Influence and Correlation in Social Networks

Aris Anagnostopoulos Ravi Kumar Mohammad Mahdian

Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089.
{aris,ravikumar,mahdian}@yahoo-inc.com

ABSTRACT

In many online social systems, social ties between users play an important role in dictating their behavior. One of the ways this can happen is through *social influence*, the phenomenon that the actions of a user can induce his/her friends to behave in a similar way. In systems where social influence exists, ideas, modes of behavior, or new technologies can diffuse through the network like an epidemic. Therefore, identifying and understanding social influence is of tremendous interest from both analysis and design points of view.

This is a difficult task in general, since there are factors such as homophily or unobserved confounding variables that can induce statistical correlation between the actions of friends in a social network. Distinguishing influence from these is essentially the problem of distinguishing correlation from causality, a notoriously hard statistical problem.

In this paper we study this problem systematically. We define fairly general models that replicate the aforementioned sources of social correlation. We then propose two simple tests that can identify influence as a source of social correlation when the time series of user actions is available.

We give a theoretical justification of one of the tests by proving that with high probability it succeeds in ruling out influence in a rather general model of social correlation. We also simulate our tests on a number of examples designed by randomly generating actions of nodes on a real social network (from Flickr) according to one of several models. Simulation results confirm that our test performs well on these data. Finally, we apply them to real tagging data on Flickr, exhibiting that while there is significant social correlation in tagging behavior on this system, this correlation cannot be attributed to social influence.

Categories and Subject Descriptors: J.4 [**Computer Applications**]:Social and Behavioral Sciences—*Sociology*

General Terms: Economics, Human Factors

Keywords: Social influence, Social networks, Correlation, Tagging

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

1. INTRODUCTION

Online social networks are playing an ever-important role in shaping the behavior of users on the web. Popular social sites such as Facebook, MySpace, Flickr, and del.icio.us, are enjoying increasing traffic and are turning into community spaces, where users interact with their friends and acquaintances. The availability of such rich data at never-before–seen scales makes it possible to analyze user actions at an individual level in order to understand user behavior at large. In particular, questions interpreting a user's action in the context of his/her online friends and correlating the actions of socially connected users, become highly interesting.

There has been some theoretical and empirical work on how a user's actions can be correlated to his/her social affiliations. Backstrom et al. [1] examined the membership problem in an online community. They observed correlation between the action of a user joining an online community and the number of friends who are already members of that community. Marlow et al. [5] considered the tag usage problem in Flickr and studied the set of tags placed by a user and those placed by the friends of the user. They exhibited a correlation between social connectivity and tag vocabulary.

While these studies have established the existence of correlation between user actions and social affiliations, they do not address the *source* of the correlation. Causes of correlation in social networks can be categorized into roughly three types. The first is *influence* (also known as *induction*), where the action of a user is triggered by one of his/her friend's recent actions. An example of this scenario is when a user buys a product because one of his/her friends has recently bought the same product. The second is *homophily*, which means that individuals often befriend others who are similar to them, and hence perform similar actions. For example, two individuals who own Xboxes are more likely to become friends due to the common interest. The third is *environment* (also known as *confounding* factors or *external influence*), where external factors are correlated both with the event that two individuals become friends and also with their actions. For example, two friends are likely to live in the same city, and therefore to post pictures of the same landmarks in an online photo sharing system.

From a practical point of view, identifying situations where social influence is the source of correlation is important. In the presence of social influence, an idea, norm of behavior, or a product *diffuses* through the social network like an epidemic. A marketing firm, for example, can use this information to design viral marketing campaigns or give out coupons to *influential* nodes in the network, or a system

designer can take advantage of this information in order to induce the users to follow a desired mode of behavior. There has already been significant research on methods for designing strategies to leverage social influence in such systems [3] and on the effect of influence on the growth pattern of new products [8]. The main idea in all viral marketing strategies is essentially that in cases that influence between users is prevalent, careful targeting can have a cascading effect on the adoption of a product/technology. Therefore, being able to identify in which cases influence prevails is an important step to strategy design.

Our contributions. Given the significance of social influence, it is important to be able to test if a given social system exhibits signs of social influence. This is a particularly difficult problem in online settings where individuals are often anonymous and therefore it is impossible to control for all potential confounding factors. We overcome this problem by taking advantage of the availability of data about the timing of actions in online settings. We propose a statistical test (called the *shuffle test*) based on the intuition that if influence is not a likely source of correlation in a system, timing of actions should not matter, and therefore reshuffling the time stamps of the actions should not significantly change the amount of correlation. We prove that in a rather general model of homophily and confounding, this test succeeds in ruling out influence as the source of social correlation.

We also show the effectiveness of our test using simulations. Our test cases are based on a large social network from Flickr. We generate the action data randomly from a model with or without social influence, and run our test on this data set to decide whether the correlation is caused by influence. Our results show that in nearly all cases our algorithm succeeds in identifying the source of correlation. We also present results for another test (called the *edge-reversal* test) inspired by a recent study on the spread of obesity in real-world social networks [2].

Finally, we apply our algorithms on real tagging data in Flickr. Our results show that even though tagging behavior in this system exhibits a considerable degree of social correlation, this cannot be attributed to social influence.

Organization. In Section 2 we detail the different forms of social correlation. In Section 3 we describe our methodology, and present a theoretical analysis in a model of homophily and confounding. We describe our data generation models and present the results of simulations in Section 4. We describe our experiments on Flickr tags in Section 5.

2. MODELS OF SOCIAL CORRELATION

We study a setting where a group of individuals (also called agents or users) are nodes of a social network G. In general, G is a directed graph and is generated from an unknown probability distribution. We are concerned with individuals performing a certain *action* for the first time, e.g., purchasing a product, visiting a web-page, or tagging a photo with a particular tag.[1] After an agent performs the action, we say that the agent has become *active*. We observe the system for a certain time period $[0, T]$. Let W denote the set of agents that are active at the end of this time period.

Social correlation, i.e., correlation between the behavior of affiliated agents in a social network is a well-known phenomenon. Formally, this means that for two nodes u and v that are adjacent in G, the events that u becomes active is correlated with v becoming active. There are three primary explanations for this phenomenon: *homophily*, the *environment* (or *confounding* factors), and *social influence*.

Homophily. Homophily is the tendency of individuals to choose friends with similar characteristics [4, 6]. This is a pervasive phenomenon, and not surprisingly, leads to correlation between the actions of adjacent nodes in a social network. For example, one plausible hypothesis for why there is social correlation in membership in an online community is that individuals might know each other and become friends after joining the community. Mathematically, in a pure homophily model, the set W of active nodes is first selected according to some distribution, and then the graph G is picked from a distribution that depends on W.

Confounding. The second explanation for correlation between actions of adjacent agents in a social network is external influence from elements in the environment (also referred to as confounding factors), which are more likely to affect individuals that are located close to each other in the social network. Mathematically, this means that there is a confounding variable X, and both the network G and the set of active individuals W come from distributions correlated with X. For example, two individuals who live in the same city are more likely to become friends than two random individuals, and they are also more likely to take pictures of similar scenery and post them on Flickr with the same tag.

Note that there is a fine distinction between this explanation and homophily: homophily refers to situations where the set W affects individuals' choices to become friends, while in confounding, both the choices of individuals to become friends and their choice to become active are affected by the same unobserved variable. It is possible to distinguish between these models by looking at the time where the edges of G are established. The focus of this paper, however, is on distinguishing social influence from other types of social correlation. Therefore, we study a common generalization of the confounding and the homophily model as follows: first, the pair (G, W) is selected according to a *joint* probability distribution, and then the time of activation for individuals in W is picked i.i.d. according to a distribution $\mathcal{T}$ on $[0, T]$. We call this model the *correlation model*. The main assumption here is that the probability that an individual is active can be affected by whether their friends become active, but not by when they become active. This is in contrast with the influence model, as defined below.

Influence. The third, and perhaps the most consequential explanation for social correlation is social influence. This refers to the phenomenon that the action of individuals can induce their friends to act in a similar way. This can be through setting an example for their friends (as in the case of fashion), informing them about the action (as in the case of viral marketing), or increasing the value of an action for them (as in the case of adoption of a technology). Mathematically, this can be modeled as follows: first, the graph G is drawn according to some distribution. Then, in each of the time steps $1, \ldots, T$, each non-active agent decides whether

[1] In many cases, e.g., purchasing certain products or using certain tags, an individual might perform the action multiple times. We focus on the first time the action is performed by each individual, since subsequent occurrences of the same action by the same individual is often more dependent on the first occurrence than on the social network.

to become active. The probability of becoming active for each agent u is a function $p(x)$ of the number x of other agents v that have an edge to u and are already active.[2] Here, $p(\cdot)$ can be any increasing function, although later in the paper we consider a special class of functions that provides a good fit with the real data and also corresponds to a commonly used statistical model for estimating the probability of binary events, namely the logistic regression.

3. METHODOLOGY

In this section we present the methodology that we use to measure social correlation and test whether influence is a source of such correlation. We start in Section 3.1 by explaining how logistic regression can be used to quantify the extent of social correlation. In Section 3.2 we define the *shuffle test* for deciding if influence is a likely source of correlation, and prove that this test successfully rules out influence as the source of correlation in the correlation (confounding/homophily) model defined in Section 2. Finally, in Section 3.3 we define another test called the *edge-reversal* test, which we evaluate experimentally.

3.1 Measuring social correlation

The first step in our analysis is to obtain a measure of social correlation between the actions of an individual and that of her friends in the network. This measure is designed to recover the activation probability, assuming that the agents follow the influence model defined in Section 2.

Recall that in the influence model, each individual flips an independent coin in every time step to decide whether or not to become active. In principle, the probability of this coin can vary from agent to agent and from time to time; in the simplest model, which is the focus of most of this paper, we measure this probability as a function of only one variable: the number of already-active friends the agent has.[3] Note that the parameter we use is the number of friends that have become active at any earlier time step, as opposed to friends who have become active immediately before. This is because in online systems like Flickr actions are stored, and might be observed by others much later.

As it turns out, for most tags in the Flickr data set, a logistic function with the logarithm of the number of friends as the explanatory variable provides a good fit for the probability. Therefore, for simplicity and to reduce the possibility of overfitting, we use the logistic function with this variable, that is, we estimate the probability $p(a)$ of activation for an agent with a already-active friends as follows:[4]

$$p(a) = \frac{e^{\alpha \ln(a+1)+\beta}}{1 + e^{\alpha \ln(a+1)+\beta}}, \tag{1}$$

where α and β are coefficients. Equivalently,

$$\ln\left(\frac{p(a)}{1-p(a)}\right) = \alpha \ln(a+1) + \beta. \tag{2}$$

The coefficient α measures social correlation: a large value of α indicates a large degree of correlation. We estimate α, β using maximum likelihood logistic regression. More precisely, let $Y_{a,t}$ be the number of users who at the beginning of time t had a active friends and started using the tag at time t. Similarly, let $N_{a,t}$ be those users who at time t were inactive, had a active friends, but did not start using the tag (at time t). Finally, let $Y_a = \sum_t Y_{a,t}$, and $N_a = \sum_t N_{a,t}$. Then we compute the values of α and β that maximize the expression

$$\prod_a p(a)^{Y_a}(1-p(a))^{N_a}, \tag{3}$$

where $p(a)$ is defined in (1). Typically, the values of Y_a and N_a decrease quickly and lose their statistical significance as a grows. Therefore, for practical reasons, we may restrict the likelihood expression (3) to only all $a \leq R$, for a carefully chosen value of R, while we accumulate all the values corresponding to $a > R$ to Y_{R+1} and N_{R+1}. While in general there is no closed form solution, there are many software packages that can solve such a problem quite efficiently; we used Matlab's statistics toolbox in our experiments.

3.2 The shuffle test

In this section we introduce the *shuffle test* for identifying social influence. It is based on the idea that if influence does not play a role, even though an agent's probability of activation could depend on her friends, the timing of such activation should be *independent* of the timing of other agents.

Let G be the social network, and $W = \{w_1, \ldots, w_\ell\}$ be the set of users that are activated during the period $[0, T]$. Recall that in the correlation model, (G, W) is drawn from an arbitrary joint distribution. Assume that user w_i is first activated at time t_i. Using the method in Section 3.1, we compute Y_a and N_a, for $a \leq R$, where R is a constant, and use the maximum likelihood method to estimate α.

Next, we create a second problem instance with the same graph G and the same set W of active nodes, by picking a random permutation π of $\{1, \ldots, \ell\}$, and setting the time of activation of node w_i to $t'_i := t_{\pi(i)}$. Again we use the method in Section 3.1 to compute Y'_a and N'_a for $a \leq R$, and the social correlation coefficient α'. The shuffle test declares that the model exhibits no social influence if the values of α and α' are close to each other.

Intuitively, the reason that the shuffle test correctly rules out social influence in instances generated according to the correlation model is the following: in an instance generated from this model, the time stamps t_i are independent, identically distributed (i.i.d.) from a distribution $\mathcal{T}$ over $[0, T]$. The second instance constructed above only permutes all time stamps, and hence the new t'_i's are still i.i.d. from the same distribution $\mathcal{T}$. Therefore, the two instances come from the exact same distribution, and hence they should lead to the same *expected* social correlation coefficient α. The only thing that remains to be proven is that this coefficient is concentrated around its expectation (where the expectation is taken over the random choice of the time stamps, conditioning on a fixed choice of G and W). In the next section, we formalize this intuition, leading to Theorem 1.

[2]This model assumes that time progresses in discrete steps. A similar model with continuous time can be defined using the Poisson distribution.

[3]We also considered using the fraction of the total population that is active as another explanatory variable in our estimation on the Flickr data set, but the results indicated that this parameter is of no value: the corresponding coefficient is insignificant for almost all tags.

[4]We have also duplicated some of our experiments using a as the explanatory variable. The results are not qualitatively different, and almost always the likelihood of the fit is better with the logarithmic variable.

3.2.1 Theoretical analysis

To aid our analysis, we make three simplifying assumptions. First, we assume that the distribution $\mathcal{T}$ of the activation times is uniform over $[0, T]$. Second, we modify the test to pick each t'_i independently from $\mathcal{T}$, instead of using a permutation of the original time stamps. Neither of these assumptions is necessary, but it simplifies the arguments without substantively changing the techniques.

The third set of assumptions ensures that there are enough data to gather statistics. Let d_i^- (d_i^+) be the indegree (outdegree) of node w_i, and let d_i^{W-} (d_i^{W+}) be the indegree (outdegree) of node w_i in the subgraph induced by W (recall that W is the set of users that became active). Also, let $W' = \{w_1, \ldots, w_{\ell'}\}$, where $\ell' \geq \ell$ be the set of nodes in W and their neighbors (note that the first ℓ nodes are those in W). Then we make the following assumptions:

1. $\ell = \Theta(n)$.
2. $d_i^-, d_i^+ \leq d^{\max}$, for $i \leq \ell'$ and for some constant $d^{\max}$.
3. $\left|\{i : d_i^{W-} \geq R+1\}\right| = \Theta(n)$.

These assumptions are not the strictest possible for our results to hold, but they are nevertheless quite natural and simple to state. In particular, we make the first assumption only to simplify the notation (otherwise the results hold with probabilities that depend on ℓ and ℓ' instead of n).

THEOREM 1. *Let $G = (V, E)$ be a directed graph on n nodes and let $W = \{w_1, \ldots, w_\ell\} \subset V$ be the set of nodes that become active during the time period $[0, T]$. Assume that the activation time t_i of the node w_i is picked i.i.d. from the uniform distribution over $\{1, \ldots, T\}$, and assume that the three assumptions hold. Let α denote the social correlation coefficient computed using the method in Section 3.1. Then, with high probability[5] the value of α is close to its expectation, where the probabilities are over random choices of the activation times.*

PROOF. The main part of the proof is Lemma 2 where we show that the values of Y_a and N_a are concentrated. This is proved using concentration inequalities for martingales. We can then show (details deferred for the full version of the work) that when we apply logistic regression with inputs that are close to each other, the social correlation values α recovered are also close to each other. Therefore, with high probability the value of α recovered is close to its expectation whp. □

LEMMA 2. *Assume the conditions of Theorem 1, and let Y_a and N_a, $a \leq R+1$, defined as in Section 3.1. Then we have that Y_a and N_a are close to their expectations whp.*

PROOF. First we calculate $\mathrm{E}[Y_a]$, for a fixed a. We introduce some notation. Let $Y_a^i = 1$ if when node w_i used the tag had a active neighbors and 0 otherwise. Notice that we have $Y_a = \sum_{i=1}^{r} Y_a^i$. The probability that exactly a of the d_i^{W-} neighbors are active when node w_i used a tag is 0 if $d_i^{W-} < a$. Otherwise, if $a \leq R$, this probability is $1/(d_i^{W-}+1)$, since node w_i and its neighbors have the same probability to be the ath node among them that used the tag. Finally, if $a = R+1$ (recall that $R+1$ corresponds to the ensemble of all the values greater than R), then the probability is $(d_i^{W-} - R)/(d_i^{W-}+1)$.

Thus, we have

$$\mathrm{E}[Y_a] = \sum_{i=1}^{\ell} \mathrm{E}[Y_a^i] = \sum_{i: d_i^{W-} \geq a} \frac{1}{d_i^{W-}+1},$$

for $a \leq R$, and

$$\mathrm{E}[Y_a] = \sum_{i=1}^{\ell} \mathrm{E}[Y_a^i] = \sum_{i: d_i^{W-} \geq R+1} \frac{d_i^{W-} - R}{d_i^{W-}+1},$$

for $a = R+1$. One can verify that from our assumptions we have that both of these quantities are $\Theta(n)$.

Note that the terms are not independent. Thus, to show concentration, we will employ Azuma's inequality [7]. For a fixed a we define the (Doob's) martingale

$$X_i = \mathrm{E}[Y_a \mid t_1, t_2, \ldots, t_i].$$

We have that $X_0 = \mathrm{E}[Y_a]$ and $X_\ell = Y_a$. Note that we have that $|X_i - X_{i-1}| \leq d_i^{W+} + 1$, since a node affects only itself the nodes for which it is a contact. Then Azuma's inequality implies that

$$\Pr(|Y_a - \mathrm{E}[Y_a]| > \lambda) = \Pr(|X_\ell - X_0| > \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum (d_i^{W+}+1)^2}},$$

which is $o(1)$ for $\lambda = \omega(\sqrt{n})$.

To compute the value of $\mathrm{E}[N_a]$ we have to be a bit more careful, since a node can contribute multiple time periods to N_a. First, note that we have to count also the neighbors of the nodes in W. Recall that $W' = \{w_1, \ldots, w_{\ell'}\}$, is the set of active nodes and their neighbors.

Let us write $N_a = \sum_{i=1}^{\ell'} N_a^i$, where N_a^i counts the number of timesteps before node w_i becoming active (if at all) and had exactly a active contacts. Let us compute $\mathrm{E}[N_a^i]$, first for $i \leq \ell$. Of course, this equals 0 if $d_i^{W-} < a$. Otherwise, the expected time until one of the $d_i^{W-}+1$ nodes (w_i and its contacts) becomes activated is $T/(d_i^{W-}+2)$, thus $\mathrm{E}[N_0^i] = T(d_i^{W-}+2)$. With probability $d_i^{W-}/(d_i^{W-}+1)$ the first node is not w_i, hence we have $\mathrm{E}[N_1^i] = \frac{d_i^{W-}}{d_i^{W-}+1} \cdot \frac{T}{d_i^{W-}+2}$. More generally we get that

$$\mathrm{E}[N_a^i] = \frac{d_i^{W-} - a + 1}{d_i^{W-}+1} \cdot \frac{T}{d_i^{W-}+2},$$

for $a \leq R$, and

$$\mathrm{E}[N_a^i] = \frac{d_i^{W-} - R}{d_i^{W-}+1} \cdot \frac{d_i^{W-}+1-R}{2} T,$$

for $a = R+1$. (The first fraction is the probability that w_i becomes activated after $R+1$ neighbors, and then it is expected to arrive in the middle of the leftover period.)

For $i > \ell$ we can show with similar arguments that $\mathrm{E}[N_a^i] = 0$ if $d_i^{W-} < a$, otherwise

$$\mathrm{E}[N_a^i] = \frac{T}{d_i^{W-}+1},$$

for $a \leq R$, and

$$\mathrm{E}[N_a^i] = \frac{R+1}{d_i^{W-}+1} T,$$

[5]The term "with high probability," abbreviated whp., refers to an event that holds with probability that tends to 1 as $n \to \infty$.

for $a = R + 1$. By our assumptions for the graph we have that $N_a = \sum_{i=1}^{\ell'} N_a^i = \Theta(Tn)$.

Again we show concentration by using the Azuma inequality. We define

$$Z_i = \mathrm{E}[N_a \mid t_1, t_2, \ldots, t_i],$$

and notice that we have $|Z_i - Z_{i-1}| \leq T(d_i^+ + 1)$, with the same reasoning as previously. So we get that

$$\Pr(|N_a - \mathrm{E}[N_a]| > \lambda) = \Pr(|Z_r - Z_0| > \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum T^2(d_i^+ + 1)^2}},$$

which is $o(1)$ for $\lambda = \omega(T\sqrt{n})$. $\square$

3.2.2 Detecting influence

We showed that the values of α that we obtain with the correlation model are close to each other with high probability with and without the timestep shuffle. Now we contrast this with the influence model and we show that in the latter case the values of α that we compute with and without the timestep shuffle are in general different. We demonstrate this fact with a simple example. Consider a line graph with $n + 1$ nodes, $v_0, v_1, v_2, \ldots, v_n$, and edge set the $\{(v_i, v_{i+1});\ i = 0, 2, \ldots, n - 1\}$. For simplicity we assume that that node v_0 is has initially used a tag; this does not change the nature of our example. For some $p \in [0, 1]$, consider now the influence model with $\alpha = \log_2(p/(1 - p))$ and $\beta = 0$, and we observe the system for T time steps (with Tp being sufficiently small, say $Tp < n/2$). During the T steps, the nodes will start to use the tags from left to right, and at each step, the probability that the leftmost inactive node will become active equals p. Then at the end of the T steps, if the number of new active nodes is denoted by L, we have $\mathrm{E}[Y_1] = \mathrm{E}[L] = Tp$ and $\mathrm{E}[N_1] = T(1 - p)$.

Assume now that we perform the shuffle test. Then for $i = 1, \ldots, L$, let Y_1^i be 1 if node v_i became active after node v_{i-1}, and N_1^i the number of time steps that node v_i did not become active although node v_{i-1} was (0 if node v_{i-1} became active after node v_i). Then we have $Y_1 = \sum_{i=1}^{L} Y_1^i$ and $\mathrm{E}[Y_1] = \mathrm{E}[L/2] = Tp/2$, since the probability that node v_{i-1} becomes active before node v_i is $1/2$. Similarly, $N_1 = \sum_{i=1}^{L} N_1^i$ and

$$\mathrm{E}[N_1^v] = \sum_{i=1}^{T} \frac{1}{T} \cdot \frac{i}{Tp} \cdot \left(\frac{i}{2} - 1\right).$$

This follows since node v_i becomes active at time step i with probability $1/T$, the probability that node v_{i-1} arrives before is i/Tp and in that case the arrival time is uniformly distributed in $[0, i]$ so the expected number of times that node v_i does not become active is $i/2 - 1$. Therefore,

$$\mathrm{E}[N_1^v] = \frac{(T + 1)(2T - 5)}{12Tp},$$

and so,

$$\mathrm{E}[N_1] = \frac{(T + 1)(2T - 5)}{12}.$$

Hence we see that the input to the regression function is in general very different and as a result the values of α will in general be very different.

3.3 The edge-reversal test

In this section we introduce the second test for distinguishing influence similar to the one used in the obesity study [2]: we reverse the direction of all the edges and run logistic regression on the data using the new graph (which we call the *reverse graph*) as well[6]. Since other forms of social correlation (other than social influence) are only based on the fact that two friends often share common characteristics or are affected by the same external variables and are independent of which of these two individuals has named the other as a friend, we intuitively expect reversing the edges not to change our estimate of the social correlation significantly. On the other hand, social influence spreads in the direction specified by the edges of the graph, and hence reversing the edges should intuitively change the estimate of the correlation. We will test this hypothesis on several classes of instances generated using probabilistic models of different forms of social correlation.

4. SIMULATIONS

4.1 Generative models

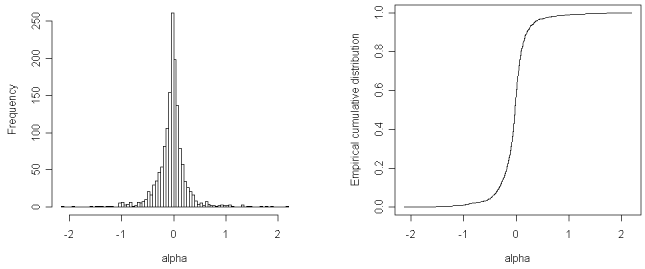
To verify the validity of the techniques described in Section 3, we define three generative models one corresponding to a setting where there is no social correlation, one corresponding to a setting that there is only social influence and one that there is social correlation but not influence. In each model, we will try to keep other aspects of the model as close to Flickr's data as possible. In particular, in all models the network (both number of users and connections) grows at the same rate as in the real Flickr data, and we will try to let the number of users that become active in each time step to follow the pattern corresponding to a tag in the real data.

The first model concerns a setting where there is no social correlation—influence or otherwise—in the pattern of activations. The second model is for a setting where influence is the only form of social correlation; this model is defined to match the logistic regression model described earlier. The third model seeks to capture situations where agents that are close to each other in the network are affected by the same external factors (the environment) that make them more likely to be activated. We now describe the models.

The no-correlation model. For every tag in the real data, we can generate a no-correlation instance as follows: the network grows exactly in the same way as in the real data. In each time step, we look at the real data to see how many new agents use the tag, and pick the same number of agents uniformly at random from the set of agents that have already joined the network and have not been picked yet.

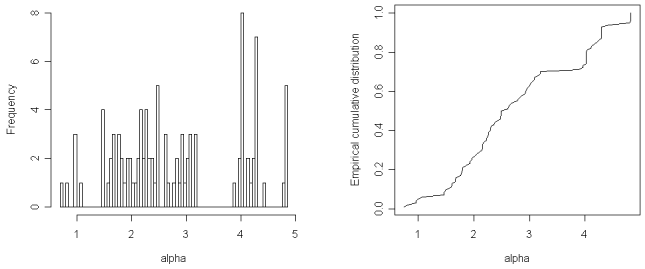
The influence model. This model is parameterized in terms of two parameters, α and β. The network, and the growth pattern of the network is kept as in the real data. In every time step, each node in the set of nodes that has joined the network but not activated yet flips a coin independently to decide if to become active in this time step. The probability of activation for this node is computed using (2), where a is the number of friends of this node that have become active in one of the previous time steps.

[6]Note that we are only able to use this test because in Flickr data set, a significant number of edges are directed.



(a) Histogram. (b) Empirical CDF.

Figure 1: Distribution of α for the no-correlation model.



(a) Histogram. (b) Empirical CDF.

Figure 2: Distribution of α for the influence model.

The correlation (no-influence) model. Again, we keep the network and the pattern of growth of the population the same as in the real data. The model is parameterized in terms of one parameter L, and follows the pattern of a given tag in the real data. Before generating the action data, we select a set S of nodes by sequentially picking a number of *centers* at random, and adding a ball of radius 2 around each to S.[7] We stop this process as soon as the size of S reaches the prespecified number L. Then, we generate the set of agents that become active in each time step in a manner similar to the one in the no-correlation model, except that in each time step we pick the set of agents to become active uniformly at random from S.

4.2 Measuring correlation

Our first set of experiments focuses on the measurement of correlation in the network. In Figure 1 we display the results of the application of logistic regression to the no-correlation model. We can see that the distribution of the values of α is centered at zero and most of the mass is around there.

In Figure 2 we can see the application of the logistic regression to the influence model. Recall from Section 3 that this model is based on the logistic function, which we are trying to fit. Not surprisingly, we recover the values of α that we set in our model. Thus, Figure 2 essentially displays those values of α.

Finally, in Figure 3 we see the results in the correlation model. Note that here as well the values of α that we recover are positive.

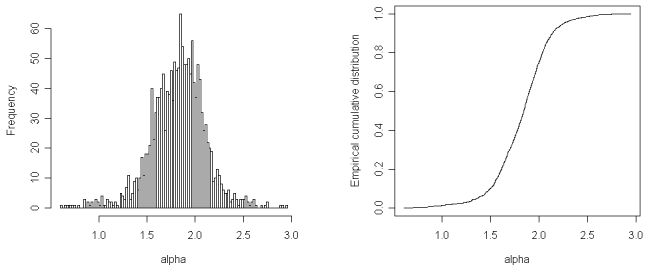
4.3 Distinguishing influence

After establishing the presence of correlation in users' behavior, we turn to tests for the source of this correlation. First we apply the shuffle test and then we turn to the edge-reversal test.

4.3.1 Shuffle test

Let us first observe the influence model, where the values of α with the original tagging times are high. From the intuition gained in Section 3.2, we expect to see those values to decrease, when we shuffle the tagging timesteps. In

[7]We have chosen a radius of 2 here since because the network is highly connected, a ball of radius 3 can become very large, while a ball of radius 1 only consists of the neighbors of a node, which is often too small.



(a) Histogram. (b) Empirical CDF.

Figure 3: Distribution of α for the correlation model.

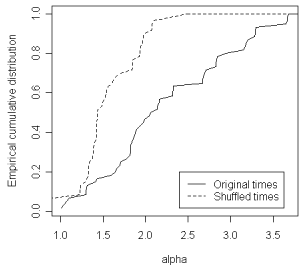
Figure 4(a) we can observe the results for some of the tags. Notice how the cumulative density function (CDF) is shifted to the left, which means that when we reverse the edges the value of α decreases. In Figure 4(b) we can see the values in absolute terms.

Now we switch to the correlation model. According to the analytical findings of Section 3.2, the values of α that we obtain with and without the shuffling should not differ with high probability. Figure 5 confirms our analytical findings and shows that for almost all tags the values of α retrieved are very close with and without the shuffle.

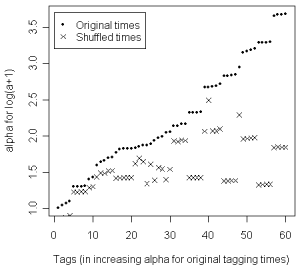
4.3.2 Edge-reversal test

Now we present the results of our second influence-detection test, the edge reversal, confirming the results of the previous section. First we apply it to the influence model, depicting the results in Figure 6. Similarly to the previous test, there is a significant difference in the values of α in the forward and backward direction.

On the contrary, in the correlation model, as seen in Figure 7, the values of α essentially coincide. In Figure 7(a) we can notice that the two CDFs essentially coincide. In Figure 7(b) we see a more detailed picture. Here every point corresponds to a tag, and the graph shows the value of α in the network versus the value of α in the network with the edges reversed. Take notice of the proximity of the points to the line $y = x$.

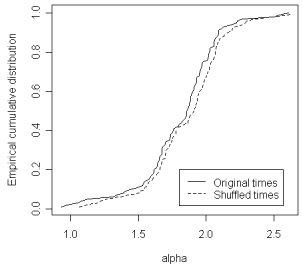


(a) Empirical probability density.

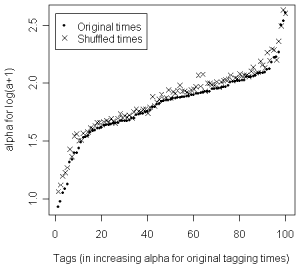


(b) α of original and shuffled tagging timesteps.

Figure 4: Shuffle test for the influence model.

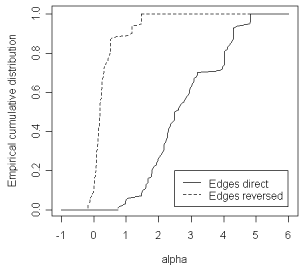


(a) Empirical probability density.

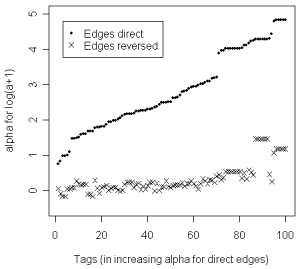


(b) α of original and shuffled tagging timesteps.

Figure 5: Shuffle test for the correlation model.

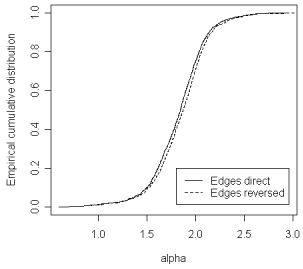


(a) Empirical probability density.

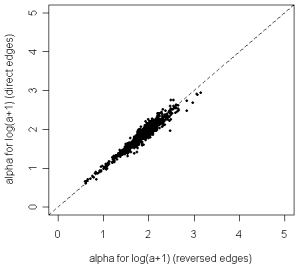


(b) α of direct vs. reversed edges.

Figure 6: Edge-reversal test for the influence model.



(a) Empirical probability density.



(b) α of direct vs. reversed edges.

Figure 7: Edge-reversal test for the correlation model.

5. EXPERIMENTS ON REAL DATA

After verifying that our techniques are effective for the simulated data, we apply them on real-world data, namely on the Flickr social network. First we describe the data set. Then we show that there is positive correlation in the users' behavior. Finally, we address the issue of the source of correlation. We apply the tests of Section 3, and we conclude that influence is not a likely source of the correlation.

5.1 The Flickr dataset

We analyzed the tagging behavior of users for a period of 16 months. The final number of users was about 800K. Since the majority of users did not exhibit any tagging behavior at all, we restricted our attention to the set of users who have tagged any photo with any tag, which is about 340K users. Looking at this subgraph at the end of the 16-month period, the size of the giant component is 160K users, the second one has size 16, and there are 165K isolated users.

The number of directed edges between the users is 2.8M and, on the average, for a given user u, the proportion of u's contacts that do not have u as a contact is 28.5%. In Figure 8 we depict the size of the subgraph that we analyze as a function of time. (The growth rate of the entire network exhibits a very similar behavior.)

Out of a collection of about 10K tags that users had used, we selected a set of $1,700$, and analyzed each of them independently. We selected tags of various types (event, colors, objects, etc.), various numbers of users (most of them were used by more than $1,000$ users), and various growth patterns: bursty (e.g., "halloween," "katrina"), smooth (e.g., "photos,") and periodic (e.g., "moon").

5.2 Measuring correlation

First we confirm the existence of correlation in the Flickr data set as expected. In Figure 9 we can see the distribution of α along the tags of Flickr. Note that for almost all the tags the value is higher than 1, suggesting that correlation is prevalent in users' tagging activities for almost all the tags. This correlation is not necessarily due to social influence; we examine this issue next.

5.3 Distinguishing influence

After establishing the presence of correlation in users' behavior, we turn to the test for the source of this correlation.

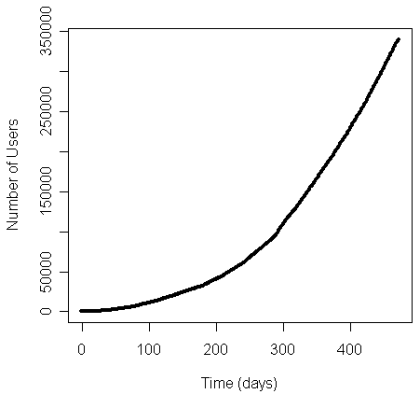
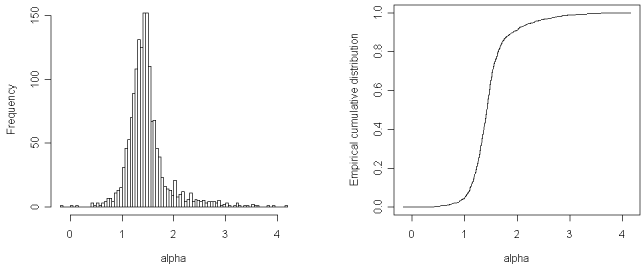
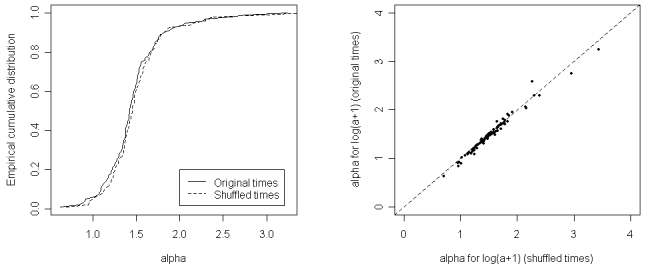


Figure 8: Growth of the Flickr network.



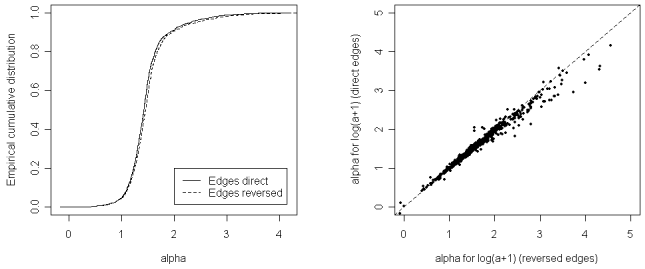
(a) Histogram. (b) Empirical CDF.

Figure 9: Distribution of α for the Flickr social network.



(a) Empirical probability density. (b) α of original timesteps vs. shuffled timesteps.

Figure 10: Shuffle test for the Flickr social network.



(a) Empirical probability density. (b) α of direct vs. reversed edges.

Figure 11: Edge-reversal test for the Flickr social network.

First we apply the shuffle test and then we turn to the edge-reversal test.

In Figure 10 we show the results of applying the shuffle test on the Flickr data set. In Figure 10(a), notice that the two cumulative distribution functions essentially coincide. It seems that the correlation that we observed in Section 5.2 cannot be attributed to influence. This indicates that either users do not tend to browse their contacts' photos to a large extent, or even when they browse, they do not tend to start using the tags they see.

In Figure 10(b) we see more details. Once again, every point corresponds to a tag, and the graph shows the value of α in the Flickr network versus the value of α in the network with the edges reversed. As before, notice the striking proximity of the points to the line $y = x$.

Finally, in Figure 11 we observe the results of applying the edge-reversal test to the Flickr network, which once again confirms all our previous observations.

5.4 Some influence in Flickr

While it is true that influence does not play an important role in users' tagging behavior in Flickr, we can actually discover that there is some limited effect by looking at the *difference* between similar tags. As a concrete example, consider the tag "graffiti"; the difference between the values of α in the two edge directions is essentially 0. A lot of users used the misspelled tag "grafitti." Here the difference turns out to

be slightly larger (still small though). It is easy to imagine that indeed there is some propagation of the misspelled version. (The analogy with the TA who grades two homeworks with the same mistakes should make this concept clear!) Finally, with a third, even less common spelling ("graffitti"), the difference increased yet more.

6. CONCLUSIONS

In this paper we applied statistical analysis on the data from a large social system in order to identify and measure social influence as a source of correlation between the actions of individuals with social ties. This is an instance of the age-old problem of distinguishing correlation from causation. This problem is very difficult in general; however, in our case, we used the availability of data about the time-step of each action, as well as asymmetric social ties between the agents in order to study this problem.

There are still many interesting open directions left for future research. First, our techniques provide only a qualitative indication of the existence of influence and not a quantitative measure. Furthermore, we do not provide any formal verification of our results. For example, is it indeed the case that in Flickr users' tagging behavior, influence has a limited role? Or, can we pinpoint social networks and behaviors where influence is indeed prevalent and verify our tests? Also, what happens when different sources of social correlation are present, as is usually the case? All these important questions might be tricky to answer and probably require the design of controlled user experiments. Furthermore, it would be very interesting to extend our theoretical model for distinguishing between social influence and other forms of correlation in social networks. Under what conditions the information about the time step of events is enough to achieve this goal? How can the pattern of the "spread" of an action be used to identify social influence even in a setting where all social ties are symmetric? How can we find an "influential" node just by looking at the data about the spread of an action? Given the great potential of viral marketing technologies to shape the future of marketing on the Internet, this and many other related questions are of tremendous practical value.

Acknowledgments

We thank Alex Jaffe, Malcolm Slaney, and Duncan Watts for invaluable discussions, as well as the anonymous reviewers for insightful comments.

7. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *12th KDD*, pages 44–54, 2006.

[2] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357(4):370–379, 2007.

[3] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *9th KDD*, pages 137–146, 2003.

[4] P. Lazarsfeld and R. K. Merton. Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel, and C. H. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, 1954.

[5] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, Flickr, academic article, to read. In *17th HYPERTEXT*, pages 31–40, 2006.

[6] M. McPherson, L. Smith-Lovin1, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[7] M. Mitzenmacher and E. Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[8] P. Young. The diffusion of innovations in social networks. In L. E. Blume and S. N. Durlauf, editors, *The Economy as a Complex Evolving System*, volume III. Oxford University Press, 2003.