Journal of Machine Learning Research 15 (2014) 565-593 Submitted 11/12; Revised 10/13; Published 2/14

# Link Prediction in Graphs with Autoregressive Features

**Emile Richard** EMILERIC@STANFORD.EDU
*Department of Electrical Engineering*
*Stanford University - Packard 239*
*Stanford, CA 94304*

**Stéphane Gaïffas** STEPHANE.GAIFFAS@CMAP.POLYTECHNIQUE.FR
*CMAP - Ecole Polytechnique*
*Route de Saclay*
*91128 Palaiseau Cedex, France*

**Nicolas Vayatis** NICOLAS.VAYATIS@CMLA.ENS-CACHAN.FR
*CMLA - ENS Cachan*
*UMR CNRS No. 8536*
*61, avenue du Président Wilson*
*94 235 Cachan cedex, France*

**Editor:** Tong Zhang

## 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. On the adjacency matrix it takes into account both sparsity and low rank properties and on the VAR it encodes the sparsity. The analysis involves oracle inequalities that illustrate the trade-offs in the choice of smoothing parameters when modeling the joint effect of sparsity and low rank. The estimate is computed efficiently using proximal methods, and evaluated through numerical experiments.

**Keywords:** graphs, link prediction, low-rank, sparsity, autoregression

## 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 by Breiman and Friedman (1997) or with autoregressive models by Tsay (2005). More recently, kernel-based regularized methods have been developed for multitask learning by Evgeniou et al. (2005) and Andreas et al. (2007). 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, see the paper by Liben-Nowell and Kleinberg (2007) and Kolar and Xing (2011). 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, see for instance the papers by Srebro et al. (2005), Candès and Tao (2009) and Abernethy et al. (2009). In the realistic setup of a time-evolving graph, matrix completion was also used and adapted by Richard et al. (2010) to take into account the dynamics of the features of the graph. The estimation of a VAR model for node degrees (that are linear graph features) has been considered by Zhang et al. (2011) and successfully applied to customer valuation, and to measure network effect in user generated content market places. Note also that sparse autoregressive models are also considered by Davis et al. (2012) and Nardi and Rinaldo (2011).

In this paper, we study the prediction problem where the observation is a sequence of graphs represented through their adjacency matrices $(A_t)_{0\leq t\leq T}$ and the goal is to predict $A_{T+1}$. This prediction problem arises in recommender systems, where the purchases or preference declarations are registered over time. 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. In this setting, fitting predictive models is a natural way for estimating evolving networks in these contexts, see the paper by Shojaie et al. (2011). A large variety of methods for link prediction only consider prediction from a single instantaneous snapshot of the graph. This includes heuristics: measures of node neighbourhoods are considered by Liben-Nowell and Kleinberg (2007), Lü and Zhou (2011) and Sarkar et al. (2010), matrix factorization by Koren (2008), diffusion by see Myers and Leskovec (2010) and probabilistic methods by Taskar et al. (2003). More recently, some works have investigated the use of sequences of observations of the graph to improve the prediction, such as regression on features extracted from the graphs by Richard et al. (2010), matrix factorization by Koren (2010), continuous-time regression by Vu et al. (2011) or non-parametric models by Sarkar et al. (2012). An hybrid approach to dynamic link prediction is considered by Huang and Lin (2009), based on a mixture of the static approach by Liben-Nowell and Kleinberg (2007) and an individual ARIMA modeling of the links evolution.

The framework of the current paper is somehow related to compressed sensing introduced by Donoho (2006) and Candès and Wakin (2008). In fact, due to stationarity assumptions, the amount of available information is very small compared to the task of predicting the quadratically many potential edges of the graph. Therefore penalization terms that encourage both sparsity and low-rank of related matrices are used to recover the edges of the graph. In the static setup, these two effects have been previously combined by Richard et al. (2012b) for the estimation of sparse and low-rank matrices, the rationale being that graphs containing cliques have block-diagonal adjacency matrices that are simultaneously sparse and low-rank. Key elements in deriving theoretical results are tools from the theory of compressed sensing, developed by Candès and Tao (2005), Bickel et al. (2009), Koltchinskii et al. (2011) and in particular the Restricted Eigenvalue of Koltchinskii (2009a), Koltchinskii (2009b) and Bickel et al. (2009). Our main results are oracle inequalities under the general

assumption that the innovation process of the VAR is a martingale increment sequence with sub-gaussian tails. These oracle inequalities prove that our procedure achieves a trade-off in the calibration of smoothing parameters that balances the sparsity and the rank of the adjacency matrix. A preliminary version of this work can be found in a previous work by Richard et al. (2012a).

The rest of this paper is organized as follows. In Section 2, we describe the general setup of this study with the main assumptions. In Section 2.3, we formulate a regularized optimization problem which aims at jointly estimating the autoregression parameters and predicting the graph. In Section 3, we provide theoretical guarantees for the joint estimation-prediction by providing oracle inequalities. In Section 4 we provide an efficient algorithm for solving the optimization problem and show empirical results that illustrate our approach. The proofs are provided in Appendix.

## 2. Modeling Low-Rank Graphs Dynamics with Autoregressive Features

We first introduce the main notations used in the paper.
*Matrix norms and entrywise matrix operations.* Denote by $A$ a matrix. In the sequel, the notations $\|A\|_F$, $\|A\|_p$, $\|A\|_\infty$, $\|A\|_*$ and $\|A\|_{\text{op}}$ stand, respectively, for the Frobenius norm of $A$, the entry-wise $\ell_p$ norm, the entry-wise $\ell_\infty$ norm, the trace-norm (or nuclear norm, given by the sum of the singular values) and operator norm (the largest singular value) of $A$. Given matrices $A$ and $B$, we denote by $\langle A, B\rangle = \text{tr}(A^\top B)$ 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 entry-wise product between $A$ and $B$ (also called Hadamard product). The matrix $|A|$ contains the absolute values of entries of $A$. The matrix $(M)_+$ is the entry-wise positive part of the matrix M, and $\text{sign}(M)$ is the sign matrix associated to $M$ with the convention $\text{sign}(0) = 0$.
*SVD and projections.* If $A$ is a $n \times n$ matrix with rank $r$, we write its Singular Value Decomposition (SVD) as $A = U\Sigma V^\top = \sum_{j=1}^r \sigma_j u_j v_j^\top$ where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$ is a $r \times r$ diagonal matrix containing the non-zero singular values of $A$ in decreasing order, and $U = [u_1, \ldots, u_r]$, $V = [v_1, \ldots, v_r]$ 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 Working Assumptions

Our approach is based on a number of beliefs which we translate into mathematical assumptions:

- Low-rank of adjacency matrices $A_t$

  This reflects the presence of highly connected groups of nodes such as communities in social networks, or product categories and groups of loyal/fanatic users in a market place data, and is sometimes motivated by the small number of factors that explain

nodes interactions. We will not make an explicit assumption in the paper but the results we obtain will be meaningful in the specific case where rank is small compared to the dimension.

- Autoregressive linear features (VAR models)

  We assume that intrinsic features of the graph can explain most of the information contained in the graph, and that these features are evolving with time. Our approach considers the simplest assumption on the dynamics over time of these features and we assume a Vector Autoregressive Linear Regression model that is described in the next subsection.

- Sub-gaussian noise process

  A probabilistic framework is considered in order to describe performance under the form of oracle inequalities and we propose to specify the distribution of the discrepancy between the VAR model and the actual observations with a a sub-gaussian tail behavior. This assumption will be formulated below in Section 3.

The first two items correspond to modeling assumptions which partly capture observations made on real-life data. The third item is a technical assumption used in the proofs.

### 2.2 An Autoregressive Linear Model for Graph Features

*Feature map.* We consider a list of graph features encoded through a linear map of the adjacency matrix with $\omega : \mathbb{R}^{n\times n} \to \mathbb{R}^d$ defined by:

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

where $\{\Omega_i\}_{1\le i\le d}$ is a set of $n \times n$ matrices. These matrices could be either deterministic or random in our theoretical analysis, but we take them deterministic for the sake of simplicity. An example of linear features is the vector of node degrees, that is, the number of edges connected to each node. The degree can be computed from the adjacency matrix using the linear function $\omega : A \mapsto A\mathbf{1}$ or $\omega : A \mapsto A^\top\mathbf{1}$ respectively for the right and left nodes degrees, where $\mathbf{1}$ denotes the vector with all coordinates equal to 1 of the appropriate length. Other (linear) measures of popularity are considered in social and e-commerce networks, such as the sum of the weights of incident edges if there is some graduation in the strength of connection between nodes. Note that nonlinear features, such as the count of the number of cycles of length $k$ ($k = 3, 4, \cdots$) through each node, may be relevant in real world applications. Such features involve, for instance, the diagonal elements of $A^k$. An extensive study of this very interesting case is beyond the scope of the present paper.
*VAR model.* We consider a linear model for the evolution of $\omega(A)$ over time.

**Assumption 1** *The vector time series $\{\omega(A_t)\}_{t\ge 0}$ has autoregressive dynamics, given by a VAR (Vector Auto-Regressive) model:*

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

*where $W_0 \in \mathbb{R}^{d\times d}$ is an unknown sparse matrix and $\{N_t\}_{t\ge 0}$ is a sequence of noise vectors in $\mathbb{R}^d$.*

In the sequel, we shall use the following compact notations:

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

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

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

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

### 2.3 Simultaneous Prediction and Estimation through Regularized Optimization

*Optimization problem formulation.* We now introduce the optimization problem which will account for both the prediction task (anticipate the appearance of new edges in the graph) and the modeling choices which are supposed to reflect phenomena observed on real data (smooth evolution of graph features). We consider that snapshots of the graph (and therefore also the corresponding features) are available at times $1, \ldots, T$ and we want to predict links which will appear at the next instant $T+1$. In order to fulfill this double objective, we combine two regularized problems in an additive fashion based on two terms:

1. First objective - data-fitting term for weight vector $W$ with sparsity-enforcing penalty

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

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

2. Second objective - data-fitting term for the features of the adjacency matrix $A$ with mixed penalty enforcing both sparsity and low-rank

$$J_2(A, W) = \frac{1}{d}\|\omega(A) - W^\top \omega(A_T)\|_2^2 + \tau\|A\|_* + \gamma\|A\|_1,$$

   where $\tau, \gamma > 0$ are smoothing parameters.

The resulting penalized criterion will be the main topic of the present paper. It is the sum of the two partial objectives $J_1$ and $J_2$, and is jointly convex with respect to $A$ and $W$:

$$\mathcal{L}(A, W) \doteq \frac{1}{T}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \kappa\|W\|_1 + \frac{1}{d}\|\omega(A) - W^\top\omega(A_T)\|_2^2 + \tau\|A\|_* + \gamma\|A\|_1. \tag{3}$$

*Rationale.* As shown by the introduction of the two functionals, our approach pursues a double goal. On the one hand, the data-fitting term on $W$ in $J_1$ aims at an estimate on the past data of the weight factor in the autoregressive modeling setup according to Assumption 1 and under a sparsity constraint. On the other hand, the link prediction goes through the estimation of a matrix $A = A_{T+1}$ which should be sparse and low-rank simultaneously. Hence, the second functional $J_2$ involves a mixed penalty of the form $A \mapsto \tau\|A\|_* + \gamma\|A\|_1$, with $\tau$, $\gamma$ smoothing parameters. Such a combination of $\ell_1$ and trace-norm was already studied by Gaïffas and Lecué (2011) for the matrix regression model, and by Richard et al. (2012b) for the prediction of an adjacency matrix. This mixed norm combines the benefits of each of the two norms and is well suited for estimating simultaneously sparse

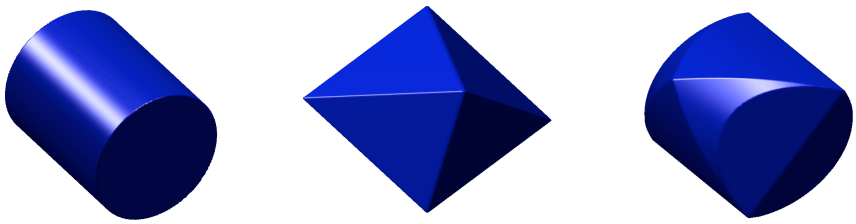

Figure 1: Unit balls for the trace norm (left), $\ell_1$ (middle) and the mixed $X \mapsto \|X\|_* + \|X\|_1$ norm (right). The norms where computed on the set of $2 \times 2$ symmetric matrices that can be identified to $\mathbb{R}^3$.

and low-rank matrices. In Figure 1 we illustrated the unit balls for the three norms $\ell_1$, trace-norm and the $\ell_1$ + trace norm. The key observation is that the ball of the mixed norm has singularities at the points where each of the two other balls are singular, but the singularities get sharper at points where both norms are singular, namely on the matrices that are sparse and low-rank at the same time.

The set of sparse and low-rank matrices obtained by minimizing an objective including this mixed norm contains matrices that can be written in a block-diagonal or overlapping block-diagonal form, up to permutations of rows and columns. These matrices can be interpreted as adjacency matrices of networks containing highly connected groups of nodes and therefore are of particular interest for prediction and denoising applications in graph data and in covariance matrix estimation. Here we extend the approach developed by Richard et al. (2012b) for the time-dependent setting by considering data-fitting measures which ensure that the features of the next graph $\omega(A_{T+1})$ are close to $W^\top \omega(A_T)$.

*Search space and general scheme of the estimation procedure.* We shall consider the case where the optimization domain consists of the cartesian product of convex cones $\mathcal{A}$ and $\mathcal{W}$ such that $\mathcal{A} \subset \mathbb{R}^{n \times n}$ and $\mathcal{W} \subset \mathbb{R}^{d \times d}$. The joint estimation-prediction procedure is then defined by

$$(\hat{A}, \hat{W}) \in \operatorname*{arg\,min}_{(A,W) \in \mathcal{A} \times \mathcal{W}} \mathcal{L}(A, W). \tag{4}$$

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 true VAR model matrix $W_0$, while the entries of the matrix $A_{T+1}$ must be positive. Table 1 summarizes the methodology in a scheme where the symbols $\downarrow_\omega$ represent the feature extraction procedure through the map $\omega : \mathbb{R}^{n \times n} \to \mathbb{R}^d$. The prediction in the feature space is represented by $\to_W$, and is handled in practice by the least squares regression on $W$. Finally, the symbol $\uparrow$ that maps the predicted feature vector $\widehat{\omega(A_{T+1})}$ to $\widehat{A_{T+1}}$ represents the inverse problem that is solved through the regression penalized by the mixed penalization.

### 2.4 An Overview of Main Results

The central contribution of our work is to provide bounds on the prediction error under a Restricted Eigenvalue (RE) assumption on the feature map. 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

| | | | | | | |
|---|---|---|---|---|---|---|
| $A_0$ | $A_1$ | $\cdots$ | $A_T$ | | $\widehat{A_{T+1}}$ | Observed adjacency matrices $\in \mathbb{R}^{n\times n}$ |
| $\downarrow\omega$ | $\downarrow\omega$ | | $\downarrow\omega$ | | $\uparrow$ | |
| $\omega(A_0)$ | $\omega(A_1)$ | $\cdots$ | $\omega(A_T)$ | $\underset{W}{\rightarrow}$ | $\widehat{\omega(A_{T+1})}$ | Features vectors $\in \mathbb{R}^d$ |

Table 1: General scheme of our method for prediction in dynamic graph sequences through a feature map $\omega$.

rank of the true adjacency matrix $A_{T+1}$, and (c) the sparsity of the true VAR model matrix $W_0$.

Namely, we prove 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

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

We point out 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 $\|\omega(\widehat{A}-A_{T+1})\|_2$.

We obtain upper bounds that are reminiscent of the bounds obtained for the Lasso by Bickel et al. (2009) for instance, and that are of the following order:

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

This upper bound, formalized in Theorem 3, exhibits 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. It indicates, in particular, that an optimal choice for the number $d$ of features is of order $T\log n$. The positive constants $C_1, C_2, C_3$ are proportional to the noise level $\sigma$. The interplay between the rank and the 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. The precise formulation of these results is given in the next section.

## 3. Oracle Inequalities

This section contains the main theoretical results ot the paper. Complete proofs and technical details are provided in the Appendix section at the end of the paper.

### 3.1 A General Oracle Inequality

We recall from subsection 2.2 that the noise sequence in the VAR model is denoted by $\{N_t\}_{t\geq 0}$. We now introduce the noise processes as

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

which are, respectively, $n \times n$ and $d \times d$ random matrices. The source of randomness comes from the noise sequence $\{N_t\}_{t \geq 0}$.

Now, if these noise processes are controlled, we can prove oracle inequalities for procedure (4). The first result is an oracle inequality of slow type, that holds in full generality.

**Theorem 1** *Under Assumption 1, let $(\hat{A}, \hat{W})$ be given by (4) and suppose that*

$$\tau \geq 2\alpha \|M\|_{\mathrm{op}}, \quad \gamma \geq 2(1-\alpha)\|M\|_{\infty} \quad \text{and} \quad \kappa \geq 2\|\Xi\|_{\infty} \tag{5}$$

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

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

### 3.2 Restricted Eigenvalue Condition and Fast Oracle Inequalities

For the proof of oracle inequalities, the *restricted eigenvalue* (RE) condition introduced by Bickel et al. (2009) and Koltchinskii (2009a,b) is of importance. As explained by van de Geer and Bühlmann (2009), this condition is acknowledged to be one of the weakest to derive fast rates for the Lasso. Matrix version of these assumptions are introduced by Koltchinskii et al. (2011). 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 (4). If $W \in \mathbb{R}^{d \times d}$, we denote by $\Theta_W = \mathrm{sign}(W) \in \{0, \pm 1\}^{d \times d}$ the signed sparsity pattern of $W$ and by $\Theta_W^{\perp} \in \{0,1\}^{d \times d}$ the complementary 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 the mixture of the terms $\|A\|_*$ and $\|A\|_1$ in procedure (4). For a fixed $A \in \mathbb{R}^{n \times n}$ and $c, \beta > 0$, we introduce

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

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 $\beta$ quantifies the interplay between these two notions.

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

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

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

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

Under this assumption, we can obtain refined oracle inequalities as shown in the next theorem.

**Theorem 2** *Under Assumption 1 and Assumption 2, let* $(\hat{A}, \hat{W})$ *be given by* (4) *and suppose that*

$$\tau \geq 3\alpha\|M\|_{\mathrm{op}}, \quad \gamma \geq 3(1-\alpha)\|M\|_{\infty} \quad \text{and} \quad \kappa \geq 3\|\Xi\|_{\infty} \tag{6}$$

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

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

*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 by Koltchinskii et al. (2011) and Bickel et al. (2009). Note that the residual term from this oracle inequality combines the sparsity of $A$ and $W$ via the terms $\operatorname{rank} A$, $\|A\|_0$ 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. To our knowledge, this is the first result of this nature to be found in literature.

### 3.3 Probabilistic Versions

We introduce the following natural hypothesis on the noise process.

**Assumption 3** *We assume that* $\{N_t\}_{t\geq 0}$ *satisfies* $\mathbb{E}[N_t|\mathcal{F}_{t-1}] = 0$ *for any* $t \geq 1$ *and that there is* $\sigma > 0$ *such that for any* $\lambda \in \mathbb{R}$ *and* $j = 1, \ldots, d$ *and* $t \geq 0$*:*

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

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

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

In the next result (Theorem 3), we obtain convergence rates for the procedure (4) by combining Theorem 2 with controls on the noise processes. We introduce the following quantities:

$$v^2_{\Omega,\mathrm{op}} = \Big\|\frac{1}{d}\sum_{j=1}^{d}\Omega_j^\top\Omega_j\Big\|_{\mathrm{op}} \vee \Big\|\frac{1}{d}\sum_{j=1}^{d}\Omega_j\Omega_j^\top\Big\|_{\mathrm{op}}, \quad v^2_{\Omega,\infty} = \Big\|\frac{1}{d}\sum_{j=1}^{d}\Omega_j\circ\Omega_j\Big\|_{\infty}, \tag{7}$$

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

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

$$\ell_T = 2\max_{j=1,\ldots,d}\log\log\left(\sigma^2_{\omega,j} \vee \frac{1}{\sigma^2_{\omega,j}} \vee e\right), \tag{8}$$

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

We consider the following calibration of smoothing parameters as a function of noise process parameters:

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

In the next Theorem 3 and Corollary 4, we fix the smoothing parameters to the latter values. These two results convey the main message of the paper as it was announced in Section 2.4.

**Theorem 3** *Under Assumption 1, Assumption 2 and Assumption 3, consider the procedure $(\hat{A},\hat{W})$ given by (4) applied with the calibration of smoothing parameters shown above for some $\alpha\in(0,1)$ and a fixed confidence level $x>0$. Then, we have, with probability larger than $1-17e^{-x}$:*

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

*where*

$$C_1 = 100e\mu_1(W)^2\sigma^2\sigma_\omega^2,\quad C_2 = 50\mu_2(A,W)^2(1-\alpha)^2\sigma^2 v_{\Omega,\infty}^2,\quad C_3 = 50\mu_2(A,W)^2\alpha^2\sigma^2 v_{\Omega,\mathrm{op}}^2,$$

*and RE constants $\mu_1(W)$ and $\mu_2(A,W)$ are taken as in Theorem 2.*

The proof of Theorem 3 follows directly from Theorem 2 together with noise control assumptions. In the next result, we propose more explicit upper bounds for both the individual estimation of $W_0$ and the prediction of $A_{T+1}$.

**Corollary 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:*

$$\begin{aligned}
\frac{1}{T}\|\mathbf{X}_T(\hat{W}-W_0)\|_F^2 \leq{}& \frac{25}{36}\kappa^2\mu_1(W_0)^2\|W_0\|_0 \\
&+ \inf_{A\in\mathcal{A}}\Big\{\frac{1}{d}\|\omega(A)-\omega(A_{T+1})\|_2^2 + \frac{25}{18}\mu_2(A,W_0)^2\big(\tau^2\,\mathrm{rank}\,A+\gamma^2\|A\|_0\big)\Big\} \quad (9)
\end{aligned}$$

- *VAR parameter estimation error:*

$$\|\hat{W} - W_0\|_1 \leq 5\kappa\mu_1(W_0)^2\|W_0\|_0 \\ +6\sqrt{\|W_0\|_0}\mu_1(W_0)\inf_{A\in\mathcal{A}}\sqrt{\frac{1}{d}\|\omega(A)-\omega(A_{T+1})\|_2^2+\frac{25}{18}\mu_2(A,W_0)^2\big(\tau^2\operatorname{rank}A+\gamma^2\|A\|_0\big)} \tag{10}$$

- *Link prediction error:*

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

## 4. Algorithms and Data Modeling

In this section, we explore how the proposed strategy of regularized optimization for simultaneously estimating the feature dynamics and predicting the forthcoming links can be implemented in practice.

### 4.1 Incremental Proximal-Gradient Algorithm for Minimizing $\mathcal{L}$

The objective to be minimized in our problem can be written as:

$$\mathcal{L} = \ell + \mathcal{R}\ ,$$

where we have set the loss function $\ell$:

$$\ell : (A, W) \mapsto \frac{1}{T}\|\mathbf{X}_T - \mathbf{X}_{T-1}W\|_F^2 + \frac{1}{d}\|\omega(A) - W^\top\omega(A_T)\|_2^2\ ,$$

and the regularizer $\mathcal{R}$:

$$\mathcal{R} : (A, W) \mapsto \kappa\ \|W\|_1 + \tau\|A\|_* + \gamma\|A\|_1\ .$$

We propose to develop an algorithm for solving this optimization problem based on proximal gradient methods. Proximal algorithms (Beck and Teboulle, 2009; Combettes and Pesquet, 2011) have been designed for solving convex optimization problems where functionals have the following structure : $\mathcal{L} = \ell + \mathcal{R}$ , where $\ell$ is convex, differentiable with a Lipschitz gradient and $\mathcal{R}$ is convex and not differentiable. This is exactly our case. In the classical setup, it is assumed that $\mathcal{R}$ has an explicit (or fast to compute) proximal operator, defined by:

$$\operatorname{prox}_{\mathcal{R}}(X) = \arg\min_Y\left\{\frac{1}{2}\|X-Y\|_F^2+\mathcal{R}(Y)\right\}\ .$$

It has been proved by Beck and Teboulle (2009) that the sequence

$$X_{k+1} = \operatorname{prox}_{\theta\mathcal{R}}(X_k - \theta\nabla\ell(X_k))$$

converges after $O(1/\epsilon)$ steps to a ball of radius $\epsilon$ of the minimizer of $\mathcal{L}$. The step size $\theta$ is usually taken of the order of magnitude of the inverse of the Lipschitz constant $L$ of $\nabla\ell$. An accelerated algorithm (FISTA) that reaches the optimal convergence rate $O(1/\sqrt{\epsilon})$ in the sense of Nesterov (2005) can be written using an auxiliary sequence, are described by Beck and Teboulle (2009) and Tseng (2008). The intuition behind the design of these algorithms relies on the linear expansion of $\ell$ around the point $X_k$ and the quadratic term $\frac{L}{2}\|X - X_k\|_F^2$ that controls the closeness of the next step point $X_{k+1}$ from $X_k$. Namely, we can write

$$\begin{aligned}\mathcal{L}(X) &\approx \ell(X_k) + \nabla\ell(X_k)^\top(X - X_k) + \mathcal{R}(X) + \frac{L}{2}\|X - X_k\|_F^2 \\ &= L\left\{\frac{1}{2}\left\|(X - X_k) + \frac{1}{L}\nabla\ell(X_k)\right\|_F^2 - \frac{1}{2L^2}\|\nabla\ell(X_k)\|_F^2 + \frac{1}{L}\ell(X_k) + \frac{1}{L}\mathcal{R}(X)\right\} \\ &= L\left\{\frac{1}{2}\left\|X - (X_k - \frac{1}{L}\nabla\ell(X_k))\right\|_F^2 + \frac{1}{L}\mathcal{R}(X)\right\} + \text{constant}.\end{aligned}$$

It follows that the point $X_{k+1} = \mathrm{prox}_{\frac{1}{L}\mathcal{R}}(X_k - \frac{1}{L}\nabla\ell(X_k))$ is a fair approximation of the minimizer of $\mathcal{L}$ around $X_k$. The detailed analysis and extensions can be found in the paper by Tseng (2008).

In our case, the presence of the sum of two simple regularizers ($\ell_1$ and trace norm) applied to the same object $A$ makes the situation slightly more complicated, since the proximal operator of this sum is non-explicit. We propose to use an incremental algorithm to address this complication. Indeed, the proximal operators of each term are available. First, it is known that the proximal operator of the trace norm is given by the spectral shrinkage operator: if $X = U\,\mathrm{diag}(\sigma_1, \cdots, \sigma_n)V^\top$ is the singular value decomposition of $X$, we have

$$\mathrm{prox}_{\tau\|\cdot\|_*}(X) = U\,\mathrm{diag}((\sigma_i - \tau)_+)V^\top.$$

For the $\ell_1$-norm, the proximal operator is the entrywise soft-thresholding defined by

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

where we recall that $\circ$ denotes the entry-wise product. The algorithm converges under very mild conditions when the step size $\theta$ is smaller than $2/L$, where $L$ is the operator norm of the joint quadratic loss.

The algorithm is described below (see Algorithm 1). It is inspired from the method proposed by Bertsekas (2011) Section 2 and conducts to the minimization our objective function. The order in which proximal mappings are performed is chosen in order to compute the SVD on a sparse matrix $Z$, for computational efficiency. If a sparse output is desired, an extra soft-thresholding step can be performed at the end. Note that this algorithm is preferable to the method previously introduced by Richard et al. (2010) as it directly minimizes $\mathcal{L}$ jointly in $(A, W)$ rather than alternately minimizing in $W$ and $A$.

### 4.2 A Generative Model for Graphs with Linearly Autoregressive Features

In order to prepare the setup for empirical evaluation of the algorithm, we now explain how synthetic data can be generated from the statistical model with linear autoregressive features. Let $V_0 \in \mathbb{R}^{n\times r}$ be a sparse matrix, $V_0^\dagger$ its pseudo-inverse such that $V_0^\dagger V_0 = V_0^\top V_0^{\top\dagger} = I_r$.

**Algorithm 1** Incremental Proximal-Gradient to Minimize $\mathcal{L}$

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

Fix two sparse matrices $K_0 \in \mathbb{R}^{r\times r}$ 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

$$U_t = U_{t-1}K_0 + N_t$$

and

$$A_t = U_t V_0^\top$$

for a sequence of i.i.d sparse noise matrices $\{N_t\}_{t\geq 0}$, which means that for any pair of indices $(i, j)$, we have $(N_t)_{i,j} = 0$ with a high probability. We consider the vectorization operator $A \mapsto \text{vec}(A)$ that stacks the columns of $A$ into a single column, and define the linear feature map

$$\omega(A) \doteq \text{vec}(A\Psi),$$

where we set for short $\Psi = (V_0^\top)^\dagger$, so that $V_0^\top \Psi = I_r$. Let us notice that

1. The sequence $\{\omega(A_t)\}_t = \{\text{vec}(U_t)\}_t$ follows the linear autoregressive relation

   $$\omega(A_t) = (K_0^\top \otimes I_n)\omega(A_{t-1}) + \text{vec}(N_t),$$

   where $\text{vec}(N_t)$ is a zero-mean noise process and $\otimes$ is the Kronecker product.

2. For any time index $t$, the matrix $A_t$ is close to $U_t V_0^\top$ that has rank at most $r$

3. The matrices $A_t$ and $U_t$ are both sparse by construction.

4. The dimension of the feature space is $d = nr \ll n^2$, so $W_0 = K_0^\top \otimes I_n \in \mathbb{R}^{nr\times nr}$. The feature map can be written in standard form, see Equation (1), after vectorization by using the design matrices

   $$\Omega_{(l-1)n+i} = e_i(\Psi^\top)_{l,\cdot}$$

   for $1 \leq l \leq r, 1 \leq i \leq n$, where the $n \times n$ design matrix $e_i(\Psi^\top)_{l,\cdot}$ contains a copy of the $l$-th column of $\Psi$ at its $i$-th row and zeros elsewhere. The standard form of the feature map is then given by the vector

   $$\omega(A) = [\langle A, \Omega_{(l-1)n+i}\rangle : 1 \leq l \leq r, 1 \leq i \leq n]^\top.$$

   As a consequence, we can compute the variance terms $v_{\Omega,\infty}$ and $v_{\Omega,\text{op}}$ from Equation (7) as functions of $\Psi$. By using

   $$e_i(\Psi^\top)_{l,\cdot}\Psi_{\cdot,l}e_i^\top = \|\Psi_{\cdot,l}\|_2^2 e_i e_i^\top \quad \text{and} \quad \Psi_{\cdot,l}e_i^\top e_i(\Psi^\top)_{l,\cdot} = \Psi_{\cdot,l}(\Psi^\top)_{l,\cdot},$$

we get respectively by summation over indices $i$ and $l$,

$$\sum_{l=1}^{r}\sum_{i=1}^{n} e_i \Psi_{l,\cdot}^\top \Psi_{l,\cdot} e_i^\top = \Big(\sum_{l=1}^{r} \|\Psi_{l,\cdot}^\top\|_2^2\Big)\Big(\sum_{i=1}^{n} e_i e_i^\top\Big) = \|\Psi\|_F^2 I_n$$

and Equation (7) gives us the values of the variance terms

$$v_{\Omega,\mathrm{op}} = \frac{1}{nr}\Big(\Big\|\sum_{l=1}^{r} \Psi_{\cdot,l}(\Psi_{\cdot,l})^\top\Big\|_{\mathrm{op}} \vee \|\Psi\|_F^2\Big) \quad \text{and} \quad v_{\Omega,\infty} = \frac{1}{n}\|\Psi\|_{\infty,2}^2,$$

where the $(\infty,2)$-norm is defined by the maximum $\ell_2$-norm of the columns, $\|X\|_{\infty,2} \doteq \max_j \|X_{\cdot,j}\|_2$.

### 4.3 Beyond the First-Order Autoregressive Model

The theory developed in Sections 2 and 4 considers the VAR model of order $p = 1$ for the sake of simplicity. However, our approach is flexible, since we may use any other time-series modelling. To give a simple illustration of this fact, we consider below an extension to the second-order VAR model. Indeed, we don't want the VAR order $p$ to be too large, since the number of parameters scales as $d^2 \times p$ (forgetting about sparsity assumptions). In our experiments (see Section 5 below), we consider and compare both first order and second order VAR models.

Let us define the $(T-1) \times d$ time-series matrices

$$\mathbf{X}_T = \big[\omega(A_2), \ldots, \omega(A_T)\big]^\top, \quad \mathbf{X}_{T-1} = \big[\omega(A_1), \ldots, \omega(A_{T-1})\big]^\top,$$
$$\mathbf{X}_{T-2} = \big[\omega(A_0), \ldots, \omega(A_{T-2})\big]^\top.$$

We consider the following order 2 extension of the features VAR model:

$$\mathbf{X}_T = \mathbf{X}_{T-1} W_1 + \mathbf{X}_{T-2} W_2 + \mathbf{N}_T,$$

where $\mathbf{N}_T = [N_2, \ldots, N_T]^\top$ denotes the centered noise vector, and $W_1, W_2$ are VAR model parameters. In this case the Lasso objective is

$$J_1(W_1, W_2) = \frac{1}{T}\Big\|\mathbf{X}_T - \begin{bmatrix}\mathbf{X}_{T-1} & \mathbf{X}_{T-2}\end{bmatrix}\begin{bmatrix}W_1 \\ W_2\end{bmatrix}\Big\|_F^2 + \kappa \left\|\begin{bmatrix}W_1 \\ W_2\end{bmatrix}\right\|_1$$

and the Lasso estimator is defined by

$$(\widehat{W_1}, \widehat{W_2}) = \operatorname*{arg\,min}_{W_1, W_2} J_1(W_1, W_2).$$

In the same spirit as in Section 2.3 we define the objective

$$\mathcal{L}(A, W_1, W_2) = J_1(W_1, W_2) + \frac{1}{d}\|\omega(A)^\top - \omega(A_T)^\top W_1 - \omega(A_{T-1})^\top W_2\|_2^2 + \tau\|A\|_* + \gamma\|A\|_1.$$

The gradients of the quadratic loss are given by

$$\frac{1}{2}\begin{bmatrix}\nabla_{W_1}\ell\\ \nabla_{W_2}\ell\end{bmatrix} = \frac{1}{T}\begin{bmatrix}\mathbf{X}_{T-1}^\top\\ \mathbf{X}_{T-2}^\top\end{bmatrix}\left\{\begin{bmatrix}\mathbf{X}_{T-1} & \mathbf{X}_{T-2}\end{bmatrix}\begin{bmatrix}W_1\\ W_2\end{bmatrix} - \mathbf{X}_T\right\} + \frac{1}{d}\begin{bmatrix}\omega(A_T)\\ \omega(A_{T-1})\end{bmatrix}\left\{\begin{bmatrix}\omega(A_T)^\top & \omega(A_{T-1})^\top\end{bmatrix}\begin{bmatrix}W_1\\ W_2\end{bmatrix} - \omega(A)^\top\right\},$$

and

$$\frac{1}{2}\nabla_A \ell^\top = \frac{1}{d}\sum_{j=1}^{d}\left\{\omega(A)_j - (\omega(A_T)^\top W_1 + \omega(A_{T-1})^\top W_2)_j\right\}\Omega_j,$$

where $\Omega_j \in \mathbb{R}^{n\times n}$ is the $j$-th design matrix. We implemented the second order autoregressive model with three different types of penalties. We used:

1. Ridge Regression using $\kappa\|W_1\|_F^2 + \kappa\|W_2\|_F^2$ as the penalty term
2. the Lasso estimator, that is, the minimizer of $J_1$
3. the estimator suggested in this work.

## 5. Empirical Evaluation

In Section 5.1 we assess our algorithms on synthetic data, generated as described in Section 4.2. In Section 5.2 we use our algorithm for the prediction of sales volume for web-marketing data

### 5.1 Experiments with Synthetic Data

*Data generator.* In our experiments, the noise matrices $M_t$ are 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 $\sigma = .5$. Since the matrix $V_0$ defining the linear map $\omega$ is unknown we consider the feature map $\omega(A) = \mathrm{vec}(AV)$ where $\widetilde{A_T} = U\Sigma V^\top$ is the SVD of $\widetilde{A_T}$.
*Competitors.* The competing methods for our problem, as considered in this paper, are:

- *Nearest Neighbors*, that scores pairs of nodes with the number of common friends between them, which is given by $A^2$ where $A$ is the cumulative graph adjacency matrix $\widetilde{A_T} = \sum_{t=0}^{T} A_t$;
- *Static sparse and low-rank*, that is the link prediction algorithm suggested by Richard et al. (2012b), which is obtained by minimizing the objective $\|X - \widetilde{A_T}\|_F^2 + \tau\|X\|_* + \gamma\|X\|_1$. It is the closest static version of our method;
- *Autoregressive low-rank* and *Static low-rank*, that are regularized using only by the trace-norm (corresponding to $\gamma = 0$);
- *Katz* scores pairs of nodes $i$ and $j$ by the sum of number of paths of length $l$ connecting $i$ and $j$, weighted by an exponentially decreasing coefficient $\beta^l$: $\sum_{l=1}^{\infty}\beta^l\left(A^l\right)_{i,j}$;

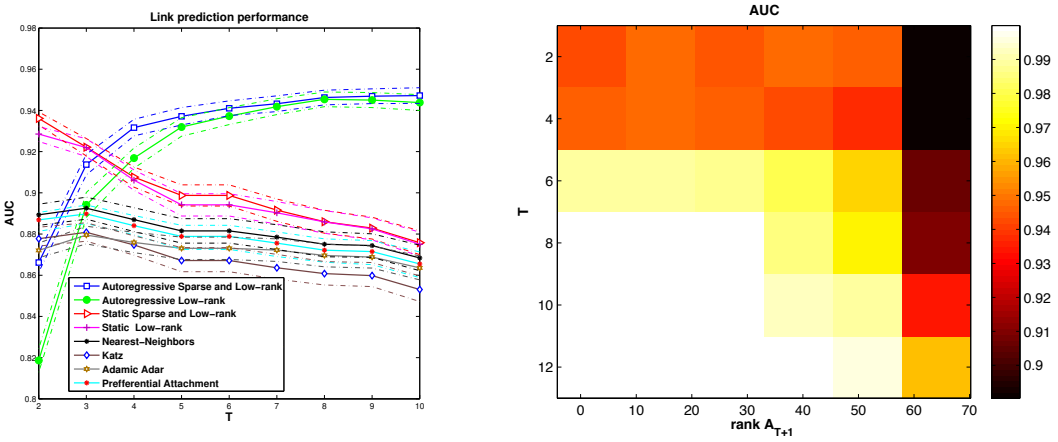


Figure 2: Left: performance of algorithms in terms of Area Under the ROC Curve, average and confidence intervals over 50 runs. Right: Phase transition diagram.

- *Adamic Adar* is the score $\sum_{\nu \in N(i) \cap N(j)} 1/\log(d_\nu)$, where $d_\nu$ is the degree of the node $\nu$ which is a common neighbor of $i$ and $j$;
- *Preferential attachment* only takes popularity into account and scores an edge $ij$ by the product of their degrees $d_i d_j$. See the papers by Liben-Nowell and Kleinberg (2007) and Lü and Zhou (2011) for details on Katz, Adamic-Adar and Preferential Attachment.

We also point out that other methods could possibly be adapted for the problem of link prediction as stated in the present paper. We mainly refer to the works by Liben-Nowell and Kleinberg (2007), Lü and Zhou (2011), Sarkar et al. (2012), Huang and Lin (2009), Nardi and Rinaldo (2011) and Davis et al. (2012). However, they were introduced either in a different setup, such as the one where multiple observations of a given edge occur, as described by Liben-Nowell and Kleinberg (2007) and Lü and Zhou (2011), or in the feature prediction problem of Nardi and Rinaldo (2011) and Davis et al. (2012), or they would involve tuning complex subroutines, such as the ones of Huang and Lin (2009), leading us far beyond the scope of the present work.

*Performance assessment for validation and test.* We compare our methods to standard baselines in link prediction by comparing predictions $\hat{A}$ to the adjacency matrix $A_{T+1} = A$, which is binary, at step $T+1$. Since the score matrix $\hat{A}$ outputs scalar values, we use a threshold parameter $t$ to build a link predictor $\mathbf{I}\{\hat{A}_{i,j} > t\}$ on the edge $(i,j)$. The quality of our estimation is then measured by considering all possible values of the threshold parameter $t$ which leads to the ROC curve as the plot of the proportion of hits (pairs $(i,j)$ such that $A_{ij} \cdot \mathbf{I}\{\hat{A}_{i,j} > t\} = 1$) versus the proportion of false detection (pairs $(i,j)$ such that $A_{ij} \cdot \mathbf{I}\{\hat{A}_{i,j} > t\} = 0$). Our criterion is the AUC for this particular definition of the ROC curve. In this approach of assessment, the size of the coefficients of $\hat{A}$ accounts for the strength of the prediction. We report empirical results averaged over 50 runs with confidence intervals in Figure 2. The parameters $\tau$ and $\gamma$ are chosen by a 10-fold cross validation for

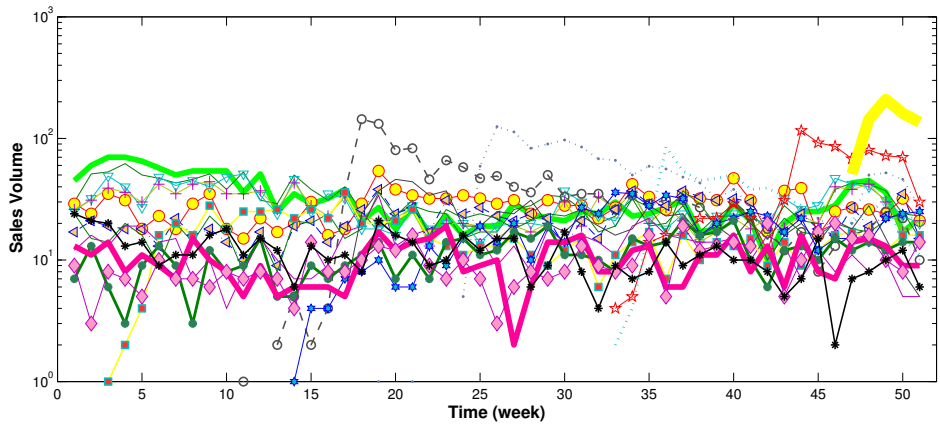


Figure 3: Sales volumes of 20 top-sold items weekly sales over the year.

each of the methods separately. The validation set is the upwards sliding time window when learning from the past. The right-hand side of Figure 2 is a phase transition diagram showing the impact of both rank and time on the accuracy of estimation of the adjacency matrix. The results are clearly better as we gain historical depth and the lower the rank of the adjacency matrix.

*Comparison with the baselines.* This experiment shows the benefits 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 to be preferred, whereas feature autoregressive approaches outperform them as soon as a *sufficient number* $T$ of graph snapshots are available (see the Phase transition diagram from Figure 2). The decreasing performance of static algorithms can be explained by the fact that they use as input a mixture of different graphs observed at different time steps, whereas they require a single simple graph as input.

### 5.2 Experiments with Real Data: Predicting Sales Volumes

*Motivations.* Predicting the popularity of products is of major interest for marketers and product managers as it allows to anticipate or even create trends that diffuse in networks. A useful observation when dealing with sales volumes is that when modeling purchases by edges in the bipartite graph of users and products, the sales volumes can be seen as the degrees of the product nodes. We use two VAR models of order 1 and 2, as described in Section 4.3, in order to show the flexibility of our approach. We consider the linear feature map $\omega(A) = A^\top \mathbf{1}$ that computes the columns degree vector. The dimension of the feature space $d$ equals the number of columns of the matrix in this case. If the input matrix $A_t$ is the *users* $\times$ *products* matrix of sales at time period $t$ then the degree of each product equals the sales volume of the product during that period, and it can be used as a fair popularity indicator for the product. It is in addition common to consider a regular evolution of such features, see the paper by Rogers (1962). Note that the suggested approach is valid for a similar activity indicator in any other network, such as users activity on a social network

| | AR(1) | | | AR(2) | | |
|---|---|---|---|---|---|---|
| Error | Ridge | Lasso | Our | Ridge | Lasso | Our |
| $T = 10$ | 0.9524 | 1.1344 | 0.9730 | 1.0037 | 1.0110 | **0.9416** |
| $T = 15$ | 0.6394 | 0.5389 | 0.5336 | 0.6010 | **0.5304** | 0.5401 |
| $T = 20$ | 0.3419 | 0.3111 | 0.4878 | 0.3310 | **0.2972** | 0.3086 |
| $T = 25$ | 0.3777 | 0.6238 | 0.5689 | 0.3356 | 0.3286 | **0.3279** |

Table 2: Relative quadratic error of the prediction of sales volume for three regularized VAR models: one based on ridge regression penalty, one base don LASSO penalty, and one based on our strategy with both sparse and low-rank regularizers.

or protein activity on a protein-protein interaction network. A last remark is that the prior knowledge in the case of e-commerce data suggests that groups of highly related items exist, which makes the adjacency matrix low-rank in addition to be sparse. In fact the adjacency matrix of a fully clustered graph would be block-diagonal, and we expect the matrix of interest to be close to such a matrix.

*Protocol and description of data sets.* We performed our experiments on the sales volumes time series of the $n = 200$ top sold books over $T = 25$ consecutive weeks excluding the Christmas period in 2009 of 31972 users.[1] The weekly sales of each book corresponds to the degree of the corresponding node. The books catalogue contains several book categories, which motivates the low-rank matrix assumption. In Figure 3 we plot the time series of the top 20 items from the catalogue. From this observation, the stationary assumption seems plausible. More precisely, we observed that the time window allowing accurate predictions (a window in which the data is stationary) is, among the range of values we used in our experiments, equal to 20 weeks.

*Comparison with other methods and performance metric.* We compare estimators of the degrees based on Ridge and Lasso penalization using the objective $J_1$ only, see Equation (2), with our procedure based on joint minimization of (3). For choosing the tuning parameters $\kappa, \tau, \gamma$ we use the data collected from the same market a year before the test set to form the training and validation sets. For testing the quality of our predictor, we used the parameters performing the best predictions over the validation set. As data are abundant we can collect past data easily for this step. On the other hand, as seasonality effects may harm the method if cross-validation is performed on data taken from a different period of the year, this is the best way to proceed for splitting the data onto training validation and test sets. We evaluated the results in terms of relative quadratic error

$$\text{Relative quadratic error} = \frac{\|\omega(\widehat{A}) - \omega(A_{T+1})\|_2}{\|\omega(A_{T+1})\|_2}$$

over the prediction of the sales volumes. The results are reported in Table 2.

*Comments on the results.* From this experiment we conclude the following. The order of the VAR is an important factor. We provided theoretical results for the VAR of order 1,

1. The data was provided by the company 1000mercis.

but fitting a higher order VAR in practice may result in better performance. This is also a parameter that should ideally be chosen using the past data in a cross-validation process. Moreover, the size of the time window $T$ should be chosen according to the data. A small value of $T$ leads to poor result due to absence of enough signal. As opposite, a too large value of $T$ harms the quality of prediction due to the nonstationary trends in too large windows of time. Note that in our synthetic data experiments only the first effect was observed: the performance is increasing as the time parameter $T$ increases. This is due to the stationarity in synthetically generated data.

### 5.3 Discussion

*Trading convexity for scalability.* In the numerical experiments, for better scalability, one can replace the penalty on $A$ by a sparsity inducing penalty on the factors of $A$. Namely if $A = UV^\top$ is a factorization of $A$, one can replace the term $\tau\|A\|_* + \gamma\|A\|_1$ by $\lambda\|U\|_1\|V\|_1$. This penalty leads to a non-convex problem, nevertheless it allows better scalability than the convex penalty both in terms of memory requirement and computational complexity, when evaluating the proximal. Another practical advantage of this change of variable is that we need to tune only one real parameter $\lambda$ instead of two ($\gamma$ and $\tau$). The maximum rank of $A = UV^\top$ (number of columns of $U$ and $V$) replaces the low-rank inducing effect of $\tau$.
*Generalization of the regression method.* In this paper, we consider only an autoregressive process of order 1 and 2. For better prediction accuracy, one could consider more 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.
*Choice of the feature map $\omega$.* In this work, we have used the projection onto the vector space of the top-$r$ singular vectors of the cumulative adjacency matrix as the linear map $\omega$, and this choice has shown empirical superiority to other choices. The question of choosing the best measurement / representation to summarize graph information as in compress sensing seems to have both theoretical and application potential. In our work the map $\omega$ was applied to a single matrix $A_t$. One can consider a mapping taking as input several successive matrices $A_t, A_{t+1}, A_{t+2}$. This idea has been used by Zhang et al. (2011) in order to distinguish the effect of new and returning customers in a marketplace. Moreover, a deeper understanding of the connections of our problem with compressed sensing, for the construction and theoretical validation of the feature map, is an important point that needs several developments. An extension to nonlinear graph features such as the distribution of triangles or other nonlinear patterns of interest is also to be considered.

## 6. Conclusion

In this work, we studied the link prediction problem under structural hypotheses on the graph generation process (sparse low-rank adjacency and autoregressive features). Our work establishes a connection between the link prediction problem and compressed sensing through the use of common tools in the model and in the theoretical analysis. Empirical experiments show the benefit of adopting such a point of view. In fact, compared to the existing heuristics, this approach offers a principled search method in the hypothesis space through the regularization and convex optimization formulation. The flexibility of our ap-

proach and its connections with several active areas of research makes it very attractive and reveals several interesting directions of investigation for future work.

## Appendix A. Proofs of the Main Results

From now on, we use the notation $\|(A,a)\|_F^2 = \|A\|_F^2 + \|a\|_2^2$ and $\langle (A,a),(B,b)\rangle = \langle A,B\rangle + \langle a,b\rangle$ for any $A,B \in \mathbb{R}^{T\times d}$ and $a,b \in \mathbb{R}^d$.

Let us introduce the linear mapping $\Phi : \mathbb{R}^{n\times n} \times \mathbb{R}^{d\times d} \to \mathbb{R}^{T\times d} \times \mathbb{R}^d$ given by

$$\Phi(A,W) = \Big(\frac{1}{\sqrt{T}}\mathbf{X}_{T-1}W, \frac{1}{\sqrt{d}}(\omega(A) - W^\top \omega(A_T))\Big).$$

Using this mapping, the objective (3) can be written in the following reduced way:

$$\mathcal{L}(A,W) = \Big\|\Big(\frac{1}{\sqrt{T}}\mathbf{X}_T, 0\Big) - \Phi(A,W)\Big\|_F^2 + \gamma\|A\|_1 + \tau\|A\|_* + \kappa\|W\|_1.$$

Recalling that the error writes, for any $A$ and $W$:

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

we have

$$\mathcal{E}(A,W)^2 = \big\|\Phi(A - A_{T+1}, W - W_0)\|_F^2.$$

Let us introduce also the empirical risk

$$R_n(A,W) = \Big\|\Big(\frac{1}{\sqrt{T}}\mathbf{X}_T, 0\Big) - \Phi(A,W)\Big\|_F^2.$$

The proofs of Theorem 1 and 2 are based on tools developed by Koltchinskii et al. (2011) and Bickel et al. (2009). However, the context considered here is very different from the setting considered in these papers, so our proofs require a different scheme.

### A.1 Proof of Theorem 1

First, note that

$$\begin{aligned} &R_n(\hat{A},\hat{W}) - R_n(A,W) \\ &\qquad = \|\Phi(\hat{A},\hat{W})\|_F^2 - \|\Phi(A,W)\|_F^2 - 2\langle(\frac{1}{\sqrt{T}}\mathbf{X}_T, 0), \Phi(\hat{A}-A, \hat{W}-W)\rangle. \end{aligned}$$

Since

$$\begin{aligned} &\|\Phi(\hat{A},\hat{W})\|_F^2 - \|\Phi(A,W)\|_F^2 \\ &\qquad = \mathcal{E}(\hat{A},\hat{W})^2 - \mathcal{E}(A,W)^2 + 2\langle\Phi(\hat{A}-A,\hat{W}-W), \Phi(A_{T+1},W_0)\rangle, \end{aligned}$$

we have

$$\begin{aligned} &R_n(\hat{A},\hat{W}) - R_n(A,W) \\ &\qquad = \mathcal{E}(\hat{A},\hat{W})^2 - \mathcal{E}(A,W)^2 + 2\langle\Phi(\hat{A}-A,\hat{W}-W), \Phi(A_{T+1},W_0) - (\frac{1}{\sqrt{T}}\mathbf{X}_T, 0)\rangle \\ &\qquad = \mathcal{E}(\hat{A},\hat{W})^2 - \mathcal{E}(A,W)^2 + 2\langle\Phi(\hat{A}-A,\hat{W}-W), (-\frac{1}{\sqrt{T}}\mathbf{N}_T, \frac{1}{\sqrt{d}}N_{T+1})\rangle. \end{aligned}$$

The next Lemma will come in handy several times in the proofs.

**Lemma 5** *For any $A \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{d\times d}$ we have*

$$\langle (\frac{1}{\sqrt{T}}\mathbf{N}_T, -\frac{1}{\sqrt{d}}N_{T+1}), \Phi(A,W)\rangle = \langle (M,\Xi),(A,W)\rangle = \langle W,\Xi\rangle + \langle A, M\rangle.$$

This Lemma follows from a direct computation, and the proof is thus omitted. This Lemma entails, together with (4), that

$$\begin{aligned}\mathcal{E}(\hat{A},\hat{W})^2 \leq \mathcal{E}(A,W)^2 &+ 2\langle \hat{W}-W,\Xi\rangle + 2\langle \hat{A}-A, M\rangle \\ &+ \tau(\|A\|_* - \|\widehat{A}\|_*) + \gamma(\|A\|_1 - \|\widehat{A}\|_1) + \kappa(\|W\|_1 - \|\widehat{W}\|_1).\end{aligned}$$

Now, using Hölder's inequality and the triangle inequality, and introducing $\alpha \in (0,1)$, we obtain

$$\begin{aligned}\mathcal{E}(\hat{A},\hat{W})^2 \leq \mathcal{E}(A,W)^2 &+ \Big(2\alpha\|M\|_{\mathrm{op}} - \tau\Big)\|\hat{A}\|_* + \Big(2\alpha\|M\|_{\mathrm{op}} + \tau\Big)\|A\|_* \\ &+ \Big(2(1-\alpha)\|M\|_\infty - \gamma\Big)\|\hat{A}\|_1 + \Big(2(1-\alpha)\|M\|_\infty + \gamma\Big)\|A\|_1 \\ &+ \Big(2\|\Xi\|_\infty - \kappa\Big)\|\hat{W}\|_1 + \Big(2\|\Xi\|_\infty + \kappa\Big)\|W\|_1,\end{aligned}$$

which concludes the proof of Theorem 1, using (5). □

### A.2 Proof of Theorem 2

Let $A \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{d\times d}$ be fixed, and let $A = U\,\mathrm{diag}(\sigma_1,\dots,\sigma_r)V^\top$ be the SVD of $A$. Recalling that $\circ$ is the entry-wise product, we have $A = \Theta_A \circ |A| + \Theta_A^\perp \circ A$, where $\Theta_A \in \{0,\pm 1\}^{n\times n}$ is the entry-wise sign matrix of $A$ and $\Theta_A^\perp \in \{0,1\}^{n\times n}$ is the orthogonal sparsity pattern of $A$.

The definition (4) of $(\hat{A},\hat{W})$ is equivalent to the fact that one can find $\hat{G} \in \partial\mathcal{L}(\hat{A},\hat{W})$ (an element of the subgradient of $\mathcal{L}$ at $(\hat{A},\hat{W})$) that belongs to the normal cone of $\mathcal{A}\times\mathcal{W}$ at $(\hat{A},\hat{W})$. This means that for such a $\hat{G}$, and any $A \in \mathcal{A}$ and $W \in \mathcal{W}$, we have

$$\langle \hat{G}, (\hat{A}-A, \hat{W}-W)\rangle \leq 0. \tag{12}$$

Any subgradient of the function $g(A) = \tau\|A\|_* + \gamma\|A\|_1$ writes

$$Z = \tau Z_* + \gamma Z_1 = \tau\Big(UV^\top + \mathcal{P}_A^\perp(G_*)\Big) + \gamma\Big(\Theta_A + G_1 \circ \Theta_A^\perp\Big)$$

for some $\|G_*\|_{\mathrm{op}} \leq 1$ and $\|G_1\|_\infty \leq 1$ (see for instance the paper by Lewis (1995)). So, if $\hat{Z} \in \partial g(\hat{A})$, we have, by monotonicity of the sub-differential, that for any $Z \in \partial g(A)$

$$\langle \hat{Z}, \hat{A}-A\rangle = \langle \hat{Z}-Z, \hat{A}-A\rangle + \langle Z, \hat{A}-A\rangle \geq \langle Z, \hat{A}-A\rangle,$$

and, by duality, we can find $Z$ such that

$$\langle Z, \widehat{A}-A\rangle = \tau\langle UV^\top, \widehat{A}-A\rangle + \tau\|\mathcal{P}_A^\perp(\widehat{A})\|_* + \gamma\langle \Theta_A, \widehat{A}-A\rangle + \gamma\|\Theta_A^\perp \circ \widehat{A}\|_1.$$

By using the same argument with the function $W \mapsto \|W\|_1$ and by computing the gradient of the empirical risk $(A, W) \mapsto R_n(A, W)$, Equation (12) entails that

$$\begin{aligned}
&2\langle \Phi(\widehat{A} - A_{T+1}, \widehat{W} - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle \\
&\quad \leq 2\langle (\frac{1}{\sqrt{T}}\mathbf{N}_T, -\frac{1}{\sqrt{d}}N_{T+1}), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle - \tau\langle UV^\top, \widehat{A} - A\rangle - \tau\|\mathcal{P}_A^\perp(\widehat{A})\|_* \\
&\qquad - \gamma\langle \Theta_A, \widehat{A} - A\rangle - \gamma\|\Theta_A^\perp \circ \widehat{A}\|_1 - \kappa\langle \Theta_W, \widehat{W} - W\rangle - \kappa\|\Theta_W^\perp \circ \widehat{W}\|_1.
\end{aligned} \tag{13}$$

Using Pythagora's theorem, we have

$$\begin{aligned}
&2\langle \Phi(\widehat{A} - A_{T+1}, \hat{W} - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle \\
&\quad = \|\Phi(\widehat{A} - A_{T+1}, \hat{W} - W_0)\|_2^2 + \|\Phi(\widehat{A} - A, \widehat{W} - W)\|_2^2 - \|\Phi(A - A_{T+1}, W - W_0)\|_2^2.
\end{aligned} \tag{14}$$

It shows that if $\langle \Phi(\widehat{A} - A_{T+1}, W - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle \leq 0$, then Theorem 2 trivially holds. Let us assume that

$$\langle \Phi(\widehat{A} - A_{T+1}, W - W_0), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle > 0. \tag{15}$$

Using Hölder's inequality, we obtain

$$\begin{aligned}
|\langle UV^\top, \hat{A} - A\rangle| &= |\langle UV^\top, \mathcal{P}_A(\hat{A} - A)\rangle| \leq \|UV^\top\|_{\mathrm{op}}\|\mathcal{P}_A(\hat{A} - A)\|_* = \|\mathcal{P}_A(\hat{A} - A)\|_*, \\
|\langle \Theta_A, \hat{A} - A\rangle| &= |\langle \Theta_A, \Theta_A \circ (\hat{A} - A)\rangle| \leq \|\Theta_A\|_\infty\|\Theta_A \circ (\hat{A} - A)\|_1 = \|\Theta_A \circ (\hat{A} - A)\|_1,
\end{aligned}$$

and the same is done for $|\langle \Theta_W, \hat{W} - W\rangle| \leq \|\Theta_W \circ (\hat{W} - W)\|_1$. So, when (15) holds, we obtain by rearranging the terms of (13):

$$\begin{aligned}
&\tau\|\mathcal{P}_A^\perp(\widehat{A} - A)\|_* + \gamma\|\Theta_A^\perp \circ (\widehat{A} - A)\|_1 + \kappa\|\Theta_W^\perp \circ (\widehat{W} - W)\|_1 \\
&\qquad \leq \tau\|\mathcal{P}_A(\hat{A} - A)\|_* + \gamma\|\Theta_A \circ (\hat{A} - A)\|_1 + \kappa\|\Theta_W \circ (\hat{W} - W)\|_1 \\
&\qquad\quad + 2\langle (\frac{1}{\sqrt{T}}\mathbf{N}_T, -\frac{1}{\sqrt{d}}N_{T+1}), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle.
\end{aligned} \tag{16}$$

Using Lemma 5, together with Hölder's inequality, we have for any $\alpha \in (0, 1)$:

$$\begin{aligned}
\langle (\frac{1}{\sqrt{T}}\mathbf{N}_T, -\frac{1}{\sqrt{d}}N_{T+1}), \Phi(\widehat{A} - A, \widehat{W} - W)\rangle &= \langle M, \hat{A} - A\rangle + \langle \Xi, \hat{W} - W\rangle \\
&\leq \alpha\|M\|_{\mathrm{op}}\|\mathcal{P}_A(\hat{A} - A)\|_* + \alpha\|M\|_{\mathrm{op}}\|\mathcal{P}_A^\perp(\hat{A} - A)\|_* \\
&\quad + (1 - \alpha)\|M\|_\infty\|\Theta_A \circ (\hat{A} - A)\|_1 + (1 - \alpha)\|M\|_\infty\|\Theta_A^\perp \circ (\hat{A} - A)\|_1 \\
&\quad + \|\Xi\|_\infty(\|\Theta_W \circ (\hat{W} - W)\|_1 + \|\Theta_W^\perp \circ (\hat{W} - W)\|_1)\ .
\end{aligned} \tag{17}$$

Now, using (16) together with (17), we obtain

$$\begin{aligned}
&\big(\tau - 2\alpha\|M\|_{\mathrm{op}}\big)\|\mathcal{P}_A^\perp(\hat{A} - A)\|_* + \big(\gamma - 2(1 - \alpha)\|M\|_\infty\big)\|\Theta_A^\perp \circ (\hat{A} - A)\|_1 \\
&\qquad + \big(\kappa - 2\|\Xi\|_\infty\big)\|\Theta_W^\perp \circ (\hat{W} - W)\|_1 \\
&\leq \big(\tau + 2\alpha\|M\|_{\mathrm{op}}\big)\|\mathcal{P}_A(\hat{A} - A)\|_* + \big(\gamma + 2(1 - \alpha)\|M\|_\infty\big)\|\Theta_A \circ (\hat{A} - A)\|_1 \\
&\qquad + \big(\kappa + 2\|\Xi\|_\infty\big)\|\Theta_W \circ (\hat{W} - W)\|_1
\end{aligned}$$

which proves, using (6), that

$$\tau\|\mathcal{P}_A^\perp(\hat{A}-A)\|_* + \gamma\|\Theta_A^\perp \circ (\hat{A}-A)\|_1 \leq 5\tau\|\mathcal{P}_A(\hat{A}-A)\|_* + 5\gamma\|\Theta_A \circ (\hat{A}-A)\|_1.$$

This proves that $\hat{A}-A \in \mathcal{C}_2(A,5,\gamma/\tau)$. In the same way, using (16) with $A=\hat{A}$ together with (17), we obtain that $\hat{W}-W \in \mathcal{C}_1(W,5)$.

Now, using together (13), (14) and (17) , and the fact that the Cauchy-Schwarz inequality entails

$$\|\mathcal{P}_A(\hat{A}-A)\|_* \leq \sqrt{\operatorname{rank} A}\|\mathcal{P}_A(\hat{A}-A)\|_F, \quad |\langle UV^\top, \hat{A}-A\rangle| \leq \sqrt{\operatorname{rank} A}\|\mathcal{P}_A(\hat{A}-A)\|_F,$$
$$\|\Theta_A \circ (\hat{A}-A)\|_1 \leq \sqrt{\|A\|_0}\|\Theta_A \circ (\hat{A}-A)\|_F, \quad |\langle \Theta_A, \hat{A}-A\rangle| \leq \sqrt{\|A\|_0}\|\Theta_A \circ (\hat{A}-A)\|_F \ .$$

and similarly for $\hat{W}-W$, we arrive at

$$\begin{aligned}
&\|\Phi(\widehat{A}-A_{T+1}, \hat{W}-W_0)\|_2^2 + \|\Phi(\widehat{A}-A, \widehat{W}-W)\|_2^2 - \|\Phi(A-A_{T+1}, W-W_0)\|_2^2 \\
&\quad\leq \big(2\alpha\|M\|_{\mathrm{op}}+\tau\big)\sqrt{\operatorname{rank} A}\|\mathcal{P}_A(\hat{A}-A)\|_F + \big(2\alpha\|M\|_{\mathrm{op}}-\tau\big)\|\mathcal{P}_A^\perp(\hat{A}-A)\|_* \\
&\qquad+ \big(2\alpha\|M\|_\infty+\gamma\big)\sqrt{\|A\|_0}\|\Theta_A \circ (\hat{A}-A)\|_F + \big(2\alpha\|M\|_\infty-\gamma\big)\|\Theta_A^\perp \circ (\hat{A}-A)\|_1 \\
&\qquad+ \big(2\alpha\|\Xi\|_\infty+\kappa\big)\sqrt{\|W\|_0}\|\Theta_W \circ (\hat{W}-W)\|_F + \big(2\alpha\|\Xi\|_\infty-\kappa\big)\|\Theta_W^\perp \circ (\hat{W}-W)\|_1,
\end{aligned}$$

which leads, using (6), to

$$\begin{aligned}
&\|\Phi(\widehat{A}-A_{T+1}, \hat{W}-W_0)\|_2^2 + \|\Phi(\widehat{A}-A, \widehat{W}-W)\|_2^2 - \|\Phi(A-A_{T+1}, W-W_0)\|_2^2 \\
&\quad\leq \frac{5\tau}{3}\sqrt{\operatorname{rank} A}\|\mathcal{P}_A(\hat{A}-A)\|_F + \frac{5\gamma}{3}\sqrt{\|A\|_0}\|\Theta_A \circ (\hat{A}-A)\|_F + \frac{5\kappa}{3}\sqrt{\|W\|_0}\|\Theta_W \circ (\hat{W}-W)\|_F.
\end{aligned}$$

Since $\hat{A}-A \in \mathcal{C}_2(A,5,\gamma/\tau)$ and $\hat{W}-W \in \mathcal{C}_1(W,5)$, we obtain using Assumption 2 and $ab \leq (a^2+b^2)/2$:

$$\begin{aligned}
&\|\Phi(\widehat{A}-A_{T+1}, \hat{W}-W_0)\|_2^2 + \|\Phi(\widehat{A}-A, \widehat{W}-W)\|_2^2 \\
&\leq \|\Phi(A-A_{T+1}, W-W_0)\|_2^2 + \frac{25}{18}\mu_2(A,W)^2\big(\operatorname{rank} A\tau^2 + \|A\|_0\gamma^2\big) \\
&\quad+ \frac{25}{36}\mu_1(W)^2\|W\|_0\kappa^2 + \|\Phi(\widehat{A}-A, \widehat{W}-W)\|_2^2,
\end{aligned}$$

which concludes the proof of Theorem 2. □

### A.3 Proof of Corollary 4

For the proof of (9), we simply use the fact that $\frac{1}{T}\|\mathbf{X}_{T-1}(\hat{W}-W_0)\|_F^2 \leq \mathcal{E}(\hat{A},\hat{W})^2$ and use Theorem 3. Then we take $W=W_0$ in the infimum over $A, W$.

For (10), we use the fact that since $\hat{W}-W_0 \in \mathcal{C}_1(W_0,5)$, we have (see the Proof of Theorem 2),

$$\begin{aligned}
\|\hat{W}-W_0\|_1 &\leq 6\sqrt{\|W_0\|_0}\|\Theta_W \circ (\hat{W}-W_0)\|_F \\
&\leq 6\sqrt{\|W_0\|_0}\|\mathbf{X}_{T-1}(\hat{W}-W_0)\|_F/\sqrt{T} \\
&\leq 6\sqrt{\|W_0\|_0}\mathcal{E}(\hat{A},\hat{W}),
\end{aligned}$$

and then use again Theorem 3. The proof of (11) follows exactly the same scheme. □

### A.4 Concentration Inequalities for the Noise Processes

The control of the noise terms $M$ and $\Xi$ is based on recent results on concentration inequalities for random matrices, developed by Tropp (2012). Moreover, the assumption on the dynamics of the features' noise vector $\{N_t\}_{t\geq 0}$ is quite general, since we only assumed that this process is a martingale increment. Therefore, our control of the noise $\Xi$ rely in particular on martingale theory.

**Proposition 6** *Under Assumption 3, the following inequalities hold for any $x > 0$. We have*

$$\Big\|\frac{1}{d}\sum_{j=1}^{d}(N_{T+1})_j\Omega_j\Big\|_{\mathrm{op}} \leq \sigma v_{\Omega,\mathrm{op}}\sqrt{\frac{2(x+\log(2n))}{d}} \tag{18}$$

*with a probability larger than $1-e^{-x}$. We have*

$$\Big\|\frac{1}{d}\sum_{j=1}^{d}(N_{T+1})_j\Omega_j\Big\|_{\infty} \leq \sigma v_{\Omega,\infty}\sqrt{\frac{2(x+2\log n)}{d}} \tag{19}$$

*with a probability larger than $1-2e^{-x}$, and finally*

$$\Big\|\frac{1}{T}\sum_{t=1}^{T}\omega(A_{t-1})N_t^\top + \frac{1}{d}\omega(A_T)N_{T+1}^\top\Big\|_{\infty} \leq \sigma\sigma_\omega\sqrt{2e(x+2\log d+\ell_T)}\Big(\frac{1}{\sqrt{T}}+\frac{1}{d}\Big) \tag{20}$$

*with a probability larger than $1-15e^{-x}$, where we recall that $\ell_T$ is given by* (8).

**Proof** For the proofs of Inequalities (18) and (19), we use the fact that $(N_{T+1})_1,\dots,(N_{T+1})_d$ are independent (scalar) sub-gaussian random variables.

From Assumption 3, we have for any $n\times n$ deterministic self-adjoint matrices $X_j$ that $\mathbb{E}[\exp(\lambda(N_{T+1})_jX_j)] \preceq \exp(\sigma^2\lambda^2X_j^2/2)$, where $\preceq$ stands for the semidefinite order on self-adjoint matrices. Using Corollary 3.7 by Tropp (2012), this leads for any $x>0$ to

$$\mathbb{P}\Big[\lambda_{\max}\Big(\sum_{j=1}^{d}(N_{T+1})_jX_j\Big)\geq x\Big] \leq n\exp\Big(-\frac{x^2}{2v^2}\Big), \quad \text{where } v^2=\sigma^2\Big\|\sum_{j=1}^{d}X_j^2\Big\|_{\mathrm{op}}. \tag{21}$$

Then, following Tropp (2012), we consider the dilation operator $\Delta:\mathbb{R}^{n\times n}\to\mathbb{R}^{2n\times 2n}$ given by

$$\Delta(\Omega)=\begin{pmatrix}0 & \Omega\\ \Omega^* & 0\end{pmatrix}.$$

We have

$$\Big\|\sum_{j=1}^{d}(N_{T+1})_j\Omega_j\Big\|_{\mathrm{op}} = \lambda_{\max}\Big(\Delta\Big(\sum_{j=1}^{d}(N_{T+1})_j\Omega_j\Big)\Big) = \lambda_{\max}\Big(\sum_{j=1}^{d}(N_{T+1})_j\Delta(\Omega_j)\Big)$$

and an easy computation gives

$$\Big\|\sum_{j=1}^{d}\Delta(\Omega_j)^2\Big\|_{\mathrm{op}} = \Big\|\sum_{j=1}^{d}\Omega_j^\top\Omega_j\Big\|_{\mathrm{op}} \vee \Big\|\sum_{j=1}^{d}\Omega_j\Omega_j^\top\Big\|_{\mathrm{op}}.$$

So, using (21) with the self-adjoint $X_j = \Delta(\Omega_j)$ gives

$$\mathbb{P}\Big[\Big\|\sum_{j=1}^d (N_{T+1})_j \Omega_j\Big\|_{\rm op} \geq x\Big] \leq 2n\exp\Big(-\frac{x^2}{2v^2}\Big) \text{ where } v^2 = \sigma^2\Big\|\sum_{j=1}^d \Omega_j^\top \Omega_j\Big\|_{\rm op} \vee \Big\|\sum_{j=1}^d \Omega_j\Omega_j^\top\Big\|_{\rm op},$$

which leads easily to (18).

Inequality (19) comes from the following standard bound on the sum of independent sub-gaussian random variables:

$$\mathbb{P}\Big[\Big|\frac{1}{d}\sum_{j=1}^d (N_{T+1})_j(\Omega_j)_{k,l}\Big| \geq x\Big] \leq 2\exp\Big(-\frac{x^2}{2\sigma^2(\Omega_j)_{k,l}^2}\Big)$$

together with an union bound on $1 \leq k, l \leq n$.

Inequality (20) is based on a classical martingale exponential argument together with a peeling argument. We denote by $\omega_j(A_t)$ the coordinates of $\omega(A_t) \in \mathbb{R}^d$ and by $N_{t,k}$ those of $N_t$, so that

$$\Big(\sum_{t=1}^T \omega(A_{t-1})N_t^\top\Big)_{j,k} = \sum_{t=1}^T \omega_j(A_{t-1})N_{t,k}.$$

We fix $j, k$ and denote for short $\varepsilon_t = N_{t,k}$ and $x_t = \omega_j(A_t)$. Since $\mathbb{E}[\exp(\lambda\varepsilon_t)|\mathcal{F}_{t-1}] \leq e^{\sigma^2\lambda^2/2}$ for any $\lambda \in \mathbb{R}$, we obtain by a recursive conditioning with respect to $\mathcal{F}_{T-1}$, $\mathcal{F}_{T-2}, \ldots, \mathcal{F}_0$, that

$$\mathbb{E}\Big[\exp\Big(\theta\sum_{t=1}^T \varepsilon_t x_{t-1} - \frac{\sigma^2\theta^2}{2}\sum_{t=1}^T x_{t-1}^2\Big)\Big] \leq 1.$$

Hence, using Markov's inequality, we obtain for any $v > 0$:

$$\mathbb{P}\Big[\sum_{t=1}^T \varepsilon_t x_{t-1} \geq x, \sum_{t=1}^T x_{t-1}^2 \leq v\Big] \leq \inf_{\theta>0}\exp(-\theta x + \sigma^2\theta^2 v/2) = \exp\Big(-\frac{x^2}{2\sigma^2 v}\Big),$$

that we rewrite in the following way:

$$\mathbb{P}\Big[\sum_{t=1}^T \varepsilon_t x_{t-1} \geq \sigma\sqrt{2vx}, \sum_{t=1}^T x_{t-1}^2 \leq v\Big] \leq e^{-x}.$$

Let us denote for short $V_T = \sum_{t=1}^T x_{t-1}^2$ and $S_T = \sum_{t=1}^T \varepsilon_t x_{t-1}$. We want to replace $v$ by $V_T$ from the previous deviation inequality, and to remove the event $\{V_T \leq v\}$. To do so, we use a peeling argument. We take $v = T$ and introduce $v_k = ve^k$ so that the event $\{V_T > v\}$ is decomposed into the union of the disjoint sets $\{v_k < V_T \leq v_{k+1}\}$. We introduce also $\ell_T = 2\log\log\Big(\frac{\sum_{t=1}^T x_{t-1}^2}{T} \vee \frac{T}{\sum_{t=1}^T x_{t-1}^2} \vee e\Big)$.

This leads to

$$\begin{aligned}\mathbb{P}\Big[S_T \geq \sigma\sqrt{2eV_T(x+\ell_T)}, V_T > v\Big] &= \sum_{k\geq 0} \mathbb{P}\big[S_T \geq \sigma\sqrt{2eV_T(x+\ell_T)}, v_k < V_T \leq v_{k+1}\Big] \\ &= \sum_{k\geq 0} \mathbb{P}\Big[S_T \geq \sigma\sqrt{2v_{k+1}(x+2\log\log(e^k\vee e))}, v_k < V_T \leq v_{k+1}\Big] \\ &\leq e^{-x}(1+\sum_{k\geq 1} k^{-2}) \leq 3.47e^{-x}.\end{aligned}$$

On $\{V_T \leq v\}$ the proof is the same: we decompose onto the disjoint sets $\{v_{k+1} < V_T \leq v_k\}$ where this time $v_k = ve^{-k}$, and we arrive at

$$\mathbb{P}\Big[S_T \geq \sigma\sqrt{2eV_T(x+\ell_T)}, V_T \leq v\Big] \leq 3.47e^{-x}.$$

This leads to

$$\mathbb{P}\Big[\sum_{t=1}^T \omega_j(A_{t-1})N_{t,k} \geq \sigma\Big(2e\sum_{t=1}^T \omega_j(A_{t-1})^2(x+\ell_{T,j})\Big)^{1/2}\Big] \leq 7e^{-x}$$

for any $1 \leq j,k \leq d$, where we introduced

$$\ell_{T,j} = 2\log\log\Big(\frac{\sum_{t=1}^T \omega_j(A_{t-1})^2}{T} \vee \frac{T}{\sum_{t=1}^T \omega_j(A_{t-1})^2} \vee e\Big).$$

The conclusion follows from an union bound on $1 \leq j,k \leq d$, and from the use of the same argument for the term $\omega(A_T)N_{T+1}^\top$. This concludes the proof of Proposition 6. ■

## References

J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

A. Andreas, M. Pontil, Y. Ying, and C. A. Micchelli. A spectral regularization framework for multi-task structure learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 25–32, 2007.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, page 85, 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

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.

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

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

E. J. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 12(51):21–30, 2008.

P. L. Combettes and J. C. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212, 2011.

R. A. Davis, P. Zang, and T. Zheng. Sparse vector autoregressive modeling. *arXiv preprint arXiv:1207.0520*, 2012.

D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4): 1289–1306, 2006.

T. Evgeniou, C. A. Micchelli, and M. Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

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

Z. Huang and D. K. J. Lin. The time-series link prediction problem with applications in communication surveillance. *INFORMS J. on Computing*, 21(2):286–303, 2009.

M. Kolar and E. P. Xing. On time varying undirected graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 407–415, 2011.

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

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

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 426–434. ACM, 2008.

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

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *J. Convex Anal.*, 2(1-2):173–183, 1995.

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.

L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

S. A. Myers and J. Leskovec. On the convexity of latent social network inference. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Y. Nardi and A. Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528–549, 2011.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

E. Richard, N. Baskiotis, T. Evgeniou, and N. Vayatis. Link discovery using graph feature tracking. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

E. Richard, S. Gaiffas, and N. Vayatis. Link prediction in graphs with autoregressive features. In *Advances in Neural Information Processing Systems (NIPS)*, 2012a.

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

E. M. Rogers. *Diffusion of Innovations*. London: The Free Press, 1962.

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.

P. Sarkar, D. Chakrabarti, and M. I. Jordan. Nonparametric link prediction in dynamic networks. In *Proceedings of 29th Annual International Conference on Machine Learning*, 2012.

A. Shojaie, S. Basu, and G. Michailidis. Adaptive thresholding for reconstructing regulatory networks from time course gene expression data. *Statistics In Biosciences*, 2011.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*. 2005.

B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Preprint*, 2008.

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.

D. Q. Vu, A. Asuncion, D. Hunter, and P. Smyth. Continuous-time regression models for longitudinal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

K. Zhang, T. Evgeniou, V. Padmanabhan, and E. Richard. Content contributor management and network effects in a ugc environment. *Marketing Science*, 2011.