Link Prediction in Graphs with Autoregressive Features

Emile Richard
CMLA UMR CNRS 8536,
ENS Cachan, France

Stéphane Gaïffas
CMAP - Ecole Polytechnique
& LSTA - Université Paris 6

Nicolas Vayatis
CMLA UMR CNRS 8536,
ENS Cachan, France

Abstract

In the paper, we consider the problem of link prediction in time-evolving graphs. We assume that certain graph features, such as the node degree, follow a vector autoregressive (VAR) model and we propose to use this information to improve the accuracy of prediction. Our strategy involves a joint optimization procedure over the space of adjacency matrices and VAR matrices which takes into account both sparsity and low rank properties of the matrices. Oracle inequalities are derived and illustrate the trade-offs in the choice of smoothing parameters when modeling the joint effect of sparsity and low rank property. The estimate is computed efficiently using proximal methods through a generalized forward-backward agorithm.

1 Introduction

Forecasting systems behavior with multiple responses has been a challenging issue in many contexts of applications such as collaborative filtering, financial markets, or bioinformatics, where responses can be, respectively, movie ratings, stock prices, or activity of genes within a cell. Statistical modeling techniques have been widely investigated in the context of multivariate time series either in the multiple linear regression setup [4] or with autoregressive models [23]. More recently, kernel-based regularized methods have been developed for multitask learning [7, 2]. These approaches share the use of the correlation structure among input variables to enrich the prediction on every single output. Often, the correlation structure is assumed to be given or it is estimated separately. A discrete encoding of correlations between variables can be modeled as a graph so that learning the dependence structure amounts to performing graph inference through the discovery of uncovered edges on the graph. The latter problem is interesting *per se* and it is known as the problem of link prediction where it is assumed that only a part of the graph is actually observed [15, 9]. This situation occurs in various applications such as recommender systems, social networks, or proteomics, and the appropriate tools can be found among matrix completion techniques [21, 5, 1]. In the realistic setup of a time-evolving graph, matrix completion was also used and adapted to take into account the dynamics of the features of the graph [18]. In this paper, we study the prediction problem where the observation is a sequence of graphs adjacency matrices $(A_t)_{0\leq t\leq T}$ and the goal is to predict A_{T+1}. This type of problem arises in applications such as recommender systems where, given information on purchases made by some users, one would like to predict future purchases. In this context, users and products can be modeled as the nodes of a bipartite graph, while purchases or clicks are modeled as edges. In functional genomics and systems biology, estimating regulatory networks in gene expression can be performed by modeling the data as graphs and fitting predictive models is a natural way for estimating evolving networks in these contexts. A large variety of methods for link prediction only consider predicting from a single static snapshot of the graph - this includes heuristics [15, 20], matrix factorization [13], diffusion [16], or probabilistic methods [22]. More recently, some works have investigated using sequences of observations of the graph to improve the prediction, such as using regression on features extracted from the graphs [18], using matrix factorization [14], continuous-time regression [25]. Our main assumption is that the network effect is a

cause and a symptom at the same time, and therefore, the edges and the graph features should be estimated simultaneously. We propose a regularized approach to predict the uncovered links and the evolution of the graph features simultaneously. We provide oracle bounds under the assumption that the noise sequence has subgaussian tails and we prove that our procedure achieves a trade-off in the calibration of smoothing parameters which adjust with the sparsity and the rank of the unknown adjacency matrix. The rest of this paper is organized as follows. In Section 2, we describe the general setup of our work with the main assumptions and we formulate a regularized optimization problem which aims at jointly estimating the autoregression parameters and predicting the graph. In Section 3, we provide technical results with oracle inequalities and other theoretical guarantees on the joint estimation-prediction. Section 4 is devoted to the description of the numerical simulations which illustrate our approach. We also provide an efficient algorithm for solving the optimization problem and show empirical results. The proof of the theoretical results are provided as supplementary material in a separate document.

2 Estimation of low-rank graphs with autoregressive features

Our approach is based on the asumption that features can explain most of the information contained in the graph, and that these features are evolving with time. We make the following assumptions about the sequence $(A_t)_{t\geq 0}$ of adjacency matrices of the graphs sequence.

Low-Rank. We assume that the matrices A_t have low-rank. This reflects the presence of highly connected groups of nodes such as communities in social networks, or product categories and groups of loyal/fan users in a market place data, and is sometimes motivated by the small number of factors that explain nodes interactions.

Autoregressive linear features. We assume to be given a linear map $\omega : \mathbb{R}^{n\times n} \to \mathbb{R}^d$ defined by

$$\omega(A) = \Big(\langle \Omega_1, A\rangle, \cdots, \langle \Omega_d, A\rangle\Big)^\top, \tag{1}$$

where $(\Omega_i)_{1\leq i\leq d}$ is a set of $n \times n$ matrices. These matrices can be either deterministic or random in our theoretical analysis, but we take them deterministic for the sake of simplicity. The vector time series $(\omega(A_t))_{t\geq 0}$ has autoregressive dynamics, given by a VAR (Vector Auto-Regressive) model:

$$\omega(A_{t+1}) = W_0^\top \omega(A_t) + N_{t+1}, \tag{2}$$

where $W_0 \in \mathbb{R}^{d\times d}$ is a unknown sparse matrix and $(N_t)_{t\geq 0}$ is a sequence of noise vectors in $\mathbb{R}^d$. An example of linear features is the degree (*i.e.* number of edges connected to each node, or the sum of their weights if the edges are weighted), which is a measure of popularity in social and commerce networks. Introducing

$$\mathbf{X}_{T-1} = (\omega(A_0), \ldots, \omega(A_{T-1}))^\top \quad \text{and} \quad \mathbf{X}_T = (\omega(A_1), \ldots, \omega(A_T))^\top,$$

which are both $T \times d$ matrices, we can write this model in a matrix form:

$$\mathbf{X}_T = \mathbf{X}_{T-1} W_0 + \mathbf{N}_T, \tag{3}$$

where $\mathbf{N}_T = (N_1, \ldots, N_T)^\top$.

This assumes that the noise is driven by time-series dynamics (a martingale increment), where each coordinates are independent (meaning that features are independently corrupted by noise), with a sub-gaussian tail and variance uniformly bounded by a constant σ^2. In particular, no independence assumption between the N_t is required here.

Notations. The notations $\|\cdot\|_F$, $\|\cdot\|_p$, $\|\cdot\|_\infty$, $\|\cdot\|_*$ and $\|\cdot\|_{\mathrm{op}}$ stand, respectively, for the Frobenius norm, entry-wise ℓ_p norm, entry-wise ℓ_∞ norm, trace-norm (or nuclear norm, given by the sum of the singular values) and operator norm (the largest singular value). We denote by $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ the Euclidean matrix product. A vector in $\mathbb{R}^d$ is always understood as a $d \times 1$ matrix. We denote by $\|A\|_0$ the number of non-zero elements of A. The product $A \circ B$ between two matrices with matching dimensions stands for the Hadamard or entry-wise product between A and B. The matrix $|A|$ contains the absolute values of entries of A. The matrix $(M)_+$ is the componentwise positive part of the matrix M, and $\mathrm{sign}(M)$ is the sign matrix associated to M with the convention $\mathrm{sign}(0) = 0$

If A is a $n \times n$ matrix with rank r, we write its SVD as $A = U\Sigma V^\top = \sum_{j=1}^r \sigma_j u_j v_j^\top$ where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ is a $r \times r$ diagonal matrix containing the non-zero singular values of A in decreasing order, and $U = [u_1, \dots, u_r]$, $V = [v_1, \dots, v_r]$ are $n \times r$ matrices with columns given by the left and right singular vectors of A. The projection matrix onto the space spanned by the columns (resp. rows) of A is given by $P_U = UU^\top$ (resp. $P_V = VV^\top$). The operator $\mathcal{P}_A : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ given by $\mathcal{P}_A(B) = P_U B + BP_V - P_U B P_V$ is the projector onto the linear space spanned by the matrices $u_k x^\top$ and $y v_k^\top$ for $1 \le j, k \le r$ and $x, y \in \mathbb{R}^n$. The projector onto the orthogonal space is given by $\mathcal{P}_A^\perp(B) = (I - P_U)B(I - P_V)$. We also use the notation $a \vee b = \max(a, b)$.

2.1 Joint prediction-estimation through penalized optimization

In order to reflect the autoregressive dynamics of the features, we use a least-squares goodness-of-fit criterion that encourages the similarity between two feature vectors at successive time steps. In order to induce sparsity in the estimator of W_0, we penalize this criterion using the ℓ_1 norm. This leads to the following penalized objective function:

$$J_1(W) = \frac{1}{T}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \kappa\|W\|_1,$$

where $\kappa > 0$ is a smoothing parameter.

Now, for the prediction of A_{T+1}, we propose to minimize a least-squares criterion penalized by the combination of an ℓ_1 norm and a trace-norm. This mixture of norms induces sparsity and a low-rank of the adjacency matrix. Such a combination of ℓ_1 and trace-norm was already studied in [8] for the matrix regression model, and in [19] for the prediction of an adjacency matrix.

The objective function defined below exploits the fact that if W is close to W_0, then the features of the next graph $\omega(A_{T+1})$ should be close to $W^\top\omega(A_T)$. Therefore, we consider

$$J_2(A, W) = \frac{1}{d}\|\omega(A) - W^\top\omega(A_T)\|_F^2 + \tau\|A\|_* + \gamma\|A\|_1,$$

where $\tau, \gamma > 0$ are smoothing parameters. The overall objective function is the sum of the two partial objectives J_1 and J_2, which is jointly convex with respect to A and W:

$$\mathcal{L}(A, W) \doteq \frac{1}{T}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \kappa\|W\|_1 + \frac{1}{d}\|\omega(A) - W^\top\omega(A_T)\|_2^2 + \tau\|A\|_* + \gamma\|A\|_1, \quad (4)$$

If we choose convex cones $\mathcal{A} \subset \mathbb{R}^{n\times n}$ and $\mathcal{W} \subset \mathbb{R}^{d\times d}$, our joint estimation-prediction procedure is defined by

$$(\hat{A}, \hat{W}) \in \operatorname*{arg\,min}_{(A,W)\in\mathcal{A}\times\mathcal{W}} \mathcal{L}(A, W). \quad (5)$$

It is natural to take $\mathcal{W} = \mathbb{R}^{d\times d}$ and $\mathcal{A} = (\mathbb{R}_+)^{n\times n}$ since there is no *a priori* on the values of the feature matrix W_0, while the entries of the matrix A_{T+1} must be positive.

In the next section we propose oracle inequalities which prove that this procedure can estimate W_0 and predict A_{T+1} at the same time.

2.2 Main result

The central contribution of our work is to bound the prediction error with high probability under the following natural hypothesis on the noise process.

Assumption 1. *We assume that $(N_t)_{t\ge0}$ satisfies $\mathbb{E}[N_t|\mathcal{F}_{t-1}] = 0$ for any $t \ge 1$ and that there is $\sigma > 0$ such that for any $\lambda \in \mathbb{R}$ and $j = 1, \dots, d$ and $t \ge 0$:*

$$\mathbb{E}[e^{\lambda(N_t)_j}|\mathcal{F}_{t-1}] \le e^{\sigma^2\lambda^2/2}.$$

Moreover, we assume that for each $t \ge 0$, the coordinates $(N_t)_1, \dots, (N_t)_d$ are independent.

The main result can be summarized as follows. The prediction error and the estimation error can be simultaneously bounded by the sum of three terms that involve homogeneously (a) the sparsity, (b) the rank of the adjacency matrix A_{T+1}, and (c) the sparsity of the VAR model matrix W_0. The tight bounds we obtain are similar to the bounds of the Lasso and are upper bounded by:

$$C_1 \frac{\log d}{T} \|W_0\|_0 + C_2 \frac{\log n}{d} \|A_{T+1}\|_0 + C_3 \frac{\log n}{d} \operatorname{rank} A_{T+1} \ .$$

The positive constants C_1, C_2, C_3 are proportional to the noise level σ. The interplay between the rank and sparsity constraints on A_{T+1} are reflected in the observation that the values of C_2 and C_3 can be changed as long as their sum remains constant.

3 Oracle inequalities

In this section we give oracle inequalities for the mixed prediction-estimation error which is given, for any $A \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{d\times d}$, by

$$\mathcal{E}(A,W)^2 \doteq \frac{1}{d}\|(W-W_0)^\top \omega(A_T) - \omega(A - A_{T+1})\|_2^2 + \frac{1}{T}\|\mathbf{X}_{T-1}(W-W_0)\|_F^2. \qquad (6)$$

It is important to have in mind that an upper-bound on $\mathcal{E}$ implies upper-bounds on each of its two components. It entails in particular an upper-bound on the feature estimation error $\|\mathbf{X}_{T-1}(\widehat{W} - W_0)\|_F$ that makes $\|(\widehat{W} - W_0)^\top \omega(A_T)\|_2$ smaller and consequently controls the prediction error over the graph edges through $\|\omega(\widehat{A} - A_{T+1})\|_2$.

The upper bounds on $\mathcal{E}$ given below exhibit the dependence of the accuracy of estimation and prediction on the number of features d, the number of edges n and the number T of observed graphs in the sequence.

Let us recall $\mathbf{N}_T = (N_1, \dots, N_T)^\top$ and introduce the noise processes

$$M = -\frac{1}{d}\sum_{j=1}^{d} (N_{T+1})_j \Omega_j \quad \text{and} \quad \Xi = \frac{1}{T}\sum_{t=1}^{T} \omega(A_{t-1}) N_t^\top + \frac{1}{d}\omega(A_T) N_{T+1}^\top,$$

which are, respectively, $n \times n$ and $d \times d$ random matrices. The source of randomness comes from the noise sequence $(N_t)_{t\geq 0}$, see Assumption 1. If these noise processes are controlled correctly, we can prove the following oracle inequalities for procedure (5). The next result is an oracle inequality of slow type (see for instance [3]), that holds in full generality.

Theorem 1. *Under Assumption 2, let $(\hat{A}, \hat{W})$ be given by* (5) *and suppose that*

$$\tau \geq 2\alpha\|M\|_{\mathrm{op}}, \quad \gamma \geq 2(1-\alpha)\|M\|_\infty \quad \textit{and} \quad \kappa \geq 2\|\Xi\|_\infty \qquad (7)$$

for some $\alpha \in (0,1)$. Then, we have

$$\mathcal{E}(\widehat{A}, \widehat{W})^2 \leq \inf_{(A,W)\in \mathcal{A}\times\mathcal{W}} \Big\{ \mathcal{E}(A,W)^2 + 2\tau\|A\|_* + 2\gamma\|A\|_1 + 2\kappa\|W\|_1 \Big\}.$$

For the proof of oracle inequalities of fast type, the *restricted eigenvalue* (RE) condition introduced in [3] and [10, 11] is of importance. Restricted eigenvalue conditions are implied by, and in general weaker than, the so-called *incoherence* or RIP (Restricted isometry property, [6]) assumptions, which excludes, for instance, strong correlations between covariates in a linear regression model. This condition is acknowledged to be one of the weakest to derive fast rates for the Lasso (see [24] for a comparison of conditions).

Matrix version of these assumptions are introduced in [12]. Below is a version of the RE assumption that fits in our context. First, we need to introduce the two restriction cones.

The first cone is related to the $\|W\|_1$ term used in procedure (5). If $W \in \mathbb{R}^{d\times d}$, we denote by $\Theta_W = \operatorname{sign}(W) \in \{0, \pm 1\}^{d\times d}$ the signed sparsity pattern of W and by $\Theta_W^\perp \in \{0,1\}^{d\times d}$ the orthogonal sparsity pattern. For a fixed matrix $W \in \mathbb{R}^{d\times d}$ and $c > 0$, we introduce the cone

$$\mathcal{C}_1(W,c) \doteq \Big\{ W' \in \mathcal{W} : \|\Theta_W^\perp \circ W'\|_1 \leq c\|\Theta_W \circ W'\|_1 \Big\}.$$

This cone contains the matrices W' that have their largest entries in the sparsity pattern of W.

The second cone is related to mixture of the terms $\|A\|_*$ and $\|A\|_1$ in procedure (5). Before defining it, we need further notations and definitions.

For a fixed $A \in \mathbb{R}^{n\times n}$ and $c, \beta > 0$, we introduce the cone

$$\mathcal{C}_2(A,c,\beta) \doteq \Big\{A' \in \mathcal{A} : \|\mathcal{P}_A^{\perp}(A')\|_* + \beta\|\Theta_A^{\perp} \circ A'\|_1 \le c\Big(\|\mathcal{P}_A(A')\|_* + \beta\|\Theta_A \circ A'\|_1\Big)\Big\}.$$

This cone consist of the matrices A' with large entries close to that of A and that are "almost aligned" with the row and column spaces of A. The parameter β quantifies the interplay between these too notions.

Assumption 2 (Restricted Eigenvalue (RE))**.** *For $W \in \mathcal{W}$ and $c > 0$, we have*

$$\mu_1(W,c) = \inf\Big\{\mu > 0 : \|\Theta_W \circ W'\|_F \le \frac{\mu}{\sqrt{T}}\|\mathbf{X}_{T-1}W'\|_F, \ \ \forall W' \in \mathcal{C}_1(W,c)\Big\}.$$

For $A \in \mathcal{A}$ and $c, \beta > 0$, we introduce

$$\mu_2(A,W,c,\beta) = \inf\Big\{\mu > 0 : \ \|\mathcal{P}_A(A')\|_F \vee \|\Theta_A \circ A'\|_F \le \frac{\mu}{\sqrt{d}}\|W'^{\top}\omega(A_T) - \omega(A')\|_2 \\ \forall W' \in \mathcal{C}_1(W,c), \forall A' \in \mathcal{C}_2(A,c,\beta)\Big\}. \quad (8)$$

The RE assumption consists of assuming that the constants μ_1 and μ_2 are finite. Now we can state the following Theorem that gives a fast oracle inequality for our procedure using RE.

Theorem 2. *Under Assumption 2 and Assumption 2, let $(\hat{A}, \hat{W})$ be given by* (5) *and suppose that*

$$\tau \ge 3\alpha\|M\|_{\mathrm{op}}, \quad \gamma \ge 3(1-\alpha)\|M\|_{\infty} \quad \textit{and} \quad \kappa \ge 3\|\Xi\|_{\infty} \quad (9)$$

for some $\alpha \in (0,1)$. Then, we have

$$\mathcal{E}(\widehat{A},\widehat{W})^2 \le \inf_{(A,W)\in\mathcal{A}\times\mathcal{W}}\Big\{\mathcal{E}(A,W)^2 + \frac{25}{18}\mu_2(A,W)^2\big(\tau^2\,\mathrm{rank}(A) + \gamma^2\|A\|_0\big) + \frac{25}{36}\kappa^2\mu_1(W)^2\|W\|_0\Big\},$$

where $\mu_1(W) = \mu_1(W,5)$ and $\mu_2(A,W) = \mu_2(A,W,5,\gamma/\tau)$ (see Assumption 2).

The proofs of Theorems 1 and 2 use tools introduced in [12] and [3].

Note that the residual term from this oracle inequality mixes the notions of sparsity of A and W via the terms $\mathrm{rank}(A)$, $\|A\|_0$ and $\|W\|_0$. It says that our mixed penalization procedure provides an optimal trade-off between fitting the data and complexity, measured by both sparsity and low-rank. This is the first result of this nature to be found in literature.

In the next Theorem 3, we obtain convergence rates for the procedure (5) by combining Theorem 2 with controls on the noise processes. We introduce

$$v_{\Omega,\mathrm{op}}^2 = \Big\|\frac{1}{d}\sum_{j=1}^{d}\Omega_j^{\top}\Omega_j\Big\|_{\mathrm{op}} \vee \Big\|\frac{1}{d}\sum_{j=1}^{d}\Omega_j\Omega_j^{\top}\Big\|_{\mathrm{op}}, \quad v_{\Omega,\infty}^2 = \Big\|\frac{1}{d}\sum_{j=1}^{d}\Omega_j \circ \Omega_j\Big\|_{\infty},$$

$$\sigma_{\omega}^2 = \max_{j=1,\dots,d}\sigma_{\omega,j}^2, \quad \text{where} \quad \sigma_{\omega,j}^2 = \Big(\frac{1}{T}\sum_{t=1}^{T}\omega_j(A_{t-1})^2 + \omega_j(A_T)^2\Big),$$

which are the (observable) variance terms that naturally appear in the controls of the noise processes. We introduce also

$$\ell_T = 2\max_{j=1,\dots,d}\log\log\left(\sigma_{\omega,j}^2 \vee \frac{1}{\sigma_{\omega,j}^2} \vee e\right),$$

which is a small (observable) technical term that comes out of our analysis of the noise process Ξ. This term is a small price to pay for the fact that no independence assumption is required on the noise sequence $(N_t)_{t\ge 0}$, but only a martingale increment structure with sub-gaussian tails.

Theorem 3. *Consider the procedure $(\hat{A}, \hat{W})$ given by* (5) *with smoothing parameters given by*

$$\tau = 3\alpha\sigma v_{\Omega,\mathrm{op}}\sqrt{\frac{2(x+\log(2n))}{d}},$$
$$\gamma = 3(1-\alpha)\sigma v_{\Omega,\infty}\sqrt{\frac{2(x+2\log n)}{d}},$$
$$\kappa = 6\sigma\sigma_{\omega}\left(\sqrt{\frac{2e(x+2\log d+\ell_T)}{T}} + \frac{\sqrt{2e(x+2\log d+\ell_T)}}{d}\right).$$

for some $\alpha \in (0,1)$ and fix a confidence level $x > 0$. Then, we have

$$\mathcal{E}(\widehat{A},\widehat{W})^2 \leq \inf_{(A,W)\in\mathcal{A}\times\mathcal{W}} \Big\{ \mathcal{E}(A,W)^2 + C_1\|W\|_0(x+2\log d+\ell_T)\Big(\frac{1}{T}+\frac{1}{d^2}\Big) + C_2\|A\|_0\frac{2(x+2\log n)}{d} + C_3\,\mathrm{rank}(A)\frac{2(x+\log(2n))}{d}\Big\}$$

where

$$C_1 = 100e\mu_1(W)^2\sigma^2\sigma_\omega^2, \quad C_2 = 25\mu_2(A,W)^2(1-\alpha)^2\sigma^2 v_{\Omega,\infty}^2, \quad C_3 = 25\mu_2(A,W)^2\alpha^2\sigma^2 v_{\Omega,\mathrm{op}}^2,$$

with a probability larger than $1-17e^{-x}$, where μ_1 and μ_2 are the same as in Theorem 2.

The proof of Theorem 3 follows directly from Theorem 2 basic noise control results. In the next Theorem, we propose more explicit upper bounds for both the indivivual estimation of W_0 and the prediction of A_{T+1}.

Theorem 4. *Under the same assumptions as in Theorem 3 and the same choice of smoothing parameters, for any $x > 0$ the following inequalities hold with probability larger than $1-17e^{-x}$:*

- *Feature prediction error:*

$$\frac{1}{T}\|\mathbf{X}_T(\hat{W}-W_0)\|_F^2 \leq \frac{25}{36}\kappa^2\mu_1(W_0)^2\|W_0\|_0 + \inf_{A\in\mathcal{A}}\Big\{\frac{1}{d}\|\omega(A)-\omega(A_{T+1})\|_2^2 + \frac{25}{18}\mu_2(A,W_0)^2\big(\tau^2\,\mathrm{rank}(A)+\gamma^2\|A\|_0\big)\Big\} \quad (10)$$

- *VAR parameter estimation error:*

$$\|\hat{W}-W_0\|_1 \leq 5\kappa\mu_1(W_0)^2\|W_0\|_0 + 6\sqrt{\|W_0\|_0}\mu_1(W_0)\inf_{A\in\mathcal{A}}\sqrt{\frac{1}{d}\|\omega(A)-\omega(A_{T+1})\|_2^2 + \frac{25}{18}\mu_2(A,W_0)^2\big(\tau^2\,\mathrm{rank}(A)+\gamma^2\|A\|_0\big)} \quad (11)$$

- *Link prediction error:*

$$\|\hat{A}-A_{T+1}\|_* \leq 5\kappa\mu_1(W_0)^2\|W_0\|_0+\mu_2(A_{T+1},W_0)(6\sqrt{\mathrm{rank}\,A_{T+1}}+5\frac{\gamma}{\tau}\sqrt{\|A_{T+1}\|_0}) \times \inf_{A\in\mathcal{A}}\sqrt{\frac{1}{d}\|\omega(A)-\omega(A_{T+1})\|_2^2 + \frac{25}{18}\mu_2(A,W_0)^2\big(\tau^2\,\mathrm{rank}(A)+\gamma^2\|A\|_0\big)}\,. \quad (12)$$

4 Algorithms and Numerical Experiments

4.1 Generalized forward-backward algorithm for minimizing $\mathcal{L}$

We use the algorithm designed in [17] for minimizing our objective function. Note that this algorithm is preferable to the method introduced in [18] as it directly minimizes $\mathcal{L}$ jointly in (S,W) rather than alternately minimizing in W and S.

Moreover we use the novel joint penalty from [19] that is more suited for estimating graphs. The proximal operator for the trace norm is given by the shrinkage operation, if $Z = U\,\mathrm{diag}(\sigma_1,\cdots,\sigma_n)V^T$ is the singular value decomposition of Z,

$$\mathrm{prox}_{\tau||.||_*}(Z) = U\,\mathrm{diag}((\sigma_i-\tau)_+)_i V^T.$$

Similarly, the proximal operator for the ℓ_1-norm is the soft thresholding operator defined by using the entry-wise product of matrices denoted by $\circ$:

$$\mathrm{prox}_{\gamma||.||_1}(Z) = \mathrm{sgn}(Z)\circ(|Z|-\gamma)_+\,.$$

The algorithm converges under very mild conditions when the step size θ is smaller than $\frac{2}{L}$, where L is the operator norm of the joint quadratic loss:

$$\Phi : (A,W) \mapsto \frac{1}{T}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \frac{1}{d}\|\omega(A)-W^\top\omega(A_T)\|_F^2\ .$$

Algorithm 1 Generalized Forward-Backward to Minimize $\mathcal{L}$

Initialize A, Z_1, Z_2, W
repeat
  Compute $(G_A, G_W) = \nabla_{A,W}\Phi(A, W)$.
  Compute $Z_1 = \text{prox}_{2\theta\tau||.||_*}(2A - Z_1 - \theta G_A)$
  Compute $Z_2 = \text{prox}_{2\theta\gamma||.||_1}(2A - Z_2 - \theta G_A)$
  Set $A = \frac{1}{2}(Z_1 + Z_2)$
  Set $W = \text{prox}_{\theta\kappa||.||_1}(W - \theta G_W)$
until convergence
return (A, W) minimizing $\mathcal{L}$

4.2 A generative model for graphs having linearly autoregressive features

Let $V_0 \in \mathbb{R}^{n\times r}$ be a sparse matrix, $V_0^\dagger$ its pseudo-inverse such, that $V_0^\dagger V_0 = V_0^\top V_0^{\top\dagger} = I_r$. Fix two sparse matrices $W_0 \in \mathbb{R}^{r\times r}$ and $U_0 \in \mathbb{R}^{n\times r}$. Now define the sequence of matrices $(A_t)_{t\geq 0}$ for $t = 1, 2, \cdots$ by

$$U_t = U_{t-1}W_0 + N_t$$

and

$$A_t = U_t V_0^\top + M_t$$

for i.i.d sparse noise matrices N_t and M_t, which means that for any pair of indices (i, j), with high probability $(N_t)_{i,j} = 0$ and $(M_t)_{i,j} = 0$. We define the linear feature map $\omega(A) = AV_0^{\top\dagger}$, and point out that

1. The sequence $\left(\omega(A_t)^\top\right)_t = \left(U_t + M_t V_0^{\top\dagger}\right)_t$ follows the linear autoregressive relation
$$\omega(A_t)^\top = \omega(A_{t-1})^\top W_0 + N_t + M_t V_0^{\top\dagger} \ .$$
2. For any time index t, the matrix A_t is close to $U_t V_0$ that has rank at most r
3. The matrices A_t and U_t are both sparse by construction.

4.3 Empirical evaluation

We tested the presented methods on synthetic data generated as in section (4.2). In our experiments the noise matrices M_t and N_t where built by soft-thresholding *i.i.d.* noise $\mathcal{N}(0, \sigma^2)$. We took as input $T = 10$ successive graph snapshots on $n = 50$ nodes graphs of rank $r = 5$. We used $d = 10$ linear features, and finally the noise level was set to $\sigma = .5$. We compare our methods to standard baselines in link prediction. We use the area under the ROC curve as the measure of performance and report empirical results averaged over 50 runs with the corresponding confidence intervals in figure 4.3. The competitor methods are the *nearest neighbors* (NN) and static sparse and low-rank estimation, that is the link prediction algorithm suggested in [19]. The algorithm NN scores pairs of nodes with the number of common friends between them, which is given by A^2 when A is the cumulative graph adjacency matrix $\widetilde{A_T} = \sum_{t=0}^{T} A_t$ and the static sparse and low-rank estimation is obtained by minimizing the objective $\|X - \widetilde{A_T}\|_F^2 + \tau\|X\|_* + \gamma\|X\|_1$, and can be seen as the closest *static* version of our method. The two methods *autoregressive low-rank* and *static low-rank* are regularized using only the trace-norm, (corresponding to forcing $\gamma = 0$) and are slightly inferior to their sparse and low-rank rivals. Since the matrix V_0 defining the linear map ω is unknown we consider the feature map $\omega(A) = AV$ where $\widetilde{A_T} = U\Sigma V^\top$ is the SVD of $\widetilde{A_T}$. The parameters τ and γ are chosen by 10-fold cross validation for each of the methods separately.

4.4 Discussion

1. *Comparison with the baselines.* This experiment sharply shows the benefit of using a temporal approach when one can handle the feature extraction task. The left-hand plot shows that if few snapshots are available ($T \leq 4$ in these experiments), then static approaches are

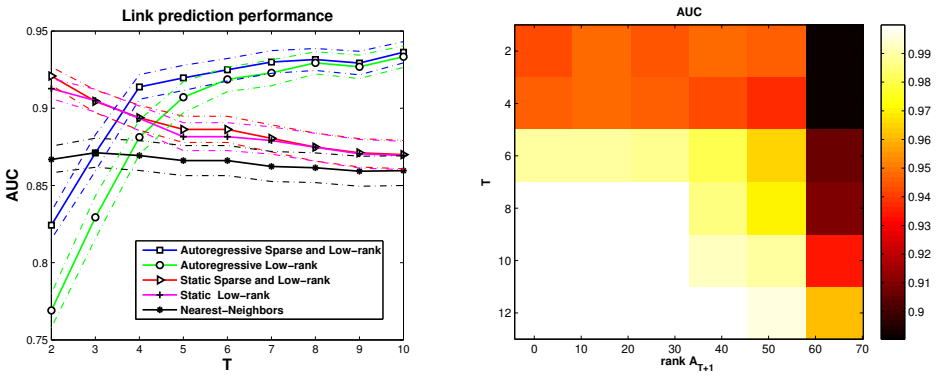

Figure 1: Left: performance of algorithms in terms of Area Under the ROC Curve, average and confidence intervals over 50 runs. Right: Phase transition diagram.

to be preferred, whereas feature autoregressive approaches outperform as soon as *sufficient number* T graph snapshots are available (see phase transition). The decreasing performance of static algorithms can be explained by the fact that they use as input a mixture of graphs observed at different time steps. Knowing that at each time step the nodes have specific latent factors, despite the slow evolution of the factors, adding the resulting graphs leads to confuse the factors.

2. *Phase transition.* The right-hand figure is a phase transition diagram showing in which part of rank and time domain the estimation is accurate and illustrates the interplay between these two domain parameters.

3. *Choice of the feature map* ω. In the current work we used the projection onto the vector space of the top-r singular vectors of the cumulative adjacency matrix as the linear map ω, and this choice has shown empirical superiority to other choices. The question of choosing the best measurement to summarize graph information as in compress sensing seems to have both theoretical and application potential. Moreover, a deeper understanding of the connections of our problem with compressed sensing, for the construction and theoretical validation of the features mapping, is an important point that needs several developments. One possible approach is based on multi-kernel learning, that should be considered in a future work.

4. *Generalization of the method.* In this paper we consider only an autoregressive process of order 1. For better prediction accuracy, one could consider mode general models, such as vector ARMA models, and use model-selection techniques for the choice of the orders of the model. A general modelling based on state-space model could be developed as well. We presented a procedure for predicting graphs having linear autoregressive features. Our approach can easily be generalized to non-linear prediction through kernel-based methods.

References

[1] J. Abernethy, F. Bach, Th. Evgeniou, and J.-Ph. Vert. A new approach to collaborative filtering: operator estimation with spectral regularization. *JMLR*, 10:803–826, 2009.

[2] A. Argyriou, M. Pontil, Ch. Micchelli, and Y. Ying. A spectral regularization framework for multi-task structure learning. *Proceedings of Neural Information Processing Systems (NIPS)*, 2007.

[3] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37, 2009.

[4] L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society (JRSS): Series B (Statistical Methodology)*, 59:3–54, 1997.

[5] E.J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5), 2009.

[6] Candès E. and Tao T. Decoding by linear programming. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2005.

[7] Th. Evgeniou, Ch. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[8] S. Gaiffas and G. Lecue. Sharp oracle inequalities for high-dimensional matrix prediction. *Information Theory, IEEE Transactions on*, 57(10):6942 –6957, oct. 2011.

[9] M. Kolar and E. P. Xing. On time varying undirected graphs. *in Proceedings of the 14th International Conference on Artifical Intelligence and Statistics AISTATS*, 2011.

[10] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[11] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.

[12] V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy matrix completion. *Annals of Statistics*, 2011.

[13] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[14] Y. Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

[15] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[16] S.A. Myers and Jure Leskovec. On the convexity of latent social network inference. In *NIPS*, 2010.

[17] H. Raguet, J. Fadili, and G. Peyré. Generalized forward-backward splitting. *Arxiv preprint arXiv:1108.4404*, 2011.

[18] E. Richard, N. Baskiotis, Th. Evgeniou, and N. Vayatis. Link discovery using graph feature tracking. *Proceedings of Neural Information Processing Systems (NIPS)*, 2010.

[19] E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low-rank matrices. In *Proceeding of 29th Annual International Conference on Machine Learning*, 2012.

[20] P. Sarkar, D. Chakrabarti, and A.W. Moore. Theoretical justification of popular link prediction heuristics. In *International Conference on Learning Theory (COLT)*, pages 295–307, 2010.

[21] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *in Proceedings of Neural Information Processing Systems 17*, pages 1329–1336. MIT Press, Cambridge, MA, 2005.

[22] B. Taskar, M.F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Neural Information Processing Systems*, volume 15, 2003.

[23] R. S. Tsay. *Analysis of Financial Time Series*. Wiley-Interscience; 3rd edition, 2005.

[24] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

[25] D.Q. Vu, A. Asuncion, D. Hunter, and P. Smyth. Continuous-time regression models for longitudinal networks. In *Advances in Neural Information Processing Systems*. MIT Press, 2011.