Autoregressive graph Volterra models and applications

Yang, Qiuling ; Coutino, Mario; Leus, G.J.T.; Giannakis, Georgios B.

DOI
10.1186/s13634-022-00960-6

Publication date
2023

Document Version
Final published version

Published in
Eurasip Journal on Advances in Signal Processing

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Autoregressive graph Volterra models and applications

Qiuling Yang[1], Mario Coutino[2], Geert Leus[3*] and Georgios B. Giannakis[4]

*Correspondence:
G.J.T.Leus@tudelft.nl

[1] Department of Mechanical Engineering, University of Alberta, Edmonton, Canada
[2] Radar Technology Department, TNO, The Hague, The Netherlands
[3] Circuits and Systems (CAS), Department of Microelectronics, Delft University of Technology, Delft, The Netherlands
[4] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA

Abstract

Graph-based learning and estimation are fundamental problems in various applications involving power, social, and brain networks, to name a few. While learning pair-wise interactions in network data is a well-studied problem, discovering higher-order interactions among subsets of nodes is still not yet fully explored. To this end, encompassing and leveraging (non)linear structural equation models as well as vector autoregressions, this paper proposes autoregressive graph Volterra models (AGVMs) that can capture not only the connectivity between nodes but also higher-order interactions presented in the networked data. The proposed overarching model inherits the identifiability and expressibility of the Volterra series. Furthermore, two tailored algorithms based on the proposed AGVM are put forth for topology identification and link prediction in distribution grids and social networks, respectively. Real-data experiments on different real-world collaboration networks highlight the impact of higher-order interactions in our approach, yielding discernible differences relative to existing methods.

Keywords: Higher-order interactions, Volterra series, Graph inference, Link prediction

1 Introduction

Full awareness of networks and networked interactions is required for understanding the behavior of complex systems. These systems are typically modeled as graphs in many applications such as financial markets, social networks, power systems, and transportation systems [1–3], to name a few. Graph structure identification (prediction) is to identify (predict) if an edge exists (will exist) between a pair of nodes of a graph, given a set of network observations in the form of node attributes at different time instances. Applications include studying the growth of social networks [1] and their dynamics in social sciences, predicting what are the most likely links between users in recommender systems, unveiling pair-wise interactions between elements of different ecological niches or predicting interactions that were not studied due to time or cost restrictions in biology [4].

Albeit pair-wise interactions have the ability to capture the dynamics of the underlying graph, a lot of the interplay among networked data occurs beyond just two nodes [5]. For instance, human interaction over social media takes place with a team rather than two individuals. Furthermore, molecules tend to show more interactions

among different groups. Moreover, in smart grids, the dependency among power variables such as voltage and current occurs in a region instead of single pairs. Early approaches to capture higher-order relations in the underlying network have mainly relied on set systems, hypergraphs, and simplicial complexes [6–10], to name a few. Furthermore, to address the existence of nonlinear connectivity, topology identification approaches leveraging partial correlations as well as kernels have recently been developed [11, 12]. A link prediction approach for evaluating higher-order data models of complex systems is proposed in [13]. While offering mathematical frameworks to study higher-order relations, these works either directly leverage the network structure, e.g., connections in a hypergraph, or make use of concepts inherited from physical phenomena, e.g., the analysis based on simplicial complexes, using cohomology [14], which might not exist for all kinds of datasets.

Aiming to modeling dynamical processes over a network, different attempts have been made. Specially, data-driven neural network solutions have attracted growing attention recently to learn the nonlinear connectivity [15–17]. Furthermore, rooted in structural equation models [18] (SEMs), and in particular combinatorial vector autoregressive models [19], several efforts leveraging either kernels or partial correlations [11, 12] have been devoted to capturing the dynamics; see [20–22], and references therein. Despite that these approaches manage to capture complex dynamics existing in networked data, they lack interpretability beyond pair-wise interactions. Another issue of the mentioned models is their poor scalability; in other words, the complexity grows exponentially along with the model order.

Volterra series and kernels have emerged as promising tools for data analysis in different applications, e.g., brain networks [23], gene data [24] and communications [25]. Leveraging the sparsity of the Volterra kernels as well as a parsimonious model description, the computational complexity can be reduced [24], especially when using an appropriate basis expansion model of the kernels. Moreover, one can retrieve the original Volterra kernels from the considered basis expansion without losing model expressibility [26]. Although the Volterra series is powerful in modeling nonlinear-temporal interactions, using it to capture the higher-order relations in the networked data is not well-explored. Furthermore, in Volterra series models, the autoregressive property is not fully considered as in SEM when interpreting the dynamics in networked data.

Building upon SEMs and Volterra series models, this work advocates an autoregressive graph Volterra model (AGVM) to capture higher-order interactions present in networked data. Different identifiability conditions for the proposed model are derived including the identifiability of the network connection based on exogenous data, sparsity (relieving sampling complexity) and the restricted isometry property in the bipolar case. The proposed model uses graph Volterra kernels to identify interactions between nodes or groups of nodes, providing a principled way to tackle higher-order interactions in networks. Furthermore, to estimate the graph Volterra kernels, two tailored AGVM algorithms for topology identification in power systems and link prediction in social networks are introduced. The proposed approaches differ from existing higher-order interaction methods [13, 27], which solely focus on extending metrics commonly used in informal scoring for classical link prediction and identification.

Paper outline. The rest of the paper is organized as follows: Sect. 2 briefly reviews the mathematical model that is used throughout the paper. Section 3 introduces the higher-order interactions model based on Volterra series. Section 4 provides identifiability guarantees for the proposed model. Section 5 presents two tailored algorithms for the application of topology identification in power systems as well as link prediction in social networks. Concluding remarks are drawn in Sect. 6.

Notation. Lower- (upper-) case boldface letters denote column vectors (matrices), and normal letters represent scalars. Calligraphic letters are reserved for sets with the exception of matrices $\mathcal{X}$ and $\mathcal{H}$. Symbol $\top$ stands for transposition. The (i, j)th entry of matrix $\boldsymbol{X}$ is denoted as $x_{i,j}$ or $[\boldsymbol{X}]_{i,j}$. The definition of the operator $\boxtimes$ is the reduced Kronecker product on two equal-size vectors, that is, $\boldsymbol{x} \boxtimes \boldsymbol{y} := [x_1y_1\ x_1y_2\ \cdots\ x_iy_j\ \cdots\ x_{N-1}y_N\ x_Ny_N]^\top$, where $i \leq j$. The operation $*$ denotes the Khatri-Rao product.

2 Preliminaries

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex (node) set with cardinality $|\mathcal{V}| = N$, and $\mathcal{E}$ is the edge set with cardinality $|\mathcal{E}| = E$, respectively. A time-series of graph signals $\{\boldsymbol{x}(t) \in \mathbb{R}^N\}_{t=1}^T$ is collected at the nodes $\mathcal{V}$. In addition, external (exogenous) observables $\{\boldsymbol{\zeta}(t) \in \mathbb{R}^N\}_{t=1}^T$ are available such as features of the nodes, inputs from different networks, and network-level snapshots or layers [28].

The classical structural equation model (SEM) [18] considering the signal $\boldsymbol{x}(t)$ over the graph and the exogenous variables $\boldsymbol{\zeta}(t)$ can be described as follows

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

where $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ is a diagonal matrix representing the mapping of the exogenous input on the node variables $\boldsymbol{x}(t)$, and $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ represents the inter-relations among those variables. Let $x_i(t)$ and $\zeta_i(t)$ denote the ith entry of $\boldsymbol{x}(t)$ and $\boldsymbol{\zeta}(t)$, respectively. The signal at the ith node $x_i(t)$ can then be obtained by a weighted combination of the signal of all other nodes and the corresponding exogenous variables as

$$x_i(t) = \sum_{j\in\mathcal{V}} a_{i,j}x_j(t) + \sum_{j\in\mathcal{V}} \gamma_{i,j}\zeta_j(t) \tag{2}$$

where $a_{i,j}$ and $\gamma_{i,j}$ are the (i, j)th entry of $\boldsymbol{A}$ and $\boldsymbol{\Gamma}$, respectively.

The SEM in (2) is able to express the relation between different node variables through the nonzero entries $a_{i,j}$ of $\boldsymbol{A}$, where $\boldsymbol{A}$ is a hollow matrix and shares the support with the adjacency matrix of the graph. However, it only accounts for the first-order dependencies (i.e., pair-wise relations) in a linear fashion. Several efforts have focused on expanding the expressive power of linear SEMs via nonlinear kernels of nodal variables; see, e.g., [11] and references therein. Although meaningful in many applications, they neglect the so-called higher-order interactions that are present in networked data through *higher-order* graph structures [5], such as subgraphs and k-cliques, which are subsets of vertices of an undirected graph where every two distinct vertices in the subset are adjacent.

In the following section, we introduce a Volterra model to capture such higher-order interactions and their descriptors.

3 Higher-order interactions in graphs

Before modeling the higher-order interactions in graphs, let us give a description of the ith node signal, $x_i(t)$, in terms of a *set of subsets* of nodes, $\mathcal{S}_i$. To model first-order interactions, these subsets of nodes are simply single nodes and the set $\mathcal{S}_i$ is nothing more than the set of neighbors of the ith node, i.e., 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 $\mathcal{S}_i$ will yield a set of subsets of nodes.

Mathematically, the set $\mathcal{S}_i^{(P)}$ which contains the subsets for defining interactions up to order P is defined as follows

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

where p denotes the order of the set, L_p the number of subsets of order p that exist, and $\mathcal{S}_i^{(l,p)} \subset \mathcal{V}$ the lth set of p nodes related to the ith node in the graph $\mathcal{G}$. For simplicity, the exogenous variable has been omitted. With (3), we can put forth the following signal model

$$x_i(t) = f(\boldsymbol{x}(t), \mathcal{S}_i^{(P)}) + \epsilon_i(t) \ \forall \ i \in \{1, \dots, N\} \tag{4}$$

where f maps the signals $\boldsymbol{x}(t)$ from $\mathcal{S}_i^{(P)}$ to $x_i(t)$. For example, considering $P = 1$ and $\mathcal{S}_i^{(l,p)}$ only containing the lth neighbor of the ith node, i.e., $\mathcal{S}_i^{(1)} = \cup_{l=1}^{L_1} \mathcal{S}_i^{(l,1)}$, and assuming a linear map for f, we retrieve the SEM without considering exogenous variables [cf. (2)].

The subsets in $\mathcal{S}_i^{(P)}$ capture the gregarious behavior of the nodes. For example, the subsets $\{\mathcal{S}_i^{(l,2)}\}_{l=1}^{L_2}$ can be viewed as the pairs of nodes that form a triad with the ith node. Similarly, the subsets $\{\mathcal{S}_i^{(l,p)}\}_{l=1}^{L_p}$ can be defined as the nodes that complete a $(p+1)$-clique when the ith node is added. In fact, this subset assignation can be done for any other graph motif [c.f [5]] that seems adequate for the data under analysis. We can now approximate the nonlinear relationship in (4) by a Volterra expansion as follows

$$x_i(t) = h_i^{(0)} + \sum_{p=1}^{P} H_i^{(p)}[\boldsymbol{x}(t)] + \epsilon_i(t) \tag{5}$$

where $h_i^{(0)}$ is a constant term, and $H_i^{(p)}[\boldsymbol{x}(t)]$ denotes the pth order Volterra module given by

$$H_i^{(p)}[\boldsymbol{x}(t)] := \sum_{l=1}^{L_p} h_i^{(l,p)} g(\{x^{(q)}(t) : q \in \mathcal{S}_i^{(l,p)}\}) \tag{6}$$

with $h_i^{(l,p)}$ the lth expansion coefficient of order p for the ith variable, g a permutation-invariant nonlinear function describing the type of interaction among the variables, and q means the nodes that complete a $(q+1)$-clique when the ith node is added. As the set $\mathcal{S}_i^{(P)}$ is generally unknown, meaning the interactions at all orders are unknown, the module (6) can be equivalently rewritten using the set of all index combinations of size p, that is

$$H_i^{(p)}[\boldsymbol{x}(t)] = \sum_{k_1=1}^{N} \cdots \sum_{k_p=k_{p-1}}^{N} h_i^{(p)}(k_1,\ldots,k_p) g(\{x^{(k_q)}(t)\}_{q=1}^{p}) \tag{7}$$

where the Volterra kernel $h_i^{(p)}(k_1,\ldots,k_p)$ denotes a $(p+1)$-clique for the index combination $\{k_1,\ldots,k_p\}$. The nonzero coefficients of (7) can be uniquely mapped to the coefficients of (6). A fundamental result of Volterra expansion is that any continuous nonlinear system can be uniformly approximated (i.e., in the L^∞-norm) to arbitrary accuracy by a Volterra series operator of sufficient but finite order if the input signals form a compact subset of the input function space [29, 30].

Notice that in the case of the absence of exogenous variables, the Volterra expansion (5) for the ith signal can be directly related to the SEM expression (2) by considering $h_i^{(0)}, h_i^{(1)}(j) = a_{i,j}$[1] and $h_i^{(p)}(k_1,\ldots,k_p) = 0,\ \forall\, p > 1$. Thus, a SEM can be seen as a special case of the Volterra expansion where the Volterra kernels are constrained and the inputs are assumed to be the signals on the graph.

Now, we are ready to postulate our AGVM that considers higher-order interactions as follows

$$x_i(t) = h_i^{(0)} + \sum_{p=1}^{P} \sum_{k_1=1}^{N} \cdots \sum_{k_p=k_{p-1}}^{N} h_i^{(p)}(k_1,\ldots,k_p) g(\{x^{(k_q)}(t)\}_{q=1}^{p}) + \epsilon_i(t). \tag{8}$$

The proposed expansion (8) captures both the autoregressive nature of SEMs and the identifiability and expressibility of Volterra series models. This aspect distinguishes the model from existing nonlinear extensions of SEMs that only consider nonlinear functions of *pair-wise* interactions. Therefore, higher-order structures in the graph are not seen as fundamental atoms to establish the behavior of the node signal. On the other hand, the expansion (8) allows identifying the existence of higher-order interactions such as triads or p-cliques by observing its nonzero coefficients.

Remark 1 For simplicity, the present work only focuses on interactions up to the second order and uses a product for g. A generalization to higher-order interactions and other permutation-invariant functions is straightforward. For other nonlinear functions, a more careful analysis should be carried out.

By stacking the signal on node i over time steps $t = (1,\ldots,T)$ in $\boldsymbol{x}_i$, i.e., $\boldsymbol{x}_i := [x_i(1), x_i(2), \ldots, x_i(T)]^\top$ (similarly stacking the modeling errors through time in $\boldsymbol{\epsilon}_i$), stacking the signals on all nodes at time step t in $\boldsymbol{x}(t)$, i.e., $\boldsymbol{x}(t) := [x_1(t), x_2(t), \ldots, x_N(t)]^\top \in \mathbb{R}^N$, and restricting ourselves to a second-order model and a product for g, we can rewrite (8) in a matrix-vector form as

$$\boldsymbol{x}_i^\top = h_i^{(0)} \mathbf{1}^\top + (\boldsymbol{h}_i^{(1)})^\top \boldsymbol{X}^{(1)} + (\boldsymbol{h}_i^{(2)})^\top \boldsymbol{X}^{(2)} + \boldsymbol{\epsilon}_i^\top \in \mathbb{R}^T. \tag{9}$$

[1] With a slight abuse of notation, the notation after h enclosed by parentheses indicates the node index.

Here, we have made the following definitions: $\boldsymbol{X}^{(1)} := [\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)] \in \mathbb{R}^{N\times T}$, $\boldsymbol{X}^{(2)} := [\boldsymbol{x}(1) \boxtimes \boldsymbol{x}(1), \ldots, \boldsymbol{x}(T) \boxtimes \boldsymbol{x}(T)] \in \mathbf{R}^{M\times T}$, $[\boldsymbol{h}_i^{(1)}]_j = h_i^{(1)}(j)$, and

$$\boldsymbol{h}_i^{(2)} = \text{vech}\left(\begin{bmatrix} h_i^{(2)}(1,1) & h_i^{(2)}(1,2) & \cdots & h_i^{(2)}(1,N) \\ h_i^{(2)}(2,1) & h_i^{(2)}(2,2) & \cdots & h_i^{(2)}(2,N) \\ \vdots & \vdots & \vdots & \vdots \\ h_i^{(2)}(N,1) & h_i^{(2)}(N,2) & \cdots & h_i^{(2)}(N,N) \end{bmatrix}\right) \in \mathbb{R}^{\mathrm{M}}$$

where $\text{vech}(\cdot)$ is the half-vectorization operation that retrieves the upper-triangular part of its matrix argument, and $M := N(N+1)/2$.

The second-order expansion of $\boldsymbol{x}_i$ can be further expressed as

$$\begin{aligned} \boldsymbol{x}_i^\top &= \left[h_i^{(0)}, (\boldsymbol{h}_i^{(1)})^\top, (\boldsymbol{h}_i^{(2)})^\top\right] \begin{bmatrix} \mathbf{1}^\top \\ \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix} + \boldsymbol{\epsilon}_i^\top \\ &= \boldsymbol{\theta}_i^\top \boldsymbol{M} + \boldsymbol{\epsilon}_i^\top \end{aligned} \tag{10}$$

where the unknown parameter $\boldsymbol{\theta}_i \in \mathbb{R}^{1+N+M}$ contains the *equivalent graph Volterra kernels* for the ith variable and $\boldsymbol{M} \in \mathbb{R}^{(1+N+M)\times T}$ is the (known) system matrix.

For all involved node signals $i \in [N]$, we finally obtain

$$\boldsymbol{X}^{(1)} = \boldsymbol{\Theta}^\top \boldsymbol{M} + \boldsymbol{E} \in \mathbb{R}^{N\times T} \tag{11}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{(1+N+M)\times N}$ collects all the unknown parameters of the system, and $\boldsymbol{E}$ is the corresponding modeling error matrix. For interpretability of (11), one can also rewrite it as

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(0)}\mathbf{1}^\top + \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} + \boldsymbol{\Gamma}\boldsymbol{Y} + \boldsymbol{E} \tag{12}$$

where $\boldsymbol{H}^{(i)}$ is the ith-order graph Volterra kernel matrix whose entries are in lexicographic order, i.e., as defined through $\boldsymbol{X}^{(i)}$. In particular, $\boldsymbol{H}^{(0)} = \boldsymbol{h}^{(0)} \in \mathbb{R}^N$ is a column vector with all constant terms stacked. Here, we present the general form of the AGVM model also accounting for the exogenous variables $\boldsymbol{Y}$.

One of the challenges to find the unknown parameters $\boldsymbol{\Theta}$ is its large dimensionality. Although symmetrized Volterra kernels can be uniquely identified [31], the order of the number of unknown parameters in (11) is $\mathcal{O}(N^3)$ which leads to high computational costs and poor sampling efficiency. Fortunately, as it is shown ahead, judicious modeling of the graph Volterra kernels leads to efficient higher-order interaction identification. But before presenting methods for estimating the graph Volterra kernels, we need to provide identifiability guarantees for the proposed model.

4 Identifiability of AGVM

This section focuses on the conditions that the input/output data should exhibit in order to uniquely identify the second-order AGVM model (12). Although asymptotic results have been obtained for sparse regression, i.e., the Lasso estimator, here we are more interested in the finite-sample regime. Therefore, borrowing tools from the compressing sensing literature and linear algebra, we are able to provide recovery guarantees in both deterministic and probabilistic settings.

Without loss of generality, we present our following results considering that the zeroth-order term $\boldsymbol{H}^{(0)}$ is projected out and the noise term $\boldsymbol{E}$ is not present or has been removed.

Our first result asserts identifiability of the network connections, represented through the graph Volterra kernels, highlighting the role of the exogenous data $\boldsymbol{Y}$. Notice that this result is a generalization of the result in [32] obtained for resolving direction ambiguities in structural equation models applied for directed graphs.

Theorem 1 *Suppose that data* $\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$ *and* $\boldsymbol{Y}$ *abide to the second-order AGVM model*

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} + \boldsymbol{\Gamma}\boldsymbol{Y}$$

for a matrix $\boldsymbol{H}^{(1)}$ *with diagonal entries* $[\boldsymbol{H}^{(1)}]_{ii} = 0$ *and diagonal matrix* $\boldsymbol{\Gamma}$ *with diagonal entries* $[\boldsymbol{\Gamma}]_{ii} \neq 0$. *If* $\mathcal{X} := [(\boldsymbol{X}^{(1)})^\top \ (\boldsymbol{X}^{(2)})^\top]^\top$ *is full row rank, then* $\boldsymbol{H}^{(1)}$, $\boldsymbol{H}^{(2)}$ *and* $\boldsymbol{\Gamma}$ *are uniquely expressible in terms of* $\mathcal{X}$ *and* $\boldsymbol{Y}$ *as*

$$\begin{aligned} \operatorname{diag}(\boldsymbol{\Gamma}) &= \operatorname{diag}(\boldsymbol{Q}_1)^{-1} \\ \boldsymbol{H}^{(1)} &= \boldsymbol{I} - \boldsymbol{\Gamma}\boldsymbol{Q}_1 \\ \boldsymbol{H}^{(2)} &= -\boldsymbol{\Gamma}\boldsymbol{Q}_2 \end{aligned}$$

where $\boldsymbol{Y}\mathcal{X}^\dagger = [\boldsymbol{Q}_1 \in \mathbb{R}^{N\times N} \ \boldsymbol{Q}_2 \in \mathbb{R}^{N\times M}]$.

Proof See Appendix. □

This result exhibits the importance of the exogenous data (perturbation), $\boldsymbol{Y}$, to uniquely identify the AGVM model. This shows, as in the classical SEM, that given a sufficiently rich perturbation, the directionality, as well as the higher-order interactions (triplets), can be uniquely determined from the measured data.

Although the result of Theorem 1 establishes the identifiability of the AGVM model, it requires a full row rank data matrix $\mathcal{X}$, which in many cases might not be possible, i.e., the number of samples must be at least $\mathcal{O}(N^2)$. Thus, in order to improve the sampling complexity for the problem, prior information is required to constrain the model. A natural assumption, arising in many networked-data applications, is the *sparse* interaction among the nodes. That is, the number of connections (edges) among nodes are much smaller than the size of the network, and therefore, the number of triads in which they participate are restricted. Before proceeding, we need the following sparsity assumptions.

A. 1 Each row of matrix $\boldsymbol{H}^{(1)}$ has at most K_1 nonzero entries, i.e., $\|\boldsymbol{h}_i^{(1)}\|_0 \leq K_1 \ \forall \ i$.

A. 2 Each row of matrix $\boldsymbol{H}^{(2)}$ has at most K_2 nonzero entries, i.e., $\|\boldsymbol{h}_i^{(2)}\|_0 \leq K_2 \ \forall \ i$.

Assumptions A.1 and A.2 on the graph Volterra kernels can be seen as restrictions on the number of *edges* and *triangles* existing in the graph. By letting $K_1 \in \mathcal{O}(1)$ and $K_2 \in \mathcal{O}(N)$, these assumptions translate into graph Volterra kernels that represent a

sparse graph, i.e., $\mathcal{O}(N)$ edges and $\mathcal{O}(N^2)$ triangles. In addition, as shown in the following, the sparsity assumptions make the identification of the system possible when only a reduced number of measurements are available.

Before stating our second result, a definition is required.

Definition 1 The Kruskal rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{N\times M}$, denoted $\mathrm{kr}(\boldsymbol{A})$, is the maximum number k such that any combination of k columns of $\boldsymbol{A}$ forms a sub-matrix with full column rank.

Although the Kruskal rank is, in general, more restrictive than the traditional rank and harder to verify, when the entries of $\boldsymbol{A}$ are drawn from a continuous distribution, its Kruskal rank equals its rank [33].

To begin with, we consider a model without exogenous inputs. That is, a *pure self-driving* system.

Theorem 2 *Let* $\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$ *abide to the second-order AGVM*

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)}$$

for sparse matrices $\boldsymbol{H}^{(1)}$ *with diagonal entries* $[\boldsymbol{H}^{(1)}]_{ii} = 0$ *and* $\boldsymbol{H}^{(2)}$ *satisfying A.1 and A.2, respectively. If* $\mathrm{kr}(\mathcal{X}^\top) \geq 2(\mathrm{K}_1 + \mathrm{K}_2)$, *where* $\mathcal{X} := [(\boldsymbol{X}^{(1)})^\top\ (\boldsymbol{X}^{(2)})^\top]^\top$, *then* $\boldsymbol{H}^{(1)}$ *and* $\boldsymbol{H}^{(2)}$ *can be uniquely identified.*

Proof See Appendix. □

This result shows that it is possible to uniquely identify both graph Volterra kernel matrices when the Kruskal condition is met even in the case that the model is self-driven, i.e., no exogenous inputs.

In the following, we present a result involving the exogenous inputs.

Theorem 3 *Let* $\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$ *and* $\boldsymbol{Y}$ *abide to the second-order AGVM*

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} + \boldsymbol{\Gamma}\boldsymbol{Y},$$

for sparse matrices $\boldsymbol{H}^{(1)}$ *with diagonal entries* $[\boldsymbol{H}^{(1)}]_{ii} = 0$ *and* $\boldsymbol{H}^{(2)}$ *satisfying A.2; and a diagonal matrix* $\boldsymbol{\Gamma}$ *with diagonal entries* $[\boldsymbol{\Gamma}]_{ii} \neq 0$. *Given a matrix* $\boldsymbol{\Pi}_1$ *such that* $\boldsymbol{X}^{(1)}\boldsymbol{\Pi}_1 = \boldsymbol{0}$ *and* $\boldsymbol{Y}\boldsymbol{\Pi}_1 \neq \boldsymbol{0}$, *if* $\mathrm{kr}(\boldsymbol{C}[\boldsymbol{X}^{(1)} * \boldsymbol{X}^{(1)}]\boldsymbol{\Pi}_1) \geq 2\mathrm{K}_2 + 1$, *where* $\boldsymbol{C}$ *is a binary selection matrix picking the appropriate rows of the Kronecker product, then the positions of the nonzero entries of* $\boldsymbol{H}^{(2)}$ *are unique.*

Proof See Appendix. □

Here, differently from the pure self-driven case, the presence of the exogenous term leads to a different identifiability condition: *structural identifiability*. That is, given that the condition on the Kruskal rank of the projected data matrix is met, the positions of the nonzero entries of the second-order graph Volterra kernels $\boldsymbol{H}^{(2)}$ can be uniquely

identified. When the graph Volterra kernels are related to the $(p+1)$-cliques [cf. (7)] in a network, the above result allows for higher-order link prediction. In addition, the support of $\boldsymbol{H}^{(1)}$ can then also be partially estimated from the nonzero entries of $\boldsymbol{H}^{(2)}$, as by assumption, in this case, their supports share a relation. More specifically, the existence of a triangle between nodes directly implies edges among the elements in the clique. However, as the nonexistence of a clique, e.g., a triangle, does not rule out the existence of an edge, not all edges, i.e., positions of the nonzeros of $\boldsymbol{H}^{(1)}$, can be identified.

To present an identifiability result using sparse recovery, we employ the following definition.

Definition 2 (*Restricted Isometry Property (RIP)*) Matrix $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ possesses the restricted isometry of order s, denoted as $\delta_s \in (0,1)$, if for all $\boldsymbol{h} \in \mathbb{R}^M$ with $\|\boldsymbol{h}\|_0 \leq s$ [34]

$$(1-\delta_s)\|\boldsymbol{h}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{h}\|_2^2 \leq (1+\delta_s)\|\boldsymbol{h}\|_2^2. \tag{13}$$

RIP is a fundamental property for providing identifiability conditions of sparse recovery. It has been shown that given $\delta_{2s} < \sqrt{2}-1$, the constrained version of the Lasso optimization problem

$$\min_{\boldsymbol{h}\in\mathbb{R}^M} \|\boldsymbol{h}\|_1 \quad \textit{subject to} \quad \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{h}\|_2 \leq \epsilon \tag{14}$$

yields $\|\boldsymbol{h}-\boldsymbol{h}^*\|_2^2 \leq c\epsilon^2$ for some constant c depending on δ_{2s} when the linear model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{h}^* + \boldsymbol{v}, \|\boldsymbol{v}\|_2 \leq \epsilon$ holds, where $\boldsymbol{h}^*$ is the solution of (14) [34, 35].

In the literature, in particular works on sparse polynomial regression [35] and Volterra series [36], several guarantees have been established for system matrices spawning from different alphabets and/or different distributions. For instance, in [24] results for Volterra system identification have been derived for signals drawn from $\{-1,0,1\}$ and in [35] signals drawn from $\mathcal{U}[-1,1]$. However, the bipolar case, e.g., $\{-1,1\}$, has not been considered and its treatment within the self-driven Volterra expansion is still missing. Therefore, in the following, we present a RIP result for the second-order AGVM, whose technical proof is detailed in Appendix.

Theorem 4 *Let $\{x_i(t)\}_{i=1}^N$ for $t \in [1,2,\ldots,T]$ be an input sequence of independent random variables drawn from the alphabet $\{-1,1\}$ with equal probability. Assume that the AGVM regression matrix is defined as*

$$\tilde{\boldsymbol{X}}^\top = \frac{1}{\sqrt{T}}[\boldsymbol{X}^l\, \boldsymbol{X}^b]^\top \in \mathbb{R}^{T\times L},$$

where $L = N + N(N-1)/2$, $\boldsymbol{X}^l := \boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^b$ is $\boldsymbol{X}^{(2)}$ with the quadratic terms removed, i.e., it only contains bilinear terms $x_i(t)x_j(t)$, $i \neq j$. Then, for any $\delta_s \in (0,1)$ and for any $\gamma \in (0,1)$, whenever $T \geq \frac{4C}{(1-\gamma)\delta_s^2} s^2 \log N$, the matrix $\tilde{\boldsymbol{X}}^\top$ possesses RIP δ_s with probability exceeding $1 - \exp\left(-\frac{\gamma\delta_s^2}{C}\cdot\frac{T}{s^2}\right)$, where $C = 2$.

Notice that the pure quadratic terms in $\boldsymbol{X}^{(2)}$ have been removed. This is due to the fact that for the bipolar signal case, these quadratic terms are constant when the alphabet is $\{-1, 1\}$ and equivalent to $\boldsymbol{X}^{(1)}$ when the alphabet is $\{0, 1\}$. Hence, in both cases its contribution can be omitted without loss of generality. Furthermore, the data matrices are normalized concerning for the number of available measurements, i.e., T. This is done in order to guarantee that the diagonal entries of the Grammian of $\tilde{\boldsymbol{X}}^{\top}$ are unity in expectation.

This theorem asserts that $T \in \mathcal{O}(s^2 \log N)$ observations suffice to recover an s-sparse vector with graph Volterra kernels. Since it is considered that the number of unknowns per row of $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ is at most $\mathcal{O}(N)$ [cf. 1-2], the bound on the sampling complexity scales as $\mathcal{O}(N^2 \log N)$ which agrees with bounds obtained for linear filtering setups [24, 37]; however, in this paper, the constant C is relatively small.

Given that under the established conditions, the proposed AGVM model is identifiable and is able to leverage sparsity to relieve its sampling complexity, in the following section, we present task-specific constraints for higher-order link inference and methods for estimating the graph Volterra kernels.

5 Real data applications

With the identifiability guarantees at hand, this section investigates how various learning tasks can benefit from the proposed AGVM. Specifically, two tailored AGVM algorithms for different applications namely topology identification in power systems and link prediction in social networks are presented.

5.1 Topology identification in distribution grids

The vertex set $\mathcal{V}$ of a graph in a distribution grid comprises the indices of the nodal buses, while the edge set $\mathcal{E}$ collects all the power distribution lines. The distribution grid is supposed to be a radial structure and thus the vertex set $\mathcal{V} := \{0, \mathcal{N}\}$ has a root (substation) bus indexed by $n = 0$. Every non-root bus $n \in \mathcal{N} = \{1, \ldots, N\}$ has a unique parent bus π_n. Naturally, the number of non-root buses is equal to the number of power lines in a radial network, that is, $|\mathcal{N}| = |\mathcal{E}| = N$[2].

In order to reveal both the edge connections as well as their higher-order interactions in power grids, we need to analyze the dependency of the signals on buses. The signal $x_n(t)$ here is the squared voltage magnitude of bus $n \in \mathcal{N}$ at time t. Based on our previous work [38], the voltage relationship among bus n and its children buses is given by

$$x_n(t) = \sum_i h_n^{(1)}(i) x_i(t) + \sum_i \sum_j h_n^{(2)}(i,j) x_i(t) x_j(t) + \epsilon_n(t),\ n \in \mathcal{N} \tag{15}$$

where $i \in \{k : k \in \mathcal{N}, \pi_k = n\}$ and $j \in \{k : k \in \mathcal{N}, \pi_k = n, k \geq i\}$; and $h_n^{(2)}(i,j)$ are the first- and second-order expansion coefficients relating bus n with the sets $\{i\}$ and $\{i,j\}$, respectively; ϵ_n comprises the modeling error and measurement noise on bus n. Collecting the data into $\{\boldsymbol{x}(t)\}_{t=1}^{T}$, and stacking the first- and second-order coefficients $h_n^{(1)}(i)$

[2] With a slight abuse of notation, we use N here to denote the number of non-root buses and hence the number of nodes in the network is $N + 1$.

and $h_n^{(2)}(i,j)$ into the vectors $\boldsymbol{h}_n^{(1)}$ and $\boldsymbol{h}_n^{(2)}$ in a lexicographic order. Similar to the derivation from (9) to (12), we can get the voltage on all buses (15) in a compact form as

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} + \boldsymbol{E} \tag{16}$$

where the t-th columns of $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ are $\boldsymbol{x}(t)$ and $\boldsymbol{x}(t) \boxtimes \boldsymbol{x}(t)$ respectively; the n-th rows of $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ are $(\boldsymbol{h}_n^{(1)})^\top$ and $(\boldsymbol{h}_n^{(2)})^\top$, respectively.

To estimate the coefficients of model (16), the following assumptions are needed.

A. 3 There is no self-interaction on a bus; therefore, the matrix $\boldsymbol{H}^{(1)}$ is a hollow matrix, i.e., $h_n^{(1)}(n) = 0, \ \forall n \in \mathcal{N}$.

A. 4 There is no second-order interactions between two buses, that is to say, the coefficients for the second-order interactions satisfy $h_n^{(2)}(j,k) = 0$, if $n = j$, or $j = k$, or $n = k$ holds.

A. 5 The second-order interactions only exist among two buses connected through a central bus. Thus, the graph Volterra coefficients obey $h_n^{(2)}(j,k) = 0$, if there exists $h_n^{(1)}(l) = 0$, $\forall\, l \in \{j,k\}$.

Assumptions A. 3 and A. 4 both entail linear constraints, and thus can be easily fit in an optimization problem. To cope with the conditional constraint in A. 5, we call for the auxiliary matrix

$$\boldsymbol{H}_n := \begin{bmatrix} h_n^{(1)}(1) & h_n^{(2)}(1,1) & \cdots & h_n^{(2)}(1,N) \\ h_n^{(1)}(2) & h_n^{(2)}(2,1) & \cdots & h_n^{(2)}(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ h_n^{(1)}(N) & h_n^{(2)}(N,1) & \cdots & h_n^{(2)}(N,N) \end{bmatrix} \tag{17}$$

where the first column equals $\boldsymbol{h}_n^{(1)}$. To guarantee that if $h_n^{(1)}(i) = 0$, then $h_n^{(2)}(i,j) = 0$, we enforce row sparsity in n by adding using $\ell_{2,1}$-regularization on $\boldsymbol{H}_n^\top$, $\forall n \in \mathcal{N}$. Based on the sparsity result from Sect. 4, we can estimate the expansion coefficients by the following sparsity-aware $\ell_{2,1}$-regularized least-squares

$$\min_{\{\boldsymbol{\theta}_n\}_{n=1}^N} \ \sum_{n\in\mathcal{N}} \|\boldsymbol{x}_n - \mathbf{M}^\top\boldsymbol{\theta}_n\|_2^2 + \lambda\|\boldsymbol{\theta}_n\|_1 + \mu\|\boldsymbol{H}_n^\top\|_{2,1} \tag{18a}$$

$$\text{s. to} \quad [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]^\top \in \mathcal{X}_h. \tag{18b}$$

where

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix}, \ \text{and} \ \boldsymbol{\theta}_n = \begin{bmatrix} \boldsymbol{h}_n^{(1)} \\ \boldsymbol{h}_n^{(2)} \end{bmatrix}$$

The set $\mathcal{X}_h$ is convex and signifies the constraints characterized by Assumptions A. 3 and A. 4. The convex optimization problem (18) can be efficiently solved (and distributed) by off-the-shelf convex programming toolboxes.

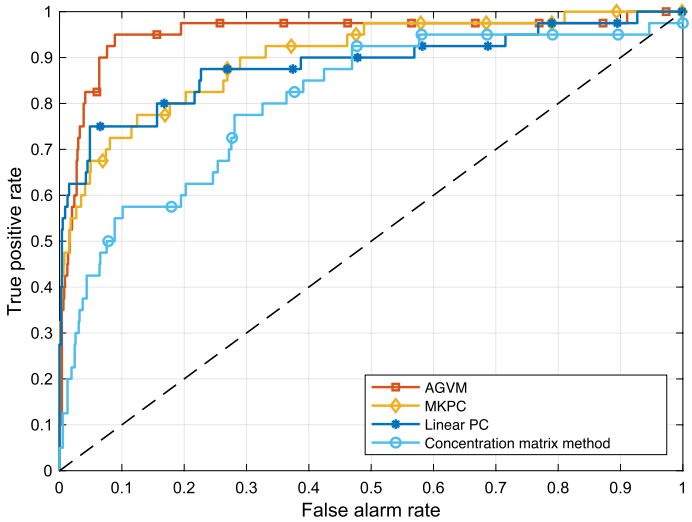

Fig. 1 ROC curves for topology inference of the SCE 47-bus distribution grid

To identify the underlying topology, the Southern California Edison (SCE) 47-bus distribution grid [3, 39] using real consumption and solar generation data from the Smart* project [40] was employed. Feeding this data to AC power flow equations [41, 42], we can obtain the voltage squared magnitude measurements $\{\boldsymbol{x}(t)\}_{t=1}^{T}$ across $T = 240$ time slots. The voltage magnitudes of the substation bus satisfies $v_0(t) = 1, \forall t \in \{1, \ldots, T\}$. The radial grid comprises 41 buses where the interactions need to be inferred, after ignoring the root bus and the buses connected to their parent buses with zero-impedance lines. With $\{\boldsymbol{x}(t)\}_{t=1}^{T}$, the graph Volterra kernels were estimated by solving (18) and constructing $\boldsymbol{H}^{(1)}$ and $\boldsymbol{H}^{(2)}$ as in (16). While removing non-significant entries by a pointwise thresholding operation, the grid topology was inferred from the support of $\boldsymbol{H}^{(1)}$ and the higher-order interactions were retrieved from $\boldsymbol{H}^{(2)}$.

To assess the performance of the proposed AGVM approach (18), we have simulated three edge connectivity recovery methods, namely: i) multi-kernel based partial correlations (MKPC)-scheme [22]; ii) linear PC-scheme [43]; and iii) concentration matrix-based scheme [44]. The results were measured by the empirical receiver operating characteristic (ROC) curves and the area under the curve (AUC) values. Figure 1 depicts the ROC curves of all methods, while the AUC for AGVM, MKPC, linear PC, and concentration matrix are 0.9483, 0.9008, 0.8836, and 0.8052, respectively. The results showcase the proposed scheme outperforms all competing alternatives by exploiting the nonlinear interactions.

The second experiment entailed the IEEE 123-bus feeder to examine the scalability and performance of the algorithm in topology identification. Voltage squared magnitude measurements $\{\boldsymbol{x}(t)\}_{t=1}^{T}$ across $T = 400$ time slots were used. The ROC curves of the AGVM, MKPC, linear PC, and concentration matrix method are shown in Fig. 2. The AUC values for the AGVM, MKPC, linear PC, and concentration matrix method are

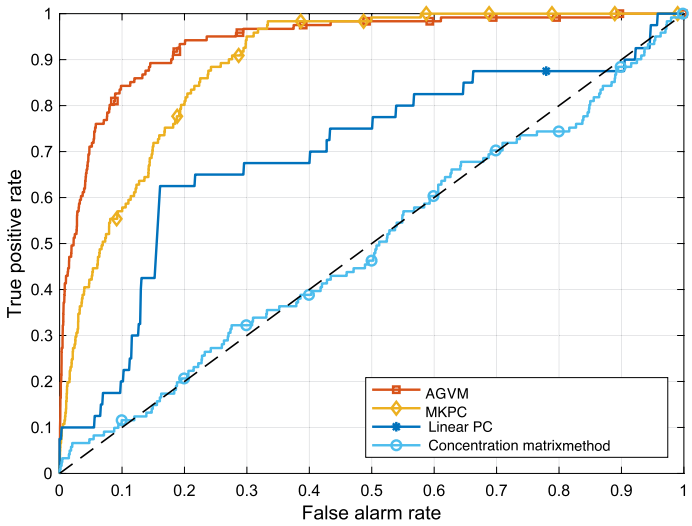

Fig. 2 ROC curves for topology inference of the IEEE 123-bus distribution grid

0.8935, 0.8024, 0.6952, and 0.4963, respectively. Evidently, these results corroborate the effectiveness of our proposed algorithm.

5.2 Link prediction in social networks

Another important domain inspiring higher-order interactions is link prediction in social and other biological networks. The goal of link prediction is to find the most likely subsets of vertices that will interact in the near future based on available observations of the activation of different nodes, e.g., song releases, email exchanges, and paper publications. Specifically, given a set of binary measurements $\boldsymbol{X} \in \{0,1\}^{N\times T}$ at time slots $t = \{1,\ldots,T\}$, one needs to predict what is the most likely set of nodes to be activated together at any $t' > T$.

In this subsection, we are considering the problem of predicting the closure of triangles, i.e., triplets of nodes that activate at the same time. Therefore, AGVMs can be restricted to order $P = 2$. Further, using the binary input data assumption, we can regard the interaction between variables as its joint activation and assume the function g in (7) is the product operation. As a result, a direct instantiation of an AGVM produces a model that constructs real-valued signals. Instead of directly modeling $x_i(t)$, we borrow an idea from binary regression methods and use a latent variable $z_i(t)$ to model the probability $P(x_i(t) = 1|z_i(t))$ as $P(x_i(t) = 1|z_i(t)) = \sigma(z_i(t))$, where $\sigma(\cdot)$ represents the sigmoid function. The latent variable $z_i(t)$ is then modeled as

$$z_i(t) = h_o^{(i)} + ((h_n^{(1)})^\top \boldsymbol{x}(t) + ((h_n^{(2)})^\top (\boldsymbol{x}(t) \boxtimes \boldsymbol{x}(t)). \tag{19}$$

Gathering the latent variables for nodes through time slots, the AGVM (12) for the link prediction task becomes

$$\boldsymbol{Z} = \boldsymbol{H}^{(0)}\mathbf{1}^\top + \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} = \boldsymbol{\Theta}^\top \boldsymbol{M} \tag{20}$$

where $[\boldsymbol{Z}]_{i,t} = z_i(t)$; $\boldsymbol{\Theta} = [\boldsymbol{H}^{(0)}\ \boldsymbol{H}^{(1)}\ \boldsymbol{H}^{(2)}]$; and $\boldsymbol{M} = [\mathbf{1}\ (\boldsymbol{X}^{(1)})^\top\ (\boldsymbol{X}^{(2)})^\top]^\top$. Different from traditional logistic regression, the binary labels in (20) are the node variables themselves.

Algorithm 1: AGVM for Link Prediction

Input: binary data: $\boldsymbol{M}$, parameter: λ, step-size: η, tolerance: ϵ, initial connectivity: $\boldsymbol{W}$

$\delta_{\bar{\boldsymbol{\theta}}} \leftarrow \infty$; $k = 0$; $\bar{\boldsymbol{\theta}}_k = \mathbf{0}$;
$\boldsymbol{B} = \texttt{build\_B\_mtx}(\boldsymbol{W})$;
$\boldsymbol{\Psi} := (\boldsymbol{M}^\top \otimes \boldsymbol{I})\boldsymbol{B}$;
while $(\delta_{\bar{\boldsymbol{\theta}}} > \epsilon)$ and $(k < k_{\max})$ **do**
  $\Delta_{\bar{\boldsymbol{\theta}}} = 0$;
  for $i = 1 : N$; $t = 1 : T$ **do**
    $\boldsymbol{\psi} = \texttt{get\_mtx\_row}(\boldsymbol{\Psi}, i, t)$;
    $\Delta_{\bar{\boldsymbol{\theta}}} = \Delta_{\bar{\boldsymbol{\theta}}} + \boldsymbol{\psi}\left([\boldsymbol{X}^{(1)}]_i(t) - \sigma(\bar{\boldsymbol{\theta}}_k^\top \boldsymbol{\psi})\right)$
  end
  $\bar{\boldsymbol{\theta}}_{k+1} = \texttt{soft\_thr}(\bar{\boldsymbol{\theta}}_k + \eta\Delta_{\bar{\boldsymbol{\theta}}}, \eta\lambda)$;
  $\delta_{\bar{\boldsymbol{\theta}}} = \|\bar{\boldsymbol{\theta}}_{k+1} - \bar{\boldsymbol{\theta}}_k\|_2$;
  $k = k + 1$;
end

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

The goal of closure prediction is to find 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 have not appeared simultaneously on a single simplex set, whose elements are the indexes of the nonzero elements of $\boldsymbol{x}(t)$ [45]. Therefore, based on the analysis in our previous work [45], we have the following conclusions. Observing the support of off-diagonal entries of $\boldsymbol{W} = \boldsymbol{X}^{(1)}(\boldsymbol{X}^{(1)})^\top$, we can obtain an initial network connectivity. From this connectivity, and enforcing $h_i^{(2)}(i,i) = 0, \forall i \in \mathcal{V}$, the set of open triangles $\mathcal{T}_O$, closed triangles $\mathcal{T}$, and the candidates for the nonzero graph Volterra coefficients $\mathcal{S}^{(2)} := \{\mathcal{S}_i^{(2)}\}_{i=1}^N$ (cf. (3)) can be obtained. Upon $\mathcal{S}^{(2)}$, it holds $\text{vec}(\boldsymbol{\Theta}) = \boldsymbol{B}\bar{\boldsymbol{\theta}}$, where $\bar{\boldsymbol{\theta}}$ captures the nonzero kernels, and $\boldsymbol{B}$ is an expansive binary matrix that relates the nonzero entries of the graph Volterra kernels in $\text{vec}(\boldsymbol{\Theta})$. These nonzero Volterra kernels are defined by the support obtained from $\boldsymbol{W}$. Noticing that we know the nonzero positions of $\text{vec}(\boldsymbol{\Theta})$, and the dimensions of $\text{vec}(\boldsymbol{\Theta})$ and $\boldsymbol{\theta}$, we can build this $\boldsymbol{B}$ matrix based on the initial network connectivity. To estimate the parameter $\bar{\boldsymbol{\theta}}$, we propose a proximal gradient ascent algorithm with sparsity regularization, which is summarized in Alg. 1. Notice that `get_mtx_row` takes the row of the input which is related to the latent variable $z_i(t)$; $\texttt{soft\_thr}(\cdot, \eta\lambda)$ entails soft threshold with respect to $\eta\lambda$. We assume that open triangles with large coefficients, which means a high level of interaction, are the most likely triangles to become closed. After obtaining $\bar{\boldsymbol{\theta}}$, we sort the entries related to $\mathcal{T}_O$ by their absolute value, and the top K entries are then the K most likely open triangles to become closed.

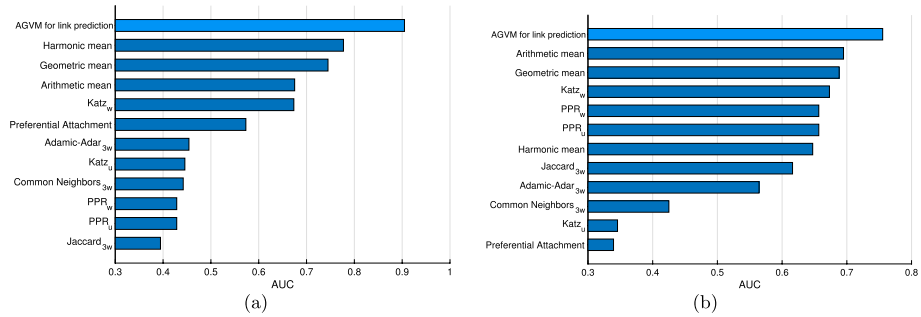

Fig. 3 AUC values for different models on dataset **a** "`Enron_email`" and **b** "`primary_school_contact`", namely harmonic, geometric, and arithmetic means of the three edge weights in the open triangle, Adamic-Adar model, preferential attachment model, dyadic link prediction Katz and personalized PageRank (PPR), Common Neighbors, and Jaccard; see [13] (and references therein) for more details. When applicable, the subscripts *w*, *u*, and 3*w* stand for weighted, unweighted and 3-way, respectively

Remark 2 The most expensive step in Alg. 1 is updating the gradient, which takes an effort of $\mathcal{O}(NTd)$, where d is the dimension of $\bar{\boldsymbol{\theta}}_k$. Thus, the complexity per iteration is $\mathcal{O}(NTd)$ as the rest of the operations incur at most a $\mathcal{O}(d)$ complexity. If we consider the worst case, that is, the algorithm runs until it hits its maximum number of iterations, $k_{\max}$, the overall worst-case complexity of the algorithm is then $\mathcal{O}(NTdk_{\max})$.

To examine the effectiveness of Alg. 1, the third experiment entailed the "`Enron_email`" [46] and "`primary_school_contact`" [47] datasets as well as several alternatives in [13] in terms of open triangle closure prediction. In the "`Enron_email`" dataset, the nodes denote the email addresses of different Enron employees, while in the "`primary_school_contact`" dataset, nodes are proximity-based contacts recorded by wearable sensors in a primary school. The training set contains the first 10% and 1% of timestamped events in the "`Enron_email`" and "`primary_school_contact`" datasets, respectively, while the testing set includes the remaining data. The proposed algorithm employs $\lambda = 10^{-3}$, $\eta = 10^{-4}$ and $k_{\max} = 500$ for both experiments. The AUC metric on the first 100 nodes of the datasets for all methods is shown in Fig. 3. The curves showcase the effectiveness of the proposed model along with the logistic regression of Alg. 1 compared with recently proposed methods based on generalizing the link prediction scores for the task of triangle closure prediction.

6 Conclusions

This paper proposes a principled manner to identify and predict the higher-order interactions in networked data. Borrowing ideas from SEMs and Volterra models, a node signal in the network is modeled as a combination of its neighbor signals and a nonlinear combination of the signals in the groups (higher-order links) it belongs to. Some identifiability guarantees of the proposed second-order AGVM are then provided under

three conditions that input/output data exhibit. Our model provides both expressibility for higher-order interactions, as well as interpretability for further understanding of the underlying network dynamics. Moreover, the proposed AGVM is particularized to handle two different applications, which are, topology identification in power systems and link prediction in social networks. The merits of the proposed algorithms relative to existing methods are corroborated through numerical tests using real data. This work also opens up interesting directions for future research, including avoiding the computational burden for large-scale networks as well as generalizations of the higher-order model to other challenging applications.

Appendix

Proof of Theorem 1 Let us consider the expansion

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} + \boldsymbol{\Gamma}\boldsymbol{Y}.$$

Rewriting the linear terms, i.e., terms related with $\boldsymbol{X}^{(1)}$, we obtain the system

$$\mathcal{H}\mathcal{X} = \boldsymbol{\Gamma}\boldsymbol{Y}$$

where $\mathcal{X} := [(\boldsymbol{X}^{(1)})^\top\ (\boldsymbol{X}^{(2)})^\top]^\top$ and $\mathcal{H} := [(\boldsymbol{I} - \boldsymbol{H}^{(1)}) \mid -\boldsymbol{H}^{(2)}]$.

Due to hypothesis that $\mathcal{X}$ is full row rank, the unique least-squares solution for the kernel matrices is obtained by $\mathcal{H}_{\text{LS}} := \boldsymbol{\Gamma}\boldsymbol{Y}\mathcal{X}^\dagger$. Defining $\boldsymbol{Q} := \boldsymbol{Y}\mathcal{X}^\dagger$ and partitioning this matrix appropriately, i.e., $\boldsymbol{Q} = [\boldsymbol{Q}_1 \in \mathbb{R}^{N\times N} \mid \boldsymbol{Q}_2 \in \mathbb{R}^{N\times M}]$,

we obtain the following relations

$$\begin{aligned}\boldsymbol{I} - \boldsymbol{H}^{(1)} &= \boldsymbol{\Gamma}\boldsymbol{Q}_1\\ -\boldsymbol{H}^{(2)} &= \boldsymbol{\Gamma}\boldsymbol{Q}_2.\end{aligned}$$

Now, let us recall that $\boldsymbol{\Gamma}$ is a diagonal matrix and that $\boldsymbol{H}^{(1)}$ is a hollow matrix, i.e., its diagonal is filled with zeros. Thus, it holds $\text{diag}(\boldsymbol{\Gamma}\boldsymbol{Q}_1) = \boldsymbol{1}$,

which implies $\text{diag}(\boldsymbol{\Gamma}_{\text{LS}}) := \text{diag}(\boldsymbol{Q}_1)^{-1}$.

Finally, the estimates for the kernel matrices are given as

$$\begin{aligned}\boldsymbol{H}^{(1)}_{\text{LS}} &:= \boldsymbol{I} - \boldsymbol{\Gamma}_{\text{LS}}\boldsymbol{Q}_1\\ \boldsymbol{H}^{(2)}_{\text{LS}} &:= -\boldsymbol{\Gamma}_{\text{LS}}\boldsymbol{Q}_2.\end{aligned}$$

The proof is completed. □

Proof of Theorem 2 First, let us rewrite the expansion as

$$\begin{aligned}(\boldsymbol{X}^{(1)})^\top &= [(\boldsymbol{X}^{(1)})^\top\ (\boldsymbol{X}^{(2)})^\top] \begin{bmatrix}(\boldsymbol{H}^{(1)})^\top \\ (\boldsymbol{H}^{(2)})^\top\end{bmatrix} \\ &= \mathcal{X}^\top \mathcal{H}^\top.\end{aligned}$$

Now, consider the ith column of both sides of the expression above, i.e., $[(\boldsymbol{X}^{(1)})^\top]_i = \mathcal{X}^\top [\mathcal{H}^\top]_i$.

Suppose there exists a vector $\boldsymbol{h}_i$, $\boldsymbol{h}_i \neq [\mathcal{H}^\top]_i$, satisfying the same relation and with $K = K_1 + K_2$ nonzero entries. This implies that $\mathbf{0} = \mathcal{X}^\top([\mathcal{H}^\top]_i - \boldsymbol{h}_i)$

must hold. As $[\mathcal{H}]_i^\top$ and $\boldsymbol{h}_i$ have at most K entries, their difference has at most $2K$ nonzero entries. Hence, if $\mathrm{kr}(\mathcal{X}^\top) \geq 2\mathrm{K}$, any possible subset of columns of $\mathcal{X}^\top$ are linearly independent. Thus, $([\mathcal{H}^\top]_i - \boldsymbol{h}_i) = \mathbf{0}$ holds. This contradicts the assumption, hence the result of the theorem holds. □

Proof of Theorem 3 Let us consider the expansion

$$\boldsymbol{X}^{(1)} = \boldsymbol{H}^{(1)}\boldsymbol{X}^{(1)} + \boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)} + \boldsymbol{\Gamma}\boldsymbol{Y}.$$

Notice that the jth column of $\boldsymbol{X}^{(2)}$ is given by $\boldsymbol{C}\left([\boldsymbol{X}^{(1)}]_j \otimes [\boldsymbol{X}^{(1)}]_j\right)$ where $[\boldsymbol{X}^{(1)}]_j$ is the jth column of $\boldsymbol{X}^{(1)}$, $\otimes$ is the Kronecker product. and $\boldsymbol{C}$ is a binary selection matrix picking the appropriate rows of the Kronecker product. Hence, we can express $\boldsymbol{X}^{(2)}$ using the Khatri-Rao product $*$, and $\boldsymbol{X}^{(1)}$ as

$$\begin{aligned}\boldsymbol{X}^{(2)} &= \boldsymbol{C}\left[[\boldsymbol{X}^{(1)}]_1 \otimes [\boldsymbol{X}^{(1)}]_1\ [\boldsymbol{X}^{(1)}]_2 \otimes [\boldsymbol{X}^{(1)}]_2\ \cdots\right] \\ &= \boldsymbol{C}(\boldsymbol{X}^{(1)} * \boldsymbol{X}^{(1)}).\end{aligned}$$

By hypothesis, we have that $\boldsymbol{X}^{(1)}\boldsymbol{\Pi}_1 = \mathbf{0}$. Thus, by right multiplying the expansion by $\boldsymbol{\Pi}_1$ and reorganizing terms, we obtain

$$\begin{aligned}-\boldsymbol{H}^{(2)}\boldsymbol{X}^{(2)}\boldsymbol{\Pi}_1 &= \boldsymbol{\Gamma}\tilde{\boldsymbol{Y}} \\ -\boldsymbol{H}^{(2)}\boldsymbol{C}(\boldsymbol{X}^{(1)} * \boldsymbol{X}^{(1)})\boldsymbol{\Pi}_1 &= \boldsymbol{\Gamma}\tilde{\boldsymbol{Y}}\end{aligned}$$

where $\tilde{\boldsymbol{Y}} := \boldsymbol{Y}\boldsymbol{\Pi}_1 \neq \mathbf{0}$ (by assumption) and the identity $\boldsymbol{X}^{(2)} = \boldsymbol{C}(\boldsymbol{X}^{(1)} * \boldsymbol{X}^{(1)})$ has been used. Now, let us consider the ith equation of the above relation, i.e.,

$$-(\tilde{\boldsymbol{X}}^{(2)})^\top \boldsymbol{h}_i^{(2)} = [\boldsymbol{\Gamma}]_{ii}\tilde{\boldsymbol{y}}_i$$

where $\tilde{\boldsymbol{X}}^{(2)} := \boldsymbol{X}^{(2)}\boldsymbol{\Pi}_1$; $\boldsymbol{h}_i^{(2)}$ and $\tilde{\boldsymbol{y}}_i$ are the ith rows of $\boldsymbol{H}^{(2)}$ and $\tilde{\boldsymbol{Y}}$ in column-vector form, respectively. As $\mathrm{kr}(\tilde{\boldsymbol{X}}^{(2)}) \geq 2\mathrm{K}_2 + 1$ and the number of nonzero elements per row in $\boldsymbol{H}^{(2)}$ is bounded above by K_2 by hypothesis, the above system has a unique solution (up to an

scalar ambiguity that does not affect the support) with such a sparsity level, i.e., sparsest solution, $\forall\, i$. □

Proof of Theorem 4 Let us consider a particular realization of $\tilde{\boldsymbol{X}}^\top := [\frac{1}{\sqrt{T}}\boldsymbol{X}^l \frac{1}{\sqrt{T}}\boldsymbol{X}^b]^\top = [\tilde{\boldsymbol{X}}^l \tilde{\boldsymbol{X}}^b] \in \{-1,1\}^{T\times L}$ and its Grammian $\tilde{\boldsymbol{G}} := \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$. By the Gersgorin disc theorem [48], if $|[\tilde{\boldsymbol{G}}]_{ii} - 1| < \delta_d$ and $|[\tilde{\boldsymbol{G}}]_{ij}| \leq \frac{\delta_o}{s}$ for every i, j with $j \neq i$ and $\delta_d + \delta_o = \delta$ for some $\delta \in (0,1)$, then $\tilde{\boldsymbol{X}}$ possesses RIP $\delta_s \leq \delta$. Therefore, we can upper bound the probability of $\tilde{\boldsymbol{X}}$ not satisfying RIP of value δ, $\Pr(\delta_s > \delta)$, as

$$\Pr\left(\bigcup_{i=1}^{L}\left\{|[\tilde{\boldsymbol{G}}]_{ii} - 1| \geq \delta_d\right\} \text{ or } \bigcup_{i=1}^{L}\bigcup_{j=1, j\neq i}^{L}\left\{|[\tilde{\boldsymbol{G}}]_{ij}| \geq \frac{\delta_o}{s}\right\}\right).$$

As $\tilde{\boldsymbol{G}}$ is symmetric, we can use the union bound only for its unique entries to upper bound $\Pr(\delta_s > \delta)$ as

$$\sum_{i=1}^{L}\Pr\left(|[\tilde{\boldsymbol{G}}]_{ii} - 1| \geq \delta_d\right) + \sum_{i=1}^{L}\sum_{j=i+1}^{L}\Pr\left(|[\tilde{\boldsymbol{G}}]_{ij}| \geq \frac{\delta_o}{s}\right).$$

To show the result of the theorem, we proceed next to bound the probabilities above. The analysis of these probabilities is similar to the one in [24]. However, here we obtain results for a different distribution and for linear and bilinear components. To simplify the notation, we introduce the following partition for the Grammian matrix, i.e.,

$$\tilde{\boldsymbol{G}} = \begin{bmatrix} \tilde{\boldsymbol{G}}^{ll} & \tilde{\boldsymbol{G}}^{lb} \\ (\tilde{\boldsymbol{G}}^{lb})^\top & \tilde{\boldsymbol{G}}^{bb} \end{bmatrix}$$

where $\tilde{\boldsymbol{G}}^{ll} := \tilde{\boldsymbol{X}}^l(\tilde{\boldsymbol{X}}^l)^\top$, $\tilde{\boldsymbol{G}}^{lb} := \tilde{\boldsymbol{X}}^l(\tilde{\boldsymbol{X}}^b)^\top$ and $\tilde{\boldsymbol{G}}^{bb} := \tilde{\boldsymbol{X}}^b(\tilde{\boldsymbol{X}}^b)^\top$.

Recalling that the raw moments for the inputs are given by

$$m_r = \begin{cases} 0 & r \text{ odd} \\ 1 & r \text{ even} \end{cases}$$

we obtain the following relations

$$[\tilde{\boldsymbol{G}}^{ll}]_{ii} = 1, \quad \mathbb{E}\{[\tilde{\boldsymbol{G}}^{ll}]_{ii}\} = 1, \quad \mathbb{E}\{[\tilde{\boldsymbol{G}}^{ll}]_{ij}\} = 0$$
$$[\tilde{\boldsymbol{G}}^{bb}]_{ii} = 1, \quad \mathbb{E}\{[\tilde{\boldsymbol{G}}^{bb}]_{ii}\} = 1, \quad \mathbb{E}\{[\tilde{\boldsymbol{G}}^{bb}]_{ij}\} = 0$$

and $\mathbb{E}\{[\tilde{\boldsymbol{G}}^{lb}]_{ij}\} = 0\ \forall\, i, j$. By a quick inspection, we notice that the terms of the first part of $\Pr(\delta_s > \delta)$ are identically zero, hence $\delta_d = 0$.

To bound the required probabilities, we make use of the following Hoeffding's inequality.

Lemma 1 (Hoeffding's Inequality) *Given $t > 0$ and independent random variables $\{x_i\}_{i=1}^N$ bounded as $a_i \leq x_i \leq b_i$ almost surely, the sum $s_N := \sum_{i=1}^N x_i$ satisfies*

$$\Pr(|\mathrm{s_N} - \mathbb{E}\{\mathrm{s_N}\}| \geq \mathrm{t}) \leq 2\exp\left(-\frac{2\mathrm{t}^2}{\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{b_i}-\mathrm{a_i})^2}\right).$$

Let us consider the off-diagonal elements of $\tilde{\boldsymbol{G}}^{ll}$. The related probability can be bounded as

$$\Pr\left(|[\tilde{\boldsymbol{G}}^{ll}]_{ij}| \geq \frac{\delta_o}{s}\right) \leq 2\exp\left(-\frac{\delta_o^2 T}{2s^2}\right)$$

as we consider that these entries are the result of a sum of T independent variables contained in $\{-1, 1\}$. Similar bounds can be found for the other entries of $\tilde{\boldsymbol{G}}$, i.e.,

$$\Pr\left(|[\tilde{\boldsymbol{G}}^{bb}]_{ij}| \geq \frac{\delta}{s^2}\right) \leq 2\exp\left(-\frac{\delta_o^2 T}{2s^2}\right),\ \Pr\left(|[\tilde{\boldsymbol{G}}^{lb}]_{ij}| \geq \frac{\delta}{s^2}\right) \leq 2\exp\left(-\frac{\delta_o^2 T}{2s^2}\right).$$

Recollecting the probabilities for all the entries, we obtain

$$\begin{aligned}\Pr(\delta_{\mathrm{s}} > \delta) :&= \sum_{i=1}^{L}\sum_{j\geq 1}^{L}\Pr\left(|[\tilde{\boldsymbol{G}}^{ll}]_{ij}| \geq \frac{\delta_o}{s}\right) + \sum_{i=1}^{L}\sum_{j\geq 1}^{L}\Pr\left(|[\tilde{\boldsymbol{G}}^{bb}]_{ij}| \geq \frac{\delta_o}{s}\right)\\ &\quad + \sum_{i=1}^{L}\sum_{j\geq 1}^{L}\Pr\left(|[\tilde{\boldsymbol{G}}^{lb}]_{ij}| \geq \frac{\delta_o}{s}\right)\\ &\leq (N^2 + \frac{1}{8}N^4 + \frac{1}{2}N^3)\exp\left(-\frac{\delta_o^2 T}{2s^2}\right)\\ &\leq N^4\exp\left(-\frac{\delta_o^2 T}{2s^2}\right)\end{aligned}$$

for $N > 2$. Considering $\delta = \delta_o$ (as $\delta_d = 0$) and setting $C = 2$, for a $\gamma \in (0, 1)$ and $T \geq \frac{4C}{(1-\gamma)\delta^2}s^2\log N$, we can simplify the above bound as

$$\Pr\left(\delta_{\mathrm{s}} > \delta\right) \leq \exp\left(-\frac{\gamma\delta^2\mathrm{T}}{\mathrm{Cs}^2}\right)$$

□

Abbreviations

AGVM	Autoregressive graph Volterra model
SEM	Structural equation model
RIP	Restricted isometry property
PGA	Proximal gradient ascend
MKPC	Multi-kernel-based partial correlations
ROC	Receiver operating characteristic
AUC	Area under the curve
PPR	Personalized PageRank similarity

Acknowledgements
Not applicable.

Author contributions
All authors contributed to this research, including the design of the simulations and analyses of the results. GL proposed the research direction. QLY and MC conceived the system model. MC completed the identifiability analysis of the paper. QLY analyzed the data and wrote this manuscript. GBG proofread the paper. All authors read and approved the final manuscript.

Funding
This work was supported in part by ASPIRE project 14926 (within the STW OTP program) financed by the Netherlands Organization for Scientific Research (NWO), NSF Grants 1509040, 1711471, and 1901134.

Availability of data and materials
All data generated or analyzed during this study are included in this paper.

Declarations

Ethics approval and consent to participate
All procedures performed in this paper were in accordance with the ethical standards of research community.

Competing interests
The authors declare that they have no competing interests.

Received: 25 March 2022 Accepted: 14 December 2022
Published online: 09 January 2023

References

1. D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Tec. **58**(7), 1019–1031 (2007)
2. S. Golshannavaz, S. Afsharnia, F. Aminifar, Smart distribution grid: optimal day-ahead scheduling with reconfigurable topology. IEEE Trans. Smart Grid **5**(5), 2402–2411 (2014)
3. Q. Yang, G. Wang, A. Sadeghi, G.B. Giannakis, J. Sun, Two-timescale voltage control in distribution grids using deep reinforcement learning. IEEE Trans. Smart Grid **11**(3), 2313–23 (2020)
4. S. Sulaimany, M. Khansari, A. Masoudi-Nejad, Link prediction potentials for biological networks. Int. J. Data Min. Bioinf. **20**(2), 161–184 (2018)
5. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
6. S. Barbarossa, S. Sardellitti, Topological signal processing over simplicial complexes. IEEE Trans. Signal Process. **68**, 2992–3007 (2020)
7. P. Frankl, V. Rödl, Extremal problems on set systems. Random Struct. Algorithm **20**(2), 131–164 (2002)
8. S. Sardellitti, S. Barbarossa, L. Testa, Topological signal processing over cell complexes, in *IEEE Asilomar Conf. Signals, Systems, Computers*, pp. 1558–1562 (2021)
9. M.T. Schaub, A.R. Benson, P. Horn, G. Lippner, A. Jadbabaie, Random walks on simplicial complexes and the normalized hodge laplacian. arXiv:1807.05044 (2018)
10. S. Zhang, Z. Ding, S. Cui, Introducing hypergraph signal processing: theoretical foundation and practical applications. IEEE Internet Things J. **7**(1), 639–660 (2020)
11. G.B. Giannakis, Y. Shen, G.V. Karanikolas, Topology identification and learning over graphs: accounting for nonlinearities and dynamics. Proc. IEEE **106**(5), 787–807 (2018)
12. M. Coutino, E. Isufi, T. Maehara, G. Leus, State-space network topology identification from partial observations. arXiv:1906.10471 (2019)
13. A.R. Benson, R. Abebe, M.T. Schaub, A. Jadbabaie, J. Kleinberg, Simplicial closure and higher-order link prediction. Proc. Natl. Acad. Sci. **115**(48), 11221–11230 (2018)
14. L. Lim, Hodge laplacians on graphs. arXiv:1507.05379 (2015)
15. S. Ebli, M. Defferrard, G. Spreemann, Simplicial neural networks. arXiv:2010.03633 (2020)
16. L. Giusti, C. Battiloro, P. Di Lorenzo, S. Sardellitti, S. Barbarossa, Simplicial attention networks. Preprint arXiv:2203.07485 (2022)
17. M. Yang, E. Isufi, G. Leus, Simplicial convolutional neural networks, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022), pp. 8847–8851
18. J.J. Hox, T.M. Bechger, An introduction to structural equation modeling. Family Sci. Rev. **11**, 354–373 (1998)
19. H. Lütkepohl, *Vector Autoregressive Models* (Springer, Berlin, 2011)
20. G.V. Karanikolas, O. Sporns, G.B. Giannakis, Multi-kernel change detection for dynamic functional connectivity graphs, in *Asilomar Conf. on Signals, Syst., and Comput., Pacific Grove, CA, USA* pp. 1555–1559 (2017)
21. E. Isufi, A. Loukas, N. Perraudin, G. Leus, Forecasting time series with VARMA recursions on graphs. IEEE Trans. Signal Process. **67**(18), 4870–4885 (2019)
22. L. Zhang, G. Wang, G.B. Giannakis, Going beyond linear dependencies to unveil connectivity of meshed grids, in *Proc. of CAMSAP, Curacao, AN*, pp. 1–5 (2017)
23. D. Song, R.H. Chan, V.Z. Marmarelis, R.E. Hampson, S.A. Deadwyler, T.W. Berger, Nonlinear dynamic modeling of spike train transformations for hippocampal-cortical prostheses. IEEE Trans. Biomed. Eng. **54**(6), 1053–1066 (2007)
24. V. Kekatos, G.B. Giannakis, Sparse volterra and polynomial regression models: recoverability and estimation. IEEE Trans. Signal Process. **59**(12), 5907–5920 (2011)
25. C. Krall, K. Witrisal, G. Leus, H. Koeppl, Minimum mean-square error equalization for second-order Volterra systems. IEEE Trans. Signal Process. **56**(10), 4729–4737 (2008)
26. V.Z. Marmarelis, Identification of nonlinear biological systems using Laguerre expansions of kernels. Ann. Biomed. Eng. **21**(6), 573–589 (1993)
27. H. Huang, J. Tang, L. Liu, J. Luo, X. Fu, Triadic closure pattern analysis and prediction in social networks. IEEE Trans. Knowl. Data Eng. **27**(12), 3374–3389 (2015)

28. M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, M.A. Porter, Multilayer networks. J. Complex Netw. **2**(3), 203–271 (2014)
29. P. Prenter, A Weierstrass theorem for real, separable Hilbert spaces. J. Approxim. Theory **3**(4), 341–351 (1970)
30. S. Boyd, L. Chua, Fading memory and the problem of approximating nonlinear operators with Volterra series. IEEE Trans. Circuits syst. **32**(11), 1150–1161 (1985)
31. M. Schetzen, The volterra and wiener theories of nonlinear systems (1980)
32. J.A. Bazerque, B. Baingana, G.B. Giannakis, Identifiability of sparse structural equation models for directed and cyclic networks, in *IEEE Glob. Conf. Signal Inf. Process*, pp. 839–842 (2013)
33. J.B. Kruskal, Rank, decomposition, and uniqueness for 3-way and N-way arrays. Multiway Data Analysis, 7–18 (1989)
34. E.J. Candes, The restricted isometry property and its implications for compressed sensing. C. R. Acad. Sci. Paris, Ser. I **346**(9–10), 589–592 (2008)
35. B. Nazer, R.D. Nowak, Sparse interactions: Identifying high-dimensional multilinear systems via compressed sensing, in *Annu. Allert. Conf. Commun. Control Comput. Allert*, pp. 1589–1596 (2010)
36. V. Kekatos, D. Angelosante, G.B. Giannakis, Sparsity-aware estimation of nonlinear volterra kernels, in *IEEE CAMSAP*, pp. 129–132 (2009)
37. J. Haupt, W.U. Bajwa, G. Raz, R. Nowak, Toeplitz compressed sensing matrices with applications to sparse channel estimation. IEEE Trans. Inf. Theory **56**(11), 5862–5875 (2010)
38. Q. Yang, M. Coutino, G. Wang, G.B. Giannakis, G. Leus, Learning connectivity and higher-order interactions in radial distribution grids, in *IEEE Int. Conf. Acoust. Speech Signal Process*, pp. 5555–5559 (2020)
39. M. Farivar, C.R. Clarke, S.H. Low, K.M. Chandy, Inverter VAR control for distribution systems with renewables, in *Proc. of IEEE SmartGridComm., Brussels, Belgium*, pp. 457–462 (2011)
40. S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J. Albrecht, Smart*: an open data set and tools for enabling research in sustainable homes. SustKDD **111**(112), 108 (2012)
41. M. Baran, F.F. Wu, Optimal sizing of capacitors placed on a radial distribution system. IEEE Trans. Power Del. **4**(1), 735–743 (1989)
42. S.H. Low, Convex relaxation of optimal power flow–Part II: exactness. IEEE Trans. Control Netw. Syst. **1**(2), 177–189 (2014)
43. S. Bolognani, N. Bof, D. Michelotti, R. Muraro, L. Schenato, Identification of power distribution network topology via voltage correlation analysis, in *Proc. of CDC, Florence, ITL*, pp. 1659–1664 (2013)
44. D. Deka, S. Talukdar, M. Chertkov, M. Salapaka, Topology estimation in bulk power grids: Guarantees on exact recovery. arXiv:1707.01596 (2017)
45. M. Coutino, G.V. Karanikolas, G. Leus, G.B. Giannakis, Self-driven graph volterra models for higher-order link prediction, in *IEEE Int. Conf. Acoust. Speech Signal Process*, pp. 3887–3891 (2020)
46. B. Klimt, Y. Yang, The enron corpus: A new dataset for email classification research, in *ECOML* (Spinger, 2004), pp. 217–226
47. 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 **6**(8), 23176 (2011)
48. R. Marsli, F.J. Hall, Geometric multiplicities and geršgorin discs. Am. Math. Mon. **120**(5), 452–455 (2013)

Publisher's Note