Information Transfer in Social Media

Greg Ver Steeg
Information Sciences Institute,
University of Southern California
Marina del Rey, California
gregv@isi.edu

Aram Galstyan
Information Sciences Institute,
University of Southern California
Marina del Rey, California
galstyan@isi.edu

ABSTRACT

Recent research has explored the increasingly important role of social media by examining the dynamics of individual and group behavior, characterizing patterns of information diffusion, and identifying influential individuals. In this paper we suggest a measure of causal relationships between nodes based on the information–theoretic notion of transfer entropy, or information transfer. This theoretically grounded measure is based on dynamic information, captures fine–grain notions of influence, and admits a natural, predictive interpretation. Networks inferred by transfer entropy can differ significantly from static friendship networks because most friendship links are not useful for predicting future dynamics. We demonstrate through analysis of synthetic and real–world data that transfer entropy reveals meaningful hidden network structures. In addition to altering our notion of who is influential, transfer entropy allows us to differentiate between weak influence over large groups and strong influence over small groups.

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory**]: Information Theory; H.3.4 [**Systems and Software**]: Information networks; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

entropy, prediction, causality, social networks, spam, point processes

1. INTRODUCTION

Recent years have witnessed an explosive growth of various social media sites such as online social networks, discussion forums and message boards, and inter-linked blogs. For researchers, social media serves as a fertile ground for examining social interactions on an unprecedented scale [7]. One important problem is the characterization and identification of *influentials*, which can be defined as users who influence the behavior of large numbers of other users. Recent work on influence propagation has used numerous characterizations of influentials based on topological centrality measures such as Pagerank score [19, 15]. To characterize influence in Twitter, researchers have suggested various measures based on number of followers, mentions, retweets [8], and Pagerank of follower network [16]. It has been observed, however, that the purely structural measures of influence can be misleading [10] and high popularity does not necessarily imply high influence [25]. More recent work has attempted to introduce dynamic information through the size of the information cascades [3] and influence–passivity score [25]. One serious drawback of existing methods is that they are based on explicit causal knowledge (i.e., A responds to B), whereas for many data sets such knowledge is not available and needs to be discovered.

Here we suggest a model–free approach to uncovering causal relationships and identifying influential users based on their capacity to *predict* the behavior of other users, through the information-theoretic notion of *transfer entropy*, interchangeably referred to as information transfer. In a nutshell, transfer entropy between two stochastic processes characterizes the reduction of uncertainty in one process due to the knowledge of the other process; a mathematical definition is given below. Transfer entropy can be thought of as a nonlinear generalization of Granger causality [5], and has been used in computational neuroscience, e.g., for examining causal relationships in cortical neurons [12]. In contrast to other correlation measures such as mutual information, transfer entropy is asymmetric and allows differentiation in the direction of information flow. Furthermore, whereas most existing studies are concerned with *aggregate* measures of influence, the approach outlined here allows more fine–grained analysis of information diffusion by analyzing information transfer on each existing link in the network. Finally, our approach is model-free. Information–theoretic measures allow us to statistically characterize our uncertainty without making assumptions about human behavior.

The rest of this paper is organized as follows. We begin by describing the basic intuition and mathematics behind the information transfer, and briefly mention computational issues of the approach. In Section 3.1 we present results of our simulation with synthetically generated data, where we thoroughly examine how the information transfer depends on various characteristics of the data generating process. In Section 3.2 we present our results on real-world data extracted from user activities on Twitter. We conclude the paper with related work in Section 4 and a discussion of results and future work in Section 5.

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

2. TRANSFER ENTROPY

2.1 Notation

For each user, X, we record the history of activity, e.g., timing of tweets, as a sequence of times as

$$S_X = \{t_j : 0 < t_1 < t_2 \ldots\}.$$

In general, we assume each user's activity is described by some stochastic point process. We are limited by finite data to consider finite temporal resolution, so we introduce a binned random variable that tells us whether an event occurred in some time interval.

$$B_X(a,b) \equiv \begin{cases} 1 & \text{if } \exists t_j \in S_X \cap (b,a], \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

If we observe the actions of a user for some long period of time T, we can define probabilities over these coarse-grained variables. Fix $\delta \in \mathcal{R}$, then

$$P(B_X(t,t-\delta) = X_t) \equiv \frac{1}{T-\delta}\int_{\delta}^{T} dt[B_i(t,t-\delta) = X_t],$$

where [] denotes an Iverson bracket.[1] Similarly, we could define a joint probability distribution over a sequence of adjacent bins,

$$P(B_X(t,t-\delta_0) = X_t, B_X(t-\delta_0, t-\delta_0-\delta_1) = X_{t-1}, \ldots),$$

with widths $\delta_0, \delta_1, \ldots, \delta_k \in \mathcal{R}$. We will omit the binning function for succinctness, $P(X_t, X_{t-1}, \ldots, X_{t-k})$. We can write this even more compactly by defining

$$X_t^{(t-k)} \equiv \{X_t, \ldots, X_{t-k}\}.$$

The dynamics of a user may depend on users they are linked to in some unknown, arbitrary way. Therefore, for two users X and Y, with activities recorded by S_X, S_Y, we define a joint probability distribution using a common set of bins denoted with widths $\delta_0, \delta_1, \ldots, \delta_k$ as $P(X_t^{(t-k)}, Y_t^{(t-k)})$.

Conditional and marginal probability distributions are defined in the usual way and we use the standard definition for conditional entropy. For discrete random variables A, B distributed according to $P(A,B)$,

$$H(A|B) = -\sum_{A,B} P(A,B)\log P(A|B). \quad (2)$$

We will use the logarithm in base two and report entropies in bits. In practice, the probability distribution $P(A,B)$ will generally be estimated according to observed frequency counts. Of course, this can lead to sampling problems which we discuss in Section 2.3.

2.2 Definition of transfer entropy

The *transfer entropy* introduced in [26] is defined as

$$T_{X\to Y} = H(Y_t|Y_{t-1}^{(t-k)}) - H(Y_t|Y_{t-1}^{(t-k)}, X_{t-1}^{(t-l)}) \quad (3)$$

The first term represents our uncertainty about Y_t given Y's history only. The second term represents the smaller uncertainty when we know X's history as well. Thus, transfer entropy explicitly describes the reduction of uncertainty in Y_t due to knowledge of X's recent activity. For simplicity, we take $l = k$ from here on.

[1]The Iverson bracket is equal to 1 when the logical condition enclosed is true and 0 otherwise.

We offer two more intuitive interpretations of the information transfer. The first fruitful comparison is with Granger causality[13], which states that X is Granger causal to Y if Y is better predicted from a model that includes X's and Y's histories than from one that includes Y's history only. In particular, linear regression models are typically used in the comparison. For Gaussian random variables, Granger causality is equivalent to information transfer [5]. In principle, conditional entropies should capture arbitrary nonlinear relationships in the signal.

Information transfer can also be written as the mutual information between Y's present and X's past, conditioned on Y's past.

$$T_{X\to Y} = H(Y_t : X_{t-1}^{(t-k)}|Y_{t-1}^{(t-k)})$$

Because of the conditioning on Y's past, the transfer entropy is asymmetric, as opposed to standard mutual information, and thus better suited for characterizing directed information transfer. This captures the intuition that we are only interested in information about Y that is explained by X but cannot be explained by Y's own history.

2.3 Sampling problems and solutions

The use of information–theoretic techniques to analyze real-world point processes has been studied almost exclusively in the context of neural activity[28]. Therefore, it is in this literature that the problems associated with estimating entropies for sparse point process data have been explored most thoroughly. The fundamental problem is that, in the absence of sufficient data, estimating entropies from probability distributions based on binned frequencies leads to systematic bias [20]. Intuitively, if we have k bins of history then we need $O(2^k)$ pieces of data in order to sample all possible histories.

A variety of remedies are available and we make use of several. Of course, we can simply pick k to be small, but if we take too coarse grain of a sample, we might omit relevant, predictive information. The most obvious solution is to restrict ourselves to situations where we have adequate data. In the subsequent analysis, we filter out users that are below a certain activity level. In practice, however, raising our activity threshold high enough to guarantee convergence of entropies would eliminate almost all users from our dataset.

The next remedy to apply is to estimate the average magnitude of the systematic bias that results from using sparse data and subtract it from our estimate. When we calculate the conditional entropies in Eq. 3, we subtract out the Panzeri-Treves bias estimate[21]. For discrete random variables A, B, the first order estimate in the bias of $H(A|B)$ due to finite sample effects is,

$$BIAS[H(A|B)] = \frac{-1}{2N\ln(2)}\sum_{b\in dom(B)}(N_b - 1), \quad (4)$$

where N is the number of joint samples of A, B and N_b is the number of unique variables $a \in dom(A)$ observed for a given $b \in dom(B)$. Therefore, when we calculate conditional entropies as in Eq. 2, we will always subtract the bias estimate in Eq. 4. Figure 2 illustrates the effect of this bias correction as a function of amount of data collected.

The definition in Eq. 3 implicitly depends on bin widths specified by the δ_i's. The simplest procedure, and the one taken in the neural spike train literature, is to set all the

bins to have equal width. We have a great deal of pre-existing empirical knowledge about human activity that can help us improve on this method. Many studies have shown that humans exhibit a heavy tail in the distribution of their response times to communications[4]. This implies that bins accounting for recent activity should be narrower while bins accounting for older activity can be wider. We can even base these bin widths on measured response times, if such data is available. Using more informative bins means we can use fewer bins, reducing the effect of sampling problems.

A final technique to reduce bias is discussed in [28] and uses a class of binless entropy estimators. These techniques carry their own mathematical difficulties and we will not consider them here. With these tools in hand, we can proceed to use information transfer to analyze user activity in social media.

3. RESULTS

In this section we report the results of our experiments with both synthetic data and real world data from Twitter. The ultimate goal is to infer information transfer between agents in the network by analyzing their patterns of activity. Patterns of activity could include many things including timing, content, and medium of messages. We focus only on the timing of activity on Twitter (tweeting of URLs). In principle, our analysis could be extended to include more complex information, but, as discussed, this would require either more data or better methods for dealing with sparse data.

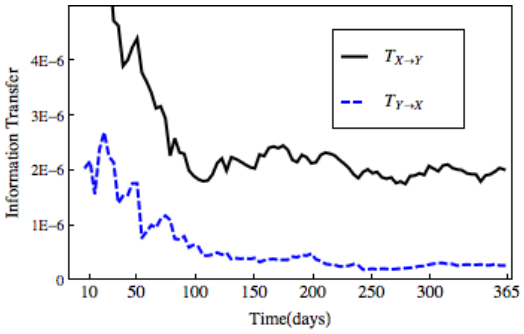

Figure 1: If we have influence from $X \to Y$ but not vice versa, the asymmetry in the information transfer correctly reflects the direction of influence. Information transfer plotted for a single pair of users.

We test and validate our ability to infer information transfer from patterns of activity in two ways. First, while our information–theoretic analysis of social network data uses only timing of activity, the data includes unique identifiers allowing us to track the flow of information through the network. On Twitter, we track specific URLs. We can use the spread of these trackable pieces of information to confirm that the information transfer inferred solely from the timing of activity corresponds to actual exchanges of information.

For the synthetic data, we dictate that an agent's activity depends on its neighbors' activity in some fixed way. This allows us to check how well information transfer recovers the hidden dependence structure from activity patterns alone. For instance, even without knowing anything about the network structure, we find that a sufficient amount of data allows perfect reconstruction of the underlying network.

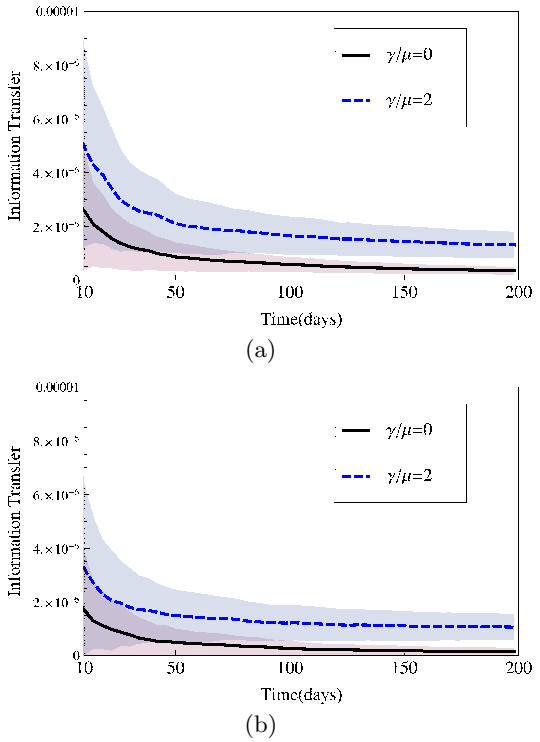

Figure 2: Mean and std for the estimate of information transfer averaging over 200 pairs of users with $\gamma/\mu = 0, 2$ as a function of time. (a) Results without correcting for bias and (b) with Panzeri-Treves bias correction[21].

3.1 Experiments with synthetic data

To form a better understanding of different factors impacting information transfer, we performed extensive experiments with synthetically generated data. Ideally, we would like our synthetic data to reflect, in a tunable way, the challenges we face with real world data. These challenges include a long tail for human response times, heterogeneous response to neighbors' activity, background noise affecting node dynamics, incorrect data, and insufficient data. We explore these challenges first for a pair of nodes, and then for an entire network.

We model user activity as a coupled, non-homogeneous Poisson point process, similar to a typical self-exciting point process [18]. Suppose that we have two nodes and a single link from $X \to Y$. We can characterize Y's activity in terms of a time-dependent rate. We define $S_X^t \equiv S_X \cap [0, t)$, that is, the activity for X until time t.

$$\lambda_Y(t|S_X^t) = \mu + \gamma \sum_{t_i \in S_X^t} g(t - t_i) \tag{5}$$

The probability of a spike in an interval of time $(t, t + dt)$ is just $\lambda_Y(t)dt$. The first term, μ, represents a constant rate of background activity. The second term represents a time-dependent increase in the rate of activity in response to activity from a neighbor. The strength of influence of X is

parametrized by γ. In practice, we will set the background rate equal to a constant and vary the relative strength γ/μ through the parameter γ. The time dependence of the influence is captured by the function g. We set

$$g(\Delta t) = \min\left(1, \left(\frac{1 \text{ hour}}{\Delta t}\right)^3\right)$$

to roughly match the observed distribution of re–tweet times in our Twitter dataset. This also agrees with the observed fact that the distribution of human response times are characterized by a long tail[4].

Along with a causal network, Eq. 5 defines a generative model for point process activity. We can efficiently generate activity according to this model using the thinning method discussed in [18]. We vary the total amount of data by fixing the background rate $\mu = 1$ event/day and varying the total amount of observation time, T. Equivalently, we could have fixed T and varied the rate of activity. After fixing the parameters, we can generate data and then use that data to infer the appropriate probabilities to calculate information transfer according to Eq. 3.

As discussed in Sec. 2.3, we take a variety of measures to ensure good estimation. In this case, we directly control the amount of data through the parameter T. For the bin widths we choose $\delta_0 = 1$ sec, fixing the finest temporal resolution. For the history we choose wider bin widths for less recent history. In the synthetic examples we take the past three hours of history into account by choosing $\delta_1 = 1 \text{ hour}, \delta_2 = 2 \text{ hours}$. Also, it should be assumed that the Panzeri-Treves bias estimate has been taken into account, except in Figure 2(a) where we compare results without bias correction.

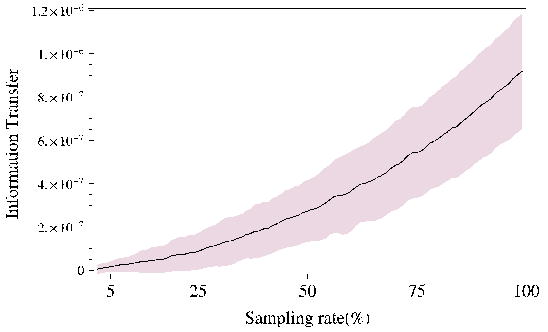

Figure 3: A summary of the mean and std of the inferred value of $T_{X\to Y}$ averaged over 200 trials as a function of the sampling rate, with $T = 500$ days and $\gamma/\mu = 2$.

Note that in the example in Eq. 5, we have allowed X to affect Y, but not vice versa. As a first test we can generate some data for a pair of users and then compare $T_{X\to Y}$ and $T_{Y\to X}$. In Figure 1, we compare these two quantities when $\gamma/\mu = 2$ as a function of the total observation time T.

In Figure 2 we examine the accuracy and convergence of information transfer estimates as a function of time both with and without bias correction. We ran 200 trials and plot the mean and standard deviation of the information transfer estimate at each time step. Clearly, there is a systematically high estimate in the low sampling regime, but, even in that case, higher influence leads to a higher information transfer

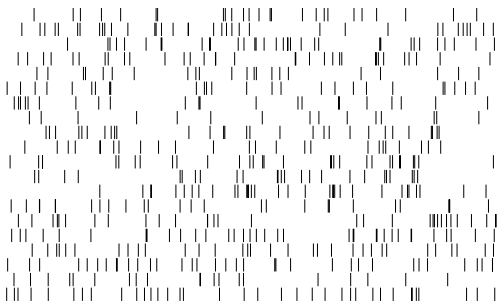

Figure 4: Each row represents a different user. Each line represents an event for that user over a time period of thirty days. With enough data we could calculate the information transfer between each pair of users and recover the unknown network structure exactly.

on average. The Panzeri-Treves bias correction drastically reduces, but does not completely eliminate, this systematic error.

Next, we consider the same scenario, where we generate X, Y according to some stochastic process, but now imagine that we do not see all activity. That is, what if we do not see every event due to limited sampling? This is often the case, for instance, with Twitter data, where researchers typically have access to only a small fraction of all tweets, ranging from $1\% - 20\%$. So we set a sampling parameter f, and say that for each $t_i \in S_X$, we only keep that event with probability f. A summary of how the final transfer entropy, $T_{X\to Y}$, depends on the sampling rate, f, is given in Figure 3. We show the results after 500 days to guarantee enough data to be very close to convergence. We see that sampling drastically reduces the inferred transfer entropy, destroying our ability to deduce flow of information.

So far, we have only considered two nodes with a single link between them. Now, we want to consider a directed, causal network of N nodes, with some arbitrary connectivity pattern. We consider a similar stochastic model as defined in Eq. 5, except now we denote the set of Y's neighbors (i.e., people who can influence Y) as $\mathcal{N}(Y)$.

$$\lambda_Y(t|S^t_{\mathcal{N}(Y)}) = \mu + \sum_{X\in\mathcal{N}(Y)} \gamma_X \sum_{t_i\in S^t_X} g(t - t_i) \qquad (6)$$

To begin we imagine $\gamma_X = \gamma$ for all neighbors, but in general a node may be affected more strongly by some neighbors than others. A sample of activity generated according to this model is given in Figure 4.

The challenge is to take the information given by the activity and recover the underlying graph structure. For each pair of nodes, X, Y, we calculate $T_{X\to Y}$. Then we pick some threshold T_0, and if $T_{X\to Y} > T_0$, we consider there to be an edge from $X \to Y$, otherwise not. We could check our true positive rate and false positive rate as a function of T_0, as shown in Figure 5(a), for $N = 20, \gamma/\mu = 1.0$ and time $= 450$ days. We show an example of the recovered versus actual network in Figure 5(b), using a threshold picked according to F-measure.

The previous example was chosen to show what kinds of errors arise given a weak signal. In general, with either enough data or strong enough influence, we can perfectly

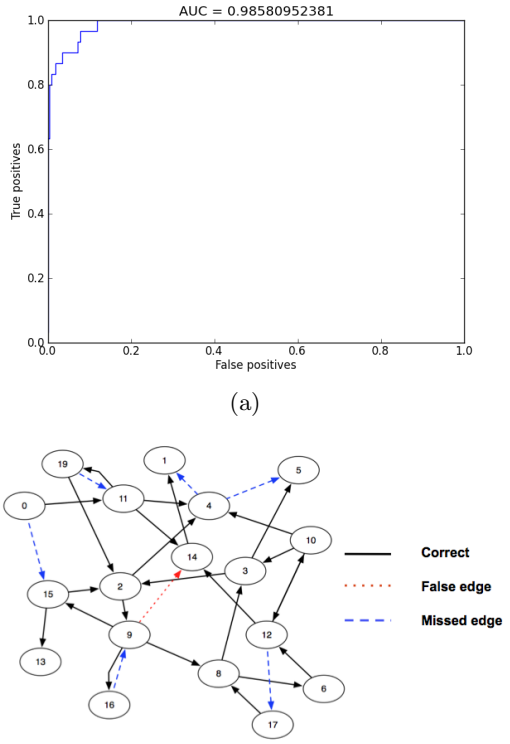

Figure 5: (a) ROC curve and (b) transfer–entropy induced graph for the synthetically generated data described in the text. Threshold is chosen according to F-measure. Black solid lines correspond to true positives, red dashed lines to false positives and blue dotted lines to false negatives.

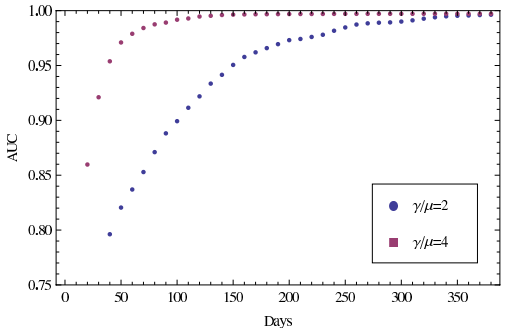

Figure 6: AUC of the network inferred using transfer entropy as a function of T, with $\gamma/\mu = 2, 4$.

recover the underlying graph structure. If we consider the area under the ROC curve (AUC), as in Figure 5(a), then an AUC of 1 corresponds to perfect reconstruction of the graph. We summarize the AUCs for random networks with $N = 20$ and $\langle k \rangle = 3$, while varying T and γ/μ in Figure 6.

As a final experiment, we can consider the effect of allowing different γ between different pairs of nodes. Again, we set $T = 500$ days to ensure that we are close to convergence. Figure 7 shows that transfer entropy is able to recover the relative influence well. However, we see that it makes more sense to consider links that have generally higher information transfer, while specific rankings of edges with similar information transfer probably has little meaning.

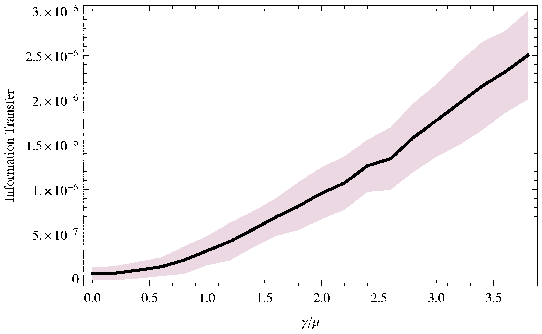

Figure 7: Information transfer between pairs of nodes for varying γ/μ with $T = 500$ days. The black line corresponds to the mean information transfer for a given γ/μ and the shaded region denotes the standard deviation after 100 trials.

In principle, there are many other effects we could have considered to make a more realistic synthetic model. Background and influence rates should vary for different individuals. There may be periodicity defined by daily, weekly, and monthly cycles. However, because information transfer makes no model assumptions, it is relatively insensitive to such details. The main constraint is data, which is why we focused on sensitivity to amount and quality of observations.

3.2 Results for Twitter dataset

Twitter is a popular micro-blogging service. As of July 2011, users send 200 million tweets per day. Twitter has become an important tool for researchers both due to the volume of activity and because of the easily available tools for data collection. Twitter's "Gardenhose" API, allows access to $20\% - 30\%$ of all tweets.

Unfortunately, as discussed in Sec. 3.1, filtering of data can lead to a drastic reduction in the measured information transfer. Instead, the Gardenhose API was used to identify URLs being tweeted. Then, the search API was used to find all mentions of these URLs in any tweets by any users. In this way, the random filtering limitation is avoided, while we restrict ourselves to the domain of URL posting. Additionally, each URL corresponds to a unique piece of information whose movement through the network can be traced. The data also includes the full social network among "active users", in this case, anyone who tweeted a URL in the three week collection period. The data we used was collected in the fall of 2010 [10]. The dataset included about 70 thousand distinct URLs, 3.5 million tweets, and 800 thousand

users. We further filtered our results to "very active" users, namely, users who tweeted at least 10 URLs during this time period.

Before we can calculate transfer entropy as presented in Eq. 3, we need to specify the relevant bin widths. We take the finest resolution to be $\delta_0 = 1$ second, the same resolution as presented by the Twitter API. For binning of the history, we used distribution of observed re-tweet response times to motivate a choice of $\delta_1 = 10$ min, $\delta_2 = 2$ hours, $\delta_3 = 24$ hours. Although we saw a long tail of re-tweet times stretching into days, our data were insufficient to include this weak effect. By limiting ourselves to only three bins, we only have to sample over 8 possible histories. Note that the activity is for any tweeting of URLs; our calculations do not make use of the information encoded in the URL. We then calculate the transfer entropy between each pair of users who are connected.

The result of this procedure is the construction of a directed, weighted graph, where each edge in the original directed graph is now labeled by the calculated transfer entropy. We can now compare standard measures of influence to measures based on this weighted graph. The simplest measure of influence on static graphs is to count the number of followers a user has. This ignores the fact that not all followers are the same, nor do followers react in the same way to different people that they follow. For instance, it may be that a recommendation from a close friend is worth more to a person than the same recommendation from five acquaintances. This problem is only exacerbated by the recent emergence of "followers for pay" services, which seek to artificially inflate the number of followers to your Twitter account. In Figure 8, we explore the comparison between out degree and transfer entropy and we find that although on average people with more followers have more transfer entropy, two people with the same number of followers may have vastly different influence as measured by transfer entropy.

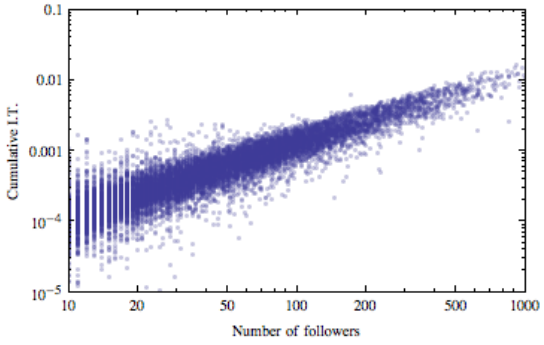

Figure 8: For each user, we compare the number of their followers to their cumulative outgoing information transfer. Note that the outgoing information transfer may differ by an order of magnitude for people with the same number of followers.

To verify that transfer entropy is a meaningful quantity, we could test how well the transfer entropy, based only on the timing of activity, matches the measured flow of information, as determined by tracing specific URLs. To that end, for each pair of connected users, $X \to Y$, we count how many specific URLs were first tweeted by X and then subsequently re-tweeted by Y. This number is compared to the transfer entropy in Figure 9. The existence of even a weak correlation is surprising considering the limited amount of data and the fact the transfer entropy is not making use of URL or re-tweet information at all. We also note that while a high number of re-tweets implies high information transfer, a low number of re-tweets is uncorrelated with information transfer. This makes sense because information transfer measures influence that is not necessarily in the form of re-tweets; we will give some examples below.

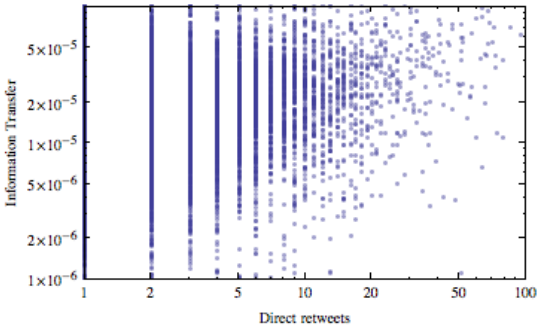

Figure 9: If the number of URLs that were first tweeted by user X and subsequently tweeted by X's follower, Y, is high, then the calculated transfer entropy between X and Y is also high, even though transfer entropy is calculated only from the timing of activity, without regard for specific URLs. Note that the converse is not true. Pearson's correlation coefficient is 0.22.

Table 1 shows the edges with the highest information transfer. These accounts are all solely for the purpose of promotion. Looking at the top example, for instance, reveals that these two accounts will tweet exactly the same message within a few seconds of each other (in a random order, hence both orderings show up in the list). In the text of their tweets neither account uses re-tweets or an "@" for attribution. Twitter specifically forbids indiscriminate automatic re-tweets and has a policy against duplicate accounts. Many of the accounts on this list have since been banned by Twitter. Figure 10 gives some examples of how the activity looks for pairs of users with high information transfer. We picked one example with high information transfer ($188 \cdot 10^{-6}$) but lower activity for comparison. Unfortunately, the differences in tweeting times are typically too small to be distinguishable on the plot.

To see more complex examples, we restrict ourselves to the top 1000 edges according to information transfer. Then we look at the largest connected components. The largest component involved 600 users in Brazil, most of whom had multiple tweets of the form "BOMBE O SEU TWITTER, COM MILHARES DE NOVOS FOLLOWERS, ATRAVES DO SITE: http://? #QueroSeguidores", where "?" was a frequently changing URL. Google translates this as "Pump up your Twitter, get thousands of new followers, link to this site: http://? #IWantFollowers." Clicking on some of these links suggests that this a "followback" service. You agree to follow previous users who have signed up and in return other users of the service follow your account. It also appears from the text that you are required to re-tweet the link to get your

User	Follower	I.T.($\cdot 10^{-6}$)
Free2BurnMusic	Free2Burn	4328
Earn_Cash_Today	income_ideas	1159
BuzTweet_com	scate	1006
Free2Burn	Free2BurnMusic	939
Kamagra_drug2	sogradrug3	929
sougolinkjp	sogolinksite	903
kcal_bot	FF_kcal_bot	902
nr1topforex	nr1forexmoney	795
wpthemeworld	wpthememarket	709
viagrakusurida	viagrakusuride	679
BoogieFonzareli	Nyce_Hunnies	668
A_tango	kobuntango	662
Kamagra_drug2	sogra_drug3	638
dti_affiliate	kekkonjyoho	630
Best_of_Deals	Orbilook_SMI	621
viagrakusurida	kamagra_100mg3	561
kcal_bot	Family_Mart	542
kamagra_100mg3	viagrakusuride	535
viagra_drug	baiagura_drug	532
kcal_bot	Seven_Eleven_	530

Table 1: List of edges with highest information transfer. All are promotional accounts and many of the accounts have been banned since the data were collected.

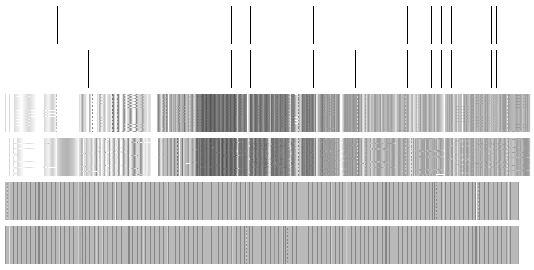

Figure 10: Timing of tweets for three pairs of users with high information transfer. From bottom to top: Free2BurnMusic, Free2Burn, Earn_Cash_Today, income_ideas, Random_Nyanko, amechihua.

followers. Some other examples of high information transfer clusters are shown in Figure 11.

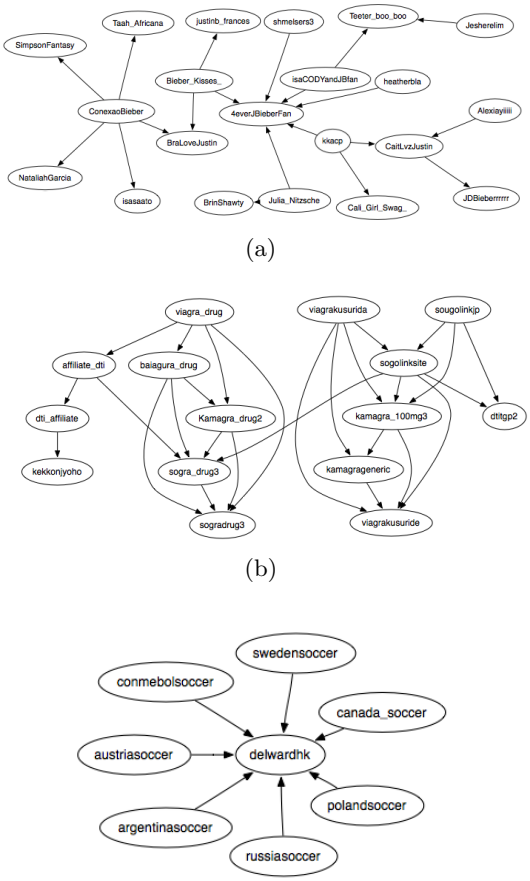

Figure 11: (a) This cluster appears to be non-automated, and revolves around fandom of singer Justin Bieber. (b) The cluster of drug spam accounts. (c) An account which aggregates soccer news by following and re-tweeting different regional soccer accounts.

We consider another advantage of measuring influence through information transfer by looking at two users who had almost the same outgoing transfer entropy (~ 0.025, in the top 20 for individuals in our dataset), but vastly different behavior of followers. The first Twitter account is SouljaBoy, a prominent American rapper who is also very active in social media. The second account is "silva_marina", the Twitter account of Marina Silva, a popular Brazilian politician. This data was taken during the run up to the Brazilian presidential election, in which Marina Silva was a candidate; she received 19.4% of the popular vote. At first it seems surprising that SouljaBoy, who has six times the followers, should have a similar outgoing transfer entropy to a politician known mostly in one country and with fewer than a million Twitter followers. On the other hand, Figure 12 reveals the reason for this disparity. Marina Silva may have fewer followers, but her effect on them tended to be much stronger. Marina Silva's activity tended to be a better pre-

dictor of her followers' behavior than Soulja Boy's activity was for his followers.

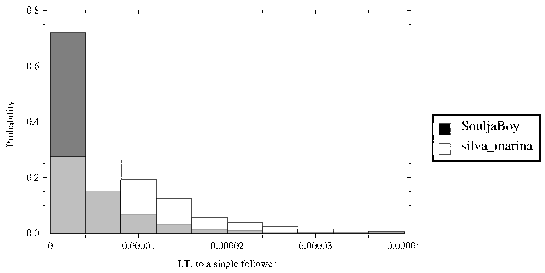

Figure 12: A histogram showing the probability distribution of outgoing transfer entropy to followers of two different Twitter accounts.

The strength of Marina Silva's influence along with the serendipitous timing before the Brazilian elections suggests another intriguing possibility. It seems likely that not only does transfer entropy vary for different followers, it may vary over time as well. This suggests that a dynamic estimate of information transfer could detect changes in the importance of individuals in the network.

4. RELATED WORK

The general problem of identifying influence in social networks has traditionally centered on topological measures like PageRank[19, 15] and other centrality measures[11]. Recent work has highlighted the insufficiency of structural measures alone to predict dynamics. For instance, in [16] they find that the number of followers and PageRank for a user on Twitter does not correlate well with the number of retweets that user can inspire. Retweets and mentions of a user were considered in [8] with the conclusion that these dynamic measures are not related to the number of followers for that user. In [3], they find that past influence on immediate neighbors is the most important predictor of how large of an information cascade will result from a user's tweet. In [25], they develop a heuristic which calculates influence taking into account the fact that many users are passive and will not be moved to retweet under any circumstances. In our framework, passivity is taken into account automatically: users who are always inactive have low entropy, which upper bounds the possible information transfer to them.

While we have chosen transfer entropy as a nonlinear, model–free approach to time series, Granger causality discovers causal relationships in time series data via linear regression [13]. It corresponds to transfer entropy for Gaussian random variables [5]. Granger causality has been successfully applied to the problem of prediction for temporal graphs [2, 17].

Using information–theoretic techniques for inference problems has a long, rich history[9], but we will briefly mention just the most mathematically analogous results. Neuroscientists have found that the electrical signal for individual neurons are characterized by sequences of spikes at specific times [6]. The goal is to "decode" the spikes to understand how the brain represents information. To that end, neuroscientists have turned to entropy as the most general way to represent information [28]. The general strategy is to present different random stimuli and then measure the mutual information between the stimuli and the pattern of neural spikes. However, applying transfer entropy in particular has also been considered in [12], which showed that transfer entropy could produce information about directed flow of information that was not captured by standard correlation measures.

Most of the results invoking transfer entropy consider continuous random processes [14], rather than stochastic point processes. For instance, an EEG produces a continuous signal by monitoring the average electrical response for thousands or millions of neurons at a time and was considered in the context of transfer entropy in [23].

Although we have taken information transfer to mean transfer entropy as defined in Eq. 3, one can certainly imagine other information–theoretic measures. Schreiber's original paper considers a comparison with time delayed mutual information[26]. A different causal measure called the directed information was used in [24] to recover causal networks among neurons. Some of the subtle differences in these measures are considered in [14].

5. CONCLUSION

We have presented a novel information–theoretic approach for measuring influence. In contrast to previous studies that focused on aggregate measures of influence, the transfer entropy used here allows us to characterize and quantify the causal information flow for any pair of users. For a small number of users, this can allow us to reconstruct the network of connections from user activity alone. For large networks, this allows us to identify the most important links in the network.

The method used here for calculating information transfer did not require any explicit causal knowledge in the form of re-tweets or other textual information. On the one hand, this may be an advantage in situations where such information is either missing or misleading, as was the case in the example for marketers on Twitter. On the other hand, we may be neglecting valuable information, and in the future we would like to incorporate textual information in more sophisticated ways but still within an information–theoretic approach. Although this should be straightforward in principle, in practice entropy based approaches require large amounts of data. More complex signals require a commensurate increase in data. Therefore, the other main thrust of future work should be towards reducing data required for entropy estimation, either through better bias correction or through binless approaches[28]. Another approach is to adaptively pick bin resolution depending on a user's activity level. This would also avoid coarse-graining of information for extremely active users.

Because this measure has a rigorous interpretation in terms of predictability, it allows us to easily understand results that might otherwise seem anomalous. For instance, in one example we found that Marina Silva, the Brazilian presidential candidate, had high information transfer both to and from a Brazilian news service. Neither Twitter account ever retweeted or explicitly mentioned a tweet of the other. However, there was an external cause, the upcoming debates and elections, that explained both of their activities. Without knowing this external cause, it is entirely consistent to say that either user's activity could help you predict the others. In fact, it may be possible to use this bi-directional predictability to identify external causes in the first place.

The possibility of hidden causes highlights a shortcoming of using transfer entropy — or Granger causality — as a causal measure. A more stringent definition of causality would use the effect of randomized interventions as a causal measure [22], but the possibilities for intervening in a social network are limited [1]. Although in some cases it may be feasible to rule out hidden causes even with purely observational studies, the requirements may be too stringent to allow results at the level of individuals [27]. Transfer entropy strikes a balance between making few implicit assumptions about the underlying process while nevertheless allowing us to make useful statements about the dynamics of specific edges in the network.

Another result that is easy to understand in the context of predictability is the high incidence of "spam" in our results. This is no surprise since a large amount of spam is produced by automated systems and these systems are intrinsically very predictable. Although identifying spam is a natural application of our analysis, some human behavior stood out as well. Diehard fandom also leads to quite predictable behavior. We also expect conversations to have a regular temporal activity pattern that could be easily identified by transfer entropy. However, only collecting tweets with URLs probably included more promotional tweets and excluded casual conversations.

Because information-theoretic measures make no assumptions about the structure of the underlying process, they are likely to be most valuable for data exploration. Looking at Fig. 10, we can instantly see that spammers have a specific temporal pattern that can likely be captured with a simple activity model. Transfer entropy in this case could be viewed as a tool to highlight important phenomena so that they can be modeled explicitly. A deeper, iterative analysis could then use information transfer to search for other patterns that are not explained by this initial model.

Many existing notions of influence are static, ad hoc, or only apply in aggregate. Information transfer is a rigorously defined, dynamic measure capable of capturing fine-grain notions of influence and admitting a straightforward predictive interpretation. Many of the mathematical techniques necessary have already been developed in the neuroscience literature and we have shown how to usefully adapt them to a social media context.

Acknowledgments

We would like to thank Armen Allahverdyan, Peter Boothe, and Winter Mason for useful discussions. This research was supported in part by the National Science Foundation under grant No. 0916534, US AFOSR MURI grant No. FA9550-10-1-0569, and DARPA grant No. W911NF–12–1–0034.

6. REFERENCES

[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS*, 106(51):21544+, December 2009.

[2] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical Granger methods. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-07)*, 2007.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[4] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.

[5] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.*, 103:238701, Dec 2009.

[6] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461, May 2004.

[7] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81(2):591–646, May 2009.

[8] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM-10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.

[9] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[10] R. Ghosh and K. Lerman. Predicting influential users in online social networks. In *Proc. KDD workshop on Social Network Analysis (SNAKDD)*, May 2010.

[11] R. Ghosh and K. Lerman. Parameterized centrality metric for network analysis. *Physical Review E*, 83(6):066118, June 2011.

[12] B. Gourevitch and J. J. Eggermont. Evaluating information transfer between auditory cortical neurons. *J Neurophysiol*, 97(3):2533–43, 2007.

[13] C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[14] K. Hlavackovaschindler, M. Palus, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, Mar. 2007.

[15] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 271–279, New York, NY, USA, 2003. ACM.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[17] A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical Granger modeling methods for temporal causal modeling. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-09)*, 2009.

[18] Y. Ogata. Seismicity analysis through point-process modeling: A review. *Pure appl. geophys.*, 155:471–507, 1999.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[20] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S.

Petersen. Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology*, 98(3):1064–1072, 2007.

[21] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107, 1996.

[22] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2009.

[23] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger. Performance of different synchronization measures in real data: A case study on electroencephalographic signals. *Phys. Rev. E*, 65:041903, Mar 2002.

[24] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.*, 30:17–44, February 2011.

[25] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. *Social Science Research Network Working Paper Series*, Aug. 2010.

[26] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461–464, Jul 2000.

[27] G. Ver Steeg and A. Galstyan. A sequence of relaxations constraining hidden variable models. In *Proc. of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

[28] J. D. Victor. Approaches to information-theoretic analysis of neural activity. *Biological Theory*, 1(3):302–316, 2006.