neighbors in C at $t_2\}$.

With these definitions we begin analyzing the difference between $p_o(k)$ and $p_s(k)$. To do this we will analyze how each of the three sets defined above contribute to $n_o(k)$, $n_s(k)$, $d_o(k)$ and $d_s(k)$.

The effect of $B(t_1)$. First we consider the joining events that occur before before the first snapshot; by definition these are the triples $(u, C, k_1) \in B(t_1)$. Of all the joining events that occur, the snapshot method only captures those which occur between the two snapshots; thus such a $(u, C, k_1) \in B(t_1)$ will not contribute in any way to p_s. On the other hand, since the ordinal time method captures all joining events, this tuple will contribute to p_o. More specifically, if u actually had $0 \leq \delta \leq k_1$ neighbors in C just before joining, then it will contribute to $n_o(\delta)$. Furthermore, since u did not join until it acquired δ neighbors, this tuple will contribute to each of $d_o(0), \ldots, d_o(\delta)$. Thus, for each value of k, $n_o(k)$ and $d_o(k)$ will be shifted upwards with respect to $n_s(k)$ and $d_s(k)$ due to the effect of triples in $B(t_1)$.

The effect of $J(t_1, t_2)$. Next we analyze the effects of the joining events that occur between the two snapshots. First we consider their effect on n_s and n_o and then we consider their effect on d_s and d_o. If a tuple (u, C, k_1, k_2) is in $J(t_1, t_2)$, then u joined C at some time between t_1 and t_2. Thus this tuple will contribute to $n_s(k_1)$. Note that u had k_1 neighbors in C at t_1 and k_2 neighbors in C at t_2. Now u could have had $0 \leq \delta \leq k_2 - k_1$ neighbors between t_1 and joining C. Thus this joining event would contribute to $n_o(k_1 + \delta)$. We can think of this as a "stretching" effect: the contribution to n_o is pushed out to a higher value of k relative to n_s. This is a result of the finer-grained observations available with the ordinal-time definition. If we assume that n_o and n_s are closely approximated by power laws, this stretching will cause the log-log slope of n_o to be greater than that of n_s.

Now we analyze how joining events that occur between the two snapshots affect $d_o(k)$ and $d_s(k)$. Again, let $(u, C, k_1, k_2) \in J(t_1, t_2)$ and suppose u actually joined C after getting $k_1 + \delta$ neighbors where $0 \leq \delta \leq k_2 - k_1$. First observe that this tuple will contribute to $d_s(k_1)$ and $d_o(k_1 + \delta)$. This is another instance of the "stretching" phenomenon that we observed occurring from n_s to n_o. In addition, this tuple will also contribute to $d_o(j)$, for all $0 \leq j < k_1 + \delta$. This is because after u acquired its j^{th} neighbor, it acquired yet another neighbor before joining C. Thus, a joining event given snapshot observations only contributes to $d_s(k_1)$, whereas in the ordinal-time measure it contributes to $d_o(0), \ldots, d_o(k_1 + \delta)$.

We can think of this as an "accumulation" phenomenon that extends the stretching phenomenon: the contribution to d_o is accumulated over multiple values of k. More concretely, we have $d_o(k) = \sum_{j \geq k} d_s(j - \delta)$. If $d_s(k)$ is approximated by a power law distribution, $d_s(k) \approx ck^{-\alpha}$, then

$$d_o(k) \approx c \int_{x=k}^{\infty} (x - \delta)^{-\alpha} \approx c' k^{-\alpha+1},$$

where c and c' are constants.

The effect of $N(t_2)$. Finally, we analyze the effect of the tuples in $N(t_2)$. Recall that if $(u, C, k_2) \in N(t_2)$ then u did not join C before t_2 and u had k_2 neighbors in C at t_2. Since u did not join C between the two snapshots, this tuple does not contribute to n_s. But u may join C after the second snapshot, which would contribute to $n_o(k)$ for some $k \geq k_2$. This would cause $n_o(k)$ to shift upwards. Since u had k_2 neighbors at t_2 and u did not join C before that, this tuple will contribute to $d_s(k_2)$. Also, u will contribute to $d_o(0), \ldots, d_o(k_2)$ for the reason just discussed above: for any $0 \leq j < k_2$, user u acquired at least one more neighbor before joining C. This again results in the accumulation phenomenon.

Combining the contributions. In summary, the sets B and N result in shifting n_o and d_o upwards with respect to n_s and d_s. Also, the set J results in stretching n_o and d_o when compared to n_s and d_s. Finally, the sets J and N result in d_o becoming an accumulation or integration of d_s.

Table 1 shows the slope and y-intercept of the linear fits on a log-log scale to the power law approximations to n_o, n_s, d_o, and d_s. The plots of these for English Wikipedia are shown in Figure 2. Since the x-axis in Figure 2 measures the number of neighbors u has in a community, we have little data for x values over 30 or so. Thus we cannot use the methods of Clauset et al. (2009) to estimate the parameters of the power law or to say with statistical confidence whether or not the data is better approximated by some other type of distribution. Instead we use the best methods available to us to measure the parameters of these distributions which appear to be power laws. We only use these distributions as examples to illustrate the relationship between snapshot and ordinal time given above which applies to general distributions.

We now see whether the data supports this analysis. First observe that the y-intercepts for n_o and d_o dominate the y-intercepts for n_s and d_s. This shows the upward shift of n_o and d_o due to B and N. Also, in the case of English, the slope of n_o is slightly greater then n_s, and in the case of French and German these slopes are almost identical. This implies that the stretching caused by the set J is very minor. Finally, one can see that the difference in slopes between d_o and d_s is close to 1. This illustrates the accumulation phenomena and also shows that the effect of the stretching is minimal. We can also see this reflected in Figure 2, where the lines for n_o and d_o in 2(a) are roughly parallel, while the lines for n_s and d_s in 2(b) are converging due to slopes that differ by approximately 1.

If, as is indicated by the data, we assume that both the stretching of the numerator in going from n_s to n_o is negligible and that d_o is approximately the integral of d_s, then we can relate the plot of p_s to the plot of p_o. Observe that a large span of p_s, shown in Figure 1(b), is roughly linear. So $p_s(k) \approx ck$ for some constant c. Since transforming n_s to n_o leaves the numerator roughly unchanged, and transforming d_s to d_o increases the exponent of the power law in the denominator by one, the roughly linear span of p_s will correspond to a roughly constant span of p_o. This is exhibited by

	n_o		n_s		d_o		d_s	
	slope	y-int.	slope	y-int.	slope	y-int.	slope	y-int.
English	-2.26	6.74	-2.50	6.12	-2.37	9.49	-3.49	9.02
German	-2.58	6.64	-2.57	5.91	-2.72	9.26	-3.59	8.68
French	-2.54	6.32	-2.52	5.44	-2.78	8.99	-3.70	8.28

Table 1: The slopes and y-intercepts of the linear fits on a log-log scale to n_o, n_s, d_o, and d_s.

Figures 1(a) and 1(b). On the other, consider an interval of k over which $p_s(k)$ grows sublinearly; that is $p_s(k) \approx ck^\alpha$, for some interval of k and $0 < \alpha < 1$. By the analogous reasoning to the linear case, $p_o(k)$ would be roughly approximated by $1/(k^{1-\alpha})$. Thus a sublinear span of $p_s(k)$ would correspond to a decreasing span of $p_o(k)$. Observe that in Figure 1(b) when $k \geq 20$ the curve begins to become sublinear, and this corresponds to a decreasing region of $p_o(k)$. Finally, by the same reasoning as above, if $p_s(k)$ exhibits a large superlinear region approximated by k^α where $\alpha > 1$, this will correspond to $p_o(k)$ being roughly approximated by $k^{\alpha-1}$ which is increasing.

These arguments do not directly explain the dramatic ramp-up of $p_o(k)$ that occurs from $k = 0$ to $k = 5$. To do this, we look at the best fit lines to $n_o(k)$ and $d_o(k)$ in Figure 2(a). We see that when k is small, $n_o(k)$ approaches the best fit line from below and $d_o(k)$ approaches the best fit line from above. This indicates that for the first few values of k, $p_o(k)$ will begin significantly below its eventually (approximately) constant value.

Simulating Ordinal Time from Snapshots

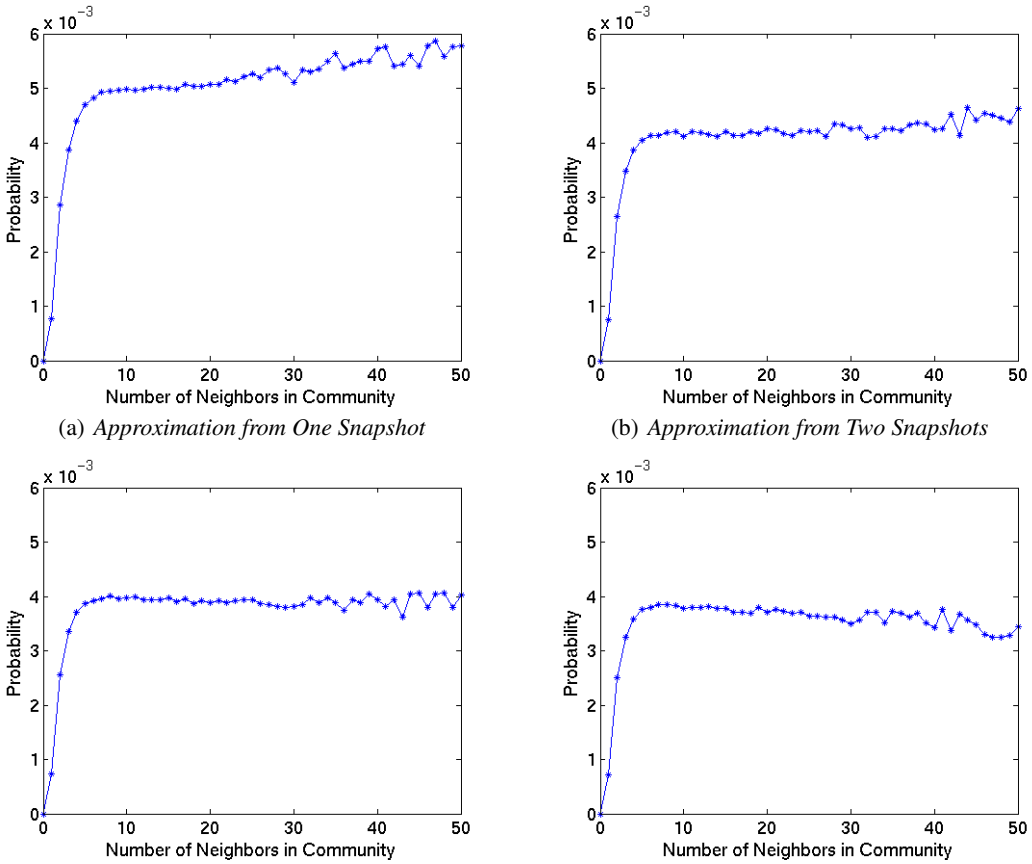
We now consider how one might take snapshot data and produce a simulated ordinal-time plot from it. This has clear potential utility in cases where only snapshots of a system are available, and one wants to make approximate comparisons with systems where ordinal-time influence measures exist.

Roughly, the simulation works by hypothesizing the number of neighbors each node had at the moment it joined a community; choosing this number from among the possible values consistent with the snapshot observations. We exploit both the $B(t_1)$ and the $J(t_1, t_2)$ sets. Recall the set $B(t_1)$ consists of triples (u, C, k_1), where u joined C before t_1, and u had k_1 neighbors in C at t_1. We choose an integer j uniformly at random in $[0, k_1]$ and assume that u had j neighbors in C at the time it joined C. Similarly, the set $J(t_1, t_2)$ consists of tuples (u, C, k_1, k_2) where u joined C between t_1 and t_2, u had k_1 neighbors in C at t_1, and u had k_2 neighbors in C at t_2. Here we construct the approximation to ordinal-time by choosing an integer j uniformly at random from $[k_1, k_2]$ and again assuming that that u had j neighbors in C at the time it joined C. Finally, we do not assume that u joins C for any tuple $(u, C, k) \in N(t_2)$. There are clearly (and necessarily) many approximations being made in this simulation, and so it is not *a priori* clear that an ordinal-time plot produced in this way from snapshot data will have a reasonable fit with the true ordinal-time plot. However, we shall see the agreement on Wikipedia data is surprisingly close at a qualitative level, even capturing the detailed structure of the ramp-up for the first few values of k.

The approximation of ordinal time data from snapshot data depends on two factors: the number of snapshots used, and the amount of time between the snapshots. We begin by considering the effect of the number of snapshots. We show how the simulation of ordinal-time depends on the number of snapshots taken for English Wikipedia in Figure 3. Figures 3(b), 3(c), and 3(d) show the results of the method described in the previous paragraph using two, three, and seven snapshots respectively. Figure 3(a) shows the results in an even more extreme situation, with only a single snapshot, when we do not have a set of the form $J(t_1, t_2)$. As one would expect, the approximation is becoming increasingly accurate with more snapshots. This is because as the number of snapshots increases the time between them goes to 0. Thus, in the limit, snapshot measurements converge to the ordinal-time measurements. Figure 3 shows that empirically just a few snapshots produce good results for these datasets which means the convergence occurs fairly rapidly as the number of snapshots increases. Repeating these simulations for the French and German Wikipedia gives qualitatively similar results.

In Figure 4 we explore the effect of changing the amount of time between snapshots, focusing on the case of two snapshots. The figures show the result of doing the approximation on English Wikipedia using two snapshots each one, three and six months apart; recall that the result of doing the approximation using two snapshots twelve months apart is in Figure 3(b). Varying the amount of time between two snapshots cannot produce effects as accurate as we saw for increasing the number of snapshots, in Figure 3, but as the time between snapshots gets longer we observe increasing accuracy in the approximation. Note in particular the more accurate values of the absolute probabilities on the y-axis compared the ground truth from Figure 1(a). This is natural because as the time between the two snapshots increases, more joining events are captured between them which causes the quality of the approximation to improve.

In general, the simulated probabilities of p_o computed using a varying number of snapshots, shown in Figure 3, are higher than the actual ordinal time probabilities shown in Figure 1(a). We believe this happens because the algorithm tends to hypothesize low values of k too often. In Wikipedia, the probabilities for $k < 5$ are below the average; a uniform distribution guesses k values between 0 and 4 too often, increasing estimates of $n_o(k)$ in this range. This also reduces all estimates of $d_o(k)$: every time the algorithm guesses a too-low k, it misses some stretching of the de-



(a) *Approximation from One Snapshot*

(b) *Approximation from Two Snapshots*

(c) *Approximation from Three Snapshots*

(d) *Approximation from Seven Snapshots*

Figure 3: Approximating ordinal time from One, Two, Three, and Seven Snapshots on English Wikipedia. In panels b, c and d the time between the first and last snapshot is one year.

nominator (e.g., by guessing $k = 1$ instead of $k = 6$, $d_o(k)$ for $2 \leq k \leq 6$ are not incremented). Both errors inflate the algorithm's estimates of $p_o(k)$ for all k. One improvement would use the snapshot data to estimate a better-than-uniform probability distribution. Direct application of this idea, however, such as assuming that the relative probabilities for each k are the number of tuples in $J(t_1, t_2)$ where u had exactly k neighbors in C at t_1 *and* t_2, divided by the number of tuples in $N(t_2)$ where u had k neighbors in C at t_2, did not improve the results, suggesting that that there is an interesting and subtle problem here for future work.

Conclusion

Our work complements and extends the existing literature around influence in online communities. Prior work has shown that how one's friends influence the groups one joins online is quite similar across a variety of domains, content types, community goals, and ways of inferring ties. We show that this type of social influence occurs in Wikipedia as well. Furthermore, our demonstration of the relationship between snapshot and ordinal-time measurements may help researchers better understand social influence by allowing them to more easily compare data gathered with different sampling procedures. The correspondence between fine-grained ordinal data and the approximation of it made from snapshots is not perfect, but it appears close enough to make useful comparisons. Future work that improves the models presented here will make snapshot approximations both more useful and more comparable, which we hope will allow researchers to better understand the similarities and differences that underlie the dynamics of influence online.

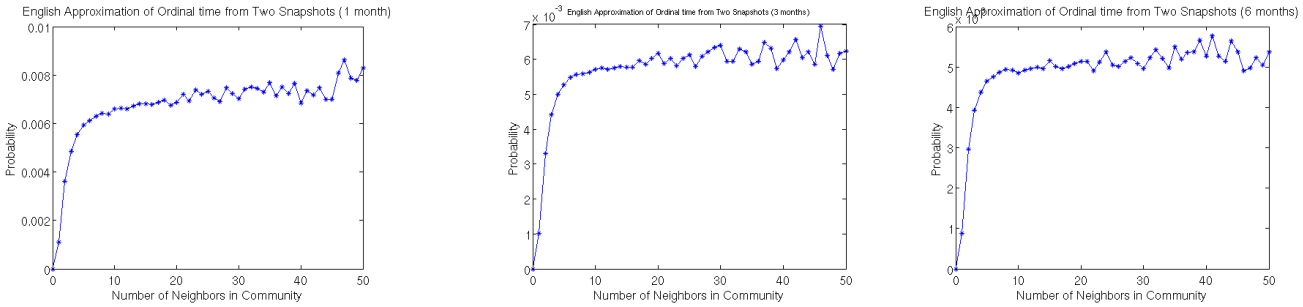
Acknowledgments

This work was partially supported by NSF grants IIS-0845351, IIS-0705774, IIS-0910664, and CCF-0910940.

References

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: Membership, growth, and evolution. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Blume, L. 1993. The statistical mechanics of strategic interaction. *Games and Economic Behavior* 5:387–424.



(a) *Two Snapshots, One Month Apart* (b) *Two Snapshots, Three Months Apart* (c) *Two Snapshots, Six Months Apart*

Figure 4: Approximating ordinal time from two snapshots taken one, three and six months apart in English Wikipedia.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4):661–703.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 57–66.

Gruhl, D.; Liben-Nowell, D.; Guha, R. V.; and Tomkins, A. 2004. Information diffusion through blogspace. In *Proc. 13th International World Wide Web Conference*.

Kempe, D.; Kleinberg, J.; and Tardos, É. 2003. Maximizing the spread of influence in a social network. In *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

Kossinets, G., and Watts, D. 2006. Empirical analysis of an evolving social network. *Science* 311:88–90.

Leskovec, J.; Adamic, L.; and Huberman, B. 2006. The dynamics of viral marketing. In *Proc. 7th ACM Conference on Electronic Commerce*.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. In *Proc. SIAM International Conference on Data Mining*.

Pfeil, U.; Zaphiris, P.; and Ang, C. 2006. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer Mediated Communication*.

Rogers, E. 1995. *Diffusion of Innovations*. Free Press, fourth edition.

Shi, X.; Zhu, J.; Cai, R.; and Zhang, L. 2009. User grouping behavior in online forums. In *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 777–786. New York, NY, USA: ACM.

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

Stvilia, B.; Twidale, M.; Smith, L.; and Gasser, L. 2005. Assessing information quality of a community-based encyclopedia. In *Proc. Int'l Conference on Information Quality*.

Viegas, F. B.; Wattenberg, M.; Kriss, J.; and van Ham, F. 2007. Talk before you type: Coordination in wikipedia. In *Proc. HICSS*.

Voss, J. 2005. Measuring wikipedia. In *Proc. International Conference of the International Society for Scientometrics and Informetrics*.

Wilkinson, D. M., and Huberman, B. A. 2007. Cooperation and quality in wikipedia. In *Proceedings of the 2007 International Symposium on Wikis*.

Young, H. P. 1998. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press.