Self-Driven Graph Volterra Models for Higher-Order Link Prediction

Coutino, Mario; Karanikolas, Georgios V; Leus, Geert; Giannakis, Georgios B.

DOI
10.1109/ICASSP40776.2020.9053655

Publication date
2020

Document Version
Final published version

Published in
ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

SELF-DRIVEN GRAPH VOLTERRA MODELS FOR HIGHER-ORDER LINK PREDICTION

Mario Coutino[†]*, Georgios V. Karanikolas*[‡]*, Geert Leus*[†]*, Georgios B. Giannakis*[‡]

[†]Delft University of Technology, Delft, The Netherlands
[‡]University of Minnesota, Minneapolis, MN, USA

ABSTRACT

Link prediction is one of the core problems in network and data science with widespread applications. While predicting pairwise nodal interactions (links) in network data has been investigated extensively, predicting higher-order interactions (higher-order links) is still not fully understood. Several approaches have been advocated to predict such higher-order interactions, but no principled method has been put forth to tackle this challenge so far. Cross-fertilizing ideas from Volterra series and linear structural equation models, the present paper introduces self-driven graph Volterra models that can capture higher-order interactions among nodal observables available in networked data. The novel model is validated for the higher-order link prediction task using real interaction data from social networks.

***Index Terms*—** higher-order interactions, link prediction, network data models, structural equation models, Volterra series

1. MOTIVATION AND CONTEXT

Link prediction is the task of predicting missing or future connections (edge weights) given a set of network observations or the network connectivity at different time instances. In other words, the goal is to infer whether an edge exists (or will exist) between a pair of nodes given side information in the form of nodal features [1]. The link prediction task finds applications in several fields. In social sciences, it is used to study the growth of social networks [2] and their dynamics, as in e.g., friendship formation. Link prediction is also the cornerstone in recommender system algorithms [3]. Recommending matches in a dating app for instance, is equivalent to predicting the most likely links between users. Likewise in biology, link prediction is used to unveil pairwise interactions between elements of different ecological niches or to predict interactions that were not studied due to time or cost constraints [4].

Although pairwise interactions capture part of the underlying dependencies and dynamics of networked data, several interactions that occur within a network involve groups comprising more than two nodes [5]. For example, research works are typically carried out by a team of authors rather than a pair of co-authors. Molecules tend to show more interactions among a small group rather than between single pairs. Finally, in digital communication applications, information exchanges, emails, text messages, and video calls, occur between a large group of people as often as from pairs.

Several approaches have been put forth to model, identify and analyze higher- than second-order relations in networked data. Most can be categorized into approaches based on sets of systems [6], hypergraphs [7], or simplicial complexes [8–10]. While each provides a mathematical framework to study higher-order interactions, they either make use of the network structure directly, as in hypergraph connections, or, they use domain-specific physics-based notions that may not hold in all datasets. For instance, simplicial complexes using cohomology rely on curl and divergence notions [8]. The need thus arises for a modeling tool based on the networked data itself to provide both expressibility in terms of higher-order relations, as well as interpretability for further understanding of the underlying network dynamics, but also for predicting higher-order links in a principled manner.

In time series analysis and for a gamut of applications, several models have been introduced to account for interactions among interconnected entities. Rooted at linear structural equation models (SEMs) [11] and vector autoregressive models (VARMs) [12], recent advances capture the interactions and dynamics across nodes and time using partial correlations and graphical kernels [13–19]. Even though these approaches capture complex dynamics present in networked data, they may fall short when interpreting interactions beyond pairs of nodes.

Another tool with well-documented merits in time series analysis is the Volterra series [20]. Volterra series has found widespread application in modeling brain data [21], gene data [22], and communication channels [23], to list a few. Despite well-appreciated challenges associated with these models, e.g., the complexity grows exponentially with the model order, it has been shown that the computational requirements can be tamed by leveraging sparsity to effect parsimony in the Volterra kernels [22]. In addition, by using an appropriate basis expansion model for the Volterra kernels, the complexity can be further reduced without losing model expressibility because the original Volterra kernels can be retrieved from the corresponding representation on the considered basis [24]. While Volterra series are indeed powerful for modeling nonlinear dependencies, they have not yet been thoroughly investigated to model higher-order interactions in networked data. In addition, Volterra series models (VSMs) only consider input/output maps but not self-driven relations as SEMs do in successfully describing spatio-temporal network dynamics.

Building upon SEMs and VSMs, this work presents a novel self-driven graph Volterra model (SD-GVM) to capture higher-order interactions present in networked data. This model uses graph Volterra kernels to identify interactions between nodes or groups of nodes, providing a principled mean of tackling higher-order link prediction. The proposed approach differs from existing higher-order link prediction methods [25, 26], which focus on extending metrics commonly used in informal scoring for classical link prediction [1].

2. MODELING INTERACTIONS OVER GRAPHS

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ with respective cardinalities $|\mathcal{V}| = N$ and $|\mathcal{E}| = E$. A set of nodal features is collected across time to form the set of endogenous vectors $\{\boldsymbol{x}(t) \in \mathbb{R}^N\}_{t=1}^T$. External (or exogenous) observables $\{\boldsymbol{\zeta}(t) \in \mathbb{R}^Q\}_{t=1}^T$ can be also available corresponding to e.g., extra

nodal features, inputs from different networks [27], and snapshots or layers of one network [28].

A linear structural equation model (SEM) comprises nodal vectors $\boldsymbol{x}(t)$ and $\boldsymbol{\zeta}(t)$, which in the noise-free case are related as [11]

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{\Gamma}\boldsymbol{\zeta}(t) \in \mathbb{R}^N \tag{1}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times Q}$ models the mapping of the exogenous input to the nodal variables in $\boldsymbol{x}(t)$; $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ corresponds to the interrelations among those variables. The per ith entry scalar counterpart of the noise-free model (1) postulates that the observable at node i is expressible through a weighted combination of signals at all other nodes along with the corresponding exogenous variables, as

$$x_i(t) = \sum_{j\in\mathcal{V}, j\neq i} a_{ij}x_j(t) + \sum_{k\in\mathcal{V}} \gamma_{ik}\zeta_k(t) \tag{2}$$

where a_{ij} and γ_{ij} are the (i,j)th entry of $\boldsymbol{A}$ and $\boldsymbol{\Gamma}$, respectively; while $\zeta_k(t)$ denotes the kth entry of $\boldsymbol{\zeta}(t)$.

Although the SEM in (2) can relate different node variables through nonzero entries a_{ij} of the adjacency matrix $\boldsymbol{A}$, it only accounts for pairwise dependencies through linear equations. Efforts to broaden the expressive power of linear SEMs have been made using nonlinear kernels of nodal variables; see e.g., [17] and references therein. Albeit meaningful in several applications, they do not directly account for the so-termed higher-order interactions present in networked data via *higher-order* graph structures [5, 26], including subgraphs and k-cliques.[1] In the ensuing section, we introduce a natural way to model such higher-order interactions, and their descriptors using the widely known Volterra kernels [20].

3. MODELING HIGHER-ORDER INTERACTIONS

Modeling higher-order interactions over graphs calls for a description of the ith nodal feature $x_i(t)$ in terms of a *set of subsets* of nodes $\mathcal{S}^{(i)}$. To model first-order interactions, these subsets of nodes are single nodes and the set $\mathcal{S}^{(i)}$ is nothing but the set of neighbors of the ith node; that is, the nodes j for which $a_{i,j}$ is nonzero in a SEM. Modeling higher-order interactions though requires the subsets to consist of more than one node, and hence the set $\mathcal{S}^{(i)}$ will comprise subsets of nodes.

To make this concrete, consider the set $\mathcal{S}_P^{(i)}$, which contains the subsets for defining interactions up to order P as

$$\mathcal{S}_P^{(i)} := \bigcup_{p=1}^{P} \mathcal{S}_{*,p}^{(i)}, \quad \text{with} \quad \mathcal{S}_{*,p}^{(i)} := \bigcup_{l=1}^{L_p} \mathcal{S}_{l,p}^{(i)} \tag{3}$$

where $\mathcal{S}_{l,p}^{(i)} \subset \mathcal{V}$ denotes the lth set of p nodes related to the ith node in the graph. Parameter p here denotes set cardinality, and L_p the number of order p subsets present.

Using these definitions and dropping the exogenous variable for simplicity, we put forth the following generic self-driven model

$$x_i(t) = f(\boldsymbol{x}(t), \mathcal{S}_P^{(i)}) \;\forall\, i \in \{1,\ldots,N\} \tag{4}$$

where f maps $\boldsymbol{x}(t)$ from $\mathcal{S}_P^{(i)}$ to $x_i(t)$. If for example $\mathcal{S}_1^{(i)} = \cup_{l=1}^{L_1}\mathcal{S}_{l,1}^{(i)}$ is considered, where $\mathcal{S}_{l,p}^{(i)}$ only contains the lth neighbor of the ith node, and assuming a linear f map, (4) boils down to the linear SEM without exogenous variables (cf. (2)).

[1] A k-clique is a subset of vertices of an undirected graph, so that every two distinct vertices in the subset are adjacent.

As most networks exhibit further structure in their node interactions, the subsets in $\mathcal{S}_P^{(i)}$ must capture the gregarious behavior of the nodes. For social networks having triangles as a building block of interactions, the subsets $\{\mathcal{S}_{l,2}^{(i)}\}_{l=1}^{L_2}$ can be defined as those pairs of nodes that form a triad with the ith node [29]. Likewise, the subsets $\{S_{l,p}^{(i)}\}_{l=1}^{L_p}$ can be defined as the nodes that complete a p-clique when the ith node is added. This subset assignment can be done for any other graph motif (cf. [5]) that appears adequate for the data under analysis. In the following, we introduce a principled way to define the functional f in (4).

3.1. Self-Driven Graph Volterra Models

A way to instantiate the abstract model (4) is through a nonlinear discrete-time relationship

$$x_i(t) = f_i(\boldsymbol{x}(t)) + \epsilon_i(t) \tag{5}$$

where the set dependency has been absorbed by the subscript of the function in (5), and $\epsilon_i(t)$ captures modeling errors as well as observation noise. Consider next that this relation is given by the series expansion

$$f_i(\boldsymbol{x}(t)) = h_o^{(i)} + \sum_{p=1}^{P} H_p^{(i)}[\boldsymbol{x}(t)] \tag{6}$$

where $h_o^{(i)}$ is a constant term, and $H_p^{(i)}[\boldsymbol{x}(t)]$ denotes the pth expansion module expressed as

$$H_p^{(i)}[\boldsymbol{x}(t)] := \sum_{l=1}^{L_p} h_{l,p}^{(i)} g(\{x_q(t) : q \in \mathcal{S}_{l,p}^{(i)}\}) \tag{7}$$

where $h_{l,p}^{(i)}$ is the lth expansion coefficient of order p for the ith variable, and g is a permutation-invariant nonlinear function describing the type of interactions among the variables. As the set $\mathcal{S}_P^{(i)}$ is generally unknown, meaning the interactions at all orders are unknown, the module (7) can be equivalently rewritten using the set of all index combinations of size p, that is

$$H_p^{(i)}[\boldsymbol{x}(t)] = \sum_{k_1=1}^{N} \cdots \sum_{k_p=k_{p-1}}^{N} h_p^{(i)}(k_1,\ldots,k_p) g(\{x_{k_q}(t)\}_{q=1}^{p}) \tag{8}$$

where $h_p^{(i)}(k_1,\ldots,k_p)$ is the expansion coefficient for index combination $\{k_1,\ldots,k_p\}$ representing a $(p+1)$-clique and the non-zero coefficients of (8) can be uniquely mapped to the coefficients of (7). Therefore, estimation of the non-zero coefficients in (8) provides understanding of the interactions (active sets) in the network. Since we have assumed g to be a permutation-invariant function, we only consider index combinations rather than permutations, therefore the upper triangular form of the module description in (8).

In the absence of exogenous variables and modeling errors, the expansion (6) for the ith signal can be directly related to the linear SEM in (2) by setting $h_o^{(i)} = 0$, $h_1^{(i)}(j) = a_{ij}$, $h_p^{(i)}(k_1,\ldots,k_p) = 0,\ \forall\ p > 1$, and g as the identity map. Thus, a linear SEM can be seen as a special case of the proposed expansion. Further, upon defining $g(\{x_{k_q}\}_{q=1}^{p}) := \Pi_{q=1}^{p} x_{k_q}$, the module (8) reduces to the pth symmetric Volterra kernel of f_i that is uniquely identifiable [30].

The proposed expansion (7) captures both the self-driven character of SEMs as well as the identifiability and expressibility properties since any continuous, nonlinear function can be uniformly approximated to arbitrary accuracy by a Volterra series under mild conditions on the VSMs [31]. These characteristics distinguish our novel

model from existing nonlinear SEMs that only consider nonlinear functions of *pairwise* interactions. Therefore, higher-order structures in the graph are now seen as fundamental atoms governing the core behavior of nodal features. On the other hand, the expansion (7) allows identifying the existence of higher-order interactions such as triads or p-cliques by observing its nonzero coefficients. Due to the close relation of (7) with VSMs, models based on (7) will be referred to as *self-defining graph Volterra models* (SD-GVMs), and their coefficients as *graph Volterra kernels*.

3.2. Constrained Graph Volterra Kernels

Although (6) can capture the instantaneous dynamics produced by higher-order interactions, it is prudent to introduce *relational* constraints derived from the interpretation of graph Volterra kernels.

We postulate that an interaction between nodes i and j occurs if $h_1^{(i)}(j) \neq 0$, meaning nodal features at i and j are dependent. Along these lines, it is meaningful to consider that higher-order interactions between a triplet $\{i, j, k\}$ only occurs if there is interaction between the pairs $\{i, j\}$, $\{j, k\}$, and $\{i, k\}$. This in turn leads to the condition

$$h_2^{(i)}(j,k) \neq 0 \implies h_1^{(l)}(m) \neq 0 \; \forall \; (l,m) \in \mathcal{P}_2(\{i,j,k\}) \quad (9)$$

where $\mathcal{P}_r(\cdot)$ denotes the permutations of size r of its argument set. Finally, $h_1^{(i)}(i) = 0$, $\forall\, i, j \in \mathcal{V}$ as in the case of linear SEMs.

In the following, we will first introduce formally the problem of higher-order link prediction from network activation data, and then argue that an SD-GVM can be devised to tackle this task.

4. HIGHER-ORDER LINK PREDICTION

Informally, higher-order link prediction amounts to finding the most likely subsets of nodes to interact in the near future [26]. Given a set of binary observations $\boldsymbol{S} \in \{0,1\}^{N \times T}$ capturing the activation of different nodes, e.g., publication of a paper, release of a song, or, sending an email at times $\{t_1, \ldots, t_T\}$, we want to predict what are the most likely sets of nodes to be activated jointly at any $t' > t_T$.

Formally, with $\boldsymbol{s}_t \in \{0,1\}^N$ denoting the tth column of $\boldsymbol{S}$, let $\boldsymbol{s}_t$ be the N-dimensional representation of a set $\mathcal{S}_t \subseteq \mathcal{V}$ whose entries are the indexes of the nonzero entries of $\boldsymbol{s}_t$, e.g., $\boldsymbol{s}_{t'} = [1,\, 0,\, 1]^T \equiv \mathcal{S}_{t'} = \{1, 3\}$. Here, we refer to such measurements as *simplex*. Higher-order link prediction for a single higher-order link can be stated as follows

(P1) *Given simplices $\{\mathcal{S}_t\}_{t=1}^T$ and a set $\mathcal{A} \subset \mathcal{V} : \mathcal{A} \not\subseteq \mathcal{S}_t$, $1 \leq t \leq T$, predict if $\exists\, \tau$, with $T < \tau \leq T_{\max} : \mathcal{A} \subseteq \mathcal{S}_\tau$.*

Given historical data, the goal of (P1) is to predict whether an unseen simplex will appear within a meaningful time interval. Despite its appeal in several applications, pursuing such a general problem might be unrealistic due to computational limitations or due to the intrinsic problem complexity. For these reasons, the higher-order link prediction task is often simplified to the *closure prediction* problem [25]. In a nutshell, closure prediction aims at finding the most likely sets of nodes, which form an *open* structure that will become *close*. Here, an open structure refers to a set of nodes, $\mathcal{A}$, that have interacted with each other, but which have not appeared simultaneously on a single simplex. Formally, a set $\mathcal{A}$ is considered *open* if $\forall\, i, j \in \mathcal{A}$, $\exists\, t' \leq T : \{i, j\} \subseteq \mathcal{S}_{t'}$. Similarly, a set $\mathcal{A}$ is considered *closed* ,if $\exists\, t' \leq T : \mathcal{A} \subseteq \mathcal{S}_{t'}$.

Using these conventions, the higher-order closure prediction problem can be formulated as

(P2) *Given simplices $\{\mathcal{S}_t\}_{t=1}^T$ and a set of candidate open sets $\{\mathcal{A}_c\}_{c=1}^C$, predict the K most likely open sets to become closed.*

Although (P2) is a simplification of the general higher-order prediction in (P1), it retains part of its appeal, while allowing for a more tractable formulation. In addition, by replacing the considered data, meaning simplices, by nodal features, (P2) generalizes the classical top-K link prediction [1]. Similar to [25, 26], the present work will focus on predicting the formation of triangles, i.e., triplets of nodes that activate at the same time, as they are the building block for higher-order interactions in social and other biological networks [5].

To solve (P2), we will show next how an instance of SD-GVMs can be employed to unveil the underlying higher-order interactions in the data. Using the nonzero expansion coefficients, we can then predict which nodes are more likely to be jointly activated in the future, as in e.g., authors publishing a paper together.

5. SD-GVM FOR TRIANGLE CLOSURE

When the goal is to predict closure of triangles, we can restrict ourselves to SD-GVMs of order $P = 2$ as we are mainly interested on interactions among 3-cliques. Also, as the input data is binary, interaction between variables can be viewed as joint activation, which implies that we can consider g as the product of its arguments.

Under these considerations, direct instantiation of an SD-GVM leads to a model that produces real-valued signals. Hence, inspired by binary regression methods, we make use of a latent variable $z_i(t)$ to model the probability $P([\boldsymbol{s}_t]_i = 1 | z_i(t))$ instead of directly modeling $[\boldsymbol{s}_t]_i$. In this work, we model this probability as $P([\boldsymbol{s}_t]_i = 1 | z_i(t)) = \sigma(z_i(t))$, where $\sigma(\cdot)$ is the sigmoid function – a choice corresponding to logistic regression. Upon defining $\boldsymbol{a} \boxtimes \boldsymbol{a} := [a_1^2,\, a_1 a_2, \ldots, a_{N-1} a_N,\, a_N^2]^T$, a latent variable model for this setup can be written as

$$z_i(t) = h_{o,i} + \boldsymbol{h}_{i,1}^T \boldsymbol{s}_t + \boldsymbol{h}_{i,2}^T (\boldsymbol{s}_t \boxtimes \boldsymbol{s}_t) \quad (10)$$

where $h_{o,i} := h_o^{(i)}$, while $\boldsymbol{h}_{i,1}$ and $\boldsymbol{h}_{i,2}$ are the graph Volterra kernels for the first and second module, respectively, with both vectors collecting the coefficients in lexicographic order. By collecting the latent variables for the different nodes through time in a matrix, i.e., $[\boldsymbol{Z}]_{i,t} := z_i(t)$, we can express the model (10) in matrix form as

$$\boldsymbol{Z} = \boldsymbol{H}^{(0)} \boldsymbol{1}^T + \boldsymbol{H}^{(1)} \boldsymbol{S} + \boldsymbol{H}^{(2)} \boldsymbol{S}^{(2)} = \bar{\boldsymbol{H}} \bar{\boldsymbol{S}} \quad (11)$$

where $\boldsymbol{H}^{(i)}$ is the ith-order graph Volterra kernel matrix; with $\boldsymbol{H}^{(0)} = \boldsymbol{h}_o \in \mathbb{R}^N$ a column vector having all constant terms stacked; $\boldsymbol{S}^{(2)} := [\boldsymbol{s}_1 \boxtimes \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T \boxtimes \boldsymbol{s}_T]$; $\bar{\boldsymbol{H}} := [\boldsymbol{H}^{(0)}\, \boldsymbol{H}^{(1)}\, \boldsymbol{H}^{(2)}]$; and $\bar{\boldsymbol{S}} := [\boldsymbol{1}\, \boldsymbol{S}^T\, (\boldsymbol{S}^{(2)})^T]^T$. Although (11) does not directly exhibit the self-driving nature of SD-GVMs, observe that this setting is different from traditional logistic regression (LR) because the *binary labels* here are the node variables $[\boldsymbol{S}]_{i,t}$. As a result, this can be seen as a self-driven version of logistic regression where the data are the labels themselves.

As we are considering the triangle closure problem, we start with knowledge of the connectivity in the training set. Based on $\boldsymbol{S}$, this means we can obtain an initial network connectivity by observing the support of the off-diagonal entries of $\boldsymbol{W} = \boldsymbol{S}^T \boldsymbol{S}$. From this connectivity, and upon enforcing $h_2^{(i)}(i,i) = 0,\; \forall\, i \in \mathcal{V}$ (due to its binary nature $[\boldsymbol{s}_t]_i = [\boldsymbol{s}_t]_i^2$) we can recover (cf. Sec. 3.2) (i) the set of open triangles $\mathcal{T}_O$; (ii) the set of closed triangles $\mathcal{T}_C$; and thus, (iii) the candidates for the nonzero graph Volterra kernels, $\mathcal{S}_2 := \{\mathcal{S}_2^{(i)}\}_{i=1}^N$. Using $\mathcal{S}_2$, we can build a binary matrix $\boldsymbol{M}$

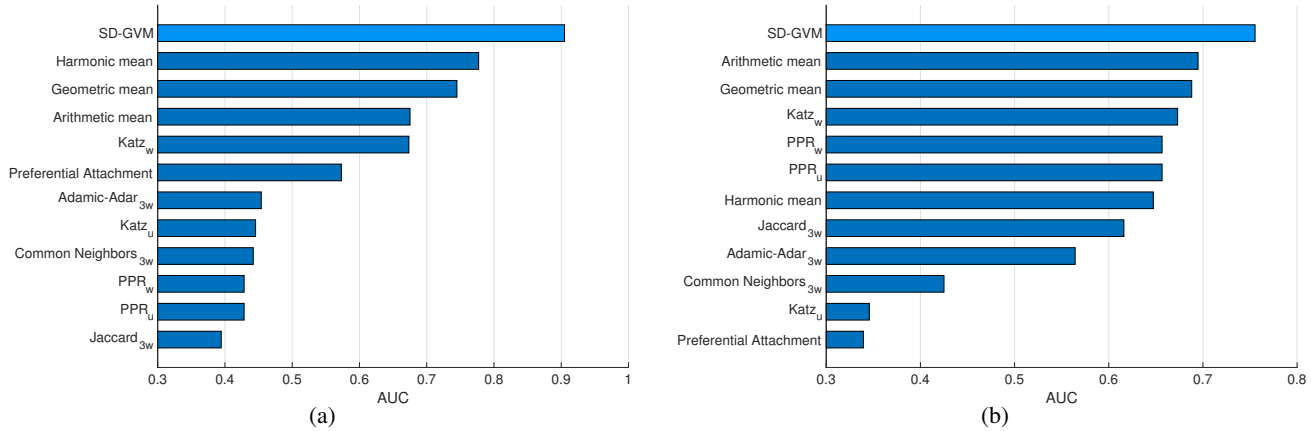

Fig. 1: Prediction performance in terms of AUC in the (a) "`Enron_email`" and (b) "`primary_school_contact`" datasets.

Algorithm 1: Sparse SD-GVM Logistic Regression

Input: binary data: $\bar{\boldsymbol{S}}$; parameter: λ; stepsize: η; tolerance: ϵ, initial connectivity: $\boldsymbol{W}$

$\delta_{\boldsymbol{\theta}} \leftarrow \infty$; $k \leftarrow 0$; $\boldsymbol{\theta}_k \leftarrow \mathbf{0}$;

$\boldsymbol{M} \leftarrow \texttt{build\_M\_mtx}(\boldsymbol{W})$;

$\boldsymbol{\Psi} \leftarrow (\bar{\boldsymbol{S}}^T \otimes \boldsymbol{I})\boldsymbol{M}$;

while $(\delta_{\boldsymbol{\theta}} > \epsilon)$ and $(k < k_{\max})$ **do**

  $\Delta_{\boldsymbol{\theta}} \leftarrow 0$;

  for $i = 1 : N$; $t = 1 : T$ **do**

    $\boldsymbol{\psi} \leftarrow \texttt{get\_mtx\_row}(\boldsymbol{\Psi}, i, t)$;

    $\Delta_{\boldsymbol{\theta}} = \Delta_{\boldsymbol{\theta}} + \boldsymbol{\psi}\Big([\boldsymbol{S}]_{i,t} - \sigma(\boldsymbol{\theta}_k^T \boldsymbol{\psi})\Big)$

  end

  $\boldsymbol{\theta}_{k+1} = \texttt{soft\_thr}(\boldsymbol{\theta}_k + \eta\Delta_{\boldsymbol{\theta}}, \eta\lambda)$;

  $\delta_{\boldsymbol{\theta}} = \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2$; $k = k + 1$;

end

Output: Graph Volterra kernels: $\boldsymbol{\theta}_k$

such that $\text{vec}(\bar{\boldsymbol{H}}) = \boldsymbol{M}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ collects the graph Volterra kernels. Considering such a parametrization, a proximal gradient ascent (PGA) algorithm with sparsity regularization can be derived to fit $\boldsymbol{\theta}$ to the data. This procedure is summarized in Alg. 1. In this algorithm, `get_mtx_row` picks the row of its input related to the $z_i(t)$ latent variable; and $\texttt{soft\_thr}(\cdot, \eta\lambda)$ applies soft thresholding with parameter $\eta\lambda$.

Once $\boldsymbol{\theta}$ has been estimated, the entries related to $\mathcal{T}_O$ are sorted by their absolute value, and the K-top ones are declared as the most likely open triangles to become closed. Our working hypothesis is that open triangles with large coefficients, capturing high level of interaction, are the most likely triangles to become closed.

6. NUMERICAL TESTS

The proposed SD-GVM approach was compared with a number of alternatives in terms of open triangle closure prediction performance, as measured by the area under the curve (AUC) in the receiver operating characteristic metric on the first 100 nodes of the "`Enron_email`" [32], and the "`primary_school_contact`" [33] datasets. In short, the former dataset consists of emails between Enron employees, with nodes representing email addresses, whereas the latter consists of proximity-based contacts as recorded by wearable sensors in a primary school. Due to space limitations we refer the interested reader to [26] (and references therein) for a comprehensive treatment of the competing alternatives. Where applicable, the subscripts w, u, and $3w$, stand for weighted, unweighted and 3-way, respectively. Finally, PPR stands for personalized PageRank similarity.

In particular, for our tests, the first 10% and 1% of timestamped events comprised the training sets in the "Enron email" and "primary school contact" datasets, respectively, with the rest of the data being used for testing. For both experiments, $\lambda = 10^{-3}$, $\eta = 10^{-4}$ and $k_{\max} = 500$ were used for SD-GVM. The merits of our approach become evident in Figs. 1a and 1b, as the proposed SD-GVM consistently outperforms all competing alternatives higher-order link prediction performance. The proposed approach is a step forward towards a model-based understanding of higher-order prediction over networks.

7. CONCLUSIONS

The present contribution put forth a novel model to capture higher-order interactions in networked data. Drawing upon linear structural equation models and Volterra series models, nodal features were expressed as combinations of features from neighboring nodes, along with nonlinear combinations of nodal features belonging to groups of nodes capturing higher-order dependencies. The novel self-driven graph Volterra model was then specialized to handle binary measurements, and was applied to the higher-order link prediction problem. Through numerical tests involving real social interaction data, the proposed model along with logistic regression, was demonstrated to outperform recently proposed methods based on generalizing the link prediction scores for the task of triangle closure prediction.

Acknowledgements. M. Coutino and G. Leus were supported in part by the ASPIRE project 14926 (within the STW OTP program) financed by the Netherlands Organization for Scientific Research (NWO); and M. Coutino in part by CONACYT. G. V. Karanikolas and G. B. Giannakis were supported in part by NSF grants 1711471 and 1901134. Emails: {m.a.coutinominguez, g.j.t.leus}@tudelft.nl; {karan029,georgios}@umn.edu

8. REFERENCES

[1] E. D. Kolaczyk and G. Csárdi, *Statistical Analysis of Network Data with R*. Springer, 2014, vol. 65.

[2] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Tec.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[3] P. Resnick and H. R. Varian, "Recommender systems," *Commun. ACM*, vol. 40, no. 3, pp. 56–59, 1997.

[4] S. Sulaimany, M. Khansari, and A. Masoudi-Nejad, "Link prediction potentials for biological networks," *Int. J. Data Min. Bioin.*, vol. 20, no. 2, pp. 161–184, 2018.

[5] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[6] P. Frankl and V. Rödl, "Extremal problems on set systems," *Random Struct. Algor.*, vol. 20, no. 2, pp. 131–164, 2002.

[7] C. Berge, *Hypergraphs: Combinatorics of Finite Sets*. Elsevier, 1984, vol. 45.

[8] L. Lim, "Hodge laplacians on graphs," *arXiv:1507.05379*, 2015.

[9] M. T. Schaub, A. R. Benson, P. Horn, G. Lippner, and A. Jadbabaie, "Random walks on simplicial complexes and the normalized hodge laplacian," *arXiv:1807.05044*, 2018.

[10] S. Barbarossa and S. Sardellitti, "Topological signal processing over simplicial complexes," *arXiv:1907.11577*, 2019.

[11] J. J. Hox and T. M. Bechger, "An introduction to structural equation modeling," 1998.

[12] H. Lütkepohl, *Vector Autoregressive Models*. Springer, 2011.

[13] X. Cai, J. A. Bazerque, and G. B. Giannakis, "Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations," *PLoS Comput. Biol.*, vol. 9, no. 5, p. e1003068, 2013.

[14] G. V. Karanikolas, G. B. Giannakis, K. Slavakis, and R. M. Leahy, "Multi-kernel based nonlinear models for connectivity identification of brain networks," in *Intl. Conf. on Acoustics Speech and Signal Process.*, Shanghai, China, March 20-25, 2016, pp. 6315–6319.

[15] G. V. Karanikolas, O. Sporns, and G. B. Giannakis, "Multi-kernel change detection for dynamic functional connectivity graphs," in *Asilomar Conf. on Signals, Syst., and Comput.*, Pacific Grove, CA, USA, October 29 - November 1, 2017, pp. 1555–1559.

[16] Y. Shen, B. Baingana, and G. B. Giannakis, "Kernel-based structural equation models for topology identification of directed networks," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2503–2516, 2017.

[17] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. of the IEEE*, vol. 106, no. 5, pp. 787–807, 2018.

[18] E. Isufi, A. Loukas, N. Perraudin, and G. Leus, "Forecasting time series with VARMA recursions on graphs," *IEEE Trans. Signal Process.*, vol. 67, no. 18, pp. 4870–4885, 2019.

[19] Y. Shen, G. B. Giannakis, and B. Baingana, "Nonlinear structural vector autoregressive models with application to directed brain networks," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5325–5339, 2019.

[20] R. W. Brockett, "Volterra series and geometric control theory," *Automatica*, vol. 12, no. 2, pp. 167–176, 1976.

[21] D. Song, R. H. Chan, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, "Nonlinear dynamic modeling of spike train transformations for hippocampal-cortical prostheses," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 6, pp. 1053–1066, 2007.

[22] V. Kekatos and G. B. Giannakis, "Sparse Volterra and polynomial regression models: Recoverability and estimation," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5907–5920, 2011.

[23] C. Krall, K. Witrisal, G. Leus, and H. Koeppl, "Minimum mean-square error equalization for second-order Volterra systems," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4729–4737, 2008.

[24] V. Z. Marmarelis, "Identification of nonlinear biological systems using Laguerre expansions of kernels," *Ann. Biomed. Eng.*, vol. 21, no. 6, pp. 573–589, 1993.

[25] H. Huang, J. Tang, L. Liu, J. Luo, and X. Fu, "Triadic closure pattern analysis and prediction in social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 12, pp. 3374–3389, Dec. 2015.

[26] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Natl. Acad. Sci.*, vol. 115, no. 48, pp. E11 221–E11 230, 2018.

[27] V. N. Ioannidis, Y. Shen, P. Traganitis, and G. B. Giannakis, "Kernel-based learning of processes over multi-layer graphs," in *Proc. of SPAWC*, 2018.

[28] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *J. Complex Netw.*, vol. 2, no. 3, pp. 203–271, 2014.

[29] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets*. Cambridge university press, Cambridge, 2010, vol. 8.

[30] M. Schetzen, "The Volterra and Wiener theories of nonlinear systems," 1980.

[31] S. Boyd and L. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Trans. Circuits Syst.*, vol. 32, no. 11, pp. 1150–1161, 1985.

[32] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *ECOML*. Springer, 2004, pp. 217–226.

[33] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina *et al.*, "High-resolution measurements of face-to-face contact patterns in a primary school," *PloS one*, vol. 6, no. 8, p. e23176, 2011.