

# Multivariate Embedding based Causality Detection with Short Time Series

Chuan Luo<sup>1</sup> Daniel Zeng<sup>1,2</sup>

<sup>1</sup>State Key Laboratory of Management and Control for Complex Systems,  
Institute of Automation, Chinese Academy of Sciences, Beijing, China

<sup>2</sup>Department of Management Information Systems, University of Arizona, Tucson, USA  
chuan.luo@ia.ac.cn, zeng@email.arizona.edu

**Abstract**—Existing causal inference methods for social media usually rely on limited explicit causal context, pre-assume certain user interaction model, or neglect the nonlinear nature of social interaction, which could lead to bias estimations of causality. Besides, they often require sufficiently long time series to achieve reasonable results. Here we propose to take advantage of multivariate embedding to perform causality detection in social media. Experimental results show the efficacy of the proposed approach in causality detection and user behavior prediction in social media.

**Keywords**—causality detection; user influence; nonlinear dynamic system; multivariate embedding

## I. INTRODUCTION

Recent years have witnessed an explosive growth of various social media sites such as online social networks, blogs, microblogs, social news websites and virtual social worlds. This gives researchers a great opportunity to study social interactions on an unprecedented scale. Since individual behaviors have an effect on the decisions of friends in a social network, the knowledge of who-influences-whom has enormous implications in security informatics [1]. Unfortunately, the influence structure among users is usually unknown and unobserved. Although some methods have been proposed to address this question, they often require sufficiently long time series, pre-assumption of certain particular user interaction model, a linear view of social interaction, or a purely stochastic view of social system. However, in many cases we only have very short time series data and the pre-assumption of particular user interaction model usually fails to capture the complexity of human behavior. Besides, nonlinearity is ubiquitous in nature [2] and user interaction may have nonlinear property.

In this paper, from the perspective of nonlinear dynamic systems, we propose a novel causality detection approach based on multivariate embedding to tackle this problem.

## II. RELATED WORK

To infer the underlying causal structure, current studies [3], [4] are built on the a priori assumptions of certain information diffusion model to describe the user interaction mechanism. However, these models may fail to capture the complexity of user interaction. Later, some researchers start to take advantage of information theory measures, such as Granger causality or its variants like Transfer Entropy [5], to infer the causal structure among users. The drawback of such works is that they usually adopt a linear view of social

interaction and a purely stochastic view of social system, which does not fit for the real world. Inspired from advances in ecology studies, a recent work [6] addressed the influence inference problem in social media from the perspective of nonlinear dynamic systems. The presented approach differs from previous work in that it can deal with the nonlinear interaction between users using very short time series and key steps in this approach can also be applied to time series prediction.

## III. PROPOSED APPROACH

Given two short time series variables, X and Y (i.e., two users' activity level time series), the goal here is to infer whether there is a causation relationship between the two users. The proposed approach is built on nonlinear state space reconstruction. In this section, we first introduce some background definitions in nonlinear dynamic systems and then demonstrate the basic idea of our approach.

### A. Background Definitions

Consider a dynamic process  $\phi$  describing the temporal evolution of points in an  $E$ -dimensional state space (e.g., the activity level of  $E$  users during time). Its trajectories converge to some  $d$ -dimensional ( $d \leq E$ ) manifold  $M$  such that  $\phi: M \rightarrow M$ . For each user  $X$ , there is a corresponding time series of length  $L$ ,  $\{X\} = [X(1), X(2), \dots, X(L)]$ , that tracks the trajectory of points in  $M$  mapped to a sequence of real numbers (e.g., the activity level time series of user  $X$ ).

A lagged-coordinate embedding uses  $E$  time-lagged values of  $\{X\}$  as coordinate axes or dimensions to reconstruct a shadow attractor manifold  $M_X$  as shown in Figure 1. The points in this manifold, denoted by  $x(t)$ , consist of the set of  $E$ -dimensional vectors  $x(t) = \langle X(t), X(t-\tau), X(t-2\tau), \dots, X(t-(E-1)\tau) \rangle$  where the time lag  $\tau$  is positive.

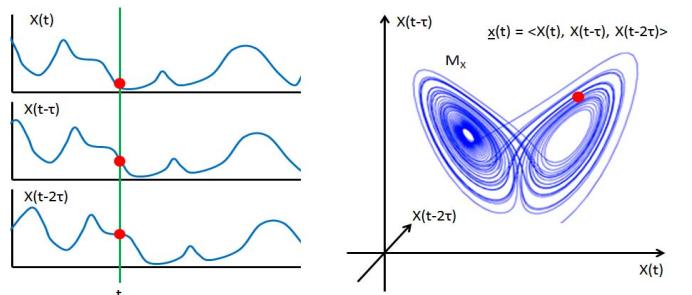


Figure 1. Construction of a shadow manifold  $M_X$ .

## B. Time Series Prediction with Univariate Embedding

The time series  $\{X\}$  can be predicted by conducting simplex projection [7] on the shadow manifold  $M_X$ , which is a nearest neighborhood based forecasting algorithm that involves tracking the forward evolution of nearby points in the embedding.

We observe that simplex projection based on univariate (i.e., variable  $X$ ) embedding can be seen as a cross map from the shadow manifold  $M_X$  to a “forward” manifold  $M_Z$  as shown in Figure 2. The points in the forward manifold  $M_Z$ , denoted by  $z(t)$ , consist of the set of  $E$ -dimensional vectors  $z(t) = \langle Z(t), Z(t-\tau), Z(t-2\tau), \dots, Z(t-(E-1)\tau) \rangle$  and  $Z(t) = X(t+\alpha)$ , where  $\alpha$  is the forward step (i.e.,  $\alpha = 1$ ).

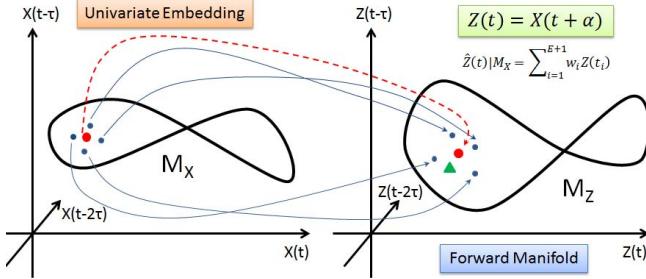


Figure 2. Cross map from the shadow manifold to the forward manifold.

### C. Smoothness of the Map

Since simplex projection is based on finding the nearest neighborhood, it requires a sufficiently long time series to achieve reasonable results. When the original time series is very short, it may fail to work. In fact, a good cross map prediction result means nearby points on the shadow manifold  $M_X$  correspond to nearby points on the forward manifold  $M_Z$ . A recent work [8] states that finding mutual neighbors is actually measuring the smoothness of the map. The critical fact is that any smooth map can be approximated by a neural network, while the approximation will fail with large training errors when the map is unsmooth.

As a result, by treating points on  $M_X$  as input and points on  $M_Z$  as output, we can train a neural network and use the training error to equivalently measure how well the time series  $\{X\}$  can be predicted based on its historical values.

### D. Multivariate Embedding based Causality Detection

In addition to univariate embedding, a recent work [9] has proven that multivariate embedding created from multiple variables can also reconstruct the original system dynamics. By multivariate embedding, we can add another time series (i.e., variable  $Y$ ) to the construction of manifold  $M_X$ . As a result, we have a set of vectors  $xy(t) = \langle X(t), X(t-\tau), X(t-2\tau), \dots, X(t-(E-2)\tau), Y(t) \rangle$ , which make up the mixed manifold  $M_{XY}$ .

Similar to the spirit of Granger causality, if the variable  $Y$  influences  $X$ , then adding  $Y$  should improve the univariate prediction of  $X$ . In other words, if  $Y$  influences  $X$ , then the cross map from mixed manifold  $M_{XY}$  to forward manifold  $M_Z$

should be smoother than the map from shadow manifold  $M_X$  to forward manifold  $M_Z$ . Consequently, as shown in Figure 3, we can train two neural networks,  $\Phi_{X:Z}$  for univariate embedding and  $\Phi_{XY:Z}$  for multivariate embedding. If the training error of neural network  $\Phi_{XY:Z}$  is smaller than training error of  $\Phi_{X:Z}$ , then we say that variable  $Y$  influences  $X$ .

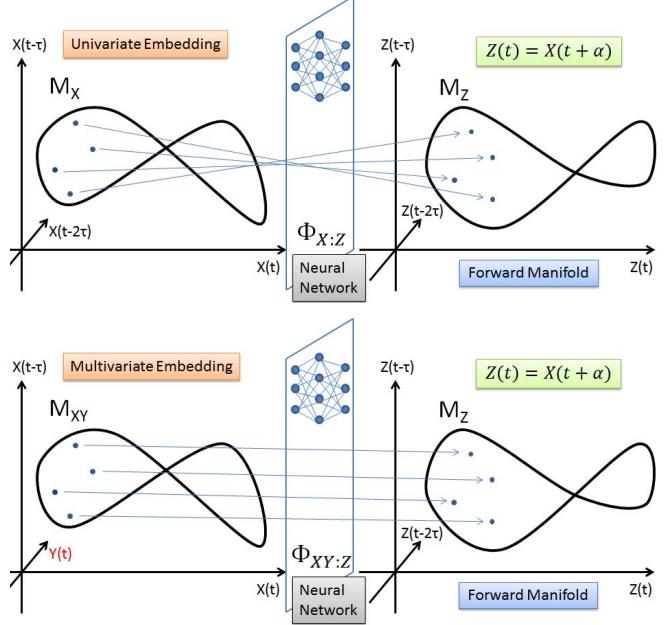


Figure 3. Training two neural networks for univariate and multivariate embedding.

We call this proposed approach described above as Multivariate Embedding based Causality (MEC) detection. In the implementation, we choose to train a widely used neural network, namely Radial Basis Function (RBF) neural network. The training error is obtained by the leave-one-out method. Specifically, if the training error  $\varepsilon(\Phi_{XY:Z})$  of  $\Phi_{XY:Z}$  is smaller than training error  $\varepsilon(\Phi_{X:Z})$  of  $\Phi_{X:Z}$ , then we calculate the causality strength from  $Y$  to  $X$  as follows:

$$MEC_{Y \rightarrow X} = \exp\left(-\frac{\varepsilon(\Phi_{XY:Z})}{\varepsilon(\Phi_{X:Z})}\right)$$

## IV. EXPERIMENT

We evaluate the proposed approach by conducting experiments on both synthetic and real world data.

### A. On synthetic data

Suppose there is a closed community containing only two users  $X$ ,  $Y$ , and the user activity level (e.g., # of posts) during time can be described as follows:

$$\begin{aligned} X_{t+1} &= X_t(r_x - r_x X_t - \gamma_{yx} Y_t) \\ Y_{t+1} &= Y_t(r_y - r_y Y_t - \gamma_{xy} X_t) \end{aligned}$$

where  $r_x = 3.8, r_y = 3.5, \gamma_{yx} = 0, \gamma_{xy} = 0, 0.01, 0.02, \dots, 2.99, 3$ .

The parameters  $\gamma_{yx}, \gamma_{xy}$  denote the influence of  $Y$  on  $X$  and

the influence of  $X$  on  $Y$ , respectively. Only 30 time points are generated for each user. Two baselines were also evaluated, which include 1) Granger causality and 2) the recently proposed CCM algorithm [6]. We examine how well these methods can detect the influence between user  $X$  and  $Y$ . The two time series (when  $\gamma_{xy} = 0.20$ ) and detection results are shown in Figure 4.

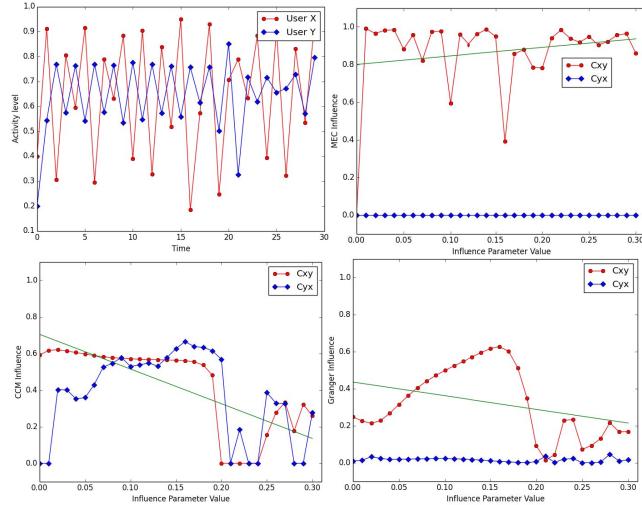


Figure 4. Illustration of time series  $X$ ,  $Y$  and detection results of MEC, CCM and Granger causality.

The results shown in Figure 4 demonstrate the proposed MEC approach outperforms other baselines and there is a positive correlation between MEC influence strength and the influence parameter value in the user activity model.

#### B. On real world data

We further evaluated the validity of MEC by predicting user behavior based on the inferred influence strength in 15M dataset [10], in which Twitter messages were collected in the period April-May 2011, related to the political events (15M movement) occurred at that time in Spain.

We take a linear threshold classifier for user behavior (i.e., adopt a hashtag) prediction in social media. Intuitively, the more friends of a user perform a behavior, the larger probability the user will also perform it. And the friends are weighted with the inferred influence strength obtained by four methods: 1) Influence-Oblivious, meaning if there is an edge from  $v$  to  $u$  in the social network, then the weight is 1, otherwise the weight is 0; 2) Granger causality; 3) CCM influence; 4) our proposed causality detection method, MEC.

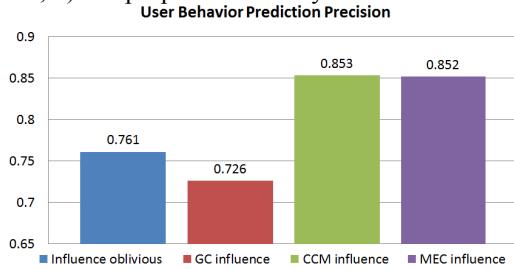


Figure 5. User behavior prediction results.

The results shown in Figure 5 demonstrate that influence inferred by MEC can significantly improve the user behavior prediction precision.

## V. CONCLUSIONS

In this paper, we have presented Multivariate Embedding based Causality detection, a novel causal inference approach in social media, from a nonlinear dynamic system perspective. This approach allows us to infer the causal relationship for any pair of users based on very short user activity level time series data alone. It does not require any explicit causal knowledge like retweeting or other content information. Furthermore, it is a model free approach which does not require the pre-assumption of any particular user activity model. In addition, key steps in this approach can be applied to time series prediction. Experimental results on both synthetic and real world data show the efficacy of the proposed approach. We believe that important security intelligence applications can be built based on this work in the future.

## ACKNOWLEDGMENT

This work was supported in part by the following grants: The National Natural Science Foundation of China under Grant No. 71025001, 71472175, 71103180, 61175040, and 61172106; The National Institutes of Health (NIH) of USA under Grant No. (Grant No.1R01DA037378-01) and the Ministry of Health under Grant No. 2012ZX10004801 and 2013ZX10004218, and by the Grant No. 2013A127.

## REFERENCES

- [1] C. Luo, K. Cui, X. Zheng, and D. Zeng, "Time Critical Disinformation Influence Minimization in Online Social Networks," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, 2014, pp. 68–74.
- [2] G. Sugihara, R. May, H. Ye, C. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting Causality in Complex Ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, Oct. 2012.
- [3] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring Networks of Diffusion and Influence," *ACM Trans Knowl Discov Data*, vol. 5, no. 4, pp. 21:1–21:37, Feb. 2012.
- [4] M. G. Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf, "Uncovering the structure and temporal dynamics of information propagation," *Netw. Sci.*, vol. 2, no. 01, pp. 26–65, 2014.
- [5] G. Ver Steeg and A. Galstyan, "Information transfer in social media," in *Proceedings of the 21st international conference on World Wide Web*, New York, NY, USA, 2012, pp. 509–518.
- [6] C. Luo, X. Zheng, and D. Zeng, "Causal Inference in Social Media Using Convergent Cross Mapping," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, 2014, pp. 260–263.
- [7] G. Sugihara and R. M. May, "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series," *Nature*, vol. 344, no. 6268, pp. 734–741, Apr. 1990.
- [8] H. Ma, K. Aihara, and L. Chen, "Detecting Causality from Nonlinear Dynamics with Short-term Time Series," *Sci. Rep.*, vol. 4, Dec. 2014.
- [9] E. R. Deyle and G. Sugihara, "Generalized Theorems for Nonlinear State Space Reconstruction," *PLoS ONE*, vol. 6, no. 3, p. e18295, Mar. 2011.
- [10] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, "The Dynamics of Protest Recruitment through an Online Network," *Sci. Rep.*, vol. 1, Dec. 2011.