# Concentration for matrix martingales in continuous time and microscopic activity of social networks

Emmanuel Bacry[1], Stéphane Gaïffas[1], and Jean-Franccois Muzy[1,2]

[1]Centre de Matheématiques Appliquées, École Polytechnique and CNRS
UMR 7641, 91128 Palaiseau, France
[2]Laboratoire Sciences Pour l'Environnement, CNRS, Université de Corse,
UMR 6134, 20250 Corté, France

December 24, 2014

**Abstract**

This paper gives new concentration inequalities for the spectral norm of matrix martingales in continuous time. Both cases of purely discountinuous and continuous martingales are considered. The analysis is based on a new supermartingale property of the trace exponential, based on tools from stochastic calculus. Matrix martingales in continuous time are probabilistic objects that naturally appear for statistical learning of time-dependent systems. We focus here on the the microscopic study of (social) networks, based on self-exciting counting processes, such as the Hawkes process, together with a low-rank prior assumption of the self-exciting component. A consequence of these new concentration inequalities is a push forward of the theoretical analysis of such models.

## 1 Introduction

Matrix concentration inequalities control the deviation of a random matrix around its mean. Until now, results in literature consider the case of sums of independent random matrices, or matrix martingales in discrete time. A first matrix version of the Chernoff bound is given in [1] and was adapted to yield matrix analogues of standard scalar concentration inequalities in [11, 35, 36, 33]. Later, these results were improved in [44, 45], by the use of a theorem due to Lieb [26], which is a deep result closely related to the joint convexity of quantum entropy in physics. See also [29] for a family of sharper and more general results based on the Stein's method. These works contain extensions to random matrices of classical concentration inequalities for sums of independent scalar random variables, such as the Bernstein's inequality for sub-exponential random variables, and Hoeffding inequality for sub-Gaussian random variables. Matrix concentration inequalities have a pletora of applications, in particular in compressed sensing and statistical estimation [23], to develop a simpler analysis of matrix completion [18, 39], for matrix regression [34, 24, 43, 7], for randomized linear algebra [16, 30], and robust PCA [8], which are some examples from a large corpus of works. On the other hand, concentration inequalities for scalar continuous-time martingales are well-known, see for instance [28, 46] and [40] for uniform versions of scalar concentration, and have a large number of applications in high-dimensional statistics [19, 15] among many others.

However, no extension of these results to random matrices, namely continuous-time matrix martingales, is available in literature: the aim of this paper is to provide such results, by combining tools from random matrix theory and from stochastic calculus [28]. The focus of this paper is on concentration inequalities for continuous-time matrix martingales: we provide a matrix version of Bernstein's inequality for purely discountinuous matrix martingales (see Theorem 1 in Section 4) and an Hoeffding's inequality for continuous martingales (see Theorem 2 in Section 5). Matrix martingales in continuous time are probabilistic objects that appear naturally in several problems, in particular in models that describe time-dependent systems, such as models used for the study of social networks at a "microscopic" time scale. The microscopic time scale stands for a recent approach for the study of social-networks: instead of studying a fixed macroscopic graph of connected nodes (users, products) that are linked by edges (click, like, friendship, retweet, etc.), it considers instead the network at a more microscopic scale, by looking at the timestamps of the users actions, in order to recover an implicit connectivity of users. This approach has known a strong development recently: cascade models based on survival analysis are developped in [42, 17, 13], and models based on self-exciting point processes, namely the Hawkes process [20] are given in [12, 4, 49, 48, 27, 14, 4, 21], among others.

The paper is organized as follows. Section 2 introduces the probabilistic background and the main notations used all along the paper. The main purpose of Section 3 is to establish a matrix supermartingale property (Proposition 1) that is essential for obtaining concentration inequalities. Section 4 provides a matrix version of Bernstein's inequality for purely discountinuous matrix martingales whereas Section 5 provides a matrix version of Hoeffding's inequality for continuous martingales. Our new concentration results allows for the theoretical study of such models, in particular for the multivariate Hawkes model, since it gives a sharp control of the noise term, which is in this case a purely discountinuous matrix martingale (see Section 4), as it comes from the compensation of a counting process. We give a quick illustration (see Section 6) of a use of our new concentration results to low-rank multivariate Hawkes processes: low-rank modeling is now a standard approach in collaborative filtering problems [25], where a very popular strategy is to use a penalization technique based on a convex relaxation of the rank, given by the trace normn (see [9, 10] and references mentioned above), which is given by the sum of singular values. For the sake of clarity, some technical proofs are gathered in Appendix A.

# 2 Probabilistic background and Notations

## 2.1 Probabilistic background

We consider a complete probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\{\mathscr{F}_t\}_{t \geq 0}$ of $\sigma$-algebras included in $\mathscr{F}$. Expectation $\mathbb{E}$ is always taken with respect to $\mathbb{P}$. We assume that the stochastic system $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t \geq 0}, \mathbb{P})$ satisfies the *usual* conditions, namely that $\mathscr{F}_0$ is augmented by the $\mathbb{P}$-null sets, and that the filtration is right continuous, namely $\mathscr{F}_t = \cap_{u > t} \mathscr{F}_u$ for any $t \geq 0$. We shall denote $\mathscr{F}_{t-}$ as the smallest $\sigma$-algebra containing all $\mathscr{F}_s$ for $s < t$.

A matrix-valued stochastic processes $\{\boldsymbol{X}_t\}_{t \geq 0}$ is a family of random matrices defined on $(\Omega, \mathscr{F}, \mathbb{P})$. We say that $\{\boldsymbol{X}_t\}_{t \geq 0}$ is adapted if for each $t \geq 0$, all the entries of $\boldsymbol{X}_t$ are $\mathscr{F}_t$-measurable. We say that it is càdlàg if the trajectories on $[0, +\infty]$ of each entries have left limits and are right continuous for all $\omega \in \Omega$. If $\{\boldsymbol{X}_t\}_{t \geq 0}$ is càdlàg, then we define its jump process $\{\Delta \boldsymbol{X}_t\}_{t \geq 0}$ where $\Delta \boldsymbol{X}_t = \boldsymbol{X}_t - \boldsymbol{X}_{t^-}$. We say that $\{\boldsymbol{X}_t\}_{t \geq 0}$ is predictable if all its entries are càdlàg and predictable.

A *matrix martingale* $\{\boldsymbol{M}_t\}_{t \geq 0}$ is a matrix-valued stochastic process with entries that are all martingales. Namely, we assume that for all possible indexes $(i, j)$ of entries, $(\boldsymbol{M}_t)_{i,j}$ is adapted, càdlàg, such that $\mathbb{E}|(\boldsymbol{M}_t)_{i,j}| < +\infty$ for all $t \geq 0$, and that

$$\mathbb{E}[\boldsymbol{M}_t | \mathscr{F}_s] = \boldsymbol{M}_s$$

for any $0 \leq s \leq t$, where the conditional expectation is applied entrywise on $\boldsymbol{M}_t$. More generally, expectations and conditional expectations are always applied entrywise on the entries.

### 2.2 Notations

We denote by $\mathbb{1}$ the column vector with all entries equal to 1 (with size depending on the context).

Let $\boldsymbol{X}$ be a real matrix and $x$ a real vector. The notations $\operatorname{diag}[x]$ stands for the diagonal matrix with diagonal equal to $x$, while if $\boldsymbol{X}$ is a square matrix, $\operatorname{diag}[\boldsymbol{X}]$ stands for the diagonal matrix with diagonal equal to the one of $\boldsymbol{X}$, $\operatorname{tr} \boldsymbol{X}$ stands for the trace of $\boldsymbol{X}$. The operator norm (largest singular value) will be denoted by $\|\boldsymbol{X}\|_{\mathrm{op}}$.

If $\boldsymbol{Y}$ is another real matrix, the notation $\boldsymbol{X} \odot \boldsymbol{Y}$ stands for the entrywise product (Hadamard product) of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with same dimensions, namely $(\boldsymbol{X} \odot \boldsymbol{Y})_{j,k} = (\boldsymbol{X})_{j,k}(\boldsymbol{Y})_{j,k}$. We shall denote by $\boldsymbol{X}^{\odot k}$ the Hadamard power, where each entry of $\boldsymbol{X}^{\odot k}$ is the $k$-th power of the corresponding entry of $\boldsymbol{X}$.

We also denote $\boldsymbol{X}_{\bullet,j}$ for the $j$-th column of $\boldsymbol{X}$ while $\boldsymbol{X}_{j,\bullet}$ stands for the $j$-th row. Moreover, in the case $\boldsymbol{X}$ is a square matrix, for $p \geq 1$, we define the norms

$$\|\boldsymbol{X}\|_{p,\infty} = \max_j \|\boldsymbol{X}_{j,\bullet}\|_p \quad \text{and} \quad \|\boldsymbol{X}\|_{\infty,p} = \max_j \|\boldsymbol{X}_{\bullet,j}\|_p,$$

where $\|\cdot\|_p$ is the vector $\ell_p$-norm.

For a self-adjoint (s.a.) matrix $\boldsymbol{X}$, the largest eigenvalue is denoted $\lambda_{\max}(\boldsymbol{X})$. Moreover, the symbol $\preccurlyeq$ stands for the positive semidefinite (p.s.d.) order on s.a. matrices, namely $\boldsymbol{X} \preccurlyeq \boldsymbol{Y}$ if $\boldsymbol{Y} - \boldsymbol{X}$ is p.s.d.

We shall also denote, when well-defined, $\int_0^t \boldsymbol{X}_s ds$ for the matrix of integrated entries of $\boldsymbol{X}_s$, namely $(\int_0^t \boldsymbol{X}_s ds)_{i,j} = \int_0^t (\boldsymbol{X}_s)_{i,j} ds$, and we denote stochastic integrals matricially, namely $\int_0^t \boldsymbol{X}_s d\boldsymbol{Y}_s$ for the matrix with entries given by the stochastic integral

$$\Big(\int_0^t \boldsymbol{X}_s d\boldsymbol{Y}_s\Big)_{i,j} = \sum_k \int_0^t (\boldsymbol{X}_s)_{i,k} d(\boldsymbol{Y}_s)_{k,j}.$$

## 3 Tools for the study of matrix martingales in continuous time

In this section we give tools for the study of matrix martingales in continuous time. We proceed by steps. The main result of this section, namely Proposition 1, proves that the trace exponential of a matrix martingale is a supermartingale, when properly corrected by terms involving quadratic covariations.

### 3.1 A result from linear algebra

We give first a simple lemma that links the largest eigenvalues of random matrices to the trace exponential of their difference.

**Lemma 1.** *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two s.a. random matrices such that*

$$\operatorname{tr} \mathbb{E}[e^{\boldsymbol{X}-\boldsymbol{Y}}] \leq k$$

*for some $k > 0$. Then, we have*

$$\mathbb{P}[\lambda_{\max}(\boldsymbol{X}) \geq \lambda_{\max}(\boldsymbol{Y}) + x] \leq ke^{-x}$$

*for any $x > 0$.*

*Proof.* Using the fact that [38]

$$\boldsymbol{A} \preccurlyeq \boldsymbol{B} \Rightarrow \operatorname{tr}\exp(\boldsymbol{A}) \leq \operatorname{tr}\exp(\boldsymbol{B}), \quad \text{for any } \boldsymbol{A}, \boldsymbol{B} \text{ s.a.}, \tag{1}$$

along with the fact that $\boldsymbol{Y} \preccurlyeq \lambda_{\max}(\boldsymbol{Y})\boldsymbol{I}$, one has

$$\operatorname{tr}\exp(\boldsymbol{X} - \boldsymbol{Y})\mathbf{1}_E \geq \operatorname{tr}\exp(\boldsymbol{X} - \lambda_{\max}(\boldsymbol{Y})\boldsymbol{I})\mathbf{1}_E,$$

where we set $E = \{\lambda_{\max}(\boldsymbol{X}) \geq \lambda_{\max}(\boldsymbol{Y}) + x\}$. Now, since $\lambda_{\max}(\boldsymbol{M}) \leq \operatorname{tr}\boldsymbol{M}$ for any s.a. matrix $\boldsymbol{M}$, we obtain

$$\begin{aligned}
\operatorname{tr}\exp(\boldsymbol{X} - \boldsymbol{Y})\mathbf{1}_E &\geq \lambda_{\max}(\exp(\boldsymbol{X} - \lambda_{\max}(\boldsymbol{Y})\boldsymbol{I}))\mathbf{1}_E \\
&= \exp(\lambda_{\max}(\boldsymbol{X}) - \lambda_{\max}(\boldsymbol{Y}))\mathbf{1}_E \\
&\geq e^{x}\mathbf{1}_E,
\end{aligned}$$

so that taking the expectation on both sides proves Lemma 1. □

### 3.2 Various definitions and Itô's Lemma for functions of matrices

In this section we describe some classical notions from stochastic calculus [28] and apply them to matrix martingales.

Let $\boldsymbol{M}_t$ be a $p \times q$ matrix whose entries are real valued square-integrable martingales. We denote by $\langle \boldsymbol{M} \rangle_t$ its (entrywise) predictable quadratic variation process, so that $\boldsymbol{M}_t^{\odot 2} - \langle \boldsymbol{M} \rangle_t$ is a martingale. The predictable quadratic covariation of $\boldsymbol{M}_t$ is defined with the help of the vectorization operator vec : $\mathbb{R}^{p \times q} \to \mathbb{R}^{pq}$ which stacks vertically the columns of $\boldsymbol{X}$, namely if $\boldsymbol{X} \in \mathbb{R}^{p \times q}$ then

$$\operatorname{vec}(\boldsymbol{X}) = \left[\boldsymbol{X}_{1,1} \cdots \boldsymbol{X}_{p,1} \boldsymbol{X}_{1,2} \cdots \boldsymbol{X}_{p,2} \cdots \boldsymbol{X}_{1,q} \cdots \boldsymbol{X}_{p,q}\right]^\top.$$

We define indeed the predictable quadratic covariation $\langle \operatorname{vec}\boldsymbol{M} \rangle_t$ of $\boldsymbol{M}_t$ as the $pq \times pq$ matrix with entries

$$(\langle \operatorname{vec}\boldsymbol{M} \rangle_t)_{i,j} = \langle (\operatorname{vec}\boldsymbol{M}_t)_i, (\operatorname{vec}\boldsymbol{M}_t)_j \rangle \tag{2}$$

for $1 \leq i, j \leq pq$, namely such that $\operatorname{vec}(\boldsymbol{M}_t)\operatorname{vec}(\boldsymbol{M}_t)^\top - \langle \operatorname{vec}\boldsymbol{M} \rangle_t$ is a martingale.

The matrix $\boldsymbol{M}_t$ can be decomposed as $\boldsymbol{M}_t = \boldsymbol{M}_t^c + \boldsymbol{M}_t^d$, where $\boldsymbol{M}_t^c$ is a continuous martingale and $\boldsymbol{M}_t^d$ is a purely discountinuous martingale. Its (entrywise) quadratic variation is defined as

$$[\boldsymbol{M}]_t = \langle \boldsymbol{M}^c \rangle_t + \sum_{0 \leq s \leq t} (\Delta \boldsymbol{M}_t)^2, \tag{3}$$

and its quadratic covariation by

$$[\operatorname{vec}\boldsymbol{M}]_t = \langle \operatorname{vec}\boldsymbol{M}^c \rangle_t + \sum_{0 \leq s \leq t} \operatorname{vec}(\Delta \boldsymbol{M}_s)\operatorname{vec}(\Delta \boldsymbol{M}_s)^\top.$$

We say that $\boldsymbol{M}$ is *purely discontinuous* if the process $\langle \operatorname{vec}\boldsymbol{M}^c \rangle_t$ is identically the zero matrix.

An important tool for our proofs is Itô's lemma, that allows to compute the differential $dF(\boldsymbol{M}_t)$ where $F : \mathbb{R}^{p\times q} \to \mathbb{R}$ is a twice diffentiable real function. We denote by $\frac{dF}{d\mathrm{vec}(\boldsymbol{X})}$ the $pq$-dimensional vector such that

$$\Big[\frac{dF}{d\mathrm{vec}(\boldsymbol{X})}\Big]_i = \frac{\partial F}{\partial(\mathrm{vec}\boldsymbol{X})_i} \quad \text{for} \quad 1 \le i \le pq.$$

The second order derivative is the $pq \times pq$ s.a. (self-adjoint) matrix given by

$$\Big[\frac{d^2F}{d\mathrm{vec}(\boldsymbol{X})d\mathrm{vec}(\boldsymbol{X})^\top}\Big]_{i,j} = \frac{\partial^2 F}{\partial(\mathrm{vec}\boldsymbol{X})_i\partial(\mathrm{vec}\boldsymbol{X})_j} \text{ for } 1 \le i,j \le pq.$$

A direct application of the multivariate Itô's Lemma ([28] Theorem 1, p. 118) writes for matrix martingales as follows.

**Lemma 2** (Itô's Lemma)**.** *Let $\{\boldsymbol{X}_t\}_{t\ge0}$ be a $p\times q$ matrix martingale and $F : \mathbb{R}^{p\times q} \to \mathbb{R}$ be a twice continuously differentiable function. Then*

$$\begin{aligned} dF(\boldsymbol{X}_t) = {} & \Big(\frac{dF}{d\mathrm{vec}(\boldsymbol{X})}(\boldsymbol{X}_{t^-})\Big)^\top \mathrm{vec}(d\boldsymbol{X}_t) + \Delta F(\boldsymbol{X}_t) - \Big(\frac{dF}{d\mathrm{vec}\boldsymbol{X}}(\boldsymbol{X}_{t^-})\Big)^\top \mathrm{vec}(\Delta\boldsymbol{X}_t) \\ & + \frac{1}{2}\,\mathrm{tr}\bigg(\Big(\frac{d^2F}{d\mathrm{vec}(\boldsymbol{X})d\mathrm{vec}(\boldsymbol{X})^\top}\Big)^\top d\langle \mathrm{vec}\boldsymbol{X}^c\rangle_t\bigg). \end{aligned}$$

As an application, let us apply Lemma 2 to the function $F(\boldsymbol{X}) = \mathrm{tr}\exp(\boldsymbol{X})$ that acts on the set of s.a. matrices. This result will be of importance for the proof of our results.

**Lemma 3** (Itô's Lemma for the trace exponential)**.** *Let $\{\boldsymbol{X}_t\}$ be a $d\times d$ s.a. matrix martingale. The Itô formula for $F(\boldsymbol{X}_t) = \mathrm{tr}\exp(\boldsymbol{X}_t)$ gives*

$$d(\mathrm{tr}\, e^{\boldsymbol{X}_t}) = \mathrm{tr}(e^{\boldsymbol{X}_{t^-}} d\boldsymbol{X}_t) + \Delta(\mathrm{tr}\, e^{\boldsymbol{X}_t}) - \mathrm{tr}(e^{\boldsymbol{X}_{t^-}}\Delta\boldsymbol{X}_t) + \sum_{i=1}^d \mathrm{tr}(e^{\boldsymbol{X}_{t^-}} d\langle \boldsymbol{X}^c_{\bullet,i}\rangle_t), \tag{4}$$

*where $\langle \boldsymbol{X}^c_{\bullet,i}\rangle_t$ denotes the $d\times d$ predictable quadratic variation of the continuous part of the $i$-th column $(\boldsymbol{X}_t)_{\bullet,i}$ of $\boldsymbol{X}_t$.*

*Proof.* An easy computation gives

$$\mathrm{tr}\, e^{\boldsymbol{X}+\boldsymbol{H}} = \mathrm{tr}\, e^{\boldsymbol{X}} + \mathrm{tr}(e^{\boldsymbol{X}}\boldsymbol{H}) + \mathrm{tr}(e^{\boldsymbol{X}}\boldsymbol{H}^2) + \text{ higher order terms in } \boldsymbol{H}$$

for any s.a. matrices $\boldsymbol{X}$ and $\boldsymbol{H}$. Note that $\mathrm{tr}(e^{\boldsymbol{X}}\boldsymbol{H}) = (\mathrm{vec}\boldsymbol{H})^\top \mathrm{vec}(e^{\boldsymbol{X}})$, and we have from [22] Exercice 25 p. 252 that

$$\mathrm{tr}(e^{\boldsymbol{X}}\boldsymbol{H}^2) = \mathrm{tr}(\boldsymbol{H}e^{\boldsymbol{X}}\boldsymbol{H}) = (\mathrm{vec}\boldsymbol{H})^\top(\boldsymbol{I}\otimes e^{\boldsymbol{X}})(\mathrm{vec}\boldsymbol{H}),$$

where the Kronecker product $\boldsymbol{I}\otimes e^{\boldsymbol{X}}$ stands for the block matrix

$$\boldsymbol{I}\otimes\boldsymbol{Y} = \begin{bmatrix} e^{\boldsymbol{X}} & 0 & \cdots & 0 \\ 0 & e^{\boldsymbol{X}} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & e^{\boldsymbol{X}} \end{bmatrix}.$$

This entails that

$$\frac{d(\mathrm{tr}\, e^{\boldsymbol{X}})}{d\mathrm{vec}(\boldsymbol{X})} = \mathrm{vec}(e^{\boldsymbol{X}}) \quad \text{and} \quad \frac{d^2(\mathrm{tr}\, e^{\boldsymbol{X}})}{d\mathrm{vec}(\boldsymbol{X})d\mathrm{vec}(\boldsymbol{X})^\top} = \boldsymbol{I}\otimes e^{\boldsymbol{X}}.$$

Hence, using Lemma 2 with $F(\boldsymbol{X}) = \operatorname{tr} e^{\boldsymbol{X}}$ we obtain

$$\begin{aligned} d(\operatorname{tr} e^{\boldsymbol{X}_t}) &= \operatorname{vec}(e^{\boldsymbol{X}_{t-}})^\top \operatorname{vec}(d\boldsymbol{X}_t) + \Delta(\operatorname{tr} e^{\boldsymbol{X}_t}) - \operatorname{vec}(e^{\boldsymbol{X}_{t-}})^\top \operatorname{vec}(\Delta \boldsymbol{X}_t) \\ &\quad + \frac{1}{2} \operatorname{tr}\big((\boldsymbol{I} \otimes e^{\boldsymbol{X}_{t-}}) d\langle \operatorname{vec} \boldsymbol{X}^c \rangle_t\big). \end{aligned}$$

Since $\operatorname{vec}(\boldsymbol{Y})^\top \operatorname{vec}(\boldsymbol{Z}) = \operatorname{tr}(\boldsymbol{Y}\boldsymbol{Z})$, one gets

$$d(\operatorname{tr} e^{\boldsymbol{X}_t}) = \operatorname{tr}(e^{\boldsymbol{X}_{t-}} d\boldsymbol{X}_t) + \Delta(\operatorname{tr} e^{\boldsymbol{X}_t}) - \operatorname{tr}(e^{\boldsymbol{X}_{t-}} \Delta \boldsymbol{X}_t) + \frac{1}{2} \operatorname{tr}\big((\boldsymbol{I} \otimes e^{\boldsymbol{X}_{t-}}) d\langle \operatorname{vec} \boldsymbol{X}^c \rangle_t\big).$$

To conclude the proof of Equation (4), it remains to prove that

$$\operatorname{tr}\big((\boldsymbol{I} \otimes e^{\boldsymbol{X}_{t-}}) d\langle \operatorname{vec} \boldsymbol{X}^c \rangle_t\big) = \sum_{i=1}^d \operatorname{tr}(e^{\boldsymbol{X}_{t-}} d\langle \boldsymbol{X}^c_{\bullet,i} \rangle_t).$$

First, let us write

$$d\langle \operatorname{vec} \boldsymbol{X}^c \rangle_t = \sum_{1 \le i,j \le d} \boldsymbol{E}^{i,j} \otimes d\langle \boldsymbol{X}^c_{\bullet,i}, \boldsymbol{X}^c_{\bullet,j} \rangle_t,$$

where $\boldsymbol{E}^{i,j}$ is the $d \times d$ matrix with all entries equal to zero excepted for the $(i,j)$-entry, which is equal to one. Since

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C}) \otimes (\boldsymbol{B}\boldsymbol{D}) \quad \text{and} \quad \operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$$

for any matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ with matching dimensions (see for instance [22]), we have

$$\operatorname{tr}\big((\boldsymbol{I} \otimes e^{\boldsymbol{X}_{t-}}) d\langle \operatorname{vec} \boldsymbol{X}^c \rangle_t\big) = \sum_{1 \le i,j \le d} \operatorname{tr}(\boldsymbol{E}^{i,j}) \operatorname{tr}(e^{\boldsymbol{X}_{t-}} d\langle \boldsymbol{X}^c_{\bullet,i}, \boldsymbol{X}^c_{\bullet,j} \rangle_t) = \sum_{i=1}^d \operatorname{tr}(e^{\boldsymbol{X}_{t-}} d\langle \boldsymbol{X}^c_{\bullet,i}, \boldsymbol{X}^c_{\bullet,i} \rangle_t)$$

since $\operatorname{tr} \boldsymbol{E}^{i,j} = 0$ for $i \neq j$ and 1 otherwise. This concludes the proof of Lemma 4. $\square$

### 3.3 A matrix supermartingale property

The next proposition is a key property that is used below for the proofs of concentration inequalities both for purely discountinuous matrix martingales and for continuous matrix martingales.

**Proposition 1.** *Let $\{\boldsymbol{Y}_t\}_{t \ge 0}$ be a $d \times d$ s.a. matrix martingale such that $\boldsymbol{Y}_0 = \boldsymbol{0}$ and $\mathbb{E} \operatorname{tr} \exp(\boldsymbol{Y}_t) < +\infty$ for all $t \ge 0$, and assume that*

$$\sum_{0 \le s \le t} \mathbb{E}[e^{(\Delta \boldsymbol{Y}_s)_{i,j}} \mathbf{1}_{|(\Delta \boldsymbol{Y}_s)_{i,j}| > 0}] < +\infty \tag{5}$$

*for all $t \ge 0$ and all $1 \le i,j \le d$. Assume that*

$$\boldsymbol{U}_t = \sum_{s \le t} \big(e^{\Delta \boldsymbol{Y}_s} - \Delta \boldsymbol{Y}_s - \boldsymbol{I}\big) \tag{6}$$

*has an entrywise predictable compensator $\langle \boldsymbol{U} \rangle_t$ which is continuous (i.e. $\boldsymbol{U}_t - \langle \boldsymbol{U} \rangle_t$ is a matrix martingale). Then, the process*

$$L_t = \operatorname{tr} \exp\Big(\boldsymbol{Y}_t - \langle \boldsymbol{U} \rangle_t - \frac{1}{2} \sum_{j=1}^d \langle \boldsymbol{Y}^c_{\bullet,j} \rangle_t\Big) \tag{7}$$

*is a supermartingale. In particular, we have $\mathbb{E} L_t \le d$ for any $t \ge 0$.*

Proposition 1 can be understood as an extension to random matrices of the exponential supermartingale property given implicitely in the proof of Lemma 2.2. in [46], of the supermartingale property for multivariate counting processes from [6], see Theorem 2, p. 165 and in Chapter 4.13 from [28].

*Proof of Proposition 1.* Define for short

$$\boldsymbol{X}_t = \boldsymbol{Y}_t - \langle \boldsymbol{U} \rangle_t - \frac{1}{2}\sum_{j=1}^d \langle \boldsymbol{Y}^c_{\bullet,j} \rangle_t.$$

We have that $\langle \boldsymbol{U} \rangle_t$ and $\langle \boldsymbol{Y}^c_{\bullet,j} \rangle_t$ for $j = 1, \ldots, d$ are continuous and finite variation processes, so

$$\langle \mathrm{vec} \boldsymbol{X}^c \rangle = \langle \mathrm{vec} \boldsymbol{Y}^c \rangle \tag{8}$$

and in particular $\langle \boldsymbol{Y}^c_{\bullet,j} \rangle = \langle \boldsymbol{X}^c_{\bullet,j} \rangle$ for any $j = 1, \ldots, d$. Using Lemma 3, one has that for all $t_1 < t_2$:

$$\begin{aligned} L_{t_2} - L_{t_1} &= \int_{t_1}^{t_2} \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\boldsymbol{X}_t) + \sum_{t_1 \le t \le t_2} \big(\Delta(\mathrm{tr}\, e^{\boldsymbol{X}_t}) - \mathrm{tr}(e^{\boldsymbol{X}_{t-}} \Delta \boldsymbol{X}_t)\big) \\ &\quad + \frac{1}{2}\int_{t_1}^{t_2} \sum_{j=1}^d \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\langle \boldsymbol{X}^c_{\bullet,j} \rangle_t) \\ &= \int_{t_1}^{t_2} \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\boldsymbol{Y}_t) - \int_{t_1}^{t_2} \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\langle \boldsymbol{U} \rangle_t) \\ &\quad + \sum_{t_1 \le t \le t_2} \big(\mathrm{tr}(e^{\boldsymbol{X}_{t-} + \Delta \boldsymbol{Y}_t}) - \mathrm{tr}(e^{\boldsymbol{X}_{t-}}) - \mathrm{tr}(e^{\boldsymbol{X}_{t-}} \Delta \boldsymbol{Y}_t)\big), \end{aligned}$$

where we used (8) together with the fact that $\Delta \boldsymbol{X}_t = \Delta \boldsymbol{Y}_t$, since $\langle \boldsymbol{U} \rangle_t$ and $\langle \boldsymbol{Y}^c_{\bullet,j} \rangle_t$ are continuous.

The Golden-Thompson's inequality, see [3], states that $\mathrm{tr}\, e^{\boldsymbol{A}+\boldsymbol{B}} \le \mathrm{tr}(e^{\boldsymbol{A}} e^{\boldsymbol{B}})$ for any s.a. matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Using this inequality we get

$$\begin{aligned} L_{t_2} - L_{t_1} &\le \int_{t_1}^{t_2} \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\boldsymbol{Y}_t) - \int_{t_1}^{t_2} \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\langle \boldsymbol{U} \rangle_t) + \sum_{t_1 \le t \le t_2} \mathrm{tr}\left(e^{\boldsymbol{X}_{t-}} (e^{\Delta \boldsymbol{Y}_t} - \Delta \boldsymbol{Y}_t - \boldsymbol{I})\right) \\ &= \int_{t_1}^{t_2} \mathrm{tr}(e^{\boldsymbol{X}_{t-}} d\boldsymbol{Y}_t) + \int_{t_1}^{t_2} \mathrm{tr}\left(e^{\boldsymbol{X}_{t-}} d(\boldsymbol{U}_t - \langle \boldsymbol{U} \rangle_t)\right). \end{aligned}$$

Since $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t - \langle \boldsymbol{U} \rangle_t$ are matrix martingales and $e^{\boldsymbol{X}_{t-}}$ is a predictable process, it results that $\mathbb{E}[L_{t_2} - L_{t_1} | \mathscr{F}_{t_1}] \le 0$, which proves that $L_t$ is a supermartingale. Using this last inequality with $t_1 = 0$ and $t_2 = t$ gives $\mathbb{E}[L_t] \le d$. This concludes the proof of Proposition 1. □

# 4 Bernstein's inequality for purely discontinuous matrix martingales

In this Section as well as in the next one, we shall consider $\boldsymbol{Z}_t$ to be a $m \times n$ matrix martingale which can be written in the form (using notations introduced in Section 2.2) :

$$\boldsymbol{Z}_t = \int_0^t \boldsymbol{A}_s (\boldsymbol{C}_s \odot d\boldsymbol{M}_s) \boldsymbol{B}_s, \tag{9}$$

where $\{\boldsymbol{A}_t\}$, $\{\boldsymbol{C}_t\}$ and $\{\boldsymbol{B}_t\}$ are predictable random matrices, respectively of size $m \times p$, $p \times q$ and $q \times n$ and where $\{\boldsymbol{M}_t\}_{t \ge 0}$ is a $p \times q$ matrix martingale. Thus the entries of $\boldsymbol{Z}_t$ are given by

$$(\boldsymbol{Z}_t)_{i,j} = \sum_{k=1}^{p} \sum_{l=1}^{q} \int_0^t (\boldsymbol{A}_s)_{i,k} (\boldsymbol{C}_s)_{k,l} (\boldsymbol{B}_s)_{l,j} (d\boldsymbol{M}_s)_{k,l}.$$

In this section, we consider that $\boldsymbol{M}_t$ is a purely discountinuous matrix martingale $\boldsymbol{M}_t = \boldsymbol{M}_t^{\mathrm{dis}}$. Moreover, we assume the following :

**Assumption 1.** *Assume that $\{\boldsymbol{M}_t^{\mathrm{dis}}\}$ is a purely discontinuous matrix-martingales with entries that do not jump at the same time, meaning that $[\mathrm{vec}\boldsymbol{M}^{\mathrm{dis}}]_t$ is a diagonal matrix. Assume also that the entrywise predictable quadratic variation is continuous and admits an intensity, i.e.*

$$\langle \boldsymbol{M}^{\mathrm{dis}} \rangle_t = \int_0^t \boldsymbol{\lambda}_s ds$$

*for a non-negative and predictable intensity process $\{\boldsymbol{\lambda}_t\}_{t \ge 0}$.*

The next Theorem is a Bernstein's concentration inequality for $\|\boldsymbol{Z}_t\|_{\mathrm{op}}$, the operator norm of $\boldsymbol{Z}_t$. Using notations introduced in Section 2.2, we define the matrix

$$\boldsymbol{V}_t = \int_0^t \|\boldsymbol{A}_s\|_{\infty,2}^2 \|\boldsymbol{B}_s\|_{2,\infty}^2 \boldsymbol{W}_s ds \tag{10}$$

where

$$\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{A}_t \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^{-1} \operatorname{diag}\left[(\boldsymbol{C}_t^{\odot 2} \odot \boldsymbol{\lambda}_t)\mathbb{1}\right] \boldsymbol{A}_t^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_t^\top \operatorname{diag}[\boldsymbol{B}_t \boldsymbol{B}_t^\top]^{-1} \operatorname{diag}\left[(\boldsymbol{C}_t^{\odot 2} \odot \boldsymbol{\lambda}_t)^\top \mathbb{1}\right] \boldsymbol{B}_t \end{bmatrix}, \tag{11}$$

and we introduce also

$$b_t = \sup_{s \in [0,t]} \|\boldsymbol{A}_s\|_{\infty,2} \|\boldsymbol{B}_s\|_{2,\infty} \|\boldsymbol{C}_s\|_{\infty}. \tag{12}$$

**Theorem 1.** *Let $\boldsymbol{Z}_t$ be given by (9) and suppose that Assumption 1 holds. Moerover, assume that for some $b > 0$ and for all $\xi \in [0,3]$ one has*

$$\mathbb{E}\left[\sum_{k=0}^{+\infty} \frac{(\xi/b)^{2k}}{(2k)!} \left(\mathrm{tr}(\boldsymbol{Z}_t \boldsymbol{Z}_t^T)^k + \mathrm{tr}(\boldsymbol{Z}_t^T \boldsymbol{Z}_t)^k\right)\right] < \infty \quad \text{and} \quad \mathbb{E}[e^{\xi (\Delta \boldsymbol{Z}_t)_{i,j}/b}] < \infty, \quad \forall i,j$$

*Then, for any $v, x > 0$, the following holds:*

$$\mathbb{P}\left[\|\boldsymbol{Z}_t\|_{\mathrm{op}} \ge \sqrt{2v(x + \log(m+n))} + \frac{b(x + \log(m+n))}{3}, \quad b_t \le b, \quad \lambda_{\max}(\boldsymbol{V}_t) \le v\right] \le e^{-x},$$

*where $\boldsymbol{V}_t$ is given by (10) and $b_t$ is given by (12).*

Theorem 1 is a first result providing a non-commutative version of Bernstein's inequality for continuous time matrix martingales, in the purely discountinous case. This Theorem has several consequences mentioned below, and applications to the modelling of the microscopic activity of social networks using Hawkes processes is given in Section 6 below. Let us first mention some simple particular cases entailed by Theorem 1.

**A first particular case.** Let us assume that when $\boldsymbol{A}_t = \boldsymbol{I}_p$ and $\boldsymbol{B}_t = \boldsymbol{I}_q$ for all $t$, so that $\boldsymbol{Z}_t = \int_0^t \boldsymbol{C}_s \odot d\boldsymbol{M}_s^{\mathrm{dis}}$. We obtain in this case that

$$\boldsymbol{V}_t = \int_0^t \begin{bmatrix} \operatorname{diag}\left[(\boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s)\mathbb{1}\right] & \boldsymbol{0} \\ \boldsymbol{0} & \operatorname{diag}\left[(\boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s)^\top \mathbb{1}\right] \end{bmatrix} ds,$$

so the largest eigenvalue is easily computed as

$$\lambda_{\max}(\boldsymbol{V}_t) = \Big\| \int_0^t \boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s ds \Big\|_{1,\infty} \vee \Big\| \int_0^t \boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s ds \Big\|_{\infty,1},$$

and $b_t = \sup_{s\in[0,t]} \|\boldsymbol{C}_s\|_\infty$. This leads easily to the following corollary.

**Corollary 1.** *Let $\{\boldsymbol{N}_t\}$ be a $p \times q$ matrix whose entries $(\boldsymbol{N}_t)_{i,j}$ are independent inhomogeneous Poisson processes with intensity $(\boldsymbol{\lambda}_t)_{i,j}$. Consider the matrix martingale $\boldsymbol{M}_t = \boldsymbol{N}_t - \boldsymbol{\Lambda}_t$, where $\boldsymbol{\Lambda}_t = \int_0^t \boldsymbol{\lambda}_s ds$ and let $\{\boldsymbol{C}_t\}$ be a $p \times q$ bounded deterministic process. We have that*

$$\Big\| \int_0^t \boldsymbol{C}_s \odot d(\boldsymbol{N}_t - \boldsymbol{\Lambda}_t) \Big\|_{\mathrm{op}} \le \sqrt{2\Big(\Big\| \int_0^t \boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s ds \Big\|_{1,\infty} \vee \Big\| \int_0^t \boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s ds \Big\|_{\infty,1}\Big)(x + \log(p+q))} + \frac{\sup_{s\in[0,t]} \|\boldsymbol{C}_s\|_\infty (x + \log(p+q))}{3}$$

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

**A second particular case.** An even more particular case is given by the case where $\boldsymbol{N}$ is a random matrix with independent Poisson random variables: take $\boldsymbol{C}_t = \boldsymbol{C}$ as the $p \times q$ matrix containing ones, take $t = 1$ and take $\boldsymbol{N}_t = \boldsymbol{N}$ with entries $\boldsymbol{N}_{i,j}$ that are homogeneous Poisson processes on $[0, 1]$, with intensity $\boldsymbol{\lambda}$. Then we get

$$\|\boldsymbol{N} - \boldsymbol{\lambda}\|_{\mathrm{op}} \le \sqrt{2(\|\boldsymbol{\lambda}\|_{1,\infty} \vee \|\boldsymbol{\lambda}\|_{\infty,1})(x + \log(p+q))} + \frac{x + \log(p+q)}{3}.$$

Such a result for random matrices with independent Poisson entries was not, up to the knowledge of the authors, explicitely exhibited in litterature. Note that the variance term depends on the maximum $\ell_1$ norm of rows and columns of $\boldsymbol{\lambda}$, which comes from the subexponentiality of the Poisson distribution. In contrast, in the sub-Gaussian case, Section 5 below, the variance term depends on the maximum $\ell_2$ norm of rows and columns.

**Proof of Theorem 1.** The process $\{\boldsymbol{Z}_t\}$ is not self-adjoint, hence following [44], we will force symmetry in our proofs by extending it in larger dimensions, using the s.a. dilation operator [37]. The s.a. dilation of a matrix $\boldsymbol{X}$ is given by

$$\mathscr{S}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{0} \end{bmatrix}.$$

Let us point out that $\mathscr{S}(\boldsymbol{X})$ is always s.a. and satisfies $\lambda_{\max}(\mathscr{S}(\boldsymbol{X})) = \|\mathscr{S}(\boldsymbol{X})\|_{\mathrm{op}} = \|\boldsymbol{X}\|_{\mathrm{op}}$.

Note that $\mathscr{S}(\boldsymbol{Z}_t)$ is purely discountinuous, so that $\langle \mathscr{S}(\boldsymbol{Z})^c_{\bullet,j}\rangle_t = \boldsymbol{0}$ for any $j$. Recall that we work on events $\{\lambda_{\max}(\boldsymbol{V}_t) \le v\}$ and $\{b_t \le b\}$. Introduce $\boldsymbol{Y}_t = \xi \mathscr{S}(\boldsymbol{Z}_t)/b$. One has

$$\exp(Y_t) = \sum_{k\ge 0} \frac{\xi^k \mathscr{S}(\boldsymbol{Z})^k / b^k}{k!},$$

with

$$\mathscr{S}(\boldsymbol{Z})^{2k} = \begin{bmatrix} (\boldsymbol{Z}\boldsymbol{Z}^\top)^k & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{Z}^\top\boldsymbol{Z})^k \end{bmatrix} \text{ and } \mathscr{S}(\boldsymbol{Z})^{2k+1} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^k \\ \boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^k & \boldsymbol{0} \end{bmatrix}.$$

Thus

$$\operatorname{tr}\exp(Y_t) = \sum_{k\ge 0} \frac{(\xi/b)^{2k}}{(2k)!} (\operatorname{tr}(\boldsymbol{Z}_t\boldsymbol{Z}_t^\top)^k + \operatorname{tr}(\boldsymbol{Z}_t^\top\boldsymbol{Z}_t)^k).$$

Thus, under the assumptions of Theorem 1, one has $\mathbb{E}\operatorname{tr}\exp(Y_t) < +\infty$, i.e., we can use Proposition 1 to obtain that

$$\mathbb{E}\operatorname{tr}\exp\Big(\frac{\xi}{b}\mathscr{S}(\boldsymbol{Z}_t) - \langle \boldsymbol{U}^{\xi/b}\rangle_t\Big) \leq m+n \tag{13}$$

for any $\xi \in [0,3]$, where we introduce for any $u > 0$

$$\boldsymbol{U}_t^u = \sum_{0\leq s\leq t}\Big(e^{u\Delta\mathscr{S}(\boldsymbol{Z}_s)} - u\Delta\mathscr{S}(\boldsymbol{Z}_s) - \boldsymbol{I}\Big). \tag{14}$$

The following technical Proposition (proved in Section A) gives an explicit upper bound, for the p.s.d. order, for $\boldsymbol{U}_t^u$.

**Proposition 2.** *Let $\boldsymbol{Z}_t$ be a martingale given by (9), suppose that Assumption 1 holds and assume that the jumps of all entries of $\boldsymbol{Z}_t$ have subexponential tails, namely we assume that*

$$\sum_{0\leq s\leq t}\mathbb{E}[e^{\xi(\Delta\boldsymbol{Z}_t)_{i,j}}] < +\infty$$

*for some $\xi > 0$ for all $i,j$ and $t \geq 0$. Then, the entrywise predictable quadratic variation $\langle \boldsymbol{U}^{\xi}\rangle_t$ of $\boldsymbol{U}_t^{\xi}$ satisfies*

$$\langle \boldsymbol{U}^{\xi}\rangle_t \preccurlyeq \int_0^t \frac{\phi\big(\xi\|\boldsymbol{A}_s\|_{\infty,2}\|\boldsymbol{B}_s\|_{2,\infty}\|\boldsymbol{C}_s\|_{\infty}\big)}{\|\boldsymbol{C}_s\|_{\infty}^2}\boldsymbol{W}_s ds,$$

*where $\phi(x) = e^x - x - 1$ and where the matrix $\boldsymbol{W}_t$ is given by (11).*

Thus, using together (13), Proposition 2 and Equation (1), one gets that

$$\mathbb{E}\Big[\operatorname{tr}\exp\Big(\frac{\xi}{b}\mathscr{S}(\boldsymbol{Z}_t) - \int_0^t \frac{\phi\big(\xi\|\boldsymbol{A}_s\|_{\infty,2}\|\boldsymbol{B}_s\|_{2,\infty}\|\boldsymbol{C}_s\|_{\infty}/b\big)}{\|\boldsymbol{C}_s\|_{\infty}^2}\boldsymbol{W}_s ds\Big)\Big] \leq m+n,$$

and using this with Lemma 1 entails

$$\mathbb{P}\Big[\frac{\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t))}{b} \geq \frac{1}{\xi}\lambda_{\max}\Big(\int_0^t \frac{\phi\big(\xi\|\boldsymbol{A}_s\|_{\infty,2}\|\boldsymbol{B}_s\|_{2,\infty}\|\boldsymbol{C}_s\|_{\infty}/b\big)}{\|\boldsymbol{C}_s\|_{\infty}^2}\boldsymbol{W}_s ds\Big) + \frac{x}{\xi}\Big] \leq (m+n)e^{-x}.$$

Note that on $\{b_t \leq b\}$ we have $\|\boldsymbol{A}_s\|_{\infty,2}\|\boldsymbol{B}_s\|_{2,\infty}\|\boldsymbol{C}_s\|_{\infty}/b \leq 1$ for any $s \in [0,t]$. The following facts on the function $\phi(x)$ hold true (cf. [31, 19]):

(i) $\phi(xh) \leq h^2\phi(x)$ for any $h \in [0,1]$ and $x > 0$

(ii) $\phi(\xi) \leq \frac{\xi^2}{2(1-\xi/3)}$ for any $\xi \in (0,3)$

(iii) $\min_{\xi\in(0,1/c)}\big(\frac{a\xi}{1-c\xi} + \frac{x}{\xi}\big) = 2\sqrt{ax} + cx$ for any $a,c,x > 0$.

Using successively (i) and (ii), one gets, on $\{b_t \leq b\}\cap\{\lambda_{\max}(\boldsymbol{V}_t) \leq v\}$, that for $\xi \in (0,3)$:

$$\begin{aligned}
&\frac{1}{\xi}\lambda_{\max}\Big(\int_0^t \frac{\phi(\xi\|\boldsymbol{A}_s\|_{\infty,2}\|\boldsymbol{B}_s\|_{2,\infty}\|\boldsymbol{C}_s\|_{\infty}/b)}{\|\boldsymbol{C}_s\|_{\infty}^2}\boldsymbol{W}_s ds\Big) + \frac{x}{\xi}\\
&\quad\leq \frac{\phi(\xi)}{\xi b^2}\lambda_{\max}\Big(\int_0^t \|\boldsymbol{A}_s\|_{\infty,2}\|\boldsymbol{B}_s\|_{2,\infty}\boldsymbol{W}_s ds\Big) + \frac{x}{\xi}\\
&\quad= \frac{\phi(\xi)}{\xi b^2}\lambda_{\max}(\boldsymbol{V}_t) + \frac{x}{\xi}\\
&\quad\leq \frac{\xi v}{2b^2(1-\xi/3)} + \frac{x}{\xi},
\end{aligned}$$

where we recall that $\boldsymbol{V}_t$ is given by (10). This gives

$$\mathbb{P}\left[\frac{\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t))}{b} \geq \frac{\xi v}{2b^2(1-\xi/3)} + \frac{x}{\xi}, \quad b_t \leq b, \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v\right] \leq (m+n)e^{-x},$$

for any $\xi \in (0,3)$. Now, by optimizing over $\xi$ using (iii) (with $a = v/2b^2$ and $c = 1/3$), one obtains

$$\mathbb{P}\left[\frac{\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t))}{b} \geq \frac{\sqrt{2vx}}{b} + \frac{x}{3}, \quad b_t \leq b, \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v\right] \leq (m+n)e^{-x}.$$

Since $\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t)) = \|\mathscr{S}(\boldsymbol{Z}_t)\|_{\text{op}}$, this concludes the proof of Theorem 1. □

# 5 Hoeffding's inequality for continuous matrix martingales

In this section, we study the matrix-martingale $\boldsymbol{Z}_t$ given by (9) with $\boldsymbol{M}_t = \boldsymbol{M}_t^{\text{con}}$ satisfying the following assumption[1] :

**Assumption 2.** *Assume that $\{\boldsymbol{M}_t^{\text{con}}\}$ is a continuous matrix-martingale whose entrywise predictable quadratic variation matrix reads:*

$$\langle \boldsymbol{M}^{\text{con}}\rangle_t = t\boldsymbol{I} \ .$$

The next Theorem provides the analog of theorem 1 in the continuous case $\boldsymbol{M}_t = \boldsymbol{M}_t^{\text{con}}$, i.e, a Hoeffding's concentration inequality for $\|\boldsymbol{Z}_t\|_{\text{op}}$, the operator norm of $\boldsymbol{Z}_t$. Thus, following the same lines as in the previous section, let us define

$$\boldsymbol{V}_t = \int_0^t \|\boldsymbol{A}_s\|_{\infty,2}^2 \|\boldsymbol{B}_s\|_{2,\infty}^2 \boldsymbol{W}_s ds \tag{15}$$

where

$$\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{A}_t \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^{-1} \operatorname{diag}\left[\boldsymbol{C}_t^{\odot 2}\mathbb{1}\right]\boldsymbol{A}_t^\top & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_t^\top \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^{-1} \operatorname{diag}\left[(\boldsymbol{C}_t^{\odot 2})^\top \mathbb{1}\right]\boldsymbol{B}_t \end{bmatrix}, \tag{16}$$

Let us point out that the so-defined quantities $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ correspond to previous definitions (10) and (11) introduced in the case of the purely discontinuous martingale $\boldsymbol{M}_t = \boldsymbol{M}_t^{\text{dis}}$ where $\boldsymbol{\lambda}_t$ has been replaced by the matrix whose entries are all equal to 1.

**Theorem 2.** *Let $\boldsymbol{Z}_t$ be given by (9) and suppose that Assumption 2 holds. Assume that $\operatorname{tr}\mathbb{E}[e^{\xi\mathscr{S}(\boldsymbol{Z}_t)}] < \infty$ and $\mathbb{E}[e^{\xi(\Delta\mathscr{S}(\boldsymbol{Z}_t))_{i,j}}] < \infty$ for any $i,j$ and any $\xi > 0$. Then, for any $v, x > 0$, the following holds:*

$$\mathbb{P}\left[\|\boldsymbol{Z}_t\|_{\text{op}} \geq \sqrt{2v(x+\ln(m+n))} \ , \quad \lambda_{\max}(\boldsymbol{V}_s) \leq v\right] \leq e^{-x}.$$

*where $\boldsymbol{V}_t$ is given by (15).*

**Proof of Theorem 2.** The proof follows the same lines as the proof of Theorem 1. We consider as before the s.a. dilation $\mathscr{S}(\boldsymbol{Z}_t)$ of $\boldsymbol{Z}_t$ (see Eq. (4)) and apply Proposition 1 with $\boldsymbol{Y}_t = \xi\mathscr{S}(\boldsymbol{Z}_t)$

[1] Let us recall that, in the previous section, we considered that $\boldsymbol{M}_t = \boldsymbol{M}_t^{\text{dis}}$ was purely discontinuous

and $d = m+n$. Since $\boldsymbol{Z}_t$ is a continuous martingale , we have $\boldsymbol{U}_t = \boldsymbol{0}$ (cf. (6)), so that $\langle \boldsymbol{U}\rangle_t = \boldsymbol{0}$ and we have $\langle \boldsymbol{Z}\rangle_t^c = \langle \boldsymbol{Z}\rangle_t$. So, Proposition 1 gives

$$\mathbb{E}\Big[\operatorname{tr}\exp\Big(\xi\mathscr{S}(\boldsymbol{Z}_t) - \frac{1}{2}\sum_{j=1}^{m+n}\xi^2\langle\mathscr{S}(\boldsymbol{Z})_{\bullet,j}\rangle_t\Big)\Big] \leq m+n. \tag{17}$$

From the definition of the dilation operator $\mathscr{S}$, it can be directly shown that:

$$\sum_{j=1}^{m+n}\langle\mathscr{S}(\boldsymbol{Z})_{\bullet,j}\rangle_t = \begin{bmatrix}\sum_{j=1}^{n}\langle\boldsymbol{Z}_{\bullet,j}\rangle_t & \boldsymbol{0}_{m,n}\\ \boldsymbol{0}_{n,m} & \sum_{j=1}^{m}\langle\boldsymbol{Z}_{j,\bullet}\rangle_t\end{bmatrix}$$

where $\langle\boldsymbol{Z}_{\bullet,j}\rangle_t$ (resp. $\langle\boldsymbol{Z}_{\bullet,j}\rangle_t$) is the $m\times m$ (resp. $n\times n$) matrix of the quadratic variation of the j-th column (resp. row) of $Z_t$.

Since $[\boldsymbol{M}^{\mathrm{con}}]_t = \langle\boldsymbol{M}^{\mathrm{con}}\rangle_t = t\boldsymbol{I}$, we have (for the sake of clarity, we omit the subscript $t$ in the matrices):

$$\begin{aligned}\sum_{j=1}^{n}(d\langle\boldsymbol{Z}_{\bullet,j}\rangle_t)_{kl} &= \sum_{j=1}^{n} d[\boldsymbol{Z}_{k,j},\boldsymbol{Z}_{l,j}]\\ &= \sum_{j=1}^{n}\sum_{a=1}^{p}\sum_{b=1}^{q}\boldsymbol{A}_{k,a}\boldsymbol{A}_{l,a}\boldsymbol{C}_{a,b}^2\boldsymbol{B}_{b,j}^2 dt\\ &= \sum_{a=1}^{p}\boldsymbol{A}_{k,a}\boldsymbol{A}_{l,a}\sum_{b=1}^{q}\boldsymbol{C}_{a,b}^2\sum_{j=1}^{n}\boldsymbol{B}_{b,j}^2 dt,\end{aligned}$$

which gives in a matrix form

$$\sum_{j=1}^{n} d\langle\boldsymbol{Z}_{\bullet,j}\rangle_t = \boldsymbol{A}\operatorname{diag}\Big[\boldsymbol{C}^{\odot 2}\Big(\operatorname{diag}[\boldsymbol{B}\boldsymbol{B}^\top]\mathbb{1}\Big)\Big]\boldsymbol{A}^\top dt$$

By inserting $\operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]\operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]^{-1}$ in the r.h.s. and noticing that $\operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t] \preccurlyeq \|\boldsymbol{A}_t\|_{\infty,2}\boldsymbol{I}$ and $\operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top] \preccurlyeq \|\boldsymbol{B}_t\|_{2,\infty}\boldsymbol{I}$, one gets:

$$\sum_{j=1}^{n} d\langle\boldsymbol{Z}_{\bullet,j}\rangle_t \preccurlyeq \|\boldsymbol{A}_t\|_{\infty,2}\|\boldsymbol{B}_t\|_{2,\infty}\boldsymbol{A}_t\operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]^{-1}\operatorname{diag}\big[\boldsymbol{C}_t^{\odot 2}\mathbb{1}\big]\boldsymbol{A}_t^\top$$

One can easily prove in the same way that

$$\sum_{j=1}^{m} d\langle\boldsymbol{Z}_{j,\bullet}\rangle_t = \boldsymbol{B}^\top\operatorname{diag}\Big[(\boldsymbol{C}^{\odot 2})^\top\Big(\operatorname{diag}[\boldsymbol{A}^\top\boldsymbol{A}]\mathbb{1}\Big)\Big]\boldsymbol{B}\,dt$$

and therefore

$$\sum_{j=1}^{m} d\langle\boldsymbol{Z}_{j,\bullet}\rangle_t \preccurlyeq \|\boldsymbol{A}_t\|_{\infty,2}\|\boldsymbol{B}_t\|_{2,\infty}\boldsymbol{B}_t^\top\operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^{-1}\operatorname{diag}\big[(\boldsymbol{C}_t^{\odot 2})^\top\mathbb{1}\big]\boldsymbol{B}_t$$

Thus,

$$\sum_{j=1}^{m+n}\langle\mathscr{S}(\boldsymbol{Z})_{\bullet,j}\rangle_t^c \preccurlyeq \boldsymbol{V}_t,$$

where $\boldsymbol{V}_t$ is given by (15). From (17), it results

$$\mathbb{E}[\operatorname{tr}\exp\Big(\xi\mathscr{S}(\boldsymbol{Z}_t) - \frac{1}{2}\xi^2\boldsymbol{V}_t\Big)] \leq m+n.$$

Then, using Lemma 1, one gets

$$\mathbb{P}\Big[\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t)) \geq \frac{\xi}{2}\lambda_{\max}(\boldsymbol{V}_t) + \frac{x}{\xi}\Big] \leq (m+n)e^{-x}. \tag{18}$$

On the event $\{\lambda_{\max}(\boldsymbol{V}_t) \leq v\}$, one gets

$$\mathbb{P}\Big[\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t)) \geq \frac{\xi}{2}v + \frac{x}{\xi}\ , \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v\Big] \leq (m+n)e^{-x}. \tag{19}$$

Optimizing on $\xi$, we apply this last result for $\xi = \sqrt{\frac{2x}{v}}$ and get

$$\mathbb{P}\Big[\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t)) \geq \sqrt{2xv}\ , \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v\Big] \leq (m+n)e^{-x}. \tag{20}$$

Since $\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t)) = \|\mathscr{S}(\boldsymbol{Z}_t)\|_{\mathrm{op}}$, this concludes the proof of Theorem 2. □

# 6 Illustrations of the newly obtained concentration inequalities

In this section we describe a statistical learning application that can be studied using our new concentration inequality given by Theorem 1, namely the use of low-rank Hawkes processes. We only explain here how matrix martingales naturally appear in such a context, a precise description of our statistical learning results are beyond the scope of the present paper, which focuses mainly on probabilistic aspects.

## 6.1 Low-rank multivariate Hawkes model

Consider a finite network with $d$ nodes (each node corresponding to a user in a social network for instance). For each node $j \in \{1, \ldots, d\}$, we observe the timestamps $\{t_{j,1}, t_{j,2}, \ldots\}$ of actions of node $j$ on the network (a message, a click, etc.). To each node $j$ is associated a counting process $N_j(t) = \sum_{i \geq 1} \mathbf{1}_{t_{j,i} \leq t}$ and we consider the $d$-dimensional counting process $N_t = [N_1(t) \ \cdots \ N_d(t)]^\top \in \mathbb{N}^d$, for $t \geq 0$. We observe this process for $t \in [0, T]$. The multivariate Hawkes model assumes that each $N_j$ has an intensity $\lambda_{j,\theta}$ given by

$$\lambda_{j,\theta}(t) = \mu_j + \int_{(0,t)} \sum_{j'=1}^{d} a_{j,j'} h_{j,j'}(t-s) dN_{j'}(s),$$

where $\theta = (\mu, \boldsymbol{A})$ with $\mu = [\mu_1, \ldots, \mu_d]^\top$ and $\boldsymbol{A} = [a_{j,j'}]_{1 \leq j,j' \leq d}$, and where $\mu_j \geq 0$ is a baseline intensity, where $a_{j,j'} \geq 0$ is a coefficient that quantifies the influence of $j'$ on $j$, and $h_{j,j'} : \mathbb{R}^+ \to \mathbb{R}^+$ are decay functions that account for the decay of influence between pairs of nodes in the network. A typical choice for $h_{j,j'}$ is the exponential kernel, i.e., $h_{j,j'}(t) = e^{-\beta_{j,j'} t}$, where $\beta_{j,j'} > 0$ is a decay coefficient. To keep this example simple, we consider these functions fixed and known. The parameter of interest is the *self-excitement* matrix $\boldsymbol{A}$, which can be viewed as a weighted asymmetrical adjacency matrix of connectivity between nodes.

Such parametrized Hawkes model is particularly revelant for the modelization of the "microscopic" activity of (social) networks, and has been considered recently a lot in literature for this kind of application (see [12, 4, 49, 48, 27, 14, 4, 21, 32], among others), with a particular emphasis on [19] that gives first theoretical results for the Lasso used with Hawkes processes with an application to neurobiology. The main point is that this simple autoregressive structure of the intensity allows to capture the direct influence of a user on to the others, based on the recurrency and the patterns of their actions, by separating the intensity into a baseline and a self-exciting component, hence allowing to filter exogeneity in the estimation of users' influences on each others.

We want to produce an estimation procedure of $\theta = (\mu, \boldsymbol{A})$ based on data from $\{N_t : t \in [0, T]\}$. A way of achieving this is to minimize the least-squares functional [19], given by

$$R_T(\theta) = \|\lambda_\theta\|_T^2 - \frac{2}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t) dN_j(t) \tag{21}$$

with respect to $\theta$, where $\|\lambda_\theta\|_T^2 = \frac{1}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t)^2 dt$ is the norm associated with the inner product

$$\langle \lambda_\theta, \lambda_{\theta'} \rangle_T = \frac{1}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t) \lambda_{j,\theta'}(t) dt. \tag{22}$$

This least-squares function is very natural, and comes from the empirical risk minimization principle [47, 31, 23, 2]: assuming that $N$ has an unknown ground truth intensity $\lambda^*$ (not necessarily following the Hawkes model), we have easily, using Doob-Meyer's decomposition that

$$\mathbb{E}[R_T(\theta)] = \mathbb{E}\|\lambda_\theta\|_T^2 - 2\mathbb{E}\langle \lambda_\theta, \lambda^* \rangle_T = \mathbb{E}\|\lambda_\theta - \lambda^*\|_T^2 - \|\lambda^*\|_T,$$

so that we expect a minimum $\hat\theta$ of $R_T(\theta)$ to lead to a good estimation $\lambda_{\hat\theta}$ of $\lambda^*$.

In addition to this goodness-of-fit criterion, we need to use a penalization that allows to reduce the dimensionality of the model. In particular, we want to reduce the dimensionality of $\boldsymbol{A}$, based on the prior assumption that latent factors explain the connectivity of users in the network. This leads to a low-rank assumption on $\boldsymbol{A}$, which is commonly used in collaborative filtering and matrix completion techniques [41]. A standard way of inducing low-rank by convex penalization is based on the trace or nuclear-norm [10]. We therefore consider this penalization here, by considering a procedure given by minimization of the least-squares penalized by the trace norm:

$$\hat\theta = (\hat\mu, \hat{\boldsymbol{A}}) \in \underset{\theta = (\mu, \boldsymbol{A}) \in \mathbb{R}_+^d \times \mathbb{R}_+^{d \times d}}{\operatorname{argmin}} \big\{ R_T(\theta) + \tau \|\boldsymbol{A}\|_* \big\}, \tag{23}$$

with $\tau > 0$ a smoothing parameter that balances goodness-of-fit and penalization, and $\|\boldsymbol{A}\|_* = \sum_{j=1}^d \sigma_j(\boldsymbol{A})$ where $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_d(\boldsymbol{A})$ are the singular values of $\boldsymbol{A}$. Note that a penalization combining the $\ell_1$ and the trace norm is considered in [49] for Hawkes processes, but focuses only on deriving a convex optimization algorithm for this problem, and do not propose theoretical guarantees.

In order to prove a rate of convergence for $\lambda_{\hat\theta}$, one exploits the fact that $\hat\theta$ is a minimum of the convex minimization problem (23). Let us endow the space $\mathbb{R}^d \times \mathbb{R}^{d \times d}$ by the inner product $\langle \theta, \theta' \rangle = \langle \mu, \mu' \rangle + \langle \boldsymbol{A}, \boldsymbol{A}' \rangle$ where $\theta = (\mu, \boldsymbol{A})$ and $\theta' = (\mu', \boldsymbol{A}')$ with $\langle \mu, \mu' \rangle = \mu^\top \mu'$ and $\langle \boldsymbol{A}, \boldsymbol{A}' \rangle = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A}')$. The parameter $\hat\theta$ satisfies first order conditions [5, 23], namely we can find $\hat{\boldsymbol{A}}_\partial \in \partial \|\hat{\boldsymbol{A}}\|_*$ such that

$$\langle \nabla R_T(\hat\theta), \hat\theta - \theta \rangle + \tau \langle \hat{\boldsymbol{A}}_\partial, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle \leq 0 \tag{24}$$

for any $\theta = (\mu, \boldsymbol{A}) \in \mathbb{R}_+^d \times \mathbb{R}_+^{d \times d}$, where $\nabla f(\theta)$ stands for the gradient of a differentiable function $f$ at $\theta$ and $\partial g(\theta)$ stands for the subdifferential of a convex function $g$ at $\theta$. Using the fact that

$dN_j(t) = \lambda^*(t)dt + dM_j(t)$ for a martingale $M_j$ by the Doob-Meyer decomposition, an easy computation gives

$$\langle \nabla R_T(\hat\theta), \hat\theta - \theta \rangle = 2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda^* \rangle_T - \frac{2}{T}\sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t))dM_j(t),$$

and the parallelogram identity gives

$$2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda^* \rangle_T = \|\lambda_{\hat\theta} - \lambda^*\|_T^2 + \|\lambda_{\hat\theta} - \lambda_\theta\|_T^2 - \|\lambda_\theta - \lambda^*\|_T^2.$$

So, an upper bound on

$$\frac{2}{T}\sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t))dM_j(t) - \tau\langle \hat\theta_\partial, \hat\theta - \theta\rangle$$

leads to an upper bound on the estimation error $\|\lambda_{\hat\theta} - \lambda^*\|_T^2$. We can decompose the noise term in the following way:

$$\frac{2}{T}\sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t))dM_j(t) = \frac{2}{T}\langle \hat\mu - \mu, \bar M_T\rangle + \frac{2}{T}\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z}_T\rangle,$$

where $\bar M_T = [\int_0^T dM_1(t) \ \cdots \ \int_0^T dM_d(t)]^\top$and where $\boldsymbol{Z}_t$ is the matrix martingale with entries

$$(\boldsymbol{Z}_t)_{j,k} = \int_0^t \int_{(0,s)} h_{j,k}(s-u)dN_k(u)dM_j(s), \tag{25}$$

that can be written in the following way:

$$\boldsymbol{Z}_t = \int_0^t \operatorname{diag}[dM_s]\boldsymbol{H}_s,$$

where $\boldsymbol{H}_t$ is the predictable process with entries $(\boldsymbol{H}_t)_{j,j'} = \int_{(0,t)} h_{j,j'}(t-s)dN_{j'}(s)$. Note that $\boldsymbol{Z}_t$ fits in the structure of the matrix martingale from Equation (9), for the choice of predictable processes $\boldsymbol{A}_t = \boldsymbol{I}_d$, $\boldsymbol{B}_t = \boldsymbol{H}_t$ and $\boldsymbol{C}_t = \boldsymbol{I}_d$. Hölder's inequality gives $\frac{2}{T}\langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z}\rangle \le \frac{2}{T}\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_* \|\boldsymbol{Z}_t\|_{\mathrm{op}}$, so that it can be proved that a control of $\|\boldsymbol{Z}_t\|_{\mathrm{op}}$ together with the use of the structure of a subgradient $\hat{\boldsymbol{A}}_\partial$ leads to an upper bound on the estimation error $\|\lambda_{\hat\theta} - \lambda^*\|_T^2$.

This motivates the use of Theorem 1, which leads to a concentration inequality for $\|\boldsymbol{Z}_t\|_{\mathrm{op}}$ that writes as follows:

$$\mathbb{P}\Big[\|\boldsymbol{Z}_t\|_{\mathrm{op}} \ge \sqrt{2v(x + \log(m+n))} + \frac{b(x+\log(m+n))}{3}, \quad b_t \le b, \quad \lambda_{\max}(\boldsymbol{V}_t) \le v\Big] \le e^{-x},$$

for any $v, x, b > 0$, where

$$\boldsymbol{V}_t = \int_0^t \|\boldsymbol{H}_s\|_{2,\infty}^2 \begin{bmatrix} \operatorname{diag}[\lambda_s^*] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_s^\top \operatorname{diag}[\boldsymbol{H}_s\boldsymbol{H}_s^\top]^{-1}\operatorname{diag}[\lambda_s^*]\boldsymbol{H}_s \end{bmatrix} ds$$

and $b_t = \sup_{s\in[0,t]} \|\boldsymbol{H}_s\|_{2,\infty}$. Note that this control of the noise term $\|\boldsymbol{Z}_t\|_{\mathrm{op}}$ gives a choice for the smoothing parameter $\tau$ that compensates the noise. However this choice depends on the matrix $\boldsymbol{V}_t$ that itself depends on the unknown intensity $\lambda^*$. Actually, an "empirical" version of Theorem 1 can be proved, that involves an observable variance term $\hat{\boldsymbol{V}}_t$ instead of $\boldsymbol{V}_t$. Such a result involves extra technicalities and will be developed elsewhere.

## 6.2 Cascading multivariate Hawkes model

Now, we give another example of modelling based on Hawkes processes, where time events are naturally organized in *cascades* $c = 1, \ldots, C$. A cascade is a subset of time events, among all time events of every nodes $j = 1, \ldots, d$, that have in common keywords, themes, tags, or have a direct connection through retweets or forwarded messages. For each node $j$ and cascade $c$ we observe time stamps $\{t_{j,c,1}, t_{j,c,2}, \ldots\}$ of actions done by user $j$ in the cascade $c$, and we consider the corresponding counting processes $\boldsymbol{N}_{j,c}(t) = \sum_{i \geq 1} \mathbf{1}_{t_{j,c,i} \leq t}$, and introduce the $d \times C$ counting process $\boldsymbol{N}_t = [\boldsymbol{N}_{j,c}(t)]_{1 \leq j \leq d, 1 \leq c \leq C}$, that we observed for $t \in [0, T]$. Once again, we are interested by the global connectivity between users, hence the connectivity matrix $\boldsymbol{A}$ does not depend on the cascade, but we exploit the timestamps patterns along the cascades to reconstruct $\boldsymbol{A}$. On the other hand, we assume that the reactivity of a node $j$ depends on the cascade $c$, since reactivity of a user typically depends on the themes of messages. This motivates the use of the intensity

$$\boldsymbol{\lambda}_{j,c,\theta}(t) = \boldsymbol{\mu}_{j,c} + \int_{(0,t)} \sum_{j'=1}^{d} a_{j,j'} h_{j',c}(t-s) d\boldsymbol{N}_{j',c}(s),$$

where $\mu_{j,c}$ is the baseline activity of node $j$ in cascade $c$, $a_{j,j'}$ is influence of node $j'$ one node $j$ and $h_{j,c} : \mathbb{R}^+ \to \mathbb{R}^+$ are decay functions that account for the decay of influence of events from cascade $c$ on node $j$ (see also the previous example).

The same can be done as in Section 6.1, and a least-squares goodness-of-fit functional for this model is given by

$$R_T(\theta) = \|\boldsymbol{\lambda}_\theta\|_T^2 - \frac{2}{T} \sum_{j=1}^{d} \sum_{c=1}^{C} \int_{[0,T]} \boldsymbol{\lambda}_{j,c,\theta}(t) d\boldsymbol{N}_{j,c}(t), \tag{26}$$

where this time $\|\boldsymbol{\lambda}_\theta\|_T^2 = \frac{1}{T} \sum_{j=1}^{d} \sum_{c=1}^{C} \int_{[0,T]} \boldsymbol{\lambda}_{j,c,\theta}(t)^2 dt$ and

$$\langle \boldsymbol{\lambda}_\theta, \boldsymbol{\lambda}_{\theta'} \rangle_T = \frac{1}{T} \sum_{j=1}^{d} \sum_{c=1}^{C} \int_{[0,T]} \boldsymbol{\lambda}_{j,c,\theta}(t) \boldsymbol{\lambda}_{j,c,\theta'}(t) dt,$$

where $\theta = (\boldsymbol{\mu}, \boldsymbol{A})$ with $\boldsymbol{\mu} = [\mu_{j,c}]$ and $\boldsymbol{A} = [a_{j,j'}]$. The same computations as in the first example (see Section 6.1) can be done, and we end-up with a matrix-martingale noise term given by

$$\frac{2}{T} \sum_{j=1}^{d} \sum_{c=1}^{C} \int_0^T (\boldsymbol{\lambda}_{j,c,\hat{\theta}}(t) - \boldsymbol{\lambda}_{j,c,\theta}(t)) d\boldsymbol{M}_{j,c}(t) = \frac{2}{T} \langle \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \bar{\boldsymbol{M}}_T \rangle + \frac{2}{T} \langle \hat{\boldsymbol{A}} - \boldsymbol{A}, \boldsymbol{Z}_T \rangle,$$

where $d\boldsymbol{M}_t = d\boldsymbol{N}_t - \boldsymbol{\lambda}_t dt$ is a matrix martingale, $\bar{\boldsymbol{M}}_T = \int_0^T d\boldsymbol{M}_t$ and $\boldsymbol{Z}_T$ is the matrix martingale with entries

$$(\boldsymbol{Z}_T)_{j,j'} = \int_0^T \sum_{c=1}^{C} \int_{(0,t)} h_{j',c}(t-s) d\boldsymbol{N}_{j',c}(s) d\boldsymbol{M}_{j,c}(t), \tag{27}$$

that can be written in the following way:

$$\boldsymbol{Z}_T = \int_0^T d\boldsymbol{M}_t \boldsymbol{H}_t,$$

where $\boldsymbol{H}_t$ is the predictable process with entries $(\boldsymbol{H}_t)_{c,j'} = \int_{(0,t)} h_{j',c}(t-s) d\boldsymbol{N}_{j',c}(s)$. This time, Theorem 1 gives the following concentration:

$$\mathbb{P}\Big[ \|\boldsymbol{Z}_t\|_{\mathrm{op}} \geq \sqrt{2v(x + \log(m+n))} + \frac{b(x + \log(m+n))}{3}, \quad b_t \leq b, \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v \Big] \leq e^{-x},$$

for any $v, x, b > 0$, where

$$\boldsymbol{V}_t = \int_0^t \|\boldsymbol{H}_s\|_{2,\infty}^2 \begin{bmatrix} \operatorname{diag}[\boldsymbol{\lambda}_s^* \mathbb{1}] & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_s^\top \operatorname{diag}[\boldsymbol{H}_s \boldsymbol{H}_s^\top]^{-1} \operatorname{diag}[(\boldsymbol{\lambda}_s^*)^\top \mathbb{1}] \boldsymbol{H}_s \end{bmatrix} ds$$

and $b_t = \sup_{s \in [0,t]} \|\boldsymbol{H}_s\|_{2,\infty}$. Note that the block diagonal structure of $\boldsymbol{V}_t$ gives

$$\begin{aligned} \lambda_{\max}(\boldsymbol{V}_t) = \max_{j=1,\dots,d} \int_0^t \|\boldsymbol{H}_s\|_{2,\infty}^2 \|\boldsymbol{\lambda}_{j,\bullet}(s)\|_1 ds \\ \bigvee \lambda_{\max}\Big( \int_0^t \|\boldsymbol{H}_s\|_{2,\infty}^2 \boldsymbol{H}_s^\top \operatorname{diag}[\boldsymbol{H}_s \boldsymbol{H}_s^\top]^{-1} \operatorname{diag}[(\boldsymbol{\lambda}_s^*)^\top \mathbb{1}] \boldsymbol{H}_s ds \Big). \end{aligned}$$

This variance term depends on the maximum $\ell_2$ norms of rows of $\boldsymbol{H}_t$ and $\ell_1$ norms of rows of $\boldsymbol{\lambda}_t$ for the first term, as well as on the $\ell_1$ norm of columns of $\boldsymbol{\lambda}_t$ via the second term. This algebraic structure comes from the fact that $\boldsymbol{H}_t$ is on the right hand side of the infinitesimal $d\boldsymbol{M}_t$. Other model structures shall lead to other matrix martingales, whose operator norms can be controlled thanks to Theorem 1.

## Acknowledgements

This research benefited from the support of the "Chair Markets in Transition", under the aegis of "Louis Bachelier Finance and Sustainable Growth" laboratory, a joint initiative of École Polytechnique, Université d'Évry Val d'Essonne and Fédération Bancaire Francaise.

## A Proof of Proposition 2

The proof of Proposition 2 is organized in the following subsections. We set (see Eq. 14)

$$\boldsymbol{U}_t^\xi = \sum_{0 \le s \le t} \boldsymbol{S}_s^\xi \quad \text{where} \quad \boldsymbol{S}_t^\xi = e^{\xi \Delta \mathscr{S}(\boldsymbol{Z}_t)} - \xi \Delta \mathscr{S}(\boldsymbol{Z}_t) - \boldsymbol{I}.$$

To simplify notations, we shall drop the index $t$ in time varying quantities, when there is no ambiguity. We recall that our aim is to compute an upper bound for the p.s.d. order of $\langle \boldsymbol{U}^\xi \rangle_t$. The proof is decomposed in the following steps.

### A.1 Computation of $\boldsymbol{S}_t^\xi$

We have

$$\exp(\xi \Delta \mathscr{S}(\boldsymbol{Z})) = \sum_{k \ge 0} \frac{\xi^k \mathscr{S}(\Delta \boldsymbol{Z})^k}{k!}, \tag{28}$$

and it is easy to check that

$$\begin{aligned} \mathscr{S}(\Delta \boldsymbol{Z})^{2k} &= \begin{bmatrix} (\Delta \boldsymbol{Z} \Delta \boldsymbol{Z}^\top)^k & \mathbf{0} \\ \mathbf{0} & (\Delta \boldsymbol{Z}^\top \Delta \boldsymbol{Z})^k \end{bmatrix} \\ \text{and} \quad \mathscr{S}(\Delta \boldsymbol{Z})^{2k+1} &= \begin{bmatrix} \mathbf{0} & \Delta \boldsymbol{Z} (\Delta \boldsymbol{Z}^\top \Delta \boldsymbol{Z})^k \\ \Delta \boldsymbol{Z}^\top (\Delta \boldsymbol{Z} \Delta \boldsymbol{Z}^\top)^k & \mathbf{0} \end{bmatrix}. \end{aligned}$$

Since $(\Delta \boldsymbol{Z} (\Delta \boldsymbol{Z}^\top \Delta \boldsymbol{Z})^k)^\top = \Delta \boldsymbol{Z}^\top (\Delta \boldsymbol{Z} \Delta \boldsymbol{Z}^\top)^k$, we need to compute three terms: $(\Delta \boldsymbol{Z} \Delta \boldsymbol{Z}^\top)^k$, $(\Delta \boldsymbol{Z}^\top \Delta \boldsymbol{Z})^k$ and $\Delta \boldsymbol{Z}^\top (\Delta \boldsymbol{Z} \Delta \boldsymbol{Z}^\top)^k$.

### A.1.1 Computation of $(\Delta\boldsymbol{Z}\Delta\boldsymbol{Z}^\top)^k$ and $(\Delta\boldsymbol{Z}^\top\Delta\boldsymbol{Z})^k$ for $k \geq 1$

One has

$$\begin{aligned}(\Delta\boldsymbol{Z}\Delta\boldsymbol{Z}^\top)^k &= \Big(\boldsymbol{A}(\boldsymbol{C}\odot\Delta\boldsymbol{M})\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\boldsymbol{A}^\top\Big)^k \\ &= \boldsymbol{A}\Big((\boldsymbol{C}\odot\Delta\boldsymbol{M})\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\boldsymbol{A}^\top\boldsymbol{A}\Big)^{k-1} \\ &\qquad \times(\boldsymbol{C}\odot\Delta\boldsymbol{M})\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\boldsymbol{A}^\top.\end{aligned}$$

From Assumption 1, one has that a.s., the entries of $\boldsymbol{M}$ cannot jump at the same time, hence

$$\begin{aligned}&\Delta\boldsymbol{M}_{i_1,j_1}\times\cdots\times\Delta\boldsymbol{M}_{i_m,j_m} \\ &\qquad = \begin{cases}\Delta\boldsymbol{M}_{i_1,j_1} & \text{if } i_1=\cdots=i_m \text{ and } j_1=\cdots=j_m \\ \boldsymbol{0} & \text{otherwise}\end{cases}\end{aligned} \tag{29}$$

a.s. for any $m \geq 2$ and any indexes $i_k \in \{1,\dots,p\}$ and $j_k \in \{1,\dots,q\}$. This entails that the $(i,j)$ element of $(\Delta\boldsymbol{Z}\Delta\boldsymbol{Z}^\top)^k$ is given by

$$\begin{aligned}&\sum_{a=1}^p\sum_{b=1}^q \Delta\boldsymbol{M}_{a,b}\boldsymbol{A}_{i,a}\boldsymbol{C}_{a,b}^{2k}(\boldsymbol{B}\boldsymbol{B}^\top)_{b,b}^k(\boldsymbol{A}^\top\boldsymbol{A})_{a,a}^{k-1}\boldsymbol{A}_{j,a} \\ &\quad = \sum_{a=1}^p \boldsymbol{A}_{i,a}\boldsymbol{A}_{j,a}(\boldsymbol{A}^\top\boldsymbol{A})_{a,a}^{k-1}\sum_{b=1}^q \boldsymbol{C}_{a,b}^{2k}\Delta\boldsymbol{M}_{a,b}(\boldsymbol{B}\boldsymbol{B}^\top)_{b,b}^k \\ &\quad = \sum_{a=1}^p \boldsymbol{A}_{i,a}\boldsymbol{A}_{j,a}(\boldsymbol{A}^\top\boldsymbol{A})_{a,a}^{k-1}\Big[\operatorname{diag}[(\boldsymbol{C}^{\odot 2k}\odot\Delta\boldsymbol{M})\operatorname{diag}[\boldsymbol{B}\boldsymbol{B}^\top]^k\mathbb{1}]\Big]_{a,a},\end{aligned}$$

a.s., where $\mathbb{1}$ is the column vector of size $q$ with all entries equal to 1 and where we recall that $\operatorname{diag}[\mathbf{V}]$ is the diagonal matrix whose entries on the diagonal are the entries of the vector $\mathbf{V}$. We assumed that $\operatorname{diag}(\boldsymbol{A}^\top\boldsymbol{A})$ and $\operatorname{diag}(\boldsymbol{B}\boldsymbol{B}^\top)$ are invertible, so we can write

$$(\Delta\boldsymbol{Z}_t\Delta\boldsymbol{Z}_t^\top)^k = \boldsymbol{A}_t\operatorname{diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t)^{-1/2}\boldsymbol{D}_{1,t}^{(k)}\operatorname{diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t)^{-1/2}\boldsymbol{A}_t^\top \tag{30}$$

a.s. where $\boldsymbol{D}_{1,t}^{(k)}$ is the diagonal matrix given by

$$\boldsymbol{D}_{1,t}^{(k)} = \operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]^k\operatorname{diag}[(\boldsymbol{C}_t^{\odot 2k}\odot\Delta\boldsymbol{M}_t)\operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^k\mathbb{1}].$$

Similar computations lead to

$$(\Delta\boldsymbol{Z}_t^\top\Delta\boldsymbol{Z}_t)^k = \boldsymbol{B}_t^\top\operatorname{diag}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1/2}\boldsymbol{D}_{2,t}^{(k)}\operatorname{diag}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1/2}\boldsymbol{B}_t$$

a.s., where $\boldsymbol{D}_{2,t}^{(k)}$ is the diagonal matrix given by

$$\boldsymbol{D}_{2,t}^{(k)} = \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^k\operatorname{diag}[(\boldsymbol{C}_t^{\odot 2k}\odot\Delta\boldsymbol{M}_t)^\top\operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]^k\mathbb{1}].$$

### A.1.2 Computation of $\Delta\boldsymbol{Z}^\top(\Delta\boldsymbol{Z}\Delta\boldsymbol{Z}^\top)^k$ for $k \geq 1$

Following the same lines as above, we obtain

$$\begin{aligned}\Delta\boldsymbol{Z}^\top(\Delta\boldsymbol{Z}\Delta\boldsymbol{Z}^\top)^k &= \boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\boldsymbol{A}^\top\Big(\boldsymbol{A}(\boldsymbol{C}\odot\Delta\boldsymbol{M})\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\boldsymbol{A}^\top\Big)^k \\ &= \boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\Big(\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{C}\odot\Delta\boldsymbol{M})\boldsymbol{B}\boldsymbol{B}^\top(\boldsymbol{C}\odot\Delta\boldsymbol{M})^\top\Big)^k\boldsymbol{A}^\top.\end{aligned}$$

Thus, using again (29), we obtain that the $(i,j)$ element of $\Delta\boldsymbol{Z}^\top(\Delta\boldsymbol{Z}\Delta\boldsymbol{Z}^\top)^k$ is given a.s. by

$$\begin{aligned}
&\sum_{a=1}^{p}\sum_{b=1}^{q}\Delta\boldsymbol{M}_{a,b}\boldsymbol{B}_{b,i}\boldsymbol{C}_{a,b}^{2k+1}(\boldsymbol{A}^\top\boldsymbol{A})_{a,a}^k(\boldsymbol{B}\boldsymbol{B}^\top)_{b,b}^k\boldsymbol{A}_{j,a}\\
&\quad=\sum_{b=1}^{q}\boldsymbol{B}_{b,i}(\boldsymbol{B}\boldsymbol{B}^\top)_{b,b}^k\sum_{a=1}^{p}\Delta\boldsymbol{M}_{a,b}\boldsymbol{C}_{a,b}^{2k+1}(\boldsymbol{A}^\top\boldsymbol{A})_{a,a}^k\boldsymbol{A}_{j,a}\\
&\quad=\sum_{b=1}^{q}\boldsymbol{B}_{b,i}(\boldsymbol{B}\boldsymbol{B}^\top)_{b,b}^k\left((\boldsymbol{C}^{\odot 2k+1}\odot\Delta\boldsymbol{M})^\top\operatorname{diag}[\boldsymbol{A}^\top\boldsymbol{A}]^k\boldsymbol{A}^\top\right)_{b,j}.
\end{aligned}$$

Thus

$$\Delta\boldsymbol{Z}_t^\top(\Delta\boldsymbol{Z}_t\Delta\boldsymbol{Z}_t^\top)^k=\boldsymbol{B}_t^\top\operatorname{diag}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1/2}\boldsymbol{H}_t^{(k)}\operatorname{diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t)^{-1/2}\boldsymbol{A}_t^\top \tag{31}$$

a.s., where

$$\boldsymbol{H}_t^{(k)}=\operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^{k+1/2}(\boldsymbol{C}_t^{\odot 2k+1}\odot\Delta\boldsymbol{M}_t)^\top\operatorname{diag}[\boldsymbol{A}_t^\top\boldsymbol{A}_t]^{k+1/2}.$$

### A.1.3 Final computation of $\boldsymbol{S}_t^\xi$

We can write now

$$\begin{aligned}
\boldsymbol{S}_t^\xi&=\sum_{k\geq 1}\left(\frac{\xi^{2k}}{(2k)!}\mathscr{S}(\Delta\boldsymbol{Z}_t)^{2k}+\frac{\xi^{2k+1}}{(2k+1)!}\mathscr{S}(\Delta\boldsymbol{Z}_t)^{2k+1}\right)\\
&=\sum_{k\geq 1}\begin{bmatrix}\frac{\xi^{2k}}{(2k)!}(\Delta\boldsymbol{Z}_t\Delta\boldsymbol{Z}_t^\top)^k & \frac{\xi^{2k+1}}{(2k+1)!}\Delta\boldsymbol{Z}_t(\Delta\boldsymbol{Z}_t^\top\Delta\boldsymbol{Z}_t)^{k+1}\\ \frac{\xi^{2k+1}}{(2k+1)!}\Delta\boldsymbol{Z}_t^\top(\Delta\boldsymbol{Z}_t\Delta\boldsymbol{Z}_t^\top)^{k+1} & \frac{\xi^{2k}}{(2k)!}(\Delta\boldsymbol{Z}_t^\top\Delta\boldsymbol{Z}_t)^k\end{bmatrix}.
\end{aligned}$$

Using Equation (30) and (31), one gets

$$\boldsymbol{S}_t^\xi=\boldsymbol{P}_t{}^\top\boldsymbol{R}_t\boldsymbol{P}_t \tag{32}$$

a.s., where

$$\boldsymbol{R}_t=\sum_{k\geq 1}\boldsymbol{R}_t^{(k)}\quad\text{with}\quad\boldsymbol{R}_t^{(k)}=\begin{bmatrix}\frac{\xi^{2k}}{(2k)!}\boldsymbol{D}_{1,t}^{(k)} & \frac{\xi^{2k+1}}{(2k+1)!}(\boldsymbol{H}_t^{(k+1)})^\top\\ \frac{\xi^{2k+1}}{(2k+1)!}\boldsymbol{H}_t^{(k+1)} & \frac{\xi^{2k}}{(2k)!}\boldsymbol{D}_{2,t}^{(k)}\end{bmatrix} \tag{33}$$

and

$$\boldsymbol{P}_t=\begin{bmatrix}\operatorname{diag}(\boldsymbol{A}_t^\top\boldsymbol{A}_t)^{-1/2}\boldsymbol{A}_t^\top & \boldsymbol{0}\\ \boldsymbol{0} & \operatorname{diag}(\boldsymbol{B}_t\boldsymbol{B}_t^\top)^{-1/2}\boldsymbol{B}_t\end{bmatrix}.$$

## A.2 Computation of $\langle\boldsymbol{U}^\xi\rangle_t$

Note that $\boldsymbol{U}_t^\xi$ depends linearly on the jumps $\Delta\boldsymbol{M}_t$. So, the predicatable quadratic variation $\langle\boldsymbol{U}^\xi\rangle_t$ of $\boldsymbol{U}_t^\xi$ is now easily obtained. Recalling that

$$\boldsymbol{U}_t^\xi=\sum_{0\leq s\leq t}\boldsymbol{S}_s^\xi=\sum_{0\leq s\leq t}\boldsymbol{P}_s^\top\boldsymbol{R}_s\boldsymbol{P}_s,$$

we obtain

$$\langle\boldsymbol{U}^\xi\rangle_t=\int_0^t\boldsymbol{P}_s^\top\boldsymbol{R}_{\lambda,s}\boldsymbol{P}_s ds, \tag{34}$$

where

$$\boldsymbol{R}_{\lambda,t} = \sum_{k\geq 1} \boldsymbol{R}^{(k)}_{\lambda,t} \quad \text{with} \quad \boldsymbol{R}^{(k)}_{\lambda,t} = \begin{bmatrix} \frac{\xi^{2k}}{(2k)!}\boldsymbol{D}^{(k)}_{1,\lambda,t} & \frac{\xi^{2k+1}}{(2k+1)!}(\boldsymbol{H}^{(k+1)}_{\lambda,t})^\top \\ \frac{\xi^{2k+1}}{(2k+1)!}\boldsymbol{H}^{(k+1)}_{\lambda,t} & \frac{\xi^{2k}}{(2k)!}\boldsymbol{D}^{(k)}_{2,\lambda,t} \end{bmatrix}, \tag{35}$$

and

$$\begin{aligned}
\boldsymbol{D}^{(k)}_{1,\lambda,t} &= \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^k \operatorname{diag}[(\boldsymbol{C}_t^{\odot 2k} \odot \boldsymbol{\lambda}_t) \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^k \mathbb{1}] \\
\boldsymbol{D}^{(k)}_{2,\lambda,t} &= \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^k \operatorname{diag}[(\boldsymbol{C}_t^{\odot 2k} \odot \boldsymbol{\lambda}_t)^\top \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^k \mathbb{1}] \\
\boldsymbol{H}^{(k+1)}_{\lambda,t} &= \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^{k+1/2} (\boldsymbol{C}_t^{\odot 2k+1} \odot \boldsymbol{\lambda}_t)^\top \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^{k+1/2}.
\end{aligned}$$

### A.3 An upper bound on $\langle \boldsymbol{U}^\xi \rangle_t$ for the p.s.d. order

Recalling (34) and using the fact that [38]

$$\boldsymbol{A} \preccurlyeq \boldsymbol{B} \;\Rightarrow\; \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}^\top \preccurlyeq \boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^\top \tag{36}$$

for any real matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ (with compatible dimensions). along with the fact that p.s.d. order is preserved by integration, we have that

$$\boldsymbol{R}_{\lambda,s} \preccurlyeq \boldsymbol{T}_{\lambda,s} \;\text{ for all } s \geq 0 \;\Rightarrow\; \langle \boldsymbol{U}^\xi \rangle_t \preccurlyeq \int_0^t \boldsymbol{P}_s^\top \boldsymbol{T}_{\lambda,s} \boldsymbol{P}_s ds \quad \text{for all } t. \tag{37}$$

We therefore construct an upper bound for the $(p+q)\times(p+q)$ matrix $\boldsymbol{R}_{\lambda,t}$. In order to do so we shall apply the next simple Lemma.

**Lemma 4.** *Let $\boldsymbol{X}$ be a $n \times n$ s.a. matrix with non-negative entries. Let $\mathscr{D}(\boldsymbol{X})$ be defined as the diagonal matrix with entries*

$$\mathscr{D}(\boldsymbol{X})_{i,i} = \sum_{j=1}^n \boldsymbol{X}_{i,j}.$$

*Then, one has $\boldsymbol{X} \preccurlyeq \mathscr{D}(\boldsymbol{X})$.*

*Proof.* This comes from the fact that $\boldsymbol{X}' = \mathscr{D}(\boldsymbol{X}) - \boldsymbol{X}$ is a diagonally dominant (i.e., $|\boldsymbol{X}'_{i,i}| \geq \sum_{j\neq i} |\boldsymbol{X}'_{i,j}|$ for all $i$). Since the diagonal entries of $\boldsymbol{X}'$ are non-negative, we know that $\boldsymbol{X}' \succcurlyeq 0$. □

For $i \in \{1, \ldots, p\}$, we have

$$(\mathscr{D}(\boldsymbol{R}_{\lambda,t}))_{i,i} = \sum_{k\geq 1} \Big( \frac{\xi^{2k}}{(2k)!} (\boldsymbol{D}^{(k)}_{1,\lambda,t})_{i,i} + \frac{\xi^{2k+1}}{(2k+1)!} \sum_{j=1}^q (\boldsymbol{H}^{(k+1)}_{\lambda,t})_{j,i} \Big).$$

Since

$$\begin{aligned}
\big(\boldsymbol{D}^{(k)}_{1,\lambda,t}\big)_{i,i} &= \Big( \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^k \operatorname{diag}\big[(\boldsymbol{C}_t^{\odot 2k} \odot \boldsymbol{\lambda}_t) \operatorname{diag}[\boldsymbol{B}_t\boldsymbol{B}_t^\top]^k \mathbb{1}\big] \Big)_{i,i} \\
&= \|(\boldsymbol{A}_t)_{\bullet,i}\|_2^{2k} \sum_{j=1}^q (\boldsymbol{C}_t)_{i,j}^{2k} \, (\boldsymbol{\lambda}_t)_{i,j} \, \|(\boldsymbol{B}_t)_{j,\bullet}\|_2^{2k} \\
&\leq \|\boldsymbol{A}_t\|_{2,\infty}^{2k} \|\boldsymbol{B}_t\|_{\infty,2}^{2k} \sum_{j=1}^q (\boldsymbol{C}_t)_{i,j}^{2k} (\boldsymbol{\lambda}_t)_{i,j}
\end{aligned}$$

and

$$\sum_{j=1}^{q} (\boldsymbol{H}_{\lambda,t}^{(k)})_{j,i} = \sum_{j=1}^{q} (\operatorname{diag}[\boldsymbol{B}_t \boldsymbol{B}_t^\top]^{k+1/2} \big(\boldsymbol{C}_t^{\odot 2k+1} \odot \boldsymbol{\lambda}_t\big) \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^{k+1/2})_{i,j}$$
$$\leq \|\boldsymbol{A}_t\|_{2,\infty}^{2k+1} \|\boldsymbol{B}_s\|_{\infty,2}^{2k+1} \sum_{j=1}^{q} (\boldsymbol{C}_t)_{i,j}^{2k+1} (\boldsymbol{\lambda}_t)_{i,j},$$

we have for $i \in \{1, \ldots, p\}$,

$$\begin{aligned}
\mathscr{D}(\boldsymbol{R}_{\lambda,t})_{i,i} &\leq \sum_{k \geq 2} \frac{\xi^k}{k!} \|\boldsymbol{A}_t\|_{2,\infty}^{k} \|\boldsymbol{B}_t\|_{\infty,2}^{k} \sum_{j=1}^{q} (\boldsymbol{C}_t)_{i,j}^{k} (\boldsymbol{\lambda}_t)_{i,j} \\
&\leq \sum_{k \geq 2} \frac{\xi^k}{k!} \|\boldsymbol{A}_t\|_{2,\infty}^{k} \|\boldsymbol{B}_t\|_{\infty,2}^{k} \|\boldsymbol{C}_t\|_{\infty}^{k-2} \sum_{j=1}^{q} (\boldsymbol{C}_t)_{i,j}^{2} (\boldsymbol{\lambda}_t)_{i,j} \\
&= \frac{\phi\big(\xi \|\boldsymbol{A}_t\|_{2,\infty} \|\boldsymbol{B}_t\|_{\infty,2} \|\boldsymbol{C}_t\|_{\infty}\big)}{\|\boldsymbol{C}_t\|_{\infty}^2} \big((\boldsymbol{C}_t^{\odot 2} \odot \boldsymbol{\lambda}_t)\mathbb{1}\big)_{i,i},
\end{aligned}$$

where we recall that $\phi(x) = e^x - 1 - x$. Similar computations lead, for $i \in \{p+1, \ldots, p+q\}$, to

$$\mathscr{D}(\boldsymbol{R}_{\lambda,t})_{i,i} \leq \frac{\phi\big(\xi \|\boldsymbol{A}_t\|_{2,\infty} \|\boldsymbol{B}_t\|_{\infty,2} \|\boldsymbol{C}_t\|_{\infty}\big)}{\|\boldsymbol{C}_t\|_{\infty}^2} \big((\boldsymbol{C}_t^{\odot 2} \odot \boldsymbol{\lambda}_t)^\top \mathbb{1}\big)_{i,i}.$$

Recalling the definition (35) of $\boldsymbol{R}_{\lambda,t}$, we note that it is a s.a. matrix with non-negative entries. So, using Lemma 4, one gets that

$$\boldsymbol{R}_{\lambda,t} \preccurlyeq \mathscr{D}(\boldsymbol{R}_{\lambda,t}) \preccurlyeq \frac{\phi\big(\xi \|\boldsymbol{A}_t\|_{2,\infty} \|\boldsymbol{B}_t\|_{\infty,2} \|\boldsymbol{C}_t\|_{\infty}\big)}{\|\boldsymbol{C}_t\|_{\infty}^2} \begin{bmatrix} \operatorname{diag}[(\boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_t)\mathbb{1}] & \boldsymbol{0} \\ \boldsymbol{0} & \operatorname{diag}[(\boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_t)^\top \mathbb{1}] \end{bmatrix}.$$

Hence, using (37), we finally get

$$\langle \boldsymbol{U}^\xi \rangle_t \preccurlyeq \int_0^t \frac{\phi\big(\xi \|\boldsymbol{A}_s\|_{\infty,2} \|\boldsymbol{B}_s\|_{2,\infty} \|\boldsymbol{C}_s\|_{\infty}\big)}{\|\boldsymbol{C}_s\|_{\infty}^2} \boldsymbol{W}_s ds,$$

where $\boldsymbol{W}_t$ is given by (11). This concludes the proof of Proposition 2

## References


[1] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *Information Theory, IEEE Transactions on*, 48(3):569–579, 2002.

[2] Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[3] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer, 1997.

[4] Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

[5] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics: Martingale Dynamics*. Springer, 1981.

[7] Florentina Bunea, Yiyuan She, Marten H Wegkamp, et al. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

[8] E. J. Candès, X. Li, Y. Ma, and W. John. Robust principal component analysis? *Journal of ACM*, 8:1–37, 2009.

[9] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 12(51):4203–4215, 2004.

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

[11] Demetres Christofides and Klas Markström. Expansion properties of random cayley graphs and vertex transitive graphs via matrix martingales. *Random Structures & Algorithms*, 32(1):88–100, 2008.

[12] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41), 2008.

[13] N. Daneshmand, M. Rodriguez, L. Song, and B. Schölkpof. Estimating diffusion network structure: Recovery conditions, sample complexity, and a soft-thresholding algorithm. *ICML*, 2014.

[14] Christopher DuBois, Carter Butts, and Padhraic Smyth. Stochastic blockmodeling of relational event dynamics. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.

[15] Stéphane Gaïffas and Agathe Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

[16] Alex Gittens. The spectral norm error of the naive nystrom extension. *arXiv preprint arXiv:1110.5305*, 2011.

[17] M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Modeling information propagation with survival theory. *ICML*, 2013.

[18] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57, 2011.

[19] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. Technical report, Arvix preprint, 2012.

[20] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[21] Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274. ACM, 2013.

[22] Charles R Johnson and Roger A Horn. Topics in matrix analysis. *Cambridge University*, 1991.

[23] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

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

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

[26] Elliott H Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973.

[27] Scott W Linderman and Ryan P Adams. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*, 2014.

[28] R. S. Liptser and A. N. Shiryayev. *Theory of martingales*. Springer, 1989.

[29] Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, Joel A Tropp, et al. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014.

[30] Lester W Mackey, Michael I Jordan, and Ameet Talwalkar. Divide-and-conquer matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1134–1142, 2011.

[31] Pascal Massart and Jean Picard. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

[32] I. Mastromatteo, E. Bacry, and J.F. Muzy. Linear processes in high-dimension: phase space and critical properties. *arXiv preprint arXiv:1412.6998*, 2014.

[33] Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *arXiv preprint arXiv:1112.5448*, 2011.

[34] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[35] Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[36] Roberto Imbuzeiro Oliveira. Sums of random hermitian matrices and an inequality by rudelson. *Electron. Commun. Probab*, 15(203-212):26, 2010.

[37] Vern I Paulsen, B Bollobás, W Fulton, A Katok, F Kirwan, and P Sarnak. *Completely bounded maps and operator algebras*, volume 78. Cambridge Univ Press, 2002.

[38] Dénes Petz. A survey of certain trace inequalities. *Functional analysis and operator theory*, 30:287–298, 1994.

[39] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[40] Patricia Reynaud-Bouret. Compensator and exponential inequalities for some suprema of counting processes. *Statistics & probability letters*, 76(14):1514–1521, 2006.

[41] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[42] M. Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. *ICML*, 2011.

[43] Angelika Rohde, Alexandre B Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

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

[45] Joel A Tropp. Freedman's inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270, 2011.

[46] Sara van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, pages 1779–1801, 1995.

[47] Sara Van De Geer and SA Van De Geer. *Empirical Processes in M-estimation*, volume 105. Cambridge university press Cambridge, 2000.

[48] S.-H. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In *ICML*, 2013.

[49] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. *AISTATS*, 2013.