Detecting hidden states in stochastic dynamical systems

Rayan Succar
Department of Mechanical and Aerospace Engineering, Tandon School of Engineering, New York University, Brooklyn, New York, 11201, USA
Center for Urban Science and Progress, New York University, Brooklyn, New York, 11201, USA

Alain Boldini
Department of Mechanical and Aerospace Engineering, Tandon School of Engineering, New York University, Brooklyn, New York, 11201, USA
Center for Urban Science and Progress, New York University, Brooklyn, New York, 11201, USA
Department of Mechanical Engineering, New York Institute of Technology, Old Westbury, New York, 11568, USA

Maurizio Porfiri*
Department of Mechanical and Aerospace Engineering, Tandon School of Engineering, New York University, Brooklyn, New York, 11201, USA
Department of Biomedical Engineering, Tandon School of Engineering, New York University, Brooklyn, New York, 11201, USA
*Center for Urban Science and Progress, New York University, Brooklyn, New York, 11201, USA**
(Dated: July 31, 2024)

Inferring the number of states of a stochastic system from partial measurements is a fundamental problem in physics, for which methodological tools remain scarce. It is difficult to distinguish the stochastic dynamical states from measurements, deceiving us into incorrect models and flawed understanding of natural phenomena. Here, we propose a model-free, statistical framework, grounded in network and control theory, to estimate the number of states of a stochastic system from perceptible dynamics. The framework extends previous techniques for deterministic systems, based on the rank of ancillary matrices. We show applications of our approach to a variety of physics domains, such as statistical mechanics, biophysics, physical chemistry, and epidemiology.

I. INTRODUCTION

The study of network dynamical systems has fascinated scientists for centuries [1, 2]. From climate networks [3] to fish schooling [4] and human mobility [5], physicists, mathematicians, biologists, and social scientists have sought to describe and understand the complex, emergent behaviors that arise from interactions of individual units.

The advent of large-scale data acquisition systems has allowed the development of new techniques for the study of network dynamical systems [6]. Starting from the time-series of the dynamics of individual units, these tools offer a potent lens through which one can reveal and detail their interactions. However, these techniques often rely on the assumption that the number of states of the considered system is known, a condition seldom verified in practice. For example, neuroscientists can reconstruct inter-neuronal connections in the brain, but may not have an accurate estimate of the number of neurons involved [7].

Several methodological advancements have been made in recent years to address the problem of inferring the number of states of a system from measurements on a subset of its units. Haehne *et al.* [8], Porfiri [9], and Tang *et al.* [10] proposed the assembly of representative matrices from the time-series of the unforced dynamics of perceptible nodes, whose rank would be related to the size of the largest observable component of the system. Tyloo and Delabays [11] reconstructed the size of a network system by probing it with sinusoidal inputs and measuring the response of selected units. These approaches are exclusively applicable to deterministic dynamics. Only recent efforts have started leveraging noise-induced stochasticity to estimate the size of an otherwise deterministic system [12], for very specific collective dynamics that are only seen in some real-world systems.

While the dichotomy between a deterministic or a random world still exists in theoretical physics, randomness is unavoidable, as the failed attempt to establish orderliness in celestial mechanics by Henri Poincaré taught us. Since the work of Poincaré, randomness has been embraced in many fields of physics [13–16], from statistical mechanics [17] to quantum mechanics [18] and nuclear physics [19], which led to greater understanding of the physical word. For example, the classical experiments of Perrin [20] that led to the estimation of the Avogadro's number involved the Brownian motion of a particle suspended in a liquid, building on the theories of Einstein [21] and Smoluchowski [22].

Here, we propose a novel statistical framework to infer the number of states of systems that are stochastic by nature and have hidden states. Our approach makes very general assumptions about the dynamics of the system

* mporfiri@nyu.edu; R.S. and A.B. contributed equally to this work.

as a hidden Markov chain [23, 24], which are satisfied by many real-world stochastic phenomena, from thermodynamics to epidemiology [25, 26]. We extend previous work on deterministic dynamics [8–10] by relating the number of states of a stochastic system to the rank of a detection matrix, assembled from realizations of the system. Due to the noisy nature of the matrix, we design a statistical test to correctly reconstruct its rank. The proposed approach can be used in denoising any matrix corrupted by noise with known structure, considerably overperforming state-of-the-art techniques for matrix denoising [27]. The proposed framework combines and builds on techniques from statistics, control theory, and perturbation theory to contribute to the fields of general Markovian processes and stochastic network systems. We first demonstrate our methodology with the classical Ehrenfest urn model of diffusion in statistical mechanics [28], and then we show applications to other domains of physics, including biophysics, physical chemistry, and epidemiology.

The remainder of this paper is organized as follows. In Section II, we present the mathematical formulation of the detection matrix and introduce a novel statistical test designed to identify the accurate rank of a noisy matrix. Moving on to practical applications, Section III A demonstrates our approach in the context of the classical Ehrenfest urn model problem. In Section III B, we showcase the application of our method to unveil hidden behavioral states of bacteria through their swimming motion. Therein, we also validate the noise model employed in our statistical test. In Section III C, we highlight how our approach can effectively detect the number of chemical compounds in an enzyme reaction using partial measurements of enzyme states. We also provide numerical evidence supporting the applicability of our approach to perturbed Markov chains. In Section III D, we illustrate the potential of our approach to identify hidden exposed states in an epidemic, even in scenarios where some properties of Markov chains are violated. Section III E presents a numerical comparison between our proposed statistical test for denoising matrices and state-of-the-art method developed by Gavish and Donoho [27]. Finally, Section IV brings the paper to a close, summarizing limitations and key findings.

II. THEORY

Let X_k be a first-order, time-homogeneous Markov chain, where k indicates the discrete time-step ($k = 1, \dots, K$). The Markov chain has $N \in \mathbb{N}_+$ states, such that X_k has a finite alphabet $s_1, \cdots, s_N$. We define the probability mass function (pmf) of X_k as $\pi_k = \left[\Pr(X_k = s_1), \cdots, \Pr(X_k = s_N)\right]^{\mathrm{T}}$. The time evolution of the pmf is governed by

$$\pi_{k+1} = \mathbf{P}^{\mathrm{T}} \pi_k, \tag{1}$$

where $\mathbf{P} \in [0,1]^{N \times N}$ is the row-stochastic transition matrix.

We do not have access to realizations of X_k, but only to an output stochastic process Y_k with $M < N$ states and finite alphabet $\bar{s}_1, \cdots, \bar{s}_M$. The perceptible dynamics Y_k is not necessarily a first-order Markov process, but its probability depends on X_k only, such that $\Pr(Y_k \mid X_k, X_{k-1}, \dots, X_1) = \Pr(Y_k \mid X_k)$. We define the pmf of Y_k as $\phi_k = \left[\Pr(Y_k = \bar{s}_1), \cdots, \Pr(Y_k = \bar{s}_M)\right]^{\mathrm{T}}$. Without loss of generality, we consider cases in which the realizations of Y_k are deterministically related to those of X_k. In this vein, we establish

$$\phi_k = \mathbf{C} \pi_k, \tag{2}$$

where $\mathbf{C} \in \{0,1\}^{M \times N}$ is a column-stochastic, Boolean emission matrix, with each column containing only one "1". We note that the emission matrix can generally have real elements ($\mathbf{C} \in [0,1]^{M \times N}$) without any modification to the approach (the proof in the Methods does not impose any conditions on $\mathbf{C}$), so that the realizations of Y_k can be stochastically related to those of X_k. Each row of $\mathbf{C}$ corresponds to a symbol of Y_k, such that the non-zero elements in the row identify the states of X_k mapped into that Y_k symbol. When more than one non-zero element is present in the row, the mapping cannot be inverted and the corresponding states of X_k are indistinguishable from each other from the measurement of Y_k.

As a prototypical example, we consider the classical Ehrenfest urn model of diffusion [28], which has long served as a benchmark for statistical mechanics concepts [29, 30]. At each time step, a ball is independently picked from one urn and moved to the other, resulting in a Markovian process. As a thought experiment, we hypothesize that there are three balls in the urns ($N = 4$) and only a simple, binary sensor in one of the urns that can reveal whether that urn is empty or not (Fig. 1a and 1b). The sensor cannot distinguish between one, two, or three balls in that urn ($M = 2$). The corresponding matrices are

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 2/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}. \tag{3}$$

The alphabet of the hidden Markov chain is $\{s_1 = 0, s_2 = 1, s_3 = 2, s_4 = 3\}$, while the alphabet of the observed process is $\{\bar{s}_1, \bar{s}_2\}$. s_1 always provide $\bar{s}_1$ as output, while s_2, s_3, and s_4 are indistinguishable from the output since they are all mapped to $\bar{s}_2$ (Fig. 1b).

The system in (1) and (2) constitutes a discrete-time, linear shift-invariant (LSI) system with unmeasured states. Taking inspiration from deterministic systems [8, 9], we assemble a detection matrix $\mathbf{T} \in \mathbb{R}_+^{MK \times L}$ of the evolution of $\phi_k^{(l)}$ from different initial pmfs $\pi_0^{(l)}$,

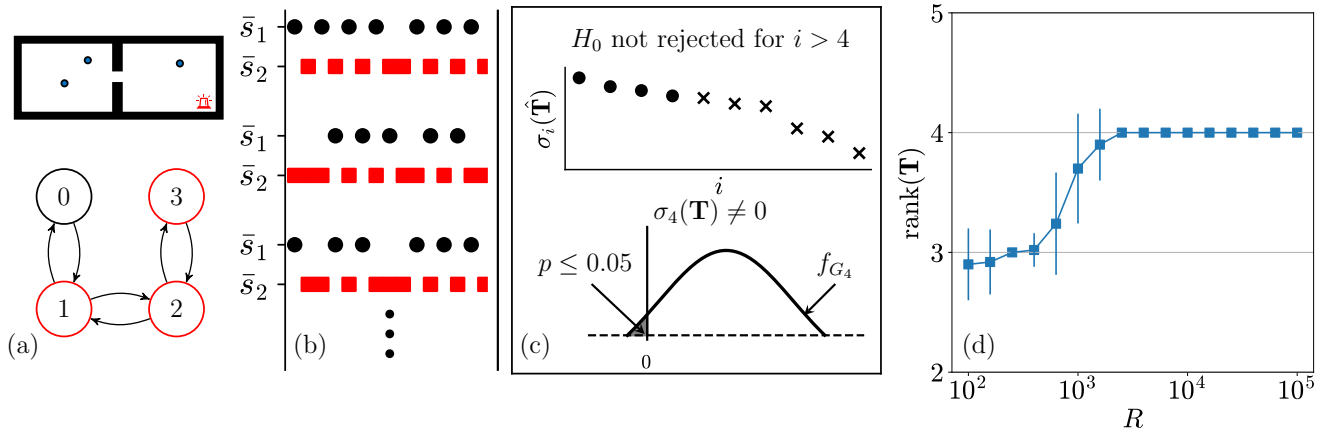

FIG. 1: Illustration of the approach. (a) Representation of the Ehrenfest urns with a binary sensor in one urn that detects whether there are balls in the urn and the corresponding hidden Markov diagram (system (3)). Subsets of states that map to the same output symbol are indicated with the same color. (b) Sampled time-series of the output process. (c) Estimate $\hat{\mathbf{T}}$ of the detection matrix, and illustration of the inference of rank($\mathbf{T}$) through the statistical test for the singular values. (d) Number of states detected from the statistical test for different values of R; the error bar represents the standard deviation over 50 trials.

$l = 1, \ldots, L$,

$$\mathbf{T} = \begin{bmatrix} \phi_1^{(1)} & \phi_1^{(2)} & \cdots & \phi_1^{(L)} \\ \phi_2^{(1)} & \phi_2^{(2)} & \cdots & \phi_2^{(L)} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_K^{(1)} & \phi_K^{(2)} & \cdots & \phi_K^{(L)} \end{bmatrix}. \tag{4}$$

Proposition 1 *Under loose assumptions on the size of* $\mathbf{T}$*, the rank of the detection matrix is equal to the size of the largest observable subspace of the LSI system [9], that is,* rank($\mathbf{T}$) = rank($\boldsymbol{\mathcal{O}}$)*,* $\boldsymbol{\mathcal{O}} \in \mathbb{R}_+^{NM \times N}$ *being the observability matrix of the LSI system [31].*

Proof *We consider the LSI system in* (1) *and* (2)*. Let us define the vector of the initial probability distribution of the hidden Markov chain as* $\boldsymbol{\Pi} = \left[\pi_1^1, \cdots, \pi_1^{(L)}\right]$*, which we assume to be full-rank. Through* (1) *and* (2)*, the detection matrix can be expressed in terms of the initial probability distributions* $\pi_1^{(l)}$ *as*

$$\mathbf{T} = \begin{bmatrix} \mathbf{C}\pi_1^{(1)} & \mathbf{C}\pi_1^{(2)} & \ldots & \mathbf{C}\pi_1^{(L)} \\ \mathbf{CA}\pi_1^{(1)} & \mathbf{CA}\pi_1^{(2)} & \ldots & \mathbf{CA}\pi_1^{(L)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{CA}^{K-1}\pi_1^{(1)} & \mathbf{CA}^{K-1}\pi_1^{(2)} & \ldots & \mathbf{CA}^{K-1}\pi_1^{(L)} \end{bmatrix}. \tag{5}$$

Thus, we can rewrite the detection matrix in the form

$$\mathbf{T} = \boldsymbol{\mathcal{O}}_K \boldsymbol{\Pi}, \tag{6}$$

where

$$\boldsymbol{\mathcal{O}}_K = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{K-1} \end{bmatrix}. \tag{7}$$

Given that $\boldsymbol{\Pi}$ *is full row-rank, we have*

$$\text{rank}\left(\mathbf{T}\right) = \text{rank}\left(\boldsymbol{\mathcal{O}}_K\right). \tag{8}$$

By invoking the Cayley–Hamilton theorem [32] and assuming $K \geq N$*, the rank of* $\boldsymbol{\mathcal{O}}_K$ *is equal to that of* $\boldsymbol{\mathcal{O}}_N$*. Hence, we determine*

$$\text{rank}\left(\mathbf{T}\right) = \text{rank}\left(\boldsymbol{\mathcal{O}}_N\right), \tag{9}$$

where

$$\boldsymbol{\mathcal{O}}_N = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{N-1} \end{bmatrix}, \tag{10}$$

which proves that the ranks of the detection matrix and the observability matrix are equal.

Contrary to the deterministic case, we do not have access to the detection matrix, but only to a noisy estimate $\hat{\mathbf{T}} = \mathbf{T} + \mathbf{E}$ from realizations of Y_k (Fig. 1b), where $\mathbf{E}$ is the noise matrix. Matrix $\hat{\mathbf{T}}$ has almost surely a higher rank than the corresponding $\mathbf{T}$ for a finite number of realizations, such that we cannot directly infer the number of states from singular values (for example, by using the largest gap between them [8]). Preliminary evidence pointing at a chief challenge in correctly assessing the rank of $\mathbf{T}$ can be found in [33].

A. Statistical test

The problem is equivalent to that of identifying the correct number of non-zero singular values of a matrix corrupted by noise. The optimal hard thresholding method to overcome such a problem was discussed in [27]. A critical assumption of the optimal threshold is the independence between the elements of the noise matrix. In our case, these elements are correlated with others in the same column. To overcome such an issue and leverage these correlations in the denoising process, we propose a statistical test based on eigen-perturbation theory.

Let $\sigma_i(\cdot)$ be the i-th singular value of a matrix, sorted in a non-increasing order. According to Weyl's additive inequality [34, 35], $\sigma_{i+j-1}(\hat{\mathbf{T}}) \leq \sigma_i(\mathbf{T}) + \sigma_j(\mathbf{E})$, for $1 \leq i, j \leq \min(MK, L)$, $i + j \leq \min(MK, L) + 1$, so that

$$\sigma_i(\mathbf{T}) \geq G_i, \text{ with } G_i \coloneqq \max_j \{\sigma_{i+j-1}(\hat{\mathbf{T}}) - \sigma_j(\mathbf{E})\}. \quad (11)$$

The probability density function f_{G_i} of G_i is numerically estimated through Monte Carlo simulations [36], using a model of the noise matrix (established in what follows) and the pertinent $\sigma_j(\hat{\mathbf{T}})$s computed from the estimate of the detection matrix [37].

We compute a p-value corresponding to the probability of G_i being non-positive,

$$p = \Pr(G_i \leq 0) = \int_{-\infty}^{0} f_{G_i}(\lambda)\, \mathrm{d}\lambda. \quad (12)$$

A small p-value (below a significance level that we set at 0.05) is used to reject H_0 – the null hypothesis that $\sigma_i(\mathbf{T}) = 0$, given the observations (Fig. 1c) – and conclude that the rank of $\mathbf{T}$ is at least i. By executing the statistical test for each i, we estimate the rank of $\mathbf{T}$.

Noise model – Let us focus on the l-th column of matrices $\hat{\mathbf{T}}$, $\mathbf{T}$, and $\mathbf{E}$, dropping the index l for ease of notation. The element $(k-1)M+m$ of the l-th column of each matrix (that is, the m-th element of the k-th time-step block) is denoted as $(\cdot)_{k_m}$. In particular, $\hat{T}_{k_m}$ is the estimate of $T_{k_m} = \Pr(Y_k = \bar{s}_m)$ from the l-th initial probability distribution. This element can be approximated through a plug-in estimator from R realizations Y_k^r of the output stochastic process.

To this end, we define an indicator variable $Z_{k_m}^r$, which is 1 when $Y_k^r = \bar{s}_m$ and 0 otherwise. $Z_{k_m}^r$ is a Bernoulli random variable with probability T_{k_m} of being 1 and $1 - T_{k_m}$ of being 0, such that $\mathrm{E}[Z_{k_m}^r] = T_{k_m}$ and $\mathrm{Var}[Z_{k_m}^r] = T_{k_m}(1-T_{k_m})$, where $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ are the expected value and variance operators, respectively. Thus, the plug-in estimator can be written as $\hat{T}_{k_m} = \sum_{r=1}^{R} Z_{k_m}^r/R$, while the noise is $E_{k_m} = \sum_{r=1}^{R} Z_{k_m}^r/R - \mathrm{E}[Z_{k_m}^r]$. Since the $Z_{k_m}^r$ are independent identically distributed random variables, $E_{k_m} \longrightarrow \mathcal{N}(0, T_{k_m}(1-T_{k_m})/R)$ as $R \longrightarrow \infty$ according to the central limit theorem [38]. Hence, the noise matrix elements are marginally Gaussian with zero means.

In practice, the elements within each column of the error matrix are correlated, as elements at the same k-th time-step should sum to zero and elements at future time-steps depend on elements at previous ones (Columns of $\mathbf{E}$ are uncorrelated when different realizations are used to estimate each column. One can utilize the same realizations for estimating multiple columns of $\hat{\mathbf{T}}$, at the price of correlating the columns of $\mathbf{E}$). The covariance between any two elements of the same column can be expressed as

$$\begin{aligned}
&\mathrm{Cov}(E_{k_m}, E_{p_q}) \\
&= \mathrm{E}[E_{k_m} E_{p_q}] \\
&= \mathrm{E}\left[\frac{(\sum_{r=1}^{R} Z_{k_m}^r - RT_{k_m})(\sum_{r=1}^{R} Z_{p_q}^r - RT_{p_q})}{R^2}\right] \\
&= \mathrm{E}\left[\frac{(\sum_{r=1}^{R} Z_{k_m}^r)(\sum_{r=1}^{R} Z_{p_q}^r)}{R^2}\right] - \mathrm{E}\left[\frac{(\sum_{r=1}^{R} Z_{k_m}^r)RT_{p_q}}{R^2}\right] \\
&\quad - \mathrm{E}\left[\frac{RT_{k_m}(\sum_{r=1}^{R} Z_{p_q}^r)}{R^2}\right] + \mathrm{E}\left[\frac{RT_{k_m}RT_{p_q}}{R^2}\right] \\
&= \frac{1}{R^2}\sum_{r=1}^{R}\sum_{\rho=1}^{R} \mathrm{E}[Z_{k_m}^r Z_{p_q}^\rho] - T_{k_m}T_{p_q}, \\
&\quad \text{(since } Z_{k_m}^r \text{ and } Z_{p_q}^\rho \text{ are independent for } r \neq \rho) \\
&= \frac{1}{R^2}\big((R^2 - R)T_{k_m}T_{p_q} + \sum_{r=1}^{R} \mathrm{E}[Z_{k_m}^r Z_{p_q}^r]\big) - T_{k_m}T_{p_q} \\
&= \frac{1}{R^2}\sum_{r=1}^{R} \mathrm{E}[Z_{k_m}^r Z_{p_q}^r] - \frac{1}{R}T_{k_m}T_{p_q} \\
&= \frac{1}{R}\Big(\Pr(Y_k^r = \bar{s}_m, Y_p^r = \bar{s}_q) - T_{k_m}T_{p_q}\Big),
\end{aligned} \quad (13)$$

where we used the fact that $\mathrm{E}[E_{k_m}] = 0$ for any k and m. As a first approximation for Monte Carlo simulations, we generate the random noise matrices by assuming that the first $M-1$ elements of each time-step k are jointly Gaussian, with covariance matrices estimated from realizations through plug-in estimators, since the covariance is a function of $\mathbf{T}$ (as given in (13)) to which we do not have access. The M-th elements are found by imposing that all elements at the same k-th time-step sum to zero. In the numerical experiment, the empirical distributions of the singular values of the noise matrix were generated from $5,000$ random noise matrices (see below for numerical evaluation of the noise model and validation of the joint normality assumption).

III. RESULTS AND DISCUSSION

A. Ehrenfest urn model

We simulate (3) to generate realizations of the output variable. We assemble an estimate $\hat{\mathbf{T}}$ of the detection matrix with a varying number of realizations R for

each initial probability distribution. When enough realizations are used to estimate $\hat{\mathbf{T}}$, we can conclude that $\text{rank}(\mathbf{T}) = 4$ according to the statistical test (Fig. 1d). The same experiment was repeated for two balls ($N = 2$) and four balls ($N = 4$) where our approach was able to detect the total number of balls in the urns, by using only binary readings from a sensor in one of the urns that tells whether that urn is empty is not. At each time step, a ball is independently picked from one urn and moved to the other. The resulting process is a Markov chain with the number of states N equal to the number of particles plus one, and reads as follows

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \frac{1}{N} & 0 & \frac{N-1}{N} & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \frac{2}{N} & 0 & \frac{N-2}{N} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{N-2}{N} & 0 & \frac{2}{N} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{N-1}{N} & 0 & \frac{1}{N} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 \end{bmatrix}. \quad (14)$$

Under the assumption that we rely on binary sensor telling us whether there exist at least one ball in one urn or not, the $\mathbf{C}$ matrix would read

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}. \quad (15)$$

The numerical experiments were performed with $L = 15$ and $K = 30$ for all three cases (Fig. 2).

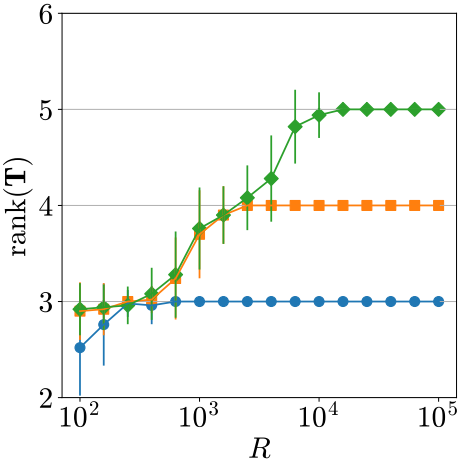

FIG. 2: Number of detected states from the statistical test as a function of the number of realizations R. Results of the Ehrenfest model with two (blue circles), three (orange square), and four (green diamonds) balls in the system. The error bars represent the standard deviation over 50 trials.

B. Biophysics

Hidden Markov chains are a fundamental mathematical model for several microscopic processes of interest in the biophysics community, including ion channels [39] and genetic sequences [40]. An example of the use of hidden Markov chains in biophysics involves the swimming behavior of *Escherichia coli* (*E. coli*), the cornerstone of our understanding of how peritrichous bacteria with flagella all above their bodies move in a fluid [41]. Recent work [42] has studied surface exploration of a pathogenic strain of *E. coli* resulting in a complex interplay between motility and transient surface adhesion events. These experimental results hint at the presence of hidden states, in addition to the two states that could be seen by the naked eye: running and stopping. The third, hidden state is suggested to be a tethered state where the bacterium use adhesion events to the surface to regulate the surface motion. The presence of a hidden state was inferred from a combination of model fitting and survival analysis; however, this approach is model-based and not scalable.

To explore the possibility of employing our approach to discover such a hidden state, we consider numerical simulations of a hidden Markov chain model where we cannot distinguish between the stopping state and the tethered state. The corresponding Markov chain consists of three states: the bacterium is not moving and is in a non-tethered state ($S := 1$); the bacterium is in a tethered state ($T := 2$); and the bacterium is running ($R := 3$). The transition probability matrix reads

$$\mathbf{P} = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ 0 & q_{22} & q_{23} \\ q_{31} & 0 & q_{33} \end{bmatrix}. \quad (16)$$

If one cannot distinguish between the non-running states, the $\mathbf{C}$ matrix is

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (17)$$

Our approach successfully detects the hidden, tethered state (Fig. 3). The numerical experiments were performed with $L = 15$, $K = 30$, $q_{11} = 0.2$, $q_{12} = 0.3$, $q_{13} = 0.5$, $q_{22} = 0.3$, $q_{23} = 0.7$, $q_{31} = 0.8$, and $q_{33} = 0.2$.

Validation of the noise model – To validate our noise model, we compared the empirical probability density function of the singular values of the true noise matrix against those of our noise model, for the biophysics example. To this end, we computed $1,000$ true noise matrices by taking the difference between $1,000$ estimates of the detection matrix from $R = 1,000$ realizations and the exact detection matrix (with $L = 4$ and $K = 2$). Second, we generated $1,000$ noise matrices from our noise model. For all of these matrices, we computed their singular values to obtain their empirical distributions. Fig. 4 compares the true noise singular values with the modeled noise singular values. Kolmogorov-Smirnov statistical tests [43] on

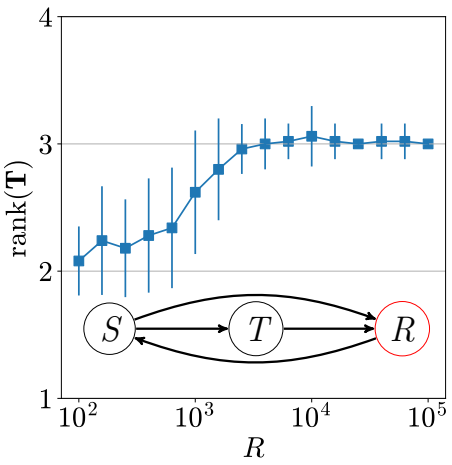

FIG. 3: Number of detected states from the statistical test as a function of the number of realizations R. Results of the inference of three behavioral states from the motion of the bacteria. The error bars represent the standard deviation over 50 trials.

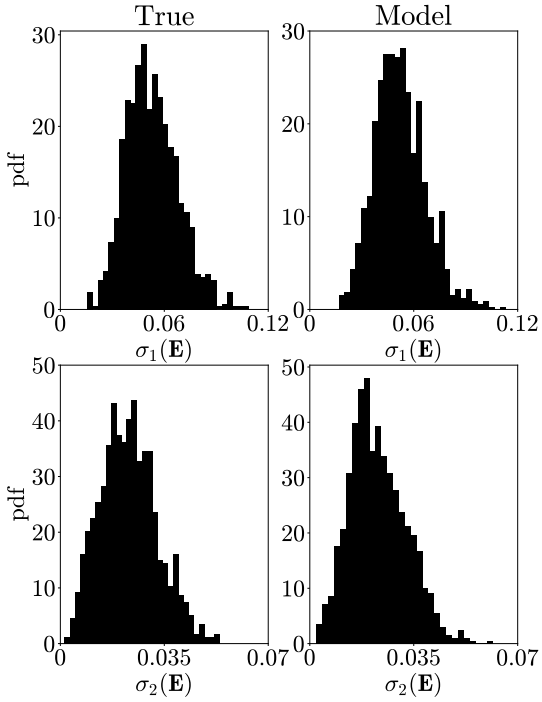

FIG. 4: Comparison between the probability density function (pdf) of the singular values of the true noise matrix and of the singular values from our noise model for the biophysics problem with $L = 4$ and $K = 2$.

each couple of distributions failed to reject the null that the empirical distributions are sampled from the same distributions ($p > 0.12$ for all pairwise comparisons).

Testing the assumption of joint Gaussianity – We showed that the elements in each column are marginally Gaussian. However, to generate realizations of the noise matrix for Monte Carlo simulations, we assumed that the elements in each column are also jointly Gaussian. Given that the elements are not independent, the accuracy of this assumption should be verified.

For a jointly Gaussian multivariate distribution of dimension d, the Manhabolis distance (between each sample of the distribution and the distribution) follows a χ^2 distribution with d degrees of freedom [44]. Hence, to test for multivariate Gaussianity, we first generated 1,000 noise samples of dimension two. We compared the exact detection matrix for the biophysics example with 1,000 estimates from a plug-in estimator, based on $R = 5,000$ realizations with $L = 1$ (since we are only interested in the multivariate distribution within one column) and $K = 2$. Then, we computed the associated Manhabolis distances and compared them with a χ^2 distribution with $d = K(M - 1) = 2$, using a Quantiles-Quantile (Q-Q) plot, a common way to quantify the similarity between two distributions, observed and theoretical.

Fig. 5 shows the Q-Q plot comparing the theoretical quantiles from the χ^2 distribution and the empirical quantiles from the measured Manhabolis distances. The quantiles match perfectly over the whole range, with small deviations toward the tails. This indicates that the assumption of joint normality is verified, at least for small errors. We acknowledge that the assumption of joint normality could be violated in other scenarios, beyond the examples considered in this work. In principle, failing to satisfy the assumption may generate random matrices that do not capture the underlying noise distribution. The extent to which an inaccurate representation of the noise would strain the algorithm is presently unknown.

C. Physical chemistry

Chemical reactions are often characterized and understood through the lens of stochastic models [45]. Enzyme reactions, for example, are modeled as Markov chains [46], where the state of an enzyme molecule varies stochastically between free enzyme and enzyme attached to different molecules, such as substrates or products. We consider the simplest example where there is only one subtract and one product, such that the Markovian states are $E \coloneqq 1$ (free enzyme), $EP \coloneqq 2$ (enzyme attached to product), and $ES \coloneqq 3$ (enzyme attached to substrate). An experimentalist can only distinguish if an enzyme is free or bonded to another molecule, so that ES and EP are indistinguishable, such that the model is written as

$$\mathbf{P} = \begin{bmatrix} r & q & p \\ p & r & q \\ q & p & r \end{bmatrix}, \tag{18}$$

and the $\mathbf{C}$ matrix is

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}. \tag{19}$$

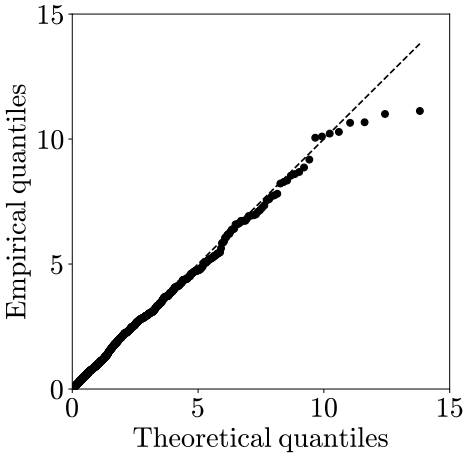

FIG. 5: Q-Q plot comparing the Manhabolis distances distribution and the χ^2 distribution for the biophysics example ($L = 1$ and $K = 2$). The Manhabolis distances were empirically computed between samples of the noise and their estimated multivariate distribution. The theoretical quantiles are the probability point functions of the χ^2 distribution with two degrees of freedom. The dashed line has a slope equal to 1.

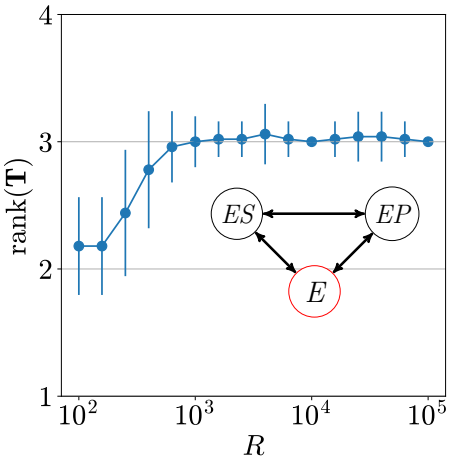

FIG. 6: Number of detected states from the statistical test as a function of the number of realizations R. Results of the inference of the presence of substrate and product within an enzyme chemical reaction. The error bars represent the standard deviation over 50 trials.

Our approach allows to infer the presence of substrate and product within the reaction (Fig. 6). The framework can be extended to other enzyme kinetics, where multiple substrates and products interact with the enzyme [47]. The numerical experiments were performed with $L = 15$, $K = 30$, $p = 0.5$, $q = 0.2$, and $r = 1 - p - q$.

Oscillating enzyme reactions – In practice, the dynamics of chemical reactions may be prone to random fluctuations that might change the transition probabilities.

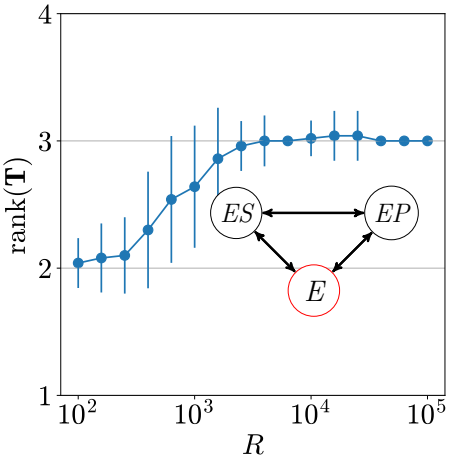

FIG. 7: Number of states detected using our approach on a time-inhomogeneous Markov chain describing a perturbed enzyme reaction. The error bar represents the standard deviation over 50 trials.

While our framework was derived for time-homogenous Markov chains, it can be applied to time-inhomogeneous chains, drawing inspiration from previous efforts on time-varying deterministic systems [9] that showed that detection matrix-based methods would work even for linear time-varying systems. We numerically tested our approach to the enzyme reaction with perturbations to the transition matrices that will render the Markov chain time-inhomogeneous. At each time step, we perturbed matrix $\mathbf{P}$ in (18) by adding noise from the uniform distribution $U(0, 0.5)$ and normalizing the rows to ensure row-stochasticity, thus re-scaling differently the different transition probabilities. Our approach is able to detect the presence of a hidden state as shown in Fig. 7. The numerical experiments were performed with $L = 15$, $K = 30$, $p = 0.5$, $q = 0.2$, and $r = 1 - p - q$.

D. Epidemiology

We seek to understand whether our methodology can unravel the presence of unobserved states in a compartmental model of a disease. This task is useful in the early stages of new epidemics, when the infectious disease is still unknown; for example, during the first wave of COVID-19, the possibility of infections from asymptomatic individuals was overlooked [48].

We focus on understanding whether a disease has an "exposed" epidemic state, where a subject is contagious but does not show symptoms, such that it is indistinguishable from a susceptible individual by only monitoring symptoms. We consider a susceptible-exposed-infected-susceptible (SEIS) [49] model, in which exposed and susceptible epidemic states map to the same output (that is, no symptoms).

The corresponding hidden Markov chain has three states ($N = 3$). Only two states are distinguishable from measurements of the symptoms ($M = 2$), such that the output process would resemble a susceptible-infected-susceptible (SIS) epidemic spreading [49]. The Markovian states of the SEIS model are defined as follows: $S \coloneqq 1$, $E \coloneqq 2$, and $I \coloneqq 3$. The corresponding transition matrix reads

$$\mathbf{P} = \begin{bmatrix} 1-\beta & \beta & 0 \\ 0 & 1-\alpha & \alpha \\ \lambda & 0 & 1-\lambda \end{bmatrix}. \tag{20}$$

When only observing the symptoms, we cannot distinguish between the susceptible and the exposed states, such that the matrix $\mathbf{C}$ is

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{21}$$

Since the system is fully observable for any $\alpha \neq 0$ ($\text{rank}(\mathcal{O}) = 3$), we can infer the presence of the exposed epidemic state. With $R \approx 10^3$ realizations, one can safely claim that the process is not an SIS and that there is some hidden state (Fig. 8). We also applied our statistical framework to a true SIS model, in which all states are distinguishable (Fig. 8), to ensure that the test would not overestimate the size of the system. The SIS Markovian states are defined as $S \coloneqq 1$ and $I \coloneqq 2$. The corresponding transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1-\beta & \beta \\ \lambda & 1-\lambda \end{bmatrix}, \tag{22}$$

and the corresponding $\mathbf{C}$ is the identity of dimension two, since both states are observable. Both numerical experiments were performed with $L = 15$, $K = 30$, $\alpha = 1/7$, $\beta = 0.3$, and $\lambda = 0.1$.

Non-geometrically distributed waiting time – The waiting time between states in a Markov chain is geometrically distributed [50]. To illustrate the robustness of the method with respect to other distributions of waiting time (and thus to other stochastic models), we consider an epidemic model where the waiting time has a Zipf distribution. The corresponding stochastic process can be regarded as a renewal process, that is, "a Markov chain whose time scale is randomly transformed" [51]. In particular, we impose that the waiting times between transitions follow a Zipf distribution. Specifically, we set

$$\Pr(WT_{(\cdot)} = N) = \frac{1}{N} \frac{1}{\sum_{n=1}^{k_{(\cdot)}} 1/n}, \tag{23}$$

where $\Pr(WT_{(\cdot)} = N)$ is the probability that the system will remain $N \in \{1, \cdots, k_{(\cdot)}\}$ time steps in state $(\cdot)$ and $k_{(\cdot)}$ is the maximum waiting time in state $(\cdot)$. For the numerical experiment, we set $L = 15, K = 30, k_{(S)} = 5, k_{(E)} = 10$ and $k_{(I)} = 15$. Similar to the regular Markov chains, we tested both SEIS and SIS models. We were able to detect a hidden state in the SEIS model while no hidden states were detected from the SIS one, as shown in Fig. 9.

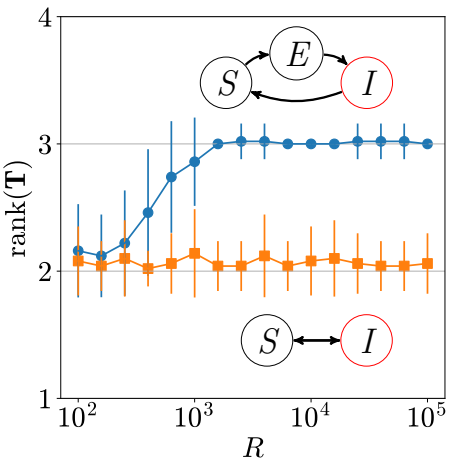

FIG. 8: Number of detected states from the statistical test as a function of the number of realizations R. Results for the inference of exposed states in epidemic models. Blue circles correspond to observing SIS from SEIS dynamics, while orange squares correspond to observing an actual SIS chain. The error bars represent the standard deviation over 50 trials.

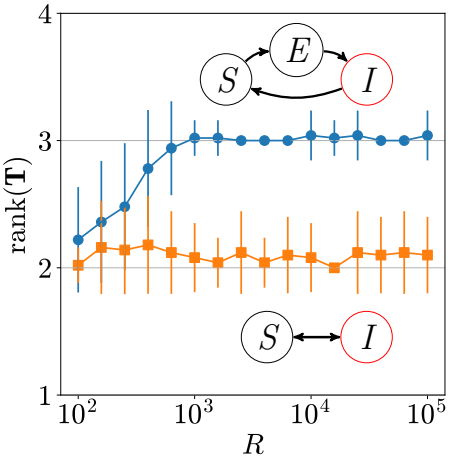

FIG. 9: Number of states detected using our approach on the epidemic models where the transition time does not follow a geometric distribution. Blue circles correspond to observing SIS from SEIS dynamics, while orange squares correspond to observing an actual SIS chain. The error bar represents the standard deviation over 50 trials.

E. Comparison of our approach and hard thresholding to denoise matrices with structured noise

In our framework, we proposed a method to detect the correct rank of a matrix corrupted by structured noise. Here, we numerically demonstrate that our method outperforms the state-of-the-art, optimal hard thresholding

method developed by Gavish and Donoho when considering Markov systems [27]. The comparison is performed by generating random square matrices $\mathbf{D} \in \mathbb{R}^{d\times d}$ of rank $\bar{N} < d$. We then corrupt them with structured noise to obtain matrices $\hat{\mathbf{D}} \in \mathbb{R}^{d\times d}$, on which we apply the two denoising techniques. To generate a random matrix of specific rank, we first sample a random matrix $\tilde{\mathbf{D}} \in \mathbb{R}^{d\times d}$ with elements from independent uniform distributions $\sim U(0,1)$, which in general has rank d. We performed the SVD to get $\tilde{\mathbf{D}} = \mathbf{U}\tilde{\Sigma}\mathbf{V}^{\mathrm{T}}$. The diagonal matrix $\tilde{\Sigma}$ contains the ordered singular values. We define a new diagonal matrix $\bar{\Sigma}$ by setting to zero all the singular values of $\tilde{\Sigma}$ after the first $\bar{N}$ ones. The resulting matrix $\mathbf{D} = \mathbf{U}\bar{\Sigma}\mathbf{V}^{\mathrm{T}}$ is a random matrix of rank $\bar{N}$.

After generating a random matrix with a specific rank, we add a noise matrix with a specific structure to obtain the final noisy matrices $\hat{\mathbf{D}}$. The columns of the additive noise matrix are sampled from a multivariate normal distribution with zero means and $\bar{\mathbf{R}}\bar{\mathbf{R}}^{\mathrm{T}}$ covariance matrix, where $\bar{\mathbf{R}} \in \mathbb{R}^{d\times d}$ is a random matrix whose elements are generated from independent uniform distributions $\sim U(-c,c)$, where c modulates the degree to which the element are correlated. To simulate the structure of noise of a hidden Markov chain with $\bar{N}$ states of which only $\bar{M}$ are distinguishable, we set subsequent blocks of length $\bar{M}$ in each column of the noise matrix to sum to zero (that is, the first $\bar{M}$ elements of each column sum to zero, the following $\bar{M}$ elements in the column sum to zero, and so on).

For our numerical experiments, we set $\bar{N} = 4$, $\bar{M} = 3$, and $d = 18$. We ran 8,000 simulations while varying c from 0.01 to 0.3 in 200 equidistant steps, such that for each c we ran 40 experiments. Out of the 8,000 experiments, our method overestimated the rank only three times and never underestimated it. The hard threshold overestimated the rank in 456 experiments (mostly in low correlation settings) and underestimated it in 5,600 experiments (in high correlation settings), see Table I. The major difference in the performance is due to the assumption of independence between the elements of the additive noise matrix for the optimal hard thresholding method, an assumption that is not valid for detection matrices.

	Our Approach	Gavish & Donoho [27]
Overestimation rate	0.037 %	5.700 %
Underestimation rate	0.000 %	70.000 %

TABLE I: Comparison of our approach and hard thresholding to denoise matrices with structured noise.

While the detection matrix we are interested in will always have a correlated noise structure, the proposed statistical test is versatile to other noise structures. We offer a fair comparison between the proposed approach and the hard thresholding method of Gavish and Donoho [27] where we do not violate their assumptions. Specifically, to generate a noisy matrix $\hat{\mathbf{D}}$, we corrupt a matrix $\mathbf{D}$ with independently sampled elements from a normal distribution $\mathcal{N}\left(0,\sigma^2\right)$. For these numerical experiments, we keep $\bar{N} = 4$ and $d = 18$, where we ran 8,000 simulations while varying σ from 0.1 to 0.3 in 200 equidistant steps, such that for each σ we have 40 experiments. The results show that our method outperforms state-of-the-art for moderate and high levels of noise ($\sigma \gtrapprox 0.15$) as illustrated in Fig. 10.

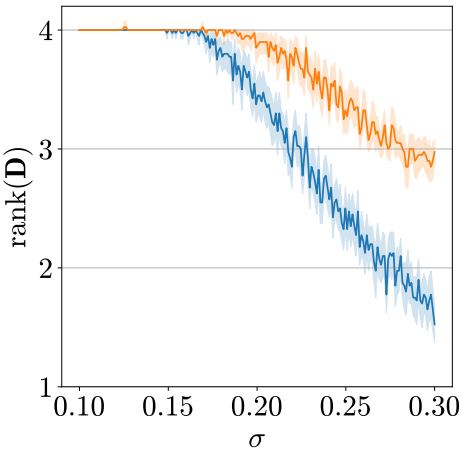

FIG. 10: Inference of the rank of random matrix $\mathbf{D}$, of true rank equals to 4, from $\hat{\mathbf{D}}$ ($\mathbf{D}$ corrupted by white noise of level σ). The orange line represents our approach while the blue line represents Gavish and Donoho's [27]. The shaded region corresponds to the 95% confidence interval.

IV. CONCLUSIONS

Inferring the number of states of a stochastic system is a fundamental problem in physics, for which methodological tools are still lacking. In this work, we propose a statistical technique to estimate the number of states of a hidden Markov chain from perceptible dynamics. This approach offers a viable framework not only to infer the number of states of a stochastic system, but also to denoise any matrix corrupted by known structured noise. Potential extensions of our work could address problems in quantum mechanics, such as quantum communication channels [52].

We showed through examples that we can detect the presence of a hidden state with about $R \approx 10^2$ realizations. To reliably infer the total number of states, more realizations are sometimes needed (typically, $R \approx 10^3$). We acknowledge that these figures might not be easy to get from observational data, but, for example, are feasible in laboratory conditions with automated experimental apparatuses that allow for generating independent realizations. We recognize that the number of realizations needed for convergence is not readily available since it depends on the degree of observability of the hidden system

(that is, even if the system is observable, some states are harder to reconstruct than others). The degree of observability can be quantitatively evaluated in different ways, such as the observability Gramian [32] and radius [53]. All the examples considered herein suggest that convergence is monotonic, so that, even if the number of needed realizations is unknown, the algorithm never overestimates the number of states. Despite these limitations, our work constitutes a critical first step toward solving a foundational issue in stochastic dynamical systems.

ACKNOWLEDGMENTS

The authors are thankful to Profs. Rastislav Levicky and Manuel Ruiz Marín for fruitful insights and acknowledge financial support from the National Science Foundation under Grants No. ECCS 1928614, No. CMMI 1932187, CMMI 1953135, and No. EF 2222418.

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, Complex networks: Structure and dynamics, Physics Reports **424**, 175 (2006).
[2] D. Ghosh, M. Frasca, A. Rizzo, S. Majhi, S. Rakshit, K. Alfaro-Bittner, and S. Boccaletti, The synchronized dynamics of time-varying networks, Physics Reports **949**, 1 (2022).
[3] N. Boers, B. Goswami, A. Rheinwalt, B. Bookhagen, B. Hoskins, and J. Kurths, Complex networks reveal global pattern of extreme-rainfall teleconnections, Nature **566**, 373 (2019).
[4] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin, Effective leadership and decision-making in animal groups on the move, Nature **433**, 513 (2005).
[5] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, A universal model for mobility and migration patterns, Nature **484**, 96 (2012).
[6] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Reviews of Modern Physics **91**, 045002 (2019).
[7] E. Bullmore and O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience **10**, 186 (2009).
[8] H. Haehne, J. Casadiego, J. Peinke, and M. Timme, Detecting hidden units and network size from perceptible dynamics, Physical Review Letters **122**, 158301 (2019).
[9] M. Porfiri, Validity and limitations of the detection matrix to determine hidden units and network size from perceptible dynamics, Physical Review Letters **124**, 168301 (2020).
[10] X. Tang, W. Huo, Y. Yuan, X. Li, L. Shi, H. Ding, and J. Kurths, Dynamical network size estimation from local observations, New Journal of Physics **22**, 093031 (2020).
[11] M. Tyloo and R. Delabays, System size identification from sinusoidal probing in diffusive complex networks, Journal of Physics: Complexity **2**, 025016 (2021).
[12] P. De Lellis and M. Porfiri, Inferring the size of a collective of self-propelled Vicsek particles from the random motion of a single unit, Communications Physics **5**, 86 (2022).
[13] S. Chandrasekhar, Stochastic problems in physics and astronomy, Reviews of Modern Physics **15**, 1 (1943).
[14] P. C. Bressloff and J. M. Newby, Stochastic models of intracellular transport, Reviews of Modern Physics **85**, 135 (2013).
[15] M. C. Wang and G. E. Uhlenbeck, On the theory of the brownian motion ii, Reviews of Modern Physics **17**, 323 (1945).
[16] K. Friston, L. Da Costa, N. Sajid, C. Heins, K. Ueltzhöffer, G. A. Pavliotis, and T. Parr, The free energy principle made simpler but not too simple, Physics Reports **1024**, 1 (2023).
[17] S. Wolfram, Statistical mechanics of cellular automata, Reviews of Modern Physics **55**, 601 (1983).
[18] L. E. Ballentine, The statistical interpretation of quantum mechanics, Reviews of Modern Physics **42**, 358 (1970).
[19] H. Weidenmüller and G. Mitchell, Random matrices and chaos in nuclear physics: Nuclear structure, Reviews of Modern Physics **81**, 539 (2009).
[20] J. Perrin, Mouvement brownien et réalité moléculaire, in *Annales de Chimie et de Physique*, Vol. 18 (1909) pp. 1–114.
[21] A. Einstein, Zur theorie der brownschen bewegung, Annalen der physik **324**, 371 (1906).
[22] V. Smoluchowski and I. D. im unbegrenzten Raum, Zusammenfassende bearbeitungen, Ann. Phys **21**, 756 (1906).
[23] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE **77**, 257 (1989).
[24] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models: estimation and control*, Vol. 29 (Springer Science & Business Media, 2008).
[25] J. Bechhoefer, Hidden Markov models for stochastic thermodynamics, New Journal of Physics **17**, 075003 (2015).
[26] B. Cooper and M. Lipsitch, The analysis of hospital infection data using hidden Markov models, Biostatistics **5**, 223 (2004).
[27] M. Gavish and D. L. Donoho, The optimal hard threshold for singular values is 4/sqrt(3), IEEE Transactions on Information Theory **60**, 5040 (2014).
[28] P. Ehrenfest and T. Ehrenfest-Afanassjewa, *Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem* (Hirzel, 1907).
[29] M. Kac, *Probability and related topics in physical sciences*, Vol. 1 (American Mathematical Soc., 1959).
[30] B. Meerson and P. Zilber, Large deviations of a long-time average in the ehrenfest urn model, Journal of Statistical Mechanics: Theory and Experiment **2018**, 053202 (2018).
[31] W. J. Rugh, *Linear system theory* (Prentice-Hall, Inc., 1996).
[32] P. J. Antsaklis and A. N. Michel, *A linear systems primer* (Springer Science & Business Media, 2007).
[33] A. Boldini and M. Porfiri, Inferring the size of stochastic

systems from partial measurements, in *European Workshop on Structural Health Monitoring* (Springer, 2022) pp. 1016–1023.

[34] H. Weyl, Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung), Mathematische Annalen **71**, 441 (1912).

[35] L. Y. Kolotilina, A generalization of Weyl's inequalities with implications, Journal of Mathematical Sciences **101**, 3255 (2000).

[36] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*, Vol. 2 (Springer, 1999).

[37] For some combinations of is and js, we have $\sigma_{i+j-1}(\hat{\mathbf{T}}) = \sigma_j(\mathbf{E}) = 0$, up to numerical precision; these cases are removed from the Monte Carlo simulations to avoid false positives.

[38] A. Papoulis and S. Unnikrishna Pillai, *Probability, random variables, and stochastic processes* (2002).

[39] N. T. Schmandt and R. F. Galán, Stochastic-shielding approximation of markov chains and its application to efficiently simulate random ion-channel gating, Physical Review Letters **109**, 118101 (2012).

[40] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge university press, 1998).

[41] H. C. Berg and D. A. Brown, Chemotaxis in escherichia coli analysed by three-dimensional tracking, Nature **239**, 500 (1972).

[42] E. Perez Ipiña, S. Otte, R. Pontier-Bres, D. Czerucka, and F. Peruani, Bacteria display optimal transport near surfaces, Nature Physics **15**, 610 (2019).

[43] V. W. Berger and Y. Zhou, Kolmogorov–smirnov test: Overview, Wiley Statsref: Statistics Reference Online (2014).

[44] F. B. Oppong and S. Y. Agbedra, Assessing univariate and multivariate normality. A guide for non-statisticians, Mathematical Theory and Modeling **6**, 26 (2016).

[45] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics **74**, 47 (2002).

[46] É. Roldán and P. Vivo, Exact distributions of currents and frenesy for markov bridges, Physical Review E **100**, 042108 (2019).

[47] D. Voet and J. G. Voet, *Biochemistry* (John Wiley & Sons, 2010).

[48] M. Gandhi, D. S. Yokoe, and D. V. Havlir, Asymptomatic transmission, the Achilles' heel of current strategies to control Covid-19, New England Journal of Medicine **382**, 2158 (2020).

[49] F. Brauer, Compartmental models in epidemiology, in *Mathematical epidemiology* (Springer, 2008) pp. 19–79.

[50] V. T. Stefanov, On some waiting time problems, Journal of Applied Probability **37**, 756 (2000).

[51] E. Cinlar, Markov renewal theory, Advances in Applied Probability **1**, 123 (1969).

[52] E. Shchukin, F. Schmidt, and P. van Loock, Waiting time in quantum repeaters with probabilistic entanglement swapping, Physical Review A **100**, 032322 (2019).

[53] G. Bianchin, P. Frasca, A. Gasparri, and F. Pasqualetti, The observability radius of networks, IEEE Transactions on Automatic Control **62**, 3006 (2016).