Use of Transfer Entropy to Infer Relationships from Behavior

Travis L. Bauer
Sandia National Laboratories
Albuquerque, NM USA
tlbauer@sandia.gov

Rich Colbaugh
Sandia National Laboratories
Albuquerque, NM USA
rcolbau@sandia.gov

Kristin Glass
New Mexico Tech
Socorro, NM USA
kglass@icasa.nmt.edu

David Schnizlein
Sandia National Laboratories
Shoreview, MN USA
dpschni@sandia.gov

ABSTRACT

This paper discusses the use of transfer entropy to infer relationships among entities. This is useful when one wants to understand relationships among entities but can only observe their behavior, but not direct interactions with one another. This is the kind of environment prevelant in network monitoring, where one can observe behavior coming into and leaving a network from many different hosts, but cannot directly observe which hosts are related to one another. In this paper, we show that networks of individuals inferred using the transfer entropy of Wikipedia editing behavior predicts observed "ground truth" social networks. At low levels of recall, transfer entropy can extract these social networks with a precision approximately 20 times higher than would be expected by chance. We'll discuss the algorithm, the data set, and various parameter considerations when attempting to apply this algorithm to a data set.

1. BACKGROUND

Studying Wikipedia data to learn about social networks and collaboration is well established. The detailed edit history that Wikipedia maintains provides a rich record of interactions. Some research tries to understand the interaction dynamics among Wikipedia editors [4]. Other research focuses on understanding group dynamics such as conflict [9]. Some research focuses on understanding individuals and their social roles[8].

Crandall et. al.[2] comes closest to this work, studying the relationship between the social networks observable in Wikipedia and edit behavior. They showed that people become more similar to each other shortly before they form a visible social connection on Wikipedia and continue to get closer after that. What this research does not show is whether the social network is predictable.

This work goes beyond current studies in Wikipedia analysis by studying if the temporal behavior of editors is predictive of whether they are also in a social network. We show that this is the case.

Transfer entropy[6] is a concept for discovering coupling among pairs of entities generating time series of events. For two entities, J and I, transfer entropy from J to I is the amount of additional information (reduction of uncertainty) about I's behavior provided by J's behavior than is provided by I's behavior alone.

Transfer Entropy has been applied successfully to a wide variety of problems. Some of these problems are biological, such as reverse engineering regulatory networks[7], examining relationships among firing in neurons[3], or inferring information transfer in calcium signaling in biochemical pathways[5]. Non biological applications of transfer entropy include stock market analysis, discovering which companies' behavior are most predictive of changes in the stock market[1].

Equation 1 shows the definition of transfer entropy.

$$T_{J->I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) log \frac{p(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})} \tag{1}$$

The notation $i_n^{(k)}$ refers to the combination of event occurrences for i over the k time steps leading up to and including time n. Transfer entropy takes into account, over every time $n+1$, whether I generated an event, which of the prior k points in time that I generated an event and which of the last l points in time that J generated an event. Transfer entropy is the sum, over all $n+1$, of the product of how probable that combination of items are by the log of the conditional probability of I's behavior is given both I and J's behavior over the probability of I's behavior alone.

2. APPLYING TRANSFER ENTROPY

To apply transfer entropy to some time series, there are four decisions to make.

1. Discretization of the time series

2. l, the window for J

3. k, the window for I

4. What constitutes an event

The algorithm assumes that one can iterate over the series in discrete steps and that one can look "back in time" some discrete number of steps in order to determine if I or J has performed some event. This decision determines the coarseness of the analysis.

The window l indicates how far back in time we look at J's behavior. If we choose a large value of l, this means that we believe I's response to J might be influenced by more of J's history. It also assumes that when J influences I it is not issuing many events unrelated to that relationships. On the other hand, choosing a small value for l indicates that when J influence I, it does so right away and that the information J is providing doesn't extend back in time.

The window k indicates how much of I's immediate history influences its own behavior. A large value for k would indicate that I's behavior coheres over a large period of time. This would be appropriate, for example, in situations where some event would generally happen in isolation or if it would start a sequence of events.

Finally, a decision must be made regarding what constitutes an event. Transfer entropy works over a symbolic time series. The events have to be discrete. In work, are expressed as binary: an event either occurs or it doesn't. Other discrete representations are possible as well.

Given the right data set, all four of these decisions can be determined empirically.

3. WIKIPEDIA DATA OVERVIEW

Wikipedia[1] is a free online encyclopedia written by volunteers. It has articles in more than 280 languages and has over three million articles in English alone. Wikipedia is freely editable and has tens of thousands of active contributors working on millions of pages.[2]

3.1 User Edits

In order to facilitate collaboration and to address vandalism, Wikipedia includes a detailed change history. Every page maintains its own change history and this change history is accessible from the web page where the article is viewed.

In addition to the content of each revision, the timestamp of each revision is recorded, along with the user who made the revision. Revisions by unregistered users are recorded by their IP address. Also, a comment is optionally included with each revision that lets the user indicate something about the changes they made. In some cases, these are standard comments that can be inserted for things like disambiguation or reversions. Finally, of course, the content of the edit itself is also recorded.

Frequently, a group of people will edit a single page in coordination with one another. The information about their interactions with one another are not recorded in the page revision history itself. Page revisions record individual people making discrete edits to single pages. Social interactions are not recorded as part of the revision record.

3.2 Social Networking

Each page in Wikipedia also has a "Talk" page where users have discussions. Talk pages work like any other page in Wikipedia. People can freely edit them. These revisions are recorded like the revisions to a regular article.

In addition to the article pages, each registered user has a "User" page. A user page is like any other Wikipedia article except that it has a prefix "User:" followed by the username. On this page, a user can write things about themselves and their contributions to Wikipedia. More importantly for this research, each User page has a "Talk" page with a revision history. On this page, other users can post comments directly to each other. Each comment is recorded in the Wikipedia history as a revision.

---

[1] http://en.wikipedia.org/wiki/Wikipedia
[2] http://en.wikipedia.org/wiki/Wikipedia:About

Number of Revisions	322M
Number of Registered Users	4.5M
Number of Pages	25.2M
Date Range	Jan 2001 - Oct 2011

Table 1: Records used in this work

3.3 Bulk Availability of Edit History

The revision information in Wikipedia is accessible through a web browser. However it is also available for bulk download from Wikipedia (http://en.wikipedia.org/wiki/Wikipedia:Database_download). The downloads include not only article content, but also all the content of every revision, including all of the metadata discussed above. These downloads come as compressed XML files that can be parsed and put into a database. The research discussed in this paper is drawn from most of the October 7, 2011 Wikipedia dump. A summary of the data extracted is show in table 1.

4. THE EXPERIMENT

The specific question we asked was whether the entity behavior of two people provide information about whether they are interacting with each other in the talk pages. This is only a partial social network so the predictions would be imperfect even if transfer entropy perfectly predicted the network. But it does provide partial ground truth against which we can measure the algorithms performance. The social network in the talk pages is a lower bound on the real social network.

4.1 Defining the Transfer Entropy parameters

As described earlier, four decisions need to be made about the data in order to apply transfer entropy. The first decision was how to discretize the time series. Timestamps on Wikipedia revisions are recorded to the millisecond. We experimented with a number of different time periods in day increments. The smallest time period we tried, one day, resulted in the best performance. So a single day is the unit of time into which the series is split.

The next decision is l, the number of time steps to look back in history of the entity that might provide the additional information on the others' behavior. This is a value that we determined empirically. We found that using small values provided the best results (we used one). We will revisit this in the discussion section.

We need to make the same decision of k, the number of time steps to look back in the behavior of the entity whose behavior is affected by the other entity. We discovered this value empirically as well. It turns out that larger values (we used 5) provided the best results. We will revision this in the discussion section as well. In general, the best values will depend on the nature of the data set.

The final parameter is what constitutes an event. Because of the richness of the data in the revision history, there is a lot that can be done with this decision. One could take into account the size of the edit, the type of edit based on the comment, or the page on which the edit occurred. For this paper, we did not take any of these into account. We determined events based on the number of edits on a particular day. We found the best values empirically, and experimented with 1, 10, 20, 40, 60, 80, and 100.

4.2 Identifying the Users to consider and constructing the graphs

We chose a candidate set of users to compare to one another based on whether or not they had ever edited a particular page. We ran the test twice, once each for a set of users identified from two

different pages. We chose users this way to find a set of candidates, some of whom are likely to have communicated. It is important to note, however, that the page was used only for selecting a pool of candidates. Once the candidates were selected, the fact that they edited a common page was not included in the transfer entropy calculations. So we will refer to the user groups for the rest of the paper in terms of what page we used to find them, but it is important to note that the fact that they edited that page played no role in the analysis other than to identify the set of users.

The two pools of users came from two different pages, Anarchism and Elvis Presley. Both pages were heavily edited over the years and have a large pool of different kinds of users. The Elvis Presley page had 2,139 distinct users who made 26,126 edits[3] between Oct 13, 2001 and Oct 7, 2011. We removed any editors that made fewer than five revisions and more than 100,000 edits to Wikipedia (not just to the Elvis Presley page). That left a pool of 1,963 users.

The other pool of users we examines were those of the page Anarchism. There were 1,335 unique registered editors of the page who made 26,462 edits between Oct 11, 2001 and Oct 8, 2011. When removing the users who made less than five edits and more than 100,000, there was a pool of 1,218 users remaining.

The edges among the users in the extracted graph were computed using transfer entropy with the parameters described above. The transfer entropy between any two users I and J was computed as

$$max(T_{I->J}, T_{J->I}) \qquad (2)$$

The "ground truth" graphs was constructed where the links between two individuals was the sum of the number of times each had edited the other's user talk page. We followed Crandall et. al.[2] in believing that the user talk page was a better indicator than whether the users had co-edited the talk page of some other part of Wikipedia, although it would be straight forward to run these same tests with different assumptions.

In the Elvis talk graph, there were 20,101 edges. Given 1,963 users, there were 1,926,684 possibilities, so the probability of randomly selecting an edge in the talk graph is about 1%. For Anarchism, there were 9,650 edges in the talk graph. Given 1,218 users, there were 741,762 possible edges. The probability of selecting one randomly is about 1.3 percent.

5. RESULTS

For each set of users, we comprehensively estimated the presence of edges among all of them using transfer entropy. We then put the edges in rank order from highest transfer entropy to lowest. This gave us a rank ordered set of potential edges against which we could compare the "ground truth" from the talk graphs.

Figures 1 and 2 show the precision/recall analysis comparing the extracted transfer entropy edges to the talk graph edges for both the Anarchism and the Elvis users.

Among the different possible threshold values for the minimum number of edits in a day that constitute an even, it appears that between 40 and 80 was providing the highest overall results for both pages. At a threshold of 100, the performance starts dropping.

In both tests, the precision is approximately 20 times better than random at low levels of recall (0.1) and stay substantially above random all way to through 0.7 recall.

These results should be seen as a lower bound on how well the algorithm is performing. Remember that User talk pages are only
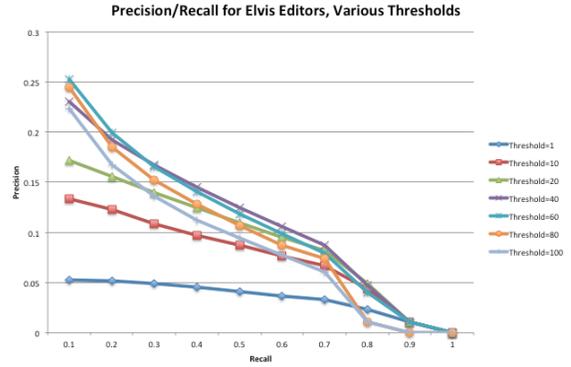

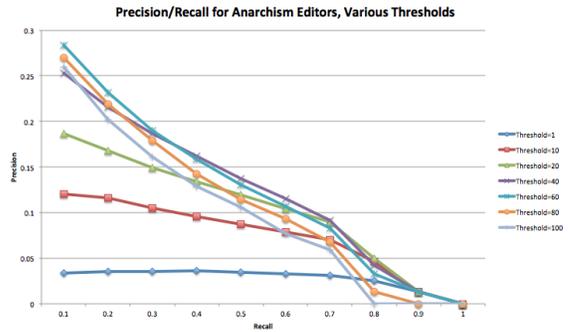**Figure 1: Precision/Recall Analysis for the Elvis Users**

Figure 2: Precision/Recall Analysis for the Anarchism Users

one way for people to communicate with one another. Some of the edges that received a high transfer entropy score were likely reflective of people communicating with one another as well, just using other means than the User talk pages.

6. DISCUSSION

The main finding from these studies is that at low levels of recall, transfer entropy extracts relationships from individual behavior with a precision of approximately 20 times better than random guessing and it performs substantially better than random at levels of recall up to 0.7. This is especially notable because the algorithm does not take into account any direct relationship information for these computations. It only considers whether a user made at least the threshold number of edits to some set of pages on a given day. When comparing the users to each other, it does not consider whether they edited the same page or not.

When computing the transfer entropy from I to J, it appears that using a small window for I is best and using a larger window for J. Intuitively, this means that if J is reacting to I, J is likely to do it quickly. Extending the window for I introduce more noise than signal. This makes sense especially because Wikipedia is page centric rather than user centric. It is not easy to for a user to know all of the edits that some other user is making at once. However, a user will obviously be aware of all their own edits. We've seen in the data that there are frequent bursts of editing activity for an individual. What this likely means is that if I have some time to edit Wikipedia, I'm likely to edit for some stretch of days. So whether I've edited in the last few days has predictive power. If some other user's behavior is going to increase the predictive power of my own

---

[3]Note that we only count here the number of edits made by registered users.

behavior, it will be because I'm responding to something specific from the user and not simply the fact that the user is making some edit at all.

However, whether this will be the case for other data sets is an empirical question. Different circumstances and the ease with which individuals can get information about their peers would affect the specific values that should be used to get maximum precision.

7. CONCLUSION

This paper has shown that transfer entropy can be used to extract social network information from behavior data at least 20 times better than random from Wikipedia. Choosing parameters for the algorithm affect performance, especially the determination of what constitutes an event. Further studies could take into account more of the rich information in the Wikipedia edit history to improve these values.

8. REFERENCES

[1] S. K. Baek, W.-S. Jung, O. Kwon, and H.-T. Moon. Transfer Entropy Analysis of the Stock Market. *ArXiv Physics e-prints*, September 2005.

[2] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 160–168, New York, NY, USA, 2008. ACM.

[3] Boris GourÂŐvitch and Jos J. Eggermont. Evaluating information transfer between auditory cortical neurons. *J Neurophysiol*, 97:2533–2543, 2007.

[4] P Massa. Social networks of wikipedia. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 221–230, New York, NY, 2011.

[5] Jurgen Pahle, Anne Green, C Jane Dixon, and Ursula Kummer. Information transfer in signaling pathways: A study using coupled simulated and experimental data. *BMC Bioinformatics*, 9(1):139, 2008.

[6] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85, Part 2:464–464, 2000.

[7] Thai Quang Tung, Taewoo Ryu, K.H. Lee, and Doheon Lee. Inferring gene regulatory networks from microarray time series data using transfer entropy. In *Computer-Based Medical Systems, 2007. CBMS '07. Twentieth IEEE International Symposium on*, pages 383 –388, june 2007.

[8] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011. ACM.

[9] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in wikipedia. *PLoS ONE*, 7(6):e38869, 06 2012.