Task-Aware Connectivity Learning for Incoming Nodes Over Growing Graphs

Das, Bishwadeep; Hanjalic, Alan; Isufi, Elvin

DOI
10.1109/TSIPN.2022.3206578

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Signal and Information Processing over Networks

Task-Aware Connectivity Learning for Incoming Nodes Over Growing Graphs

Bishwadeep Das, *Graduate Student Member, IEEE*, Alan Hanjalic, *Fellow, IEEE*, and Elvin Isufi, *Member, IEEE*

***Abstract*—Data processing over graphs is usually done on graphs of fixed size. However, graphs often grow with new nodes arriving over time. Knowing the connectivity information of these nodes, and thus, the expanded graph is crucial for processing data over the expanded graph. In its absence, its inference and the subsequent data processing become essential. This paper provides contributions along this direction by considering task-driven data processing for incoming nodes without connectivity information. We model the incoming node attachment as a random process dictated by the parameterized vectors of probabilities and weights of attachment. The attachment is driven by the existing graph topology, the corresponding graph signal, and an associated processing task. We consider two such tasks, one of interpolation at the incoming node, and that of graph signal smoothness. We show that the model bounds implicitly the spectral perturbation between the nominal topology of the expanded graph and the drawn realizations. In the absence of connectivity information our topology, task, and data-aware stochastic attachment performs better than purely data-driven and topology driven stochastic attachment rules, as is confirmed by numerical results over synthetic and real data.**

***Index Terms*—Graph signal interpolation, graph signal processing, graph smoothness, graph topology identification, incoming nodes, spectral perturbation.**

I. Introduction

GRAPH topology identification is a crucial step preceding the analysis of relationships of users in social networks [2], proteins in biological networks [3], and entities in recommender systems [4], to name just a few. Typical approaches infer a topology with a fixed number of nodes [5], [6] but graphs often grow with new nodes attaching to the existing ones [7]. This attachment is often unknown, making the downstream tasks more challenging. One such task is the cold start recommendation in collaborative filtering [8]. Here, a new user enters the system but cannot attach to the existing ones in the absence of prior information, thereby affecting the subsequent recommendation. In another scenario, a new online political blog becomes available and we want to know how it associates with the existing political blogs without knowing its affiliation and its influence on the overall network [9].

We want to process data at the incoming node in situations when the true connectivity is unknown. This may be the case when side information in collaborative filtering is unavailable or when new blogs are not associated with a particular category. However, at the same time, we want to process data in the most informed way possible. One way is to consider how the previous nodes connect and apply that rule for the incoming node. Differently, we want to handle the attachment in the context of a data-processing task over the network, focused at the new node. Taking the task into consideration will lead to more relevant attachments, possibly improving upon the performance of other attachment rules.

Existing graph identification approaches infer the full [5] or partial [6] connectivity of a fixed graph but do not consider incoming nodes, while stochastic approaches model the connectivity with a known attachment model [7], [10], [11], [12] but ignore the existing data and how it ought to be processed. Some recent works that process data over expanding graphs require the connectivity knowledge [13], [14], which is often unavailable. Thus, information processing for incoming nodes in such situations is challenging.

One way to overcome this challenge is to consider a data-driven stochastic model, where the available data is leveraged to learn the mapping between the incoming nodes and the existing graph for the task at hand. When combined with a prediction mechanism, such a hybrid approach can overcome the limitations of purely stochastic or data-oriented predictions We develop a stochastic attachment model for incoming nodes given a graph signal processing (GSP) task [15]. We combine preferential attachment with GSP and topology identification for modelling the incoming node connectivity. The mechanisms of such a framework come with the following three contributions.

1) We formulate a task-driven attachment model for incoming nodes without connectivity information. Such a model is parameterized by the probabilities of attachment and the edge-weights. We specialize the model to two GSP tasks. The first task is graph signal interpolation, where the aim is to predict the signal value at the incoming node. The second task is to learn the attachment of the incoming node such that the graph signal is smooth over the expanded graph.

Manuscript received 21 March 2022; revised 7 August 2022; accepted 28 August 2022. Date of publication 23 September 2022; date of current version 31 October 2022. The work of Bishwadeep Das was supported by TU Delft. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wei Hu. *(Corresponding author: Bishwadeep Das.)*

The authors are with the Multimedia Computing Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: b.das@tudelft.nl; a.hanjalic@tudelft.nl; e.isufi-1@tudelft.nl).

Digital Object Identifier 10.1109/TSIPN.2022.3206578

2) We propose an empirical risk minimization problem to estimate the model parameters and solve them using alternating projected gradient descent. We discuss the convergence properties of this approach to a local minimum.
3) We study the learned connectivity from a perturbation viewpoint. We look at the small perturbation analysis of the eigenvalues of the nominal graph relative to the model realizations. Using this result, we link the task-specific costs with the expected squared perturbation and show that the proposed algorithm learns a topology that keeps this perturbation in check. We corroborate the proposed approach with numerical results on synthetic and real data from recommender systems and blog networks.

The remainder of this paper is organized as follows: Section II elaborates on the related works; Section III formulates our problem; Section IV contains the task-aware connectivity modelling for incoming nodes and discusses the algorithm's convergence; Section V discusses the perturbation analysis; Section VI contains the numerical results. Section VII concludes the paper. All proofs are collected in the appendix.

II. Related Works

Inferring node connections has been approached from different viewpoints ranging from GSP to statistical models. Here, we cast our work w.r.t. these existing frameworks.

Graph Signal Processing (GSP): Topology identification via GSP estimates a fixed topology from data by leveraging different priors, such as signal smoothness [16], [17], [18], realizations of a diffusion process [19], [20], [21], [22], or a Gaussian process [23], [24], [25], to name a few. These priors have also been used to estimate time-varying topologies where a fixed topology is estimated per batch of data [26], [27], [28]. More recently, *online methods* avoid batch processing and estimate the topology on the fly from time-varying signals [29], [30], [31], [32], [33]. Differently, we will learn a stochastic model for incoming nodes rather than a fixed topology. And differently from the online methods, we consider an expanding topology but with a fixed time-invariant signal. As in these approaches, we will also consider the smoothness criterion, which is typically encountered in practice because of homophily (i.e., connected nodes share similar values) [34].

Statistical Methods: Connectivity of incoming nodes is commonly approached in network science via stochastic models, such as the *Erdős-Rényi* (ER) and the *preferential attachment*. The ER model assumes each incoming node connects uniformly at random with any of the existing nodes [10], while preferential attachment assumes each incoming node connects with a probability proportional to a node's degree [7]. More complex models include a competition factor between nodes [11], [12]. Altogether, these methods focus on the existing topology and do not account for the data over it. Accounting for the data is paramount to solving network learning tasks because of the implicit data-topology coupling. Therefore, we propose to estimate the attachment model parameters, i.e., attachment probabilities and edge weights of the incoming nodes by incorporating both the data and the task into the learning.

Link Prediction: Modelling the connectivity of incoming nodes can also be seen as a link prediction task given some topological and nodal features [6], [35]. Link prediction techniques can be grouped into three categories: i) *probabilistic approaches* that use random models to predict links using, for example, hierarchical graphs and stochastic block models [36], [37]; ii) *similarity-based approaches* that predict a link between any two nodes based on their common neighborhood features [38], [39] or global graph features [40], [41]; and iii) *classifier-based approaches* that train a machine learning model based on node features. However, most of these approaches fail in the incoming node scenario because we have no topological information about the incoming nodes and, in the absence of node features, classifier-based approaches are also inapplicable.

Learning on Expanding Graphs: Lastly, recent works consider specific expanding graph models or solve a specific task with the knowledge of the connectivity. The works in [42], [43] focus on estimating node connectivity for ER and Bollobás-Riordan models by observing auto-regressive signals on some nodes. Differently, we propose a data-driven approach that is agnostic to the graph and signal model. The works in [13], [14] solve regression tasks over expanding graphs but assume known connectivity. Instead, we consider unknown connectivity. All in all, the proposed method stands at the intersection of preferential attachment and data-driven topology estimation to learn the stochastic model parameters w.r.t. a task-specific cost function.

III. Problem Formulation

Consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of N nodes in set $\mathcal{V} = \{v_1, \ldots, v_N\}$ and E edges in set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Let $\mathbf{A}$ be the graph adjacency matrix with $A_{ij} > 0$ only if $(v_i, v_j) \in \mathcal{E}$ and $\mathbf{L} = \text{diag}(\mathbf{A1}) - \mathbf{A}$ be the graph Laplacian. When an incoming node v_+ connects to $\mathcal{G}$, it forms the expanded graph $\mathcal{G}_+ = (\mathcal{V}_+, \mathcal{E}_+)$ with node set $\mathcal{V}_+ = \mathcal{V} \cup v_+$ and edge set $\mathcal{E}_+ = \mathcal{E} \cup (v_+, v_i)$ for all new connections (v_+, v_i). The attachment of node v_+ is characterized by vector $\mathbf{a}_+ \in \mathbb{R}^N$ such that $[\mathbf{a}_+]_i = w_i$ if v_+ attaches to v_i with edge-weight w_i, and zero otherwise. This leads to the respective expanded adjacency and Laplacian matrices

$$\mathbf{A}_+ = \begin{bmatrix} \mathbf{A} & \mathbf{a}_+ \\ \mathbf{a}_+^\top & 0 \end{bmatrix}, \ \mathbf{L}_+ = \begin{bmatrix} \mathbf{L} + \text{diag}(\mathbf{a}_+) & -\mathbf{a}_+ \\ -\mathbf{a}_+^\top & \mathbf{a}_+^\top \mathbf{1} \end{bmatrix} \quad (1)$$

in which the last row and column represent the connectivity of v_+ with the nodes in $\mathcal{V}$.[1]

We consider v_+ connects independently to each existing $v_i \in \mathcal{V}$ with probability p_i. Thus, the attachment vector $\mathbf{a}_+$ is random with each entry being a weighted Bernoulli random variable; i.e,

$$[\mathbf{a}_+]_i = \begin{cases} w_i & \text{with probability } p_i \\ 0 & \text{with probability } (1 - p_i) \end{cases} \quad (2)$$

for $i = 1, \ldots, N$. The expected value of $\mathbf{a}_+$ is $\mathbb{E}[\mathbf{a}_+] = \mathbf{p} \circ \mathbf{w}$ where $\mathbf{p} = [p_1, \ldots, p_N]^\top$, $\mathbf{w} = [w_1, \ldots, w_N]^\top$, and $\circ$ is

[1]Like in the fundamental studies about growing networks [7], [10], we consider for simplicity of exposition the attachment of a single node. However, our findings extend to multiple incoming nodes with appropriate modifications. E.g., making vector $\mathbf{a}_+$ in (1) a matrix, in which each column corresponds to one incoming node.

the Hadamard product. Likewise, the variance of $[\mathbf{a}_+]_i$ is $\text{var}([\mathbf{a}_+]_i) = w_i^2 p_i(1-p_i)$ and the covariance matrix of $\mathbf{a}_+$ is

$$\mathbf{\Sigma}_+ = \text{diag}(\mathbf{w}^{\circ 2} \circ \mathbf{p} \circ (\mathbf{1} - \mathbf{p})) \tag{3}$$

where $\mathbf{a}^{\circ 2} := \mathbf{a} \circ \mathbf{a}$. The expected topology of $\mathcal{G}_+$ has the deterministic adjacency matrix

$$\mathbb{E}[\mathbf{A}_+] = \begin{bmatrix} \mathbf{A} & \mathbf{p} \circ \mathbf{w} \\ (\mathbf{p} \circ \mathbf{w})^\top & 0 \end{bmatrix}. \tag{4}$$

Thus, we can write the attachment vector of a new realization as $\mathbf{a}_+ = \mathcal{S}(\mathbf{p}) \circ \mathbf{w}$, where $\mathcal{S}(\mathbf{p})$ is a binary vector obtained by sampling $\mathbf{p}$ element-wise.

On the vertices of the existing graph $\mathcal{G}$, we have a graph signal $\mathbf{x} = [x_1, \ldots, x_N]^\top$ in which entry x_i is the value on node v_i. Processing this signal by accounting for its coupling with $\mathcal{G}$ is key to several network data tasks. For instance, we use such coupling to predict the rating of a specific item in nearest-neighbour collaborative filtering [4]. In the incoming node setting, this translates to identifying the signal value x_+ for node v_+, e.g., the rating of a new user. Since we do not have the connectivity of v_+, we rely on stochastic models governed by the attachment probabilities $\mathbf{p}$ and weights $\mathbf{w}$, which are in turn unknown. Not knowing the connectivity results from not knowing the existing preference of v_+, so we rely on the existing users and their connections to predict ratings. This works when the ratings obey some distribution over the sample space of users and the existing graph. To identify a task-specific connectivity for the incoming nodes, we merge data-driven solutions with statistical models. Given a fixed graph $\mathcal{G}$ and a training set of incoming nodes $\mathcal{T} = \{(v_{t+}, x_{t+}, \mathbf{a}_{t+}, \mathbf{b}_{t+})\}_t$, we infer the attachment probabilities $\mathbf{p}$ and weights $\mathbf{w}$ in an empirical risk minimization fashion. Each element in $\mathcal{T}$ comprises an incoming node v_{t+}, its signal x_{t+}, the attachment vector $\mathbf{a}_{t+}$, and its binary form $\mathbf{b}_{t+}$. We define a task-specific loss $f_{\mathcal{T}}(\mathbf{p}, \mathbf{w}, \mathbf{a}_{t+}, \mathbf{x}_{t+})$ measuring the incoming nodes performance. E.g., in collaborative filtering with cold starters, we build a user-user graph $\mathcal{G}$ with some existing users and treat some other users as cold starters with known connectivity and ratings in $\mathcal{T}$. Estimating the task-aware connectivity translates into solving the statistical optimization problem

$$\begin{aligned} &\min_{\mathbf{p},\mathbf{w}} \ \mathbb{E}\left[f_{\mathcal{T}}(\mathbf{p}, \mathbf{w}, \mathbf{a}_{t+}, \mathbf{x}_{t+})\right] + g_{\mathcal{T}}(\mathbf{p}, \mathbf{b}_{t+}) + h_{\mathcal{T}}(\mathbf{w}, \mathbf{a}_{t+}) \\ &\text{subject to} \quad \mathbf{p} \in [0,1]^N, \mathbf{w} \in \mathcal{W} \end{aligned} \tag{5}$$

where $g_{\mathcal{T}}(\mathbf{p}, \mathbf{b}_{t+})$ and $h_{\mathcal{T}}(\mathbf{w}, \mathbf{a}_{t+})$ are regularizers imposing a prior between $\mathbf{p}$ and the binary attachment $\mathbf{b}_+$, and between $\mathbf{w}$ and training attachment $\mathbf{a}_+$, respectively; and $\mathcal{W}$ is a convex set constraining the edge-weights, e.g., non-negative or finite. Upon estimating the probabilities $\mathbf{p}^*$ and weights $\mathbf{w}^*$ from (5), we generate realizations for $v_+ \notin \mathcal{T}$.

Problem Statement: Given graph $\mathcal{G}$ with signal $\mathbf{x}$ and a training set $\mathcal{T}$ of incoming nodes, our goal is to estimate the attachment probabilities $\mathbf{p}$ and weights $\mathbf{w}$ of a preferential attachment model w.r.t. a task-specific cost function $f_{\mathcal{T}}(\cdot)$ by solving problem (5).

We will particularize the cost function in (5) to the signal interpolation error at the incoming node and to the graph signal smoothness [44]. Since such problems are in general jointly non-convex in $\mathbf{p}$ and $\mathbf{w}$, we develop an alternating projected-gradient descent and discuss its marginal convexity and convergence (Section IV). And since each connectivity realization perturbs the graph from its nominal form, we conduct a statistical perturbation analysis [16], [17] to show the effects of the attachment model on the nominal spectrum (Section V).

IV. Task-Aware Connectivity Learning

In this section, we consider first the task of signal reconstruction at the incoming node through percolation via graph filtering [45]. Graph filters facilitate data processing at each node locally through a combination of successive shift operations and compare well with alternatives in these problems [46]. Second, we consider the task of estimating a topology such that the signal is smooth on the expanded graph.

A. Signal Interpolation

Consider the graph signal $\mathbf{x}_+ = [\mathbf{x}^\top, 0]^\top$ for $\mathcal{G}_+$, where zero is the unknown signal at v_+. The output $\mathbf{y}_+$ of an order L graph filter is

$$\mathbf{y}_+ = \mathbf{H}(\mathbf{A}_+)\mathbf{x}_+ = \sum_{l=1}^{L} h_l \mathbf{A}_+^l \mathbf{x}_+ \tag{6}$$

where $\mathbf{h} = [h_1, \ldots, h_L]^\top$ are the filter coefficients and $\mathbf{H}(\mathbf{A}_+) = \sum_{l=1}^{L} h_l \mathbf{A}_+^l$ is the filtering matrix. The filter order L implies that nodes up to L-hops away contribute to the interpolated signal of v_+. Also, the direct term $l = 0$ is ignored in (6) because it does not contribute to the output at v_+. Given the percolated signal $[\mathbf{y}_+]_{N+1}$ at node v_+ is random, the following proposition quantifies the signal reconstruction MSE as a function of the model parameters $\mathbf{p}$ and $\mathbf{w}$.

Proposition 1: Given graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with adjacency matrix $\mathbf{A}$ and signal $\mathbf{x}$, let matrix $\mathbf{A}_x = [\mathbf{x}, \ldots, \mathbf{A}^{L-1}\mathbf{x}]$ contain the first $L-1$ shifted versions of $\mathbf{x}$. Let the incoming node v_+ attach to $\mathcal{G}$ with probability vector $\mathbf{p}$ and edge weight vector $\mathbf{w}$, forming graph $\mathcal{G}_+$ with the expanded adjacency matrix $\mathbf{A}_+$ $[cf.(1)]$. The MSE of the interpolated signal $\mathbf{y}_+$ at node v_+ by an order L graph filter $\mathbf{H}(\mathbf{A}_+)$ $[cf.(6)]$ is approximately

$$\text{MSE}(\mathbf{p}, \mathbf{w}) \approx ((\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} - x_+^\star)^2 + \mathbf{h}^\top \mathbf{A}_x^\top \mathbf{\Sigma}_+ \mathbf{A}_x \mathbf{h} \tag{7}$$

where $\mathbf{h} = [h_1, \ldots, h_L]^\top$ are the filter coefficients and $x_+^\star$ is the true signal at v_+.

Proof: See Appendix A. □

Proposition 1 provides insights on the role of $\mathbf{p}$ and $\mathbf{w}$ on the signal interpolation MSE. The first term on the RHS of (7) captures the model bias w.r.t. the true signal $x_+^\star$. Essentially, the model output is the dot product between the filter output $\mathbf{A}_x\mathbf{h}$ with the expected attachment vector $\mathbf{w} \circ \mathbf{p}$. Minimizing the bias implies selecting a pair $(\mathbf{p}, \mathbf{w})$ that combines the signal at each $v \in \mathcal{V}$ to predict $x_+^\star$ accurately. The second term $\mathbf{h}^\top \mathbf{A}_x^\top \mathbf{\Sigma}_+ \mathbf{A}_x \mathbf{h} = \|\mathbf{A}_x\mathbf{h}\|^2_{\mathbf{\Sigma}_+}$ measures the percolated signal norm w.r.t. the uncertainty of the new connections, which is also the prediction variance. Minimizing this term might give solutions such as $\mathbf{p} = \mathbf{1}$ where incoming nodes connect to all $v \in \mathcal{G}$ and $\mathbf{p} = \mathbf{0}$ which prevents any connections. So, regularizers are needed for each variable. In the MSE expression, we remark that the Lth shift $\mathbf{A}^L\mathbf{x}_+$ does not appear in (7) because of

the structure of matrix $\mathbf{A}_+$ in (1). We also remark that the MSE is (7) is only an approximation because for filter order $L \geq 3$ the MSE expression becomes intractable due to the higher order moments of $\mathbf{a}_+$. Instead, if the filter order is smaller, expression (7) holds also with equality. The MSE in (7) holds also with equality for any order L when the expanded graph has directed edges landing at v_+.

Corollary 1: If each incoming node v_+ forms directed edges leaving from the nodes in $\mathcal{G}$ and landing on v_+, the MSE in (7) holds with equality for any filter of order L.

Proof: See Appendix A. □

Applications with directed links on incoming nodes include collaborative filtering [4] and a new user in a social network interacting with the existing ones.

Regularizers: The MSE plays the role of $f_{\mathcal{T}}(\cdot)$ in (5). Functions $g_{\mathcal{T}}(\cdot)$ and $h_{\mathcal{T}}(\cdot)$ regularize the problem with priors on $\mathbf{p}$ and $\mathbf{w}$, respectively. While there are several choices for the latter, we focus on the following two.

- For the probability attachment $\mathbf{p}$, we consider

$$g_{\mathcal{T}}(\mathbf{p}, \mathbf{b}_{t+}) = \mu_p \sum_{t=1}^{|\mathcal{T}|} ||\mathbf{p} - \mathbf{b}_{t+}||_q^q \tag{8}$$

  where $q \in \{1, 2\}$ and $\mu_p > 0$ is a scalar. For $q = 1$, (8) enforces sparsity on the attachment probabilities $\mathbf{p}$, i.e., the incoming node will connect only with a few of the nodes in $\mathcal{V}$. This is intuitive as graphs are sparse. However, if only a few entries in $\mathbf{p}$ are nonzero, this may lead to no connections. Using $q = 2$ may overcome this as it allows v_+ to connect in expectation to any other node but with a small probability.
- Likewise, for the weights $\mathbf{w}$ we consider

$$h_{\mathcal{T}}(\mathbf{w}) = \mu_w \sum_{t=1}^{|\mathcal{T}|} ||\mathbf{w} - \mathbf{a}_{t+}||_q^q \tag{9}$$

  where $\mu_w > 0$. Imposing sparsity on $\mathbf{w}$ results in zero weights for many edges. This implies even if the attachment probability is one, it may incur a zero edge weight. So, we prefer a two-norm penalty.

Alternatively, another approach is to consider a joint regularizer $g_{\mathcal{T}}(\mathbf{p}, \mathbf{w}) = ||\mathbf{w} \circ \mathbf{p} - \mathbf{a}_+||_q^q$. However, this might limit our control over the connectivity and the edge weights.[2] We may also consider $\mathbf{w}$ to be a random variable drawn from a normal distribution $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$. In this case, we need priors for the mean $\boldsymbol{\mu}_w$ and covariance matrix $\boldsymbol{\Sigma}_w$. The proposed approach is modular to such choices and we leave their evaluation to interested readers.

Optimization Problem: With this in place, we can formulate problem (5) as

$$\min_{\mathbf{p},\mathbf{w}} \quad C_I(\mathbf{p}, \mathbf{w}) = \text{MSE}_{\mathcal{T}}(\mathbf{p}, \mathbf{w}) + \sum_{t=1}^{|\mathcal{T}|} \left(\mu_p ||\mathbf{p} - \mathbf{b}_{t+}||_q^q + \mu_w ||\mathbf{w} - \mathbf{a}_{t+}||_q^q\right)$$

$$\text{subject to} \quad \mathbf{p} \in [0, 1]^N, \mathbf{w} \in \mathcal{W} \tag{10}$$

where $\text{MSE}_{\mathcal{T}}(\mathbf{p}, \mathbf{w})$ [cf. (7)] is the empirical MSE over the training set $\mathcal{T}$.

Problem (10) is non-convex in $\mathbf{w}$ and $\mathbf{p}$, but it is marginally convex in $\mathbf{w}$ and not always in $\mathbf{p}$ due to the variance term in (7). We solve (10) with alternating projected gradient descent. Algorithm 1 summarizes the main steps. The gradient of $C_I(\mathbf{p}, \mathbf{w})$ w.r.t. $\mathbf{p}$ and $\mathbf{w}$ for $q = 2$ are respectively.

$$\nabla_p C_I(\mathbf{p}, \mathbf{w}) = 2 \sum_{t=1}^{|\mathcal{T}|} ((\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} - x_{t+})(\mathbf{w} \circ \mathbf{A}_x \mathbf{h}) + |\mathcal{T}|(\mathbf{A}_x \mathbf{h})^{\circ 2} \circ (\mathbf{w}^{\circ 2}) \circ (\mathbf{1} - 2\mathbf{p}) + 2\mu_p \sum_{t=1}^{|\mathcal{T}|} (\mathbf{p} - \mathbf{b}_{t+}) \tag{11}$$

$$\nabla_w C_I(\mathbf{p}, \mathbf{w}) = 2 \sum_{t=1}^{|\mathcal{T}|} ((\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} - x_{t+})(\mathbf{p} \circ \mathbf{A}_x \mathbf{h}) + 2|\mathcal{T}|(\mathbf{A}_x \mathbf{h})^{\circ 2} \circ \mathbf{w} \circ \mathbf{p} \circ (\mathbf{1} - \mathbf{p}) + 2\mu_w \sum_{t=1}^{|\mathcal{T}|} (\mathbf{w} - \mathbf{a}_{t+}). \tag{12}$$

Instead, for $q = 1$, we replace terms $2\mu_p(\mathbf{p} - \mathbf{b}_{t+})$ and $2\mu_w(\mathbf{w} - \mathbf{a}_{t+})$ with $\text{sign}(\mathbf{p} - \mathbf{b}_{t+})$ and $\text{sign}(\mathbf{w} - \mathbf{a}_{t+})$ respectively, where $\text{sign}(x) = 1$ for $x \geq 0$ and -1 otherwise. To select an appropriate μ_p and μ_w, one can perform cross validation over a range of candidate values. The complexity of the algorithm is of the order $\mathcal{O}(TKE + NT)$, where K is the filter order, E the number of edges in the existing graph, N the number of existing nodes, and T the number of update steps in each of the variables. The complexity $\mathcal{O}(TKE)$ is due to the term $\mathbf{A}_x \mathbf{h}$, which incurs a complexity of the output of an order K FIR graph filter, equal to $\mathcal{O}(KE)$. The term $\mathcal{O}(NT)$ is due the projection operation over N elements.

While we can use Algorithm 1 to solve the general non-convex case of problem (10), the following corollary provides a sufficient condition for problem (10) to be marginally convex also in $\mathbf{p}$.

Algorithm 1: Alternating Projected Gradient Descent for Problems (10) and (16).

Input: Graph $\mathcal{G}$, training set $\mathcal{T}$, graph signal $\mathbf{x}$, adjacency matrix $\mathbf{A}$, number of iterations K, cost $C \in \{C_I, C_S\}$, learning rates λ_p, λ_w.
Initialization: $\mathbf{p} = \mathbf{p}^0$, $\mathbf{w} = \mathbf{w}^0$ randomly, $k = 0$.
for $k \leq K$ **do**
  $\mathbf{p}$ gradient: $\tilde{\mathbf{p}}^{k+1} = \mathbf{p}^k - \lambda_p \nabla_{\mathbf{p}} C(\mathbf{p}^k, \mathbf{w}^k)$;
  Projection: $\mathbf{p}^{k+1} = \prod_{[0,1]^N} (\tilde{\mathbf{p}}^{k+1})$;
  $\mathbf{w}$ gradient: $\mathbf{w}^{k+1} = \mathbf{w}^k - \lambda_w \nabla_w C(\mathbf{p}^{k+1}, \mathbf{w}^k)$;
  Projection: $\mathbf{w}^{k+1} = \prod_{\mathcal{W}} (\tilde{\mathbf{w}}^{k+1})$;
end for

[2]Imposing joint sparsity, we have $w_i p_i = 0$ for some $v_i \in \mathcal{V}$. If $p_i \approx 1$, i.e., the incoming node connects to v_i with a high probability, w_i would have to be zero, which will make the connection meaningless.

Corollary 2: Problem (10) is marginally convex in $\mathbf{p}$ if the regularization weight μ_p satisfies

$$\mu_p \geq w_h^2 \max_i ([\mathbf{A}_x \mathbf{h}]_i)^2 - ||\mathbf{w} \circ \mathbf{A}_x \mathbf{h}||_2^2. \tag{13}$$

Proof: See Appendix B. □

While ensuring convexity, hence a guarantee for a minima, condition (13) may lead to an optimum that is worse than the local optima of its non-convex counterpart. This is because of the greater focus on the training attachment patterns than on the cost. When $\mathbf{A}_x$, $\mathbf{h}$ and w_h are known, we can evaluate the R.H.S. of (13) and set μ_p so that we avoid this condition. We shall corroborate this in Section VI-A.

B. Signal Smoothness

We now learn the connectivity of the incoming nodes such that expanded graph signal is smooth. The smoothness of a graph signal $\mathbf{x}$ w.r.t. $\mathcal{G}$ is $S_{\mathcal{G}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{L} \mathbf{x}$, where a lower value implies connected nodes have similar signals and vice-versa. Upon attachment of v_+ with signal $x_+^\star$, the smoothness of the new graph signal $\mathbf{x}_+ = [\mathbf{x}^\top, x_+^\star]^\top$ w.r.t. the expanded graph Laplacian $\mathbf{L}_+$ is

$$\begin{aligned} S_{\mathcal{G}_+}(\mathbf{a}_+) &= \mathbf{x}^\top (\mathbf{L} + \operatorname{diag}(\mathbf{a}_+))\mathbf{x} - 2x_+^\star \mathbf{x}^\top \mathbf{a}_+ + x_+^{\star^2} \mathbf{a}_+^\top \mathbf{1} \\ &= \mathbf{a}_+^\top \hat{\mathbf{x}} + S_{\mathcal{G}}(\mathbf{x}) \end{aligned} \tag{14}$$

where $\hat{\mathbf{x}} = \mathbf{x}^{\circ^2} - 2x_+^\star \mathbf{x} + x_+^{\star^2} \mathbf{1}$. I.e., the smoothness of the expanded graph signal is linked to the connectivity of the incoming node. We use this relationship to learn a connectivity model that ensures the expanded graph signal is smooth.

Let $\mathbf{a}_+^\star$ be the true connectivity of v_+ and $\mathbf{a}_+$ be the connectivity pattern obeying the model. We are interested in the expected squared smoothness error between the model smoothness and the true smoothness, i.e., $\mathbb{E}[S_{\mathcal{G}_+}(\mathbf{a}_+) - S_{\mathcal{G}_+}(\mathbf{a}_+^\star)]^2$. The following proposition quantifies the latter.

Proposition 2: Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with Laplacian $\mathbf{L}$ and signal $\mathbf{x}$. Let an incoming node v_+ with signal $x_+^\star$ attach to $\mathcal{G}$ forming graph $\mathcal{G}_+ = (\mathcal{V}_+, \mathcal{E}_+)$ with attachment probability $\mathbf{p}$, weight $\mathbf{w}$, and covariance matrix $\mathbf{\Sigma}_+$ [cf. (3)]. Let $\mathbf{a}_+^\star$ be the true attachment. The expected squared smoothness error (ESSE) for signal $\mathbf{x}_+ = [\mathbf{x}, x_+^\star]^\top$ is

$$\begin{aligned} \operatorname{ESSE}(\mathbf{p}, \mathbf{w}) = \hat{\mathbf{x}}^\top \mathbf{\Sigma}_+ \hat{\mathbf{x}} + \hat{\mathbf{x}}^\top (\mathbf{w} \circ \mathbf{p}) \big((\mathbf{w} \circ \mathbf{p})^\top \hat{\mathbf{x}} \\ - 2\mathbf{a}_+^{\star\top} \hat{\mathbf{x}} \big) + \hat{\mathbf{x}}^\top \mathbf{a}_+^\star \mathbf{a}_+^{\star\top} \hat{\mathbf{x}} \end{aligned} \tag{15}$$

where $\hat{\mathbf{x}} = (\mathbf{x} - x_+^\star \mathbf{1})^{\circ 2}$.

Proof: See Appendix C. □

Result (15) shows the relationship between the attachment model, the existing graph signal, and the signal at the incoming node w.r.t. the overall smoothness. The first term is the quadratic norm of $\hat{\mathbf{x}}$ w.r.t. the covariance matrix $\mathbf{\Sigma}_+$. It contributes to a lower ESSE when the variance of the attachment is low at nodes with a high signal difference. The second term is the alignment between the expected attachment pattern $\mathbf{w} \circ \mathbf{p}$ and the squared difference signal $\hat{\mathbf{x}}$; the ESSE reduces when this is smaller than twice the alignment between the true attachment and the difference signal. Thus, the ESSE reduces when $\hat{\mathbf{x}}^\top (\mathbf{w} \circ \mathbf{p})$ is small.

Optimization Problem: The ESSE plays the role of $f_{\mathcal{T}}(\cdot)$ in (5) as the MSE did in problem (10). Differently though from (10), the ESSE captures the interaction between $\mathbf{p}$ and the true connectivity $\mathbf{a}_+^\star$ in the second term in (15). Thus, we drop the regularizer $g_{\mathcal{T}}(\cdot)$ on $\mathbf{p}$. Particularizing then problem (5) w.r.t. the smoothness cost, we get

$$\begin{aligned} \min_{\mathbf{p},\mathbf{w}} \quad & C_S(\mathbf{p}, \mathbf{w}) = \operatorname{ESSE}_{\mathcal{T}}(\mathbf{p}, \mathbf{w}) + \mu_w \sum_{t=1}^{|\mathcal{T}|} ||\mathbf{w} - \mathbf{a}_{t+}||_q^q \\ \text{subject to} \quad & \mathbf{p} \in [0,1]^N, \mathbf{w} \in \mathcal{W} \end{aligned} \tag{16}$$

where $\operatorname{ESSE}_{\mathcal{T}}(\mathbf{p}, \mathbf{w})$ is the empirical expression of (15) averaged over $\mathcal{T}$. The cost $C_S(\mathbf{p}, \mathbf{w})$ in (16) is non-convex and also marginally non-convex in $\mathbf{p}$ because of the first ESSE term. We again apply the alternating projected gradient in Algorithm 1. The gradients for $q = 2$ are

$$\begin{aligned} \nabla_p C_S(\mathbf{p}, \mathbf{w}) = \sum_{t=1}^{|\mathcal{T}|} \Big(& \mathbf{w}^{\circ 2} \circ (\mathbf{1} - 2\mathbf{p}) \circ \hat{\mathbf{x}}_{t+}^{\circ 2} + 2 \left((\mathbf{w} \circ \mathbf{p} \right. \\ & \left. - \mathbf{a}_{t+})^\top \hat{\mathbf{x}}_{t+} \right) \mathbf{w} \circ \hat{\mathbf{x}}_{t+} \Big) \end{aligned} \tag{17}$$

$$\begin{aligned} \nabla_w C_S(\mathbf{p}, \mathbf{w}) = \sum_{t=1}^{|\mathcal{T}|} \Big(& 2\mathbf{w} \circ \mathbf{p} \circ (\mathbf{1} - \mathbf{p}) \circ \hat{\mathbf{x}}_{t+}^{\circ 2} + 2 \left((\mathbf{w} \circ \mathbf{p} \right. \\ & \left. - \mathbf{a}_{t+})^\top \hat{\mathbf{x}}_{t+} \right) \mathbf{p} \circ \hat{\mathbf{x}}_{t+} + 2\mu_w (\mathbf{w} - \mathbf{a}_{t+}) \Big). \end{aligned} \tag{18}$$

We considered Algorithm 1 also for (10) to provide a unified approach for both problems despite problem (10) being also marginally non-convex. But we could also simply choose a joint stochastic gradient method. The choice of alternating descent approach is rather standard, as seen in [47], [48] but the alternating one allows us to characterize the convergence for both costs [cf. Appendix D].

C. Convergence

To comment on the convergence of the alternating projected gradient descent approach, we assume the following.

Assumption 1: The Hessians of the costs (10) and (16) w.r.t. the variables $\mathbf{w}$ and $\mathbf{p}$ are upper bounded by

$$\nabla_{\mathbf{p}}^2 C(\mathbf{p}, \mathbf{w}) \preceq L_p \mathbf{I}, \quad \nabla_{\mathbf{w}}^2 C(\mathbf{p}, \mathbf{w}) \preceq L_w \mathbf{I} \tag{19}$$

This implies that the maximum eigenvalue of the Hessian is upper-bounded for both costs w.r.t. both variables. This can be easily verified for (10) and (16).

Theorem 1: Given costs (10) and (16) satisfy Assumption 1 and given Algorithm 1 runs with step-sizes λ_p and λ_w. Then, it holds that:

1) *Claim:* If the step sizes satisfy $0 < \lambda_p \leq \frac{3}{4L_p}, 0 < \lambda_w \leq \frac{3}{4L_w}$, the cost is non-increasing over the iterations, i.e., $C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) \leq C(\mathbf{p}^k, \mathbf{w}^k)$.

2) *Claim:* If there exist a feasible local minimum $(\mathbf{p}^*, \mathbf{w}^*)$ and the step sizes satisfy $0 < \lambda_p \leq \frac{1}{2L_p}, 0 < \lambda_q \leq \frac{1}{2L_w}$, Algorithm 1 reaches this local minimum $(\mathbf{p}^\star, \mathbf{w}^\star)$ with convergence rate $\mathcal{O}(1/K)$, where K is the number of iterations.

Proof: See Appendix D. □

Since the choice of the local minima is arbitrary, this shows that Algorithm 1 can converge to any of the local minima. One way to deal with this is to run Algorithm 1 for multiple initializations and select the pair $(\mathbf{p}, \mathbf{w})$ that gives the lowest training cost. We show in Section VI-A that this may not always be needed as we have seen consistently that different initializations lead to similar costs.

V. Perturbation Analysis

During testing, we draw samples from the learned $\mathbf{p}$ to obtain expanded graph realizations. Such realizations are edge sampled versions of the nominal graph that contains all possible edges to sample. This leads to differences both in the vertex and spectral domain [49], which have an impact on the task of interest. To characterize such an impact, we look at the spectral difference between the realized and nominal graphs and link it with our cost functions.

Given $\mathbf{p}$ and $\mathbf{w}$, we have a nominal graph $\overline{\mathcal{G}}_+$ with adjacency matrix $\overline{\mathbf{A}}_+ = \left[\begin{smallmatrix} \mathbf{A} & \mathbf{w} \\ \mathbf{w}^\top & 0 \end{smallmatrix}\right]$ and Laplacian $\overline{\mathbf{L}}_+$. I.e., realization $\mathbf{A}_+$ is a probabilistic edge-sampled version of $\overline{\mathbf{A}}_+$ where edge (v_+, v_i) in $\overline{\mathbf{A}}_+$ is removed with probability $(1 - p_i)$. To study the difference of each realization $\mathbf{A}_+$ from its nominal version $\overline{\mathbf{A}}_+$, we analyze the nominal matrices in the spectral domain via a perturbation analysis [50]. Consider their eigendecompositions

$$\overline{\mathbf{A}}_+ = \mathbf{U}\mathbf{\Lambda}_A\mathbf{U}^\top, \;\; \overline{\mathbf{L}}_+ = \mathbf{V}\mathbf{\Pi}_L\mathbf{V}^\top \tag{20}$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$, $\mathbf{\Lambda}_A = \text{diag}(\gamma_1, \ldots, \gamma_N)$ and $(\gamma_i, \mathbf{u}_i)$ is the ith eigenpair. Let also $[\mathbf{u}_i]_{1:N}$ be the first N elements of $\mathbf{u}_i$ and $[u_i]_j$ be the j-th entry of vector u_i. Similarly, for $\overline{\mathbf{L}}_+$, define $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$, $\mathbf{\Pi}_L = \text{diag}(\pi_1, \ldots, \pi_N)$, $[\mathbf{v}_i]_{1:N}$, and $[v_i]_j$.

Assumption 2: There exists finite positive constants c_1 and c_2 such that

$$||2[\mathbf{u}_i]_{N+1}[\mathbf{u}_i]_{1:N})||^2 \leq c_1, ||([\mathbf{v}_i]_{1:N} - [\mathbf{v}_i]_{N+1}\mathbf{1}_N)^{\circ 2}||^2 \leq c_2. \tag{21}$$

These constants depend on eigenvectors $\mathbf{U}$ and $\mathbf{V}$ which are deterministic as $\overline{\mathbf{A}}_+$ and $\overline{\mathbf{L}}_+$ are in turn fixed. So, we can always evaluate c_1 and c_2. Generating realizations $\mathbf{A}_+$ and $\mathbf{L}_+$ leads to the perturbations

$$\triangle\mathbf{A}_+ = \mathbf{A}_+ - \overline{\mathbf{A}}_+, \quad \triangle\mathbf{L}_+ = \mathbf{L}_+ - \overline{\mathbf{L}}_+. \tag{22}$$

We then assume the following to study the spectral effect of the perturbation.

Assumption 3: The perturbation is small in nature, i.e.,

$$||\triangle\mathbf{A}_+||_F \ll ||\overline{\mathbf{A}}_+]||_F, \;\; ||\triangle\mathbf{L}_+||_F \ll ||\overline{\mathbf{L}}_+]||_F \tag{23}$$

where $||\cdot||_F$ *denotes the Frobenius norm.*

This is a standard assumption for robustness in the graph spectral domain [49], [50]. For this assumption to hold, $\mathbf{p}$ should have high values for most nodes in $\mathcal{V}$ or be sparse, which can be set during the training phase. Then, the spectral deviations in the ith eigenvalue γ_i of the nominal adjacency and π_i of the nominal Laplacian due to the sampling perturbation are given by

$$\triangle\gamma_i = \mathbf{u}_i^\top\triangle\mathbf{A}_+\mathbf{u}_i, \;\; \triangle\pi_i = \mathbf{v}_i^\top\triangle\mathbf{L}_+\mathbf{v}_i. \tag{24}$$

I.e., they are dictated to how aligned are the respective eigenvectors to the perturbed graph. With this in place, we claim the following.

Proposition 3: Given $\overline{\mathbf{A}}_+, \overline{\mathbf{L}}_+$ and their eigendecompositions in (20). Let Assumption 2 hold with constants c_1 and c_2. Let the covariance matrix of attachment be $\mathbf{\Sigma}_+$ and vector $\mathbf{p}_+$ be such that $[\mathbf{p}_+]_i = 1/\sqrt{p_i}$, if $p_i \neq 0$, and zero otherwise. When v_+ joins, the expected squared deviation in the ith eigenvalues of $\overline{\mathbf{A}}_+$ and $\overline{\mathbf{L}}_+$ are respectively upper bounded as

$$\mathbb{E}[\triangle^2\gamma_i] \leq c_1\mathbf{p}_+^\top\mathbf{\Sigma}_+\mathbf{p}_+, \tag{25}$$

$$\mathbb{E}[\triangle^2\pi_i] \leq c_2\mathbf{p}_+^\top\mathbf{\Sigma}_+\mathbf{p}_+. \tag{26}$$

Proof: See Appendix E. □

Proposition 3 shows that both bounds have the common term $\mathbf{p}_+^\top\mathbf{\Sigma}_+\mathbf{p}_+$, which is similar to the variance term in the MSE, i.e., $\frac{1}{2}\mathbf{h}^\top\mathbf{A}_x^\top\mathbf{\Sigma}_+\mathbf{A}_x\mathbf{h}$ [cf. (7)] and the ESSE term of $\hat{\mathbf{x}}^\top\mathbf{\Sigma}_+\hat{\mathbf{x}}$ [cf. (15)]. If $\mathbf{A}_x\mathbf{h} = \sqrt{2c_1}\mathbf{p}_+$ and $\hat{\mathbf{x}} = \sqrt{c_2}\mathbf{p}_+$, we have equality in (25) and (26), respectively. Since the MSE and the ESSE already contain similar terms, minimizing these over $\mathbf{w}$ and $\mathbf{p}$ helps minimizing the expected squared eigenvalue perturbation. So, the optimization problems (10) and (16) account implicitly for minimizing this measure of the spectral perturbation. In the next section, we contrast the perturbation achieved with other attachment methods.

VI. Numerical Results

In this section, we evaluate our approach and compare it with related methods with synthetic and real data. For comparison, we consider three attachment rules:

i) *uniform attachment:* The node attaches uniformly, i.e., $\mathbf{p}_{\text{rd}} = \frac{1}{N}\mathbf{1}$.

ii) *preferential attachment:* The node attaches to v_i with probability $p_i \propto d_i$ with d_i the degree of v_i, and $\mathbf{p}_{\text{pf}} = \frac{\mathbf{d}}{\mathbf{1}^\top\mathbf{d}}$ where $\mathbf{d} = [d_1, \ldots, d_N]^\top$ is the degree vector.

iii) *training attachment only:* It relies only on the attachment patterns available during training to build $\mathbf{p}$ and $\mathbf{w}$, i.e., we ignore the MSE and ESSE costs in their respective cost formulations. They are given by $\mathbf{p}_g = \frac{1}{|\mathcal{T}|}\sum_{t=1}^{|\mathcal{T}|}\mathbf{b}_{t+}$ and $\mathbf{w}_g = \frac{1}{|\mathcal{T}|}\sum_{t=1}^{|\mathcal{T}|}\mathbf{a}_{t+}$.

The first two rules serve as baselines to assess how the proposed data-driven stochastic model compares with conventional statistical models, while the latter assesses the importance of the task-specific cost.

A. Synthetic Data

We build two synthetic graphs of $N = 100$ nodes following the Erdős Rényi (ER) and the Barabasi-Albert (BA) model. We consider the tasks of signal interpolation at an incoming node and the prediction of the ESSE for an incoming node.

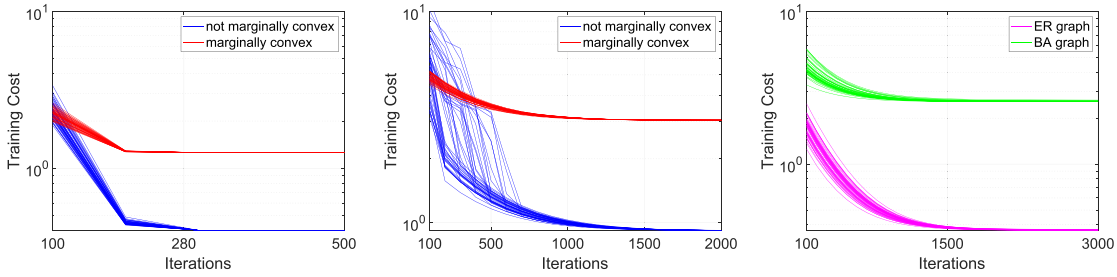

Fig. 1. Training error vs. iterations semilog plots over 50 initializations each for $(left)$: ER graph with MSE; $(centre)$: BA graph with MSE; $(right)$: ER and BA graphs with ESSE. For the MSE, the blue line represents non-convexity in $\mathbf{p}$ while the red represents otherwise. For all initializations, the proposed algorithm converges to the same training cost.

TABLE I
MSE PERFORMANCE COMPARISON BETWEEN PROPOSED, PREFERENTIAL AND RANDOM ATTACHMENT OVER THE ERDŐS RÉNYI AND BARABASI-ALBERT GRAPH FOR TRAINING OVER (LEFT TWO COLUMNS) BOTH AND (RIGHT TWO COLUMNS) ONE VARIABLE(S)

MSE	Erdős-Rényi graph			Barabasi-Albert graph			Erdős-Rényi graph			Barabasi-Albert graph		
Rule	Prop.	Pref.	Rand.	Prop.	Pref.	Rand.	both $\mathbf{p}$,$\mathbf{w}$	only $\mathbf{p}$	only $\mathbf{w}$	both $\mathbf{p}$,$\mathbf{w}$	only $\mathbf{p}$	only $\mathbf{w}$
Mean Error	**0.37**	0.64	0.73	**0.84**	1.72	1.35	**0.34**	0.94	0.37	**0.84**	6.50	0.84
Std Dev.	**0.04**	0.04	0.04	**0.11**	0.11	0.11	**0.04**	0.07	0.04	**0.10**	5.55	0.10

Experimental Setup: For each graph, we generated the graph signal $\mathbf{x}$ by combining the first 30 eigenvectors of its Laplacian matrix with weights from a normal distribution. Then, we normalized the signal to be zero mean. The edge formation probabilities for these graphs were set as $\mathbf{p}_{\text{rd}}$ and $\mathbf{p}_{\text{pf}}$, respectively. We select $\mathbf{w}$ to be the vector of all ones for both graphs. We use a filter of order one with $h_1 = 1$ to percolate the signal.

The training set comprises 1000 data points divided into an 80-20 train-test split. The regularization weights μ_p, μ_w for problems (10) and (16) were selected via ten-fold cross validation from the set $[10^{-5}, 10^{0}]$. We performed $K = 3000$ iterations of alternating projected descent [cf. Algorithm 1]. The learning rates λ_p, λ_w are 10^{-5}. We averaged the performance over 100 realizations and 50 train-test splits to get the error for each test node.

Algorithm Convergence: We solved (10) for each graph under two scenarios, one where μ_p satisfies the convexity criterion (13), and one where it does not. Fig. 1 shows the training costs as a function of the number of iterations for 50 random initializations. The non-convex cost (blue) and the marginally convex cost in both variables (red) for $\mu_p = 30$ (satisfying (13)) converge following Algorithm 1. Most importantly, optimizing over the non-convex cost yields a lower training cost because a higher weight μ_p on the regularizer $\sum_{t=1}^{|\mathcal{T}|} ||\mathbf{p} - \mathbf{b}_{t+}||^2$ results in $\mathbf{p}$ fitting the binary training attachments $\mathbf{b}_{t+}$ than reducing the MSE. Fig. 1 (right) shows the ESSE training error for both graphs. We see multiple local minima and that they all lead to the same training cost.

MSE: Here, we evaluate the signal interpolation performance. We choose $\mu_p = 1, \mu_w = 1$ for the ER graph and $\mu_p = 1, \mu_w = 0.1$ for the BA graph. Table I compares the different methods on the left two columns. The proposed approach outperforms the others in both settings in expectation and has a comparable standard deviation.

To investigate the effect of jointly training $\mathbf{p}$ and $\mathbf{w}$, we train for each of them separately while keeping the other fixed to the true value used for data generation. In Table I (right two columns), we find that training the weights, given $\mathbf{p}_{tr}$ provides a performance comparable to the proposed for the ER graph and similar to that of the BA graph. On the other hand, training $\mathbf{p}$ given $\mathbf{w}_{tr}$ degrades the performance appreciably, performing worse than $\mathbf{p}_{\text{rd}}$ and $\mathbf{p}_{\text{pf}}$. This is because when we train on $\mathbf{w}$, we deal with a convex function and reach a global optimum, whereas training on $\mathbf{p}$ leads to local minima, thus affecting the performance. However, these results show that the proposed approach is able to learn both p and w to solve the task.

ESSE: Now, we look at the ESSE performance in the same setting with $\mu_w = 0.1$ in Problem (16). Table II shows in the left two columns that the data-driven attachment outperforms the random and preferential attachment, with a lower standard deviation. Table II also shows in the right two columns the ESSE for training with one parameter fixed, as done for the MSE. We see a similar trend as before when we train only $\mathbf{w}$ for the ER graph.

Choice of regularizer: Table III highlights the role of q in estimating the attachment model, through the MSE and ESSE for the two synthetic graphs. We observe that for $q = 2$, the MSE and ESSE are lower than for $q = 1$, which promotes a sparse $\mathbf{p}$ and $\mathbf{w}$. For a sparse $\mathbf{p}$, the model restricts attachment to some nodes, and for a sparse $\mathbf{w}$, even an attachment results in zero weights, thus incurring a higher error in the inference.

Perturbation: We now analyse the mean squared deviation for each eigenvalue over multiple realizations. To give a graph-wide representative metric we report the average taken over all eigenvalues and compare with the *uniform* and *preferential attachment* with $\mathbf{w} = \mathbf{1}$. We focus on the effect of edge perturbation only.

Table IV showcases this deviation. The proposed approach achieves the lowest perturbation for both graphs while training

TABLE II
MSE PERFORMANCE COMPARISON BETWEEN PROPOSED, PREFERENTIAL AND RANDOM ATTACHMENT OVER THE ERDŐS RÉNYI AND BARABASI-ALBERT GRAPH FOR TRAINING OVER (LEFT TWO COLUMNS) BOTH AND (RIGHT TWO COLUMNS) ONE VARIABLE(S)

ESSE	Erdős-Rényi graph			Barabasi-Albert graph			Erdős-Rényi graph			Barabasi-Albert graph		
Rule	Prop.	Pref.	Rand.	Prop.	Pref.	Rand.	both $\mathbf{p}$,$\mathbf{w}$	only $\mathbf{p}$	only $\mathbf{w}$	both $\mathbf{p}$,$\mathbf{w}$	only $\mathbf{p}$	only $\mathbf{w}$
Mean Error	**0.43**	1.83	2.20	**3.20**	14.00	10.63	**0.43**	2.20	0.56	**3.20**	7.48.	4.86
Std Dev.	**0.17**	0.23	0.25	**0.81**	1.83	1.68	**0.17**	0.12	0.17	**0.81**	2.17	2.82

TABLE III
MSE AND ESSE COMPARISON FOR THE PROPOSED METHOD OVER THE ERDŐS RÉNYI AND BARABASI-ALBERT GRAPH FOR $q = 2$ AND $q = 1$

	Erdős-Rényi graph		Barabasi-Albert graph	
MSE	$q = 2$	$q = 1$	$q = 2$	$q = 1$
MSE	0.37	0.67	0.84	1.01
Std. dev.	0.04	0.06	0.11	0.13
ESSE	0.43	0.82	3.2	3.5
Std. dev.	0.17	0.31	0.81	1.5

TABLE IV
SQUARED EIGENVALUE PERTURBATION

MSE	Proposed	Random	Preferential
Erdős Rényi	3.2×10^{-4}	4.2×10^{-4}	6.3×10^{-4}
Barabasi-Albert	1.05×10^{-4}	4.1×10^{-4}	6.4×10^{-4}
ESSE	Proposed	Random	Preferential
Erdős Rényi	5.5×10^{-4}	4.8×10^{-4}	4.4×10^{-4}
Barabasi-Albert	6.59×10^{-4}	3.3×10^{-4}	5.2×10^{-4}

for MSE. However, for the ESSE, the edge perturbation is higher for the ER graph due to more links being formed. In turn, more links changes affect more eigenvalues, thereby causing a higher perturbation.

B. Collaborative Filtering

We consider the task of cold start rating prediction on the Movielens 100 K data-set. This amounts to rating prediction for unknown users, i.e., we start with some existing users as nodes of a user-user graph and interpolate the rating of a new user when joining the network. We use the graph collaborative filter in [4] to percolate the ratings of the existing users and the learnt attachment to predict the rating at the cold starter.

Experimental Setup: We retained all users and items with more than ten interactions, giving 943 users and 1152 items. We considered 50 existing users and build the adjacency matrix based on the Pearson correlation between their ratings. Next, for each item i we built the corresponding adjacency matrix by 1) retaining all outgoing links from users who rated that item; 2) building its 35 nearest neighbour graph following [4]. The remaining users were treated as cold starters and were divided into train (793) and test (100). We used an order five FIR graph filter obtained by optimally solving the rating prediction problem over the existing users and items [4]. For the interpolation cost (10), we impose an l_1-norm constraint on $\mathbf{p}$ an l_2-norm constraint on $\mathbf{w}$. We applied Algorithm 1 for $K = 1000$ iterations with learning rates $\lambda_p, \lambda_w = 10^{-4}$. We predicted the ratings for the test users and for each item we averaged the performance over 100 connectivity realizations drawn from $\mathbf{p}$. As a baseline, we considered the *mean prediction* which uses the mean of the item ratings in training to predict how cold start users will rate the item.

TABLE V
ITEM DETAILS. ITEMS 1 AND 48 HAVE HIGH, 459 AND 550 HAVE MEDIUM, 57 AND 877 HAVE LOW TRAINING SAMPLES

Item	**1**	**48**	**459**	**550**	**57**	**877**
Training Samples	362	457	196	139	64	42
Test Samples	44	65	29	22	8	8

Item-specific Learning: First, we learnt $(\mathbf{p}, \mathbf{w})$ for each item separately, which is preferred for personalized recommendations. We focused on three categories of items with a high, medium and a low number of training samples. For each category, report in Table V two examples. We evaluated the performance via the Root Median Square Error (RMedSE) between the predicted and the true ratings, which is more robust than the mean to outliers that are inevitably present in all stochastic approaches.

In Table VI we show the relative performance difference – worse (red) or better (green) between the proposed approach and the alternatives. The proposed method outperforms the alternatives convincingly for items with high training samples (1 and 48) and does well even with low training samples. When compared with the training attachment only, it is clear that including the graph structure and ratings along with the attachment patterns is more beneficial. The training only and the preferential attachment strategies prioritise users who have rated the item as only those users have links directed outwards. The performance of our approach suggests that such attachments are not always optimal for the cold start. The poor performance of the uniform and the preferential attachment (except item 459) shows the importance of using a task-aware connectivity approach. The mean prediction performance is dependent on the quality and quantity of the available ratings. For example, it is considerably worse off for all items except for item 877, even though it has few training samples.

Learning for all Items: Second, we learnt a common $(\mathbf{w}, \mathbf{p})$ across all the 1152 items in the data set. We select $\mu_p = 1$ and $\mu_w = 10^{-3}$. The results are in the last column of Table VI. We notice that with all the item data, even though the proposed performs the best, the performance gap reduces, which is somewhat expected as we are not personalizing recommendations. This suggests that to improve the cold start performance, we should approach each item individually following the spirit of nearest neighbour collaborative filtering. The attachment only method performs well because the attachment rule is cognizant of a diverse set of node attachments over many graphs and ratings. The proposed still does substantially well compared to alternatives.

TABLE VI
RMEDSE OF ALL APPROACHES. IN BRACKETS WE SHOW RELATIVE PERFORMANCE DIFFERENCE IN % TO THE PROPOSED

Item	1	48	459	550	57	877	All
Proposed	**0.494**	**0.492**	0.462	**0.512**	**0.049**	**0.99**	**0.799**
Attachment Only	0.537(+8.7)	0.611(+24.2)	0.49(+6)	0.678(+32.4)	0.057(+16.3)	1.04(+5)	0.802(+0.38)
Random	0.5417(+9.7)	0.53(+7.7)	0.52(+12.5)	0.66(+29)	0.41(+736)	1.07(+8)	0.821(+2.75)
Preferential	0.527(+6.7)	0.62(+26)	**0.40(+12.5)**	0.692(+35.2)	0.20(+308)	1.05(+6)	0.820(+2.63)
Mean Prediction	0.669(+35.4)	0.55(+11.8)	1.07(+131)	0.643(+25.6)	0.32(+553)	1.01(+1.72)	0.832(+4.1)

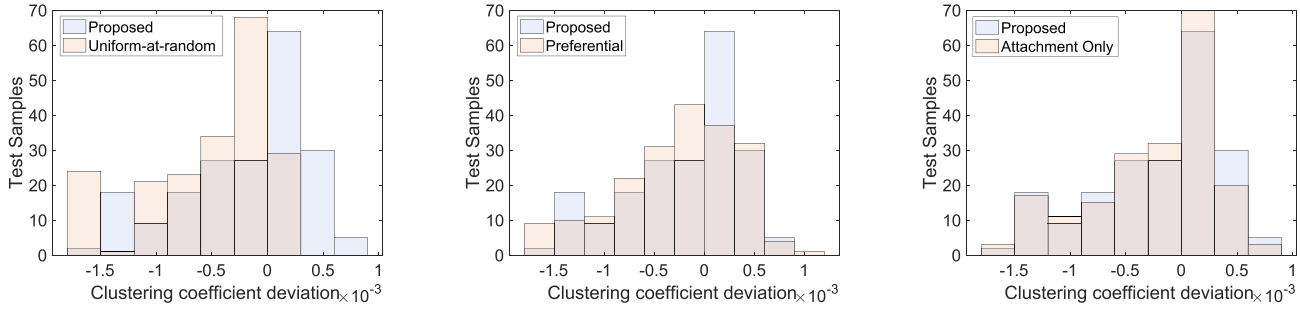

Fig. 2. Histogram of the clustering coefficient deviation of the proposed approach compared with the (a) :*uniform*; (b) :*preferential attachment*; (c) :*training attachment only*. The proposed approach causes positive deviation for the most test nodes, while causing the fewest negative deviations as well.

C. Blog Network

We consider a political blog network with blogs as nodes and their political orientation (liberal vs. conservative) as signals [9]. We study how nodes attaching based on the ESSE influence the structure of the existing graph. This is because the ESSE is low when the signal varies slowly within a cluster than arbitrarily between clusters [16], [17].

Experimental Setup: We extracted a connected sub-graph from the main network with $N = 600$ blogs such that this graph retains the clustering. The remaining 622 nodes are split into train (400) and test (200). The existing adjacency matrix is binary, so we set $\mathbf{w} = \mathbf{1}$, i.e., we train to minimise the ESSE only w.r.t. $\mathbf{p}$. We apply Algorithm 1 for 500 iterations with learning rates $\lambda_p = 10^{-5}$ and $\lambda_w = 10^{-6}$. We consider the clustering coefficient of a graph as a measure of how well it is clustered [51]. A large value implies a more clustered graph. Upon learning $\mathbf{p}$, we calculate the clustering coefficient of the expanded graph formed upon its attachment. We contrast this with the clustering coefficient with that of the true attachment.

Fig. 2 compares the histograms of the clustering coefficient difference between the realization and the true attachment between the proposed and other approaches. A positive deviation improves upon the clustered nature of the graph, while a negative deviation reduces it. Ideally, we want more positive and fewer negative deviations. In Fig. 2(a) the proposed approach outperforms the random attachment, which is likelier to make an incoming node connect to both clusters and incur a negative deviation. In Fig. 2(b), the preferential attachment incurs negative deviation for more test nodes but also reports the highest positive deviation for a few nodes. In Fig. 2(c) the proposed approach influences positive deviation for more test nodes in the two furthest bins and fewer negative deviations than training attachment only. By minimizing the ESSE, new nodes attach in a way that is likelier to retain/ improve upon the overall clustering for unknown nodes. Hence, the data-driven attachment follows the true network properties if the cost function matches with the task; here, preserve its clustering structure.

VII. CONCLUSION

We proposed an approach to learn the random connectivity model for incoming nodes by solving a signal processing task over the expanding graph. Incorporating the data-processing task to determine the attachment is beneficial compared to relying on the knowledge of previous node attachments and standard stochastic attachments. We formulated a stochastic optimization problem w.r.t. the attachment parameters for graph signal interpolation and signal smoothness. The problem is solved via an alternating projected descent approach with provable convergence to local minima. By conducting a perturbation analysis, we show that our method implicitly controls the spectral perturbation caused by such nodes.

For undirected graphs, the higher-order statistics limit the MSE analysis to be an approximation of the true one. This might be addressed by learning two graphs with incoming and outgoing directed attachments. Thus said, this method lays the foundation for approaching signal processing on expanding graphs by relying only on its stochastic connectivity model. Throughout, we consider the addition of only one node to an existing graph. To extend this approach to a continuously expanding graph, one has to generate or make available a corresponding training set. Moreover, with an increase in the dimension of the existing graph, the dimensions of $\mathbf{p}$ and $\mathbf{w}$ will also grow, which requires a treatment outside of the scope of this paper.

APPENDIX A

Proof of Proposition 1: The output of an order L filter at node v_+ is [cf. (6)] $[\mathbf{y}_+]_{N+1} = \mathbf{a}_+^\top \sum_{l=1}^{L} h_l \mathbf{A}_+^l \mathbf{x}_+$. The MSE at the incoming node is $\mathbb{E}[([\mathbf{y}_+]_{N+1} - x_+^\star)^2]$. Expanding the MSE for $L \geq 3$ leads to terms of the form $\mathbb{E}[\mathbf{a}_+^\top \mathbf{a}_+ \mathbf{a}_+^\top \mathbf{x}]$ which involve the third order statistics of $\mathbf{a}_+$. Computing the latter is notoriously challenging. We approximate the MSE up to second order statistics. Then, by substituting $\mathbf{A}_+$ [cf. (1)] into the filtering expression we get

$$[\mathbf{y}_+]_{N+1} \approx \mathbf{a}_+^\top \sum_{l=1}^{L} h_l \mathbf{A}^{l-1} \mathbf{x} = \mathbf{a}_+^\top \mathbf{A}_x \mathbf{h} \tag{27}$$

where $\mathbf{A}_x = [\mathbf{x}, \ldots, \mathbf{A}^{L-1}\mathbf{x}]$ and $\mathbf{h} = [h_1, \ldots, h_L]^\top$. The MSE is approximately

$$\text{MSE}(\mathbf{p}, \mathbf{w}) \approx \mathbb{E}[(\mathbf{a}_+^\top \mathbf{A}_x \mathbf{h} - x_+^\star)^2]. \tag{28}$$

Adding and subtracting $(\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h}$ within the expectation, we get

$$\begin{aligned}\text{MSE}(\mathbf{p}, \mathbf{w}) \approx \mathbb{E}[(&\mathbf{a}_+^\top \mathbf{A}_x \mathbf{h} - (\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} \\ &+ (\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} - x_+^\star)^2]\end{aligned} \tag{29}$$

which by expanding becomes

$$\begin{aligned}&\mathbb{E}[(\mathbf{a}_+^\top \mathbf{A}_x \mathbf{h} - (\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h})^2] + \mathbb{E}[(\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} - x_+^\star)^2] \\ &+ 2\mathbb{E}[(\mathbf{a}_+^\top \mathbf{A}_x \mathbf{h} - (\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h})((\mathbf{w} \circ \mathbf{p})^\top \mathbf{A}_x \mathbf{h} - x_+^\star)].\end{aligned} \tag{30}$$

In the first term, we expand the square, take $\mathbf{A}_x \mathbf{h}$ common and take the expectation inside to get $(\mathbf{A}_x \mathbf{h})^\top \mathbf{\Sigma}_+ \mathbf{A}_x \mathbf{h} = (\mathbf{A}_x \mathbf{h})^\top \mathbb{E}[(\mathbf{a}_+ - \mathbf{w} \circ \mathbf{p})(\mathbf{a}_+ - \mathbf{w} \circ \mathbf{p})^\top] \mathbf{A}_x \mathbf{h}$. The second term is deterministic, thus we can drop the expectation. The third term instead is zero because $\mathbb{E}[\mathbf{a}_+] = \mathbf{w} \circ \mathbf{p}$. Combining these results we get (7). □

Proof of Corollary 1: When node v_+ only forms directed edges landing on itself, the expanded adjacency matrix $\mathbf{A}_+$ and its lth power become

$$\mathbf{A}_+ = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{a}_+^\top & 0 \end{bmatrix} \text{ and } \mathbf{A}_+^l = \begin{bmatrix} \mathbf{A}^l & \mathbf{0} \\ \mathbf{a}_+^\top \mathbf{A}^{l-1} & 0 \end{bmatrix}. \tag{31}$$

Thus, the output of an order L graph filter is

$$[\mathbf{y}]_{N+1} = \mathbf{a}_+^\top \sum_{l=1}^{L} \mathbf{A}^{l-1} \mathbf{x} = \mathbf{a}_+^\top \mathbf{A}_x \mathbf{h} \tag{32}$$

which is identical to (27). Then, the proof follows similarly as for Proposition 1. □

APPENDIX B

Proof of Corollary 2: To find the convexity condition, we analyze when the Hessian of the function in (11) is positive semi-definite. The gradient of (10) w.r.t. $\mathbf{p}$ is shown in (11). The Hessian w.r.t. $\mathbf{p}$ is

$$\begin{aligned}\nabla_p^2 C_I(\mathbf{p}, \mathbf{w}) = &2(\mathbf{w} \circ \mathbf{A}_x \mathbf{h})(\mathbf{w} \circ \mathbf{A}_x \mathbf{h})^\top \\ &- 2\text{diag}((\mathbf{w} \circ \mathbf{A}_x \mathbf{h})^{\circ 2}) + 2\mu_p \mathbf{I}_N.\end{aligned} \tag{33}$$

The first term $(\mathbf{w} \circ \mathbf{A}_x \mathbf{h})(\mathbf{w} \circ \mathbf{A}_x \mathbf{h})^\top$ is a rank-one matrix with one non-zero eigenvalue $2||\mathbf{w} \circ \mathbf{A}_x \mathbf{h}||^2$ and $N-1$ zero eigenvalues. The second matrix is a diagonal matrix with eigenvalues $\{-2(w_1[\mathbf{A}_x \mathbf{h}]_1)^2, \ldots, -2(w_N[\mathbf{A}_x \mathbf{h}]_N)^2\}$. The third matrix is also diagonal but with each eigenvalue $2\mu_p$. The Hessian is the sum of a rank one matrix and two diagonal matrices. Its eigenvalues are the sum of the eigenvalues of these matrices [52]. By the semi-definite convexity condition [53], each of these eigenvalues now must be greater than or equal to zero. The condition

$$\mu_p \geq \max_i (w_i [\mathbf{A}_x \mathbf{h}]_i)^2 - ||\mathbf{w} \circ \mathbf{A}_x \mathbf{h}||_2^2 \tag{34}$$

is sufficient in this case. Since all $w_i \leq w_h$ from the constraint set [cf. (10)], we get (13) by substituting the them with the upper-bound. □

APPENDIX C

Proof of Proposition 2: The ESSE cost is

$$\text{ESSE}(\mathbf{p}, \mathbf{w}) = \mathbb{E}\left[(S_{\mathcal{G}_+}(\mathbf{a}_+) - S_{\mathcal{G}_+}(\mathbf{a}_+^\star))^2\right]. \tag{35}$$

Substituting expression (14) for $S_{\mathcal{G}_+}(\cdot)$ in (35), we get

$$\begin{aligned}\text{ESSE}(\mathbf{p}, \mathbf{w}) &= \mathbb{E}[(\mathbf{a}^\top \hat{\mathbf{x}} + S_{\mathcal{G}}(\mathbf{x}) - \mathbf{a}_+^{\star\top} \hat{\mathbf{x}} - S_{\mathcal{G}}(\mathbf{x}))^2] \\ &= \mathbb{E}\left[(\hat{\mathbf{x}}^\top (\mathbf{a}_+ - \mathbf{a}_+^\star)(\mathbf{a}_+ - \mathbf{a}_+^\star)^\top \hat{\mathbf{x}})\right]\end{aligned} \tag{36}$$

where $\hat{\mathbf{x}} = \mathbf{x}^{\circ 2} - 2x_+^\star \mathbf{x} - x_+^{\star 2}\mathbf{1}$. Taking the expectation operator inside, (36) becomes

$$\begin{aligned}\text{ESSE}(\mathbf{p}, \mathbf{w}) = \hat{\mathbf{x}}^\top \Big(&\mathbb{E}[\mathbf{a}_+ \mathbf{a}_+^\top] - (\mathbf{w} \circ \mathbf{p}) \mathbf{a}_+^{\star\top} \\ &- \mathbf{a}_+^\star (\mathbf{w} \circ \mathbf{p})^\top + \mathbf{a}_+^\star \mathbf{a}_+^{\star\top}\Big) \hat{\mathbf{x}}\end{aligned} \tag{37}$$

where we utilized $\mathbb{E}[\mathbf{a}_+] = \mathbf{w} \circ \mathbf{p}$. The term $\mathbb{E}[\mathbf{a}_+ \mathbf{a}_+^\top]$ is related to the covariance matrix of $\mathbf{a}_+$ as $\mathbf{\Sigma}_+ = \mathbb{E}[\mathbf{a}_+ \mathbf{a}_+^\top] - (\mathbf{w} \circ \mathbf{p})(\mathbf{w} \circ \mathbf{p})^\top = \text{diag}(\mathbf{w}^{\circ 2} \circ \mathbf{p} \circ (\mathbf{1} - \mathbf{p}))$ [cf.(3)]. Thus, by direct substitution we get (15). □

APPENDIX D

For the proof, we will need the following lemma.

Lemma 1: Consider a cost function $C(\mathbf{s})$ in some variable $\mathbf{s} \in \mathbb{R}^N$ satisfying Assumption 1. Let variable $\mathbf{s}$ be constrained to the convex set $\mathcal{S} = [s_l, s_h]^N$. Let also $\mathbf{s}^k$ and $\mathbf{s}^{k+1}$ be the kth and the $(k+1)$th iterations of a projected gradient descent approach on $\mathbf{s}$ for cost $C(\cdot)$ and let $\tilde{\mathbf{s}}^{k+1}$ be the output of the gradient update step

$$\tilde{\mathbf{s}}^{k+1} = \mathbf{s}^k - \lambda \nabla_{\mathbf{s}} C(\mathbf{s}^k) \tag{38}$$

with $\lambda > 0$. Then, for the projected vector update $\mathbf{s}^{k+1} = \prod_{\mathcal{S}}(\tilde{\mathbf{s}}^{k+1})$, the following holds:

$$||\mathbf{s}^{k+1} - \mathbf{s}^k|| \leq 2\lambda ||\nabla_{\mathbf{s}} C(\mathbf{s}^k)||. \tag{39}$$

Proof of Lemma 1: Consider vectors $\mathbf{s}^k$, $\mathbf{s}^{k+1}$, and $\tilde{\mathbf{s}}^{k+1}$ as points in $\mathbb{R}^N$ and $||\mathbf{s}^{k+1} - \mathbf{s}^k||$, $||\mathbf{s}^{k+1} - \tilde{\mathbf{s}}^{k+1}||$, and $||\tilde{\mathbf{s}}^{k+1} - \mathbf{s}^k||$ denote the Euclidean distances between them. The triangle inequality gives

$$||\mathbf{s}^{k+1} - \mathbf{s}^k|| \leq ||\mathbf{s}^{k+1} - \tilde{\mathbf{s}}^{k+1}|| + ||\mathbf{s}^k - \tilde{\mathbf{s}}^{k+1}||. \tag{40}$$

Since $\mathbf{s}^{k+1}$ is the Euclidean projection of $\tilde{\mathbf{s}}^{k+1}$, we have $||\mathbf{s}^{k+1} - \tilde{\mathbf{s}}^{k+1}|| \leq ||\mathbf{s}^{k} - \tilde{\mathbf{s}}^{k+1}||$ and inequality (40) becomes

$$||\mathbf{s}^{k+1} - \mathbf{s}^{k}|| \leq 2||\mathbf{s}^{k} - \tilde{\mathbf{s}}^{k+1}|| = 2|\lambda|||\nabla_s C(\mathbf{s}^k)||. \quad \square$$

Note that the lemma holds for $\mathbf{s}$ being $\mathbf{p}$ or $\mathbf{w}$ given the other is fixed, λ being λ_p or λ_w and $\mathcal{S}$ being $[0,1]^N$ or $[w_l, w_h]^N$, respectively.

Proof of Proposition 3:

Claim 1. Non-increasing cost: Let $C(\mathbf{p}^{k+1}, \mathbf{w}^k)$ be the cost function evaluated at $\mathbf{p}^{k+1}$ and $\mathbf{w}^k$ [cf. step 5, Algorithm 1]. Taking the Taylor expansion at this point, we get

$$C(\mathbf{p}^{k+1}, \mathbf{w}^k) = C(\mathbf{p}^{k}, \mathbf{w}^k) + \nabla_p^\top C(\mathbf{p}^k)(\mathbf{p}^{k+1} - \mathbf{p}^k) + \frac{1}{2}(\mathbf{p}^{k+1} - \mathbf{p}^k)^\top \nabla_p^2 C(\mathbf{p}^k, \mathbf{w}^k)(\mathbf{p}^{k+1} - \mathbf{p}^k). \tag{41}$$

Under Assumption 1 the Hessian is upper bounded as $\nabla^2_{\mathbf{p}} C(\mathbf{p}, \mathbf{w}) \preceq L_p \mathbf{I}$, thus we have

$$C(\mathbf{p}^{k+1}, \mathbf{w}^k) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) + \nabla_p^\top C(\mathbf{p}^k)(\mathbf{p}^{k+1} - \mathbf{p}^k) + \frac{L_p}{2}||\mathbf{p}^{k+1} - \mathbf{p}^k||^2. \tag{42}$$

We then substitute $\nabla_p C(\mathbf{p}^k, \mathbf{w}^k) = -\lambda_p^{-1}(\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k)$ for the gradient step to write (42) as

$$C(\mathbf{p}^{k+1}, \mathbf{w}^k) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) - \frac{1}{\lambda_p}(\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k)^\top (\mathbf{p}^{k+1} - \mathbf{p}^k) + \frac{L_p}{2}||\mathbf{p}^{k+1} - \mathbf{p}^k||^2. \tag{43}$$

Next, we use the cosine rule identity

$$(\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k)^\top (\mathbf{p}^{k+1} - \mathbf{p}^k) = \frac{1}{2}\left(||\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k||^2 + ||\mathbf{p}^{k+1} - \mathbf{p}^k||^2 - ||\mathbf{p}^{k+1} - \tilde{\mathbf{p}}^{k+1}||^2\right) \tag{44}$$

in the second term of (43) to get

$$C(\mathbf{p}^{k+1}, \mathbf{w}^k) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) - \frac{1}{2\lambda_p}||\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k||^2 - \frac{1}{2\lambda_p}||\mathbf{p}^{k+1} - \mathbf{p}^k||^2 + \frac{1}{2\lambda_p}||\mathbf{p}^{k+1} - \tilde{\mathbf{p}}^{k+1}||^2 + \frac{L_p}{2}||\mathbf{p}^{k+1} - \mathbf{p}^k||^2 \tag{45}$$

The second term on the RHS of (45) is lesser than or equal to zero, so we drop it. Since $\mathbf{p}^{k+1}$ is the Euclidean projection of $\tilde{\mathbf{p}}^{k+1}$ onto the constraint set, we have $||\mathbf{p}^{k+1} - \tilde{\mathbf{p}}^{k+1}||^2 \leq ||\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k||^2$. Then, we write (45) as

$$C(\mathbf{p}^{k+1}, \mathbf{w}^k) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) + \left(\frac{L_p}{2} - \frac{1}{2\lambda_p}\right)||\mathbf{p}^{k+1} - \mathbf{p}^k||^2 + \frac{1}{2\lambda_p}||\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k||^2. \tag{46}$$

Using Lemma 1, we substitute $||\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^k||^2 \leq 4\lambda_p^2||\nabla_p C(\mathbf{p}^k, \mathbf{w}^k)||^2$ in (46) and obtain

$$C(\mathbf{p}^{k+1}, \mathbf{w}^k) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) + \alpha||\nabla_p C(\mathbf{p}^k, \mathbf{w}^k)||^2 \tag{47}$$

where $\alpha = ((\frac{L_p}{2} - \frac{1}{2\lambda_p})4\lambda_p^2 + \frac{\lambda_p}{2}) = 2\lambda_p^2 L_p - \frac{3\lambda_p}{2}$. For $\alpha \leq 0$, the cost reduces in $\mathbf{p}$, i.e., $C(\mathbf{p}^{k+1}, \mathbf{w}^k) \leq C(\mathbf{p}^k, \mathbf{w}^k)$, thus, the step size must satisfy $0 < \lambda_p \leq \frac{3}{4L_p}$. Hence, the cost is non-increasing with each update in $\mathbf{p}$ and $\mathbf{w}^k$ fixed.

For the $\mathbf{w}$ update, we follow the same approach but we perform the Taylor expansion around the point $(\mathbf{p}^{k+1}, \mathbf{w}^k)$. Following similar derivations, it can be shown that

$$C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) \leq C(\mathbf{p}^{k+1}, \mathbf{w}^k) + \beta||\nabla_w C(\mathbf{p}^{k+1}, \mathbf{w}^k)||^2 \tag{48}$$

where $\beta = ((\frac{L_w}{2} - \frac{1}{2\lambda_w})4\lambda_w^2 + \frac{\lambda_w}{2})$, and $C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) \leq C(\mathbf{p}^{k+1}, \mathbf{w}^k)$ if the step size satisfies $0 < \lambda_w \leq \frac{3}{4L_w}$. Then, combining the two inequalities we have

$$C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) \tag{49}$$

which shows the alternating projected gradient descent step has a non-increasing cost. $\square$

Claim 2. Local minima: Let least one local minima $(\mathbf{p}^\star, \mathbf{w}^\star)$ exists in $\mathcal{S}$. Substituting (47) in (48) gives

$$C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) \leq C(\mathbf{p}^{k}, \mathbf{w}^k) + \alpha||\nabla_p C(\mathbf{p}^k, \mathbf{w}^k)||^2 + \beta||\nabla_w C(\mathbf{p}^{k+1}, \mathbf{w}^k)||^2. \tag{50}$$

We denote $\nabla_p C(\mathbf{p}^k, \mathbf{w}^k)$ and $\nabla_w C(\mathbf{p}^{k+1}, \mathbf{w}^k)$ as $\nabla_p C(\mathbf{p}^k)$ and $\nabla_w C(\mathbf{w}^k)$ to further ease the notation. We denote by $C(\mathbf{p}^\star, \mathbf{w}^\star)$ the cost at the local minima. Due to the non-increasing cost, after k iterations, we have a condition where the algorithm will be near the feasible local optima. Using then the first order Taylor expansion at this point

$$C(\mathbf{p}^{k}, \mathbf{w}^k) = C(\mathbf{p}^\star, \mathbf{w}^\star) - \nabla_p^\top C(\mathbf{p}^k)(\mathbf{p}^\star - \mathbf{p}^k) - \nabla_w^\top C(\mathbf{w}^k)(\mathbf{w}^\star - \mathbf{w}^k). \tag{51}$$

and substituting it in (50), we get

$$C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) - C(\mathbf{p}^\star, \mathbf{w}^\star) \leq \nabla_p^\top C(\mathbf{p}^k)(\mathbf{p}^k - \mathbf{p}^\star) + \alpha||\nabla_p C(\mathbf{p}^k)||^2 + \nabla_w^\top C(\mathbf{w}^k)(\mathbf{w}^k - \mathbf{w}^\star) + \beta||\nabla_w C(\mathbf{w}^k)||^2. \tag{52}$$

We then substitute the cosine rule

$$\nabla_p^\top C(\mathbf{p}^k)(\mathbf{p}^k - \mathbf{p}^\star) = \frac{\lambda_p}{2}||\nabla_p C(\mathbf{p}^k)||^2 + \frac{1}{2\lambda_p}||\mathbf{p}^k - \mathbf{p}^\star||^2 - \frac{1}{2\lambda_p}||\mathbf{p}^k - \mathbf{p}^\star - \lambda_p \nabla_p C(\mathbf{p}^k)||^2 \tag{53}$$

and its equivalent form in $\mathbf{w}$ in (52) to get

$$C(\mathbf{p}^{k+1}, \mathbf{w}^{k+1}) - C(\mathbf{p}^\star, \mathbf{w}^\star) \leq \left(\frac{\lambda_p}{2} + \alpha\right)||\nabla_p C(\mathbf{p}^k)||^2 + \left(\frac{\lambda_w}{2} + \beta\right)||\nabla_w C(\mathbf{w}^k)||^2 + \frac{1}{2\lambda_p}(||\mathbf{p}^k - \mathbf{p}^\star||^2 - ||\tilde{\mathbf{p}}^{k+1} - \mathbf{p}^\star||^2) + \frac{1}{2\lambda_w}(||\mathbf{w}^k - \mathbf{w}^\star||^2 - ||\tilde{\mathbf{w}}^{k+1} - \mathbf{w}^\star||^2). \tag{54}$$

Now, if $(\frac{\lambda_p}{2}+\alpha)\leq 0$ and $(\frac{\lambda_w}{2}+\beta)\leq 0$, we can ignore the first two terms in the RHS of (54). Substituting for α and β in these conditions, we get $\lambda_p \leq \frac{1}{2L_p}$ and $\lambda_w \leq \frac{1}{2L_w}$. To prove convergence to the local minima $(\mathbf{p}^\star, \mathbf{w}^\star)$ we utilize the inequality $||\mathbf{p}^{k+1}-\mathbf{p}^\star||^2 \leq ||\tilde{\mathbf{p}}^{k+1}-\mathbf{p}^\star||^2$, i.e., the gradient update is closer to the optima than the projection update, which holds under the assumption of the local minima being feasible. By using this inequality for both variables, we get

$$C(\mathbf{p}^{k+1},\mathbf{w}^{k+1})-C(\mathbf{p}^\star,\mathbf{w}^\star)\leq \frac{1}{2\lambda_p}(||\mathbf{p}^k-\mathbf{p}^\star||^2 - ||\mathbf{p}^{k+1}-\mathbf{p}^\star||^2)+\frac{1}{2\lambda_w}(||\mathbf{w}^k-\mathbf{w}^\star||^2-||\mathbf{w}^{k+1}-\mathbf{w}^\star||^2). \tag{55}$$

Summing from $k=0$ to K, we get a telescoping sum and can write

$$\sum_{k=0}^{K}\big(C(\mathbf{p}^{k+1},\mathbf{w}^{k+1})-C(\mathbf{p}^\star,\mathbf{w}^\star)\big)\leq \frac{1}{2\lambda_p}(||\mathbf{p}^0-\mathbf{p}^\star||^2 - ||\mathbf{p}^{K+1}-\mathbf{p}^\star||^2)+\frac{1}{2\lambda_w}(||\mathbf{w}^0-\mathbf{w}^\star||^2-||\mathbf{w}^{K+1}-\mathbf{w}^\star||^2). \tag{56}$$

We divide both sides of (56) by $(K+1)$ and use the inequality $C(\mathbf{p}^{K+1},\mathbf{w}^{K+1})\leq \frac{1}{K+1}\sum_{k=0}^{K}C(\mathbf{p}^{k+1},\mathbf{w}^{k+1})$ which holds because $C(\mathbf{p}^{K+1},\mathbf{w}^{K+1})$ is lesser than or equal to all the terms from $k=0,\dots K$ and get

$$\begin{aligned} &C(\mathbf{p}^{K+1},\mathbf{w}^{K+1})-C(\mathbf{p}^\star,\mathbf{w}^\star)\\ &\leq \frac{1}{2(K+1)\lambda_p}(||\mathbf{p}^0-\mathbf{p}^\star||^2-||\mathbf{p}^{K+1}-\mathbf{p}^\star||^2)\\ &+\frac{1}{2(K+1)\lambda_w}(||\mathbf{w}^0-\mathbf{w}^\star||^2-||\mathbf{w}^{K+1}-\mathbf{w}^\star||^2). \end{aligned} \tag{57}$$

As iteration index $K\to\infty$, $C(\mathbf{p}^{K+1},\mathbf{w}^{K+1})\to C(\mathbf{p}^\star,\mathbf{w}^\star)$. Thus, convergence to a local minima is possible with rate of convergence $\mathcal{O}(1/K)$. □

APPENDIX E

Proof of Proposition 4: The perturbation between the realization adjacency matrix $\mathbf{A}_+$ and its nominal $\overline{\mathbf{A}}_+$, $\triangle\mathbf{A}_+ = \mathbf{A}_+ - \overline{\mathbf{A}}_+$ is

$$\triangle\mathbf{A}_+ = \begin{bmatrix} \mathbf{0} & \mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1}) \\ \mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1})^\top & 0 \end{bmatrix}. \tag{58}$$

Invoking Assumption 3, we substitute $\triangle\mathbf{A}_+$ and $\mathbf{u}_i$ in (24) to get

$$\triangle\gamma_i = \mathbf{u}_i^\top \begin{bmatrix} \mathbf{0} & \mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1}) \\ \mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1})^\top & 0 \end{bmatrix}\mathbf{u}_i. \tag{59}$$

Then, denoting by $[\mathbf{u}_i]_{1:N}$ the vector containing the first N elements of $\mathbf{u}_i$, (59) can be written as

$$\triangle\gamma_i = (2[\mathbf{u}_i]_{N+1}[\mathbf{u}_i]_{1:N})^\top\mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1}). \tag{60}$$

Now, we apply the Cauchy-Schwartz inequality on (60) and square both sides to get

$$\triangle^2\gamma_i \leq 4||[\mathbf{u}_i]_{N+1}[\mathbf{u}_i]_{1:N}||^2||\mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1})||^2 \tag{61}$$

By taking the expectation, we get

$$\mathbb{E}[\triangle^2\gamma_i] \leq \mathbb{E}\left[4||[\mathbf{u}_i]_{N+1}[\mathbf{u}_i]_{1:N}||^2||\mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1})||^2\right] \tag{62}$$

which writes as

$$\mathbb{E}[\triangle^2\gamma_i] \leq c_1\mathbb{E}[||\mathbf{w}\circ(\mathcal{S}(\mathbf{p})-\mathbf{1})||^2] \tag{63}$$

The expectation operates on the sum $\sum_{i=1}^{N} w_i^2(\mathcal{S}(p_i)-1)^2$. Given w_i is fixed, and utilizing $\mathbb{E}[\mathcal{S}(p_i)^2]=p_i$, and $\mathbb{E}[\mathcal{S}(p_i)]=p_i$, the result of the expectation is $\sum_{i=1}^{N} w_i^2(1-p_i)$, which writes as $\mathbf{p}_+^\top\mathbf{\Sigma}_+\mathbf{p}_+$, where $\mathbf{\Sigma}_+$ is the covariance matrix of $\mathbf{a}_+$, and $[\mathbf{p}_+]_i = 1/\sqrt{p_i}$, if $p_i\neq 0$, and zero otherwise. The same steps hold for the perturbed nominal Laplacian $\overline{\mathbf{L}}_+$ and its ith eigenvector $\mathbf{v}_i$ to prove (26). □

REFERENCES

[1] B. Das and E. Isufi, "Learning expanding graphs for signal interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022. [Online]. Available: https://arxiv.org/pdf/2203.07966v1.pdf

[2] C. C. Aggarwal, "An introduction to social network data analytics," in *Social Network Data Analytics*. Berlin, Germany: Springer, 2011, pp. 1–15.

[3] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási, "Functional and topological characterization of protein interaction networks," *Proteomic.*, vol. 4, no. 4, pp. 928–942, 2004.

[4] W. Huang, A. G. Marques, and A. R. Ribeiro, "Rating prediction via graph signal processing," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5066–5081, Oct. 2018.

[5] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019.

[6] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 1–33, 2016.

[7] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Sci.*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[8] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, New York, NY, USA, 2002, pp. 253–260.

[9] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: Divided they blog," in *Proc. 3rd Int. Workshop Link Discov.*, 2005, pp. 36–43.

[10] P. Erdos, "On the evolution of random graphs," *Bull. Inst. Int. Statist.*, vol. 38, pp. 343–347, 1961.

[11] G. Bianconi and A.-L. Barabási, "Competition and multiscaling in evolving networks," *Europhysics Lett.*, vol. 54, no. 4, May 2001, Art. no. 436.

[12] V. N. Zadorozhnyi and E. B. Yudin, "Growing network: Models following nonlinear preferential attachment rule," *Physica A: Stat. Mechanics Appl.*, vol. 428, pp. 111–132, Jun. 2015.

[13] Y. Shen, G. Leus, and G. B. Giannakis, "Online graph-adaptive learning with scalability and privacy," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2471–2483, May 2019.

[14] A. Venkitaraman, S. Chatterjee, and B. Wahlberg, "Recursive prediction of graph signals with incoming nodes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 5565–5569.

[15] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," in *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[16] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2016, pp. 920–929.

[17] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 44–63, May 2019.

[18] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, Dec. 2016.
[19] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, Sep. 2017.
[20] R. Shafipour, S. Segarra, A. Marques, and G. Mateos, "Identifying the topology of undirected networks from diffused non-stationary graph signals," *IEEE Open J. Signal Process.*, vol. 2, pp. 171–189, 2021.
[21] D. Thanou, X. Dong, D. Kressner, and P. Frossard, "Learning heat diffusion graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 484–499, Sep. 2017.
[22] M. Coutino, E. Isufi, T. Maehara, and G. Leus, "State-space network topology identification from partial observations," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 211–225, Feb. 2020.
[23] H. E. Egilmez, E. Pavez, and A. Ortega, "Graph learning from data under laplacian and structural constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 825–841, Sep. 2017.
[24] P. Ravikumar et al., "High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.
[25] N. Meinshausen et al., "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.
[26] V. Kalofolias, A. Loukas, D. Thanou, and P. Frossard, "Learning time varying graphs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2826–2830.
[27] K. Yamada, Y. Tanaka, and A. Ortega, "Time-varying graph learning with constraints on graph temporal variation," 2020, *arXiv:2001.03346*.
[28] D. Hallac, Y. Park, S. Boyd, and J. Leskovec, "Network inference via the time-varying graphical lasso," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 205–213.
[29] S. Vlaski, H. P. Maretić, R. Nassif, P. Frossard, and A. H. Sayed, "Online graph learning from sequential data," in *Proc. IEEE Data Sci. Workshop*, 2018, pp. 190–194.
[30] M. Moscu, R. Borsoi, and C. Richard, "Online graph topology inference with kernels for brain connectivity estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 1200–1204.
[31] R. Shafipour and G. Mateos, "Online proximal gradient for learning graphs from streaming signals," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 865–869.
[32] B. Zaman, L. M. L. Ramos, D. Romero, and B. Beferull-Lozano, "Online topology identification from vector autoregressive time series," *IEEE Trans. Signal Process.*, vol. 69, pp. 210–225, Dec. 2020.
[33] A. Natali, M. Coutino, E. Isufi, and G. Leus, "Online time-varying topology identification via prediction-correction algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5400–5404.
[34] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annu. Rev. Sociol.*, vol. 27, no. 1, pp. 415–444, 2001.
[35] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statist. Mechanics Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.
[36] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.
[37] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 52, pp. 22073–22078, 2009.
[38] F. Tan, Y. Xia, and B. Zhu, "Link prediction in complex networks: A mutual information perspective," *PLoS One*, vol. 9, no. 9, 2014, Art. no. e107056.
[39] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, 2009.
[40] W. Liu and L. Lü, "Link prediction based on local random walk," *Europhysics Lett.*, vol. 89, no. 5, 2010, Art. no. 58007.
[41] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Phys. Rev. E*, vol. 73, no. 2, 2006, Art. no. 026120.
[42] V. Matta, A. Santos, and A. H. Sayed, "Graph learning with partial observations: Role of degree concentration," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 1312–1316.
[43] M. Cirillo, V. Matta, and A. H. Sayed, "Learning bollobás-riordan graphs under partial observability," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5360–5364.
[44] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.
[45] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.
[46] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," in *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.
[47] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," 2015, *arXiv:1509.03025*.
[48] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, "CNN-based projected gradient descent for consistent CT image reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1440–1453, Jun. 2018.
[49] E. Ceci and S. Barbarossa, "Graph signal processing in the presence of topology uncertainties," *IEEE Trans. Signal Process.*, vol. 68, pp. 1558–1573, Feb. 2020.
[50] E. Ceci and S. Barbarossa, "Small perturbation analysis of network topologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4194–4198.
[51] A.-L. Barabási et al., *Network Science*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[52] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, U.K.: Cambridge Univ. Press, 2018.
[53] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.