Tackling the subsampling problem to infer collective properties from limited data

Anna Levina1,2, Viola Priesemann3,4, & Johannes Zierenberg3

Abstract

Despite the development of large-scale data-acquisition techniques, experimental observations of complex systems are often limited to a tiny fraction of the system under study. This spatial subsampling is particularly severe in neuroscience, in which only a tiny fraction of millions or even billions of neurons can be individually recorded. Spatial subsampling may lead to substantial systematic biases when inferring the collective properties of the entire system naively from a subsampled part. To overcome such biases, powerful mathematical tools have been developed. In this Perspective, we give an overview of some issues arising from subsampling and review approaches developed in recent years to tackle the subsampling problem. These approaches enable one to correctly assess phenomena such as graph structures, collective dynamics of animals, neural network activity or the spread of disease from observing only a tiny fraction of the system. However, existing approaches are still far from having solved the subsampling problem in general, and we also outline what we believe are the main open challenges. Solving these challenges alongside the development of large-scale recording techniques will enable further fundamental insights into the workings of complex and living systems.

1Eberhard Karls University of Tübingen, Tübingen, Germany. 2Max Planck Institute for Biological Cybernetics, Tübingen, Germany. 3Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany. 4Institute for the Dynamics of Complex Systems, University of Göttingen, Göttingen, Germany. e-mail: viola.priesemann@ds.mpg.de
Perspective

Introduction

Complex systems, and in particular living systems, are composed of many interacting units. Particles collide, cells grow, move and divide, agents communicate and neurons send signals over long distances. These interactions give rise to emergent collective phenomena, including pattern formation, synchronization, flocking, percolation and self-organized criticality. Understanding such collective phenomena and their origin is necessary to understand, predict and influence complex systems.

However, in most complex systems of interest, one cannot achieve this goal by simply observing all the units with full resolution, because they are far too numerous. Therefore, in experiments, one must rely on either coarse sampling or spatial subsampling (see Box 1 for a formal definition). In both cases, the choice of the sampling ansatz can strongly affect and even bias the observables, and, hence, presents a challenge for understanding the collective properties. When sampling the full system at coarse resolution, one essentially assesses the emergent large-scale properties of the system directly. However, such coarse, indirect measures provide only macroscopic ‘observables’ and, thereby, might fail to achieve sufficient insight into the microscopic interactions between units that are essential to explain emergent complex phenomena. One is then left with the challenge to fill in these missing microscopic parameters using temporally and spatially precise measurements that are, however, only available as severely subsampled observations. It is therefore necessary to understand the challenges that come with limited sampling and then bridge the gap between the sampled and the full system to achieve a complete understanding of the emergent phenomena.

For systems containing a huge number of interacting units, high-resolution sampling of individual components typically comes at the cost of an incomplete sample. For example, when studying disease spread, a substantial fraction of infections might be unreported.

Box 1

Definition of spatial subsampling

The problem of spatial subsampling is that of having access to only a fraction of a high-dimensional system (figure part a). This problem is distinct from that of measuring a complete system at only a discrete subset of times (figure part b), which is a common problem addressed in signal processing. To formalize what we refer to under spatial subsampling, assume a full system of $M$ components with trajectory $\mathbf{x}(t) = (x_1(t), \ldots, x_M(t), t \in [0, T])$ that is generated from a stationary $M \times T$-dimensional distribution $P(\mathbf{x})$. Samples from this distribution give the momentary realization of the system; in a subsample, one observes a subset of all dimensions, that is, for $N \ll M$, one has access to $\mathbf{x}_{\text{obs}}(t) = (x_{1}(t), \ldots, x_{N}(t))$ that can be described by an $N \times T$-dimensional distribution $Q(\mathbf{x}_{\text{obs}})$. The relation between $P$ and $Q$ is not necessarily straightforward and, often, one can access it only by using a low-dimensional summary statistics $\xi$. The aim thus is to derive a general mapping between $P$ and $Q$ (or at least $\xi(P)$ and $\xi(Q)$). The desired mapping will, however, strongly depend on the type of subsampling, examples of which are shown in figure parts c and d in comparison with temporal subsampling (figure part e).
It has been long known that finite data can lead to severe biases when estimating covariances of correlated variables. To illustrate this, we construct each sample as a sequence of correlated Gaussian variables of finite length \( n = \{ x_1, x_2, \ldots, x_n \} \), with zero mean, unit variance \( \sigma^2 = \langle x^2 \rangle = 1 \) and non-vanishing (temporal) correlation \( \langle x_k x_l \rangle = e^{-\rho k / \tau} \) with \( \tau = 10 \). Practically, we construct a bivariate time series \( x^i \) and \( \eta^i \), in which \( \eta^i \) is drawn independently from a Gaussian distribution with zero mean and unit variance. Explicitly, \( x^i = \rho^i x_0 + \sqrt{1 - \rho^2} \sum_{l=1}^{n} \eta^l \) such that \( \langle x_i x_0 \rangle = \rho \). An example of such a time series is shown in part (a) of the figure.

We demonstrate the effect by considering a large number of incomplete samples (each with a finite duration \( n \) out of the whole correlated infinite time series) and focusing on the dependence of mean and variance estimates on sample duration \( n \). Specifically, for each sample, we estimate the mean \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \) and variance \( \sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \). As expected, we find that the mean computed from individual samples scatters broadly around the asymptotic expectation value \( \langle x \rangle = 0 \) with the interquartile range (shaded green area in part (b) of the figure) decreasing with \( n \), showing that the estimate of the mean has a large variance but is unbiased. By contrast, the estimate of the variance is biased. For small \( n \), it clearly deviates from the true variance \( \sigma^2 = 1 \) (part (c) of the figure). The reason for this bias is easy to see from the example trace. Because excursions are correlated, the fluctuations about the local mean are much smaller than the true variance. Dotted lines represent analytical results calculated as expectation values for an infinite number of incomplete samples using the geometric series.

Box 2

Incomplete sampling can lead to biased estimation

Studying collective animal dynamics such as navigation or ecological interactions in the field, one has a limited number of tracking devices to use. Furthermore, because such systems are time-evolving, it is not sufficient to gather data sequentially from different parts of a system, even if doing so would make it possible to observe the whole system at different time steps. In addition, sometimes even observing individual components can have associated challenges. For example, recording a single neuron typically involves some distortion or modification of brain tissue or cells, and observing social systems or animals in the wild may affect their behaviour. In general, such perturbations are expected to be stronger when more of the units are observed, and often there is a natural upper limit on how many units can be sampled. Recording neurons with electrodes in the brain, for example, is strongly limited by the space required for each electrode, such that only a tiny fraction of the millions or billions of neurons can be individually measured. Typically, one thus has to rely on simultaneous observations of a subset of individual units. Such spatial subsampling can severely bias the inference about the systems’ collective properties.

Spatial subsampling can lead to systematic mis-estimations of collective properties, in a way that typically cannot be overcome by more sampling. As a tutorial illustration, recall that estimates of the variance from short sequences of a single correlated process can be severely biased, and this bias remains when sampling more sequences (see Box 2). Likewise, when observing only a few units from a large system, the statistical fluctuations of the estimates of collective properties decrease with longer recordings, but the estimates themselves may still be systematically biased. An appropriate choice of the spatial sampling approach, however, can make inference about the full system feasible. To do so, the construction of a representative sample is crucial, as it is in most statistical estimates, for example, when representative surveys serve as empirical data for sociophysics models. However, the core challenge for spatial subsampling is that forming such representative sampling is not always possible, and even when using a sufficiently representative sampling, the inference about the full system is often still not trivial.

Overcoming biases from spatial subsampling requires the development of systematic mathematical approaches. Even when the sampling is, in principle, representative, naive extrapolations can incur severe biases. The bias depends on the system properties, the sampling approach and the observable of interest. Dismissing this bias without proper consideration can lead to contradictory statements, such as about the prevalence of scale-free networks in nature or about the dynamical state of the brain. To date, there does not exist a unifying theory to overcome the subsampling challenge in general. However, for several fairly generic settings, systematic mathematical and empirical solutions have been developed in recent years. In this
Perspective, we summarize recently developed approaches to infer the collective properties of an entire system from a spatially subsampled observation. We discuss applications mainly using examples from neuroscience and from scale-free networks, cases for which subsampling is a particularly strong hurdle, and, hence, many of these techniques originate; however, these methods are much more general and have potential applications to morphogenesis in living systems, the spread of a disease, news propagation in social networks and beyond.

What is spatial subsampling?
Depending on the experimental setup, spatial subsampling (see Box 1) can be subject to different constraints that determine which units can be accessed to collect observations. These constraints, in turn, affect how subsampling distorts the data.

We first illustrate three different subsampling approaches using the example of a static, spatially extended network (Fig. 1). In a typical experimental sampling ansatz, one might observe a random subset of nodes and connections between observed nodes (Fig. 1d). An example of such random subsampling is randomly selecting members of a social group to construct a contact network. Although such schemes are designed to sample units in a representative way, it has been shown that even simple statistics such as the degree distribution may not be preserved under subsampling, for example, for scale-free distributions.

A complementary experimental sampling ansatz is to record a complete set of nodes within a certain observation window (Fig. 1e). This windowed sampling is a common sampling scheme for optical-based—in the wider sense—observations. Examples include imaging a focus area smaller than the system of interest or considering traffic data from a particular urban area. The advantage of such a method is that all the connections or interactions in the field of view can be observed in detail; the disadvantage is that the sampling window might not be representative, and even if the sampling window is representative, it is subject to finite-size effects.

In some setups, researchers can control the sampling in a snowball-like mode: starting with a selected node, one follows the connections to its nearest neighbours, potentially the next-nearest neighbours, and so on (Fig. 1f). Such sampling is obtained when using tracer molecules injected in one neuron (or brain area) and observing which neurons are reached by the tracer through physical (dendritic or axonal) connections; likewise, contact tracing in social networks belongs to this class of sampling. This approach preserves many local properties of the network. However, errors in tracing can affect the results, tracing can be quite laborious and technically intensive, and the method is not applicable to all systems.

Whereas the study of sampling static graphs and their structure has been advancing for the past decades, methods to capture collective properties of stochastic dynamics in subsampled complex systems are only starting to be developed. In principle, investigating how a dynamical system unfolds over time provides valuable insights about its dynamics, even if only a fraction of units or dimensions are observed. For example, the classical Takens or embedding theorem states that, for deterministic, chaotic systems with a $D$-dimensional attractor, it is sufficient to observe just a single unit (or dimension). Embedding its temporal dynamics based on the past $2D + 1$ observations, or even less, enables one to reconstruct a map of the attractor. With this temporal delay embedding, one can not only predict the evolution of a single observed unit but can further infer certain system properties, such as Lyapunov exponents, at least approximately. Although the theorem was derived for deterministic systems without observational or system noise, delay embedding is also used to study the probabilistic evolution of stochastic systems. However, when it comes to inferring collective properties from stochastic dynamics, or to make a full reconstruction of the entire system, one requires more

---

**Fig. 1 | Classes of subsampling.** a–c, Sketches of typical structures of spatially embedded networks. d–f, Three approaches to sampling from a spatially extended system, illustrated using a scale-free network. Random sampling (part d) has the advantage of drawing a representative set of nodes. Windowed sampling (part e) provides a good local resolution, but the sample might not be representative of the full system. Snowball sampling (part f) follows the links of a selected node and thereby reveals the connectivity of that one node.
information about the temporal dynamics of sampled units, their correlations and their interactions.

As is the case for static networks, the specific subsampling ansatz can severely affect the observation of collective dynamics from the sampled units. For example, when sampling clusters of activity or cascades such as avalanches from a spatially structured system (such as the brain), the observed spatiotemporal activity patterns can strongly deviate from the patterns in the whole system (Fig. 2a–e).

![Diagram](image-url)

**Fig. 2 | Sampling of dynamical observables.** a. Schematics of activity in the xy plane at each time step and the corresponding projection onto the y axis. For each time t, the colour of the column denotes the summed activity, that is, a(t) = ∑_x a_x(t), as indicated in the legend, which also applies to parts b–e. b. Example of a single activity propagation (avalanche) projected onto the y axis, observed in a complete 32 × 32 Abelian sandpile model after a single unit is triggered at time t = 0. c. Same avalanche but projecting time into the xy plane. d. Same avalanche but from 64 randomly sampled units (blue crosses in e). e. Same avalanche but in a window of 8 × 8 units (orange area in c). f, g. Distribution of avalanche sizes P(s) from a larger 128 × 128 model sampled with different numbers of units N. For random choice of the units, the shape of the avalanche size distribution is approximately (but not perfectly) preserved (part f). Under windowed sampling, the distribution of avalanche sizes from the same model can show multiple peaks at N and multiples of N (part g). Distributions in f and g were obtained from 10^6 avalanches on a 2D network for each condition.
Distributions of avalanche sizes reveal a statistical difference between sampling types, as can be illustrated using relatively simple models, such as the Abelian sandpile model\(^\text{44}\) on a square lattice (Fig. 2f,g). Under random sampling, the overall shape of the observed data distribution is primarily preserved\(^\text{43,45}\). However, windowed sampling can generate statistical artefacts that were not present in the original data. In this case, the avalanche size distribution can exhibit characteristic peaks if one uses a small observation window\(^\text{22,25,26}\). These peaks are not present under random sampling or when sampling the full system; hence, they clearly emerge owing to the chosen sampling scheme. The exact impact of sampling strategies on observables is still to be determined. In the following, we outline known sampling biases together with routes that enable one to infer collective properties from limited data.

Solutions using scaling theory

Building on the achievements of statistical physics, an intuitive approach to overcome the subsampling problem is to adapt approaches from finite-size scaling theory. Finite-size scaling is a phenomenological theory that has been developed to probe scale invariance in complex systems. A common approach in such scaling theories is to formulate a multivariate function that describes the dependence of a system property on different parameters; identify relevant scales of such parameters; and rescale these parameters such that they generate a universal function that no longer depends on microscopic details of the underlying system. For systems that feature scale invariance, such transformations are expected to be described by so-called critical exponents. As we discuss in the following, such phenomenological scaling theories can naturally be extended to overcome the subsampling problem.

In the following, note that many observables of interest are discrete (for instance, the node degree \(k\) is an integer for unweighted networks). As a result, there is an additional minimal scale, and claims about scale-free distributions are approximate\(^\text{46}\). Nonetheless, we stick to the common integral description of continuous variables for convenience but keep in mind that the integrals become sums for discrete variables.

Phenomenological approach for scale-free degree networks

When testing scale invariance of real networks, one is faced with the constraint that real datasets represent a (potentially small) subsample of an inherently finite network. Both the finite size of the system, and the finite sample thereof, set a natural limit to the maximum extent of collective properties and thereby typically introduce a cutoff. In other words, observing avalanche sizes or node degrees above a characteristic maximal scale \(s_0\) becomes very unlikely.

To draw conclusions about collective properties from subsampled systems, one can make use of finite-size scaling tools developed for critical phenomena\(^\text{23,47}\). For example, a heuristic approach to infer scale-freeness of subsampled networks\(^\text{21,48}\) operates on the survival function \(S(k)\) (that is, the probability to observe a degree of at least \(k\)), related to the cumulative distribution function \(CDF(k)\) as \(S(k) = 1 - CDF(k)\). For scale-free networks with a power-law degree distribution \(P(k) = k^{-\gamma}\), \(\gamma > 1\), the survival function becomes \(S(k) = \int_0^k P(q) dq = k^{1-\gamma}\). Being an integral over the original degree distribution, the survival function is much smoother than \(P(k)\) in the relevant scaling region of large \(k\).

Assuming that finite-size effects lead to a crossover behaviour, in which for \(k < k_0\) the survival function follows a power law and for \(k > k_0\) the distribution falls more rapidly, then one can reformulate the finite-size scaling hypothesis to incorporate finite sample size \(N\) (ref.\(^\text{49}\))

\[
S(k, N) = k^{-\gamma} f(kN^\delta),
\]

with the (unknown) exponents \(\gamma > 0\) and \(\delta < 0\), and a universal scaling function \(f\). Instead of using different system sizes as in common finite-size scaling techniques, one can verify this scaling hypothesis on a single empirical dataset by artificial subsampling\(^\text{23,47}\). For this, one first generates multiple subnetworks by randomly selecting \(N\) of the original \(M\) nodes, then determines \(S(k, N)\) for all subnetworks and determines the optimal parameters \(\gamma\) and \(\delta\) by seeking to collapse plots of \(S(k, N)k^\delta\) versus \(kN^{\gamma}\) (Fig. 3a). The goodness of the collapse plot provides a measure for the scale-free behaviour. Because of deviations from the scale-free behaviour for small degrees, the collapse should only be performed for \(k > k_{min}\).

This approach has revealed that many empirical networks satisfy the finite-size scaling hypothesis of collapse\(^\text{47}\), arguing against the observation that scale-free networks are rare\(^\text{43}\) but supporting claims that scale-free properties are a common feature of real complex networks\(^\text{46,47,23,48}\).

Subsampling-scaling theory for random subsampling

Random subsampling presents a particularly suitable sampling ansatz for systems in which the collective properties of interest are expected to affect every unit. For random subsampling, one can go beyond heuristic approaches and also beyond scale-free distributions (extending the applicability of the method above) and explicitly calculate scaling transformations that make it possible to collapse distributions of collective properties from subsampled systems\(^\text{47}\). Such calculations provide exact results for exponential and negative binomial distributions and good approximations for the particularly interesting case of power-law distributions. In fact, for power laws, the calculations yield precise predictions for universal subsampling-scaling exponents that are independent of the underlying power-law exponent.

The resulting subsampling-scaling theory can be directly applied to collective properties that follow a power-law distribution \(P(x) = x^{-\gamma}\), \(\gamma > 1\). The power-law distribution can represent static properties, such as degree distributions, or dynamic ones, such as avalanche size distributions. For power-law distributions, it can be shown mathematically that, when sampling \(N\) uniformly, the resulting distribution of a collective property \(x\) is described by a universal scaling function:

\[
P(x, N) = N^{-\alpha} f(xN^{-\beta}), \ \text{for any } \alpha, \beta \in \mathbb{R} \ \text{with} \ \alpha - \gamma \beta = 1 - \gamma.
\]

Here, the quantity \(x\) can be any collective property computed for the observed part of the system, including the degree of a network \(x = k\) (see also the comparison with heuristic scaling in Fig. 3b) or the size of an avalanche \(x = s\) (Fig. 3c). The most convenient possibility is to select a solution \(\alpha = \beta = 1\) that does not require any knowledge about the underlying scaling exponent \(\gamma\), and it can be further shown that the scaling function \(f(x/N) = P(x, N)/N\) takes the shape of the asymptotic distribution with the correct exponent \(\gamma\). This solution for subsampling scaling allows one to extrapolate to larger sampling sizes. This is relevant if one applies artificial subsampling to an already subsampled system and wishes to extrapolate to the full system.

The finite size of real-world systems typically leads to a finite-size-related cutoff in the power law, thus properties \(x\) larger than a cutoff size \(x_{\text{cutoff}}\) are observed more rarely. The scaling of the cutoff location...
Perspective

with the number of sampled units is commonly expected to be linear ($x_{\text{cutoff}}(N) \sim N^\eta$, with $\eta = 1$, such as for neural avalanches from microelectrode array recordings\textsuperscript{60} and models\textsuperscript{69}) and, thus, the subsampling scaling introduced above with $\alpha = \beta = 1$ collapses the cutoff onsets. If $\eta \neq 1$, the scaling should still be recovered by taking $\beta = \eta$ and $\alpha = 1 + (\beta - 1)\gamma$.

One particular application of this method is to evaluate whether the collective dynamics of dissociated neuronal networks in vitro develop scale-free avalanches with maturation (Fig. 3c). However, from the tens of thousands of neurons, only about 60 could be recorded. In this scenario, subsampling scaling of neuronal avalanche size distributions reveals that the data collapse into power-law distributions with maturation of the neuronal networks. In particular, within three weeks of maturation, the neuronal networks develop towards exhibiting more and more scale-free avalanche dynamics that cover up to four orders of magnitude. This result indicates that these neural networks self-organize towards the vicinity of a critical point\textsuperscript{1,29,60,70–72}.

Scaling analyses for windowed subsampling

When using windowed sampling, the correlations that emerge from local interactions may cause systematic deviations in observed collective quantities\textsuperscript{25–27}. We showcase here an example of avalanche size distributions from a recording window that is considerably smaller than the system size\textsuperscript{25} (Fig. 2g). Applying windowed sampling to the classical Bak–Tang–Wiesenfeld model\textsuperscript{14,25} leads to characteristic effects resulting in the hard cutoff in the degree distribution). c, Using data from developing neuronal cultures, one observes that, in early stages (day 6), the network has not yet developed a scale-free avalanche size distribution $P(s, N)$ and, hence, the distributions obtained by further subsampling (colour) do not collapse under subsampling; upon maturation (day 21), the subsampling scaling collapses distributions, indicating underlying scale-free avalanche distributions. This finding allows to extrapolate the distribution to the system size. The normalization constant $c$ is the total number of recorded avalanches. Panel c adapted from ref.\textsuperscript{33} under a Creative Commons licence CC BY 4.0.
because the avalanches are proven to be spatially compact. Thus, in a compact observation window, there is an increased probability that the avalanche activates all observed units. As a result, the probability to observe avalanches of the size of the observation window — or multiples thereof — is increased, leading to a multipeaked distribution (Fig. 2g). Moreover, even beyond these peaks, the slope of the power law distribution is altered. This example may be extreme, but it illustrates how spatial sampling features (that is, sampling data from a compact window) can strongly distort the observables, and that fitting avalanche distributions may then turn into a multivariable scaling analysis with two or more control parameters that requires systematic analyses, such as through Bayesian inference.

However, finite observation windows have the advantage that they directly provide detailed spatial information within each window. In many systems, one can thus again adapt finite-size scaling approaches to extrapolate to the full system from artificial subsampling. For example, the so-called box-scaling procedure considers the instantaneous connected correlation function \( C(r,t) \) as a suitable indicator to distinguish critical dynamical states from non-critical ones. \( C(r) \) measures the average correlation of units separated by a distance \( r \) for a specific snapshot in time. Mathematically, this average can be expressed as

\[
C(r, t) = \frac{1}{c_0} \sum_{i,j} (s_i(t) - \langle s(t) \rangle) (s_j(t) - \langle s(t) \rangle) \delta(r_{ij} - r),
\]

in which the ‘smoothed’ Dirac \( \delta \) function \( \delta(r_{ij} - r) \) considers pairs with approximate distance \( r_{ij} = r \) and the normalization constant \( c_0 \) ensures that \( C(0, t) = 1 \). Note that this definition only considers the correlation with respect to the instantaneous mean \( \langle s(t) \rangle \) and thereby neglects temporal correlations of the global system. The zero-crossing of the correlation function is then argued to provide an estimate of the spatial correlation length \( r^* \) (refs. 5,26).

From the scaling of \( r^* \) in artificially subsampled observation windows (or box scaling), one can distinguish critical from non-critical systems. At a critical point, at which the asymptotic correlation length \( \xi \) diverges, \( r^* \) is bounded by the linear size \( W \) of an observation window such that \( r^* = W \) (when \( \xi = L \approx W \)). Away from the critical point, \( \xi \) approaches a finite value and one can show that \( r^* \sim -\log(W/\xi) \) (when \( \xi \ll W \)). This approach can be illustrated for model systems on 2D lattices (Fig. 4a,b). Moreover, this formalism makes it possible to study how well the correlation function can be collapsed under rescaling, showing good data collapse for critical dynamics but no collapse for non-critical dynamics, signalling the absence of scale invariance captured in a single scaling function. Applied to human brain activity from functional magnetic resonance imaging data, this method confirms linear scaling compatible with neural dynamics showing scale-free properties.

**Phenomenological renormalization group**

Complementary statistical-physics-inspired approaches to study scale invariance from subsampled activity involve phenomenological coarse-graining procedures that are motivated by the renormalization group. However, in many systems, the detailed neighbourhood information for the coarse-graining is inaccessible, requiring empirical approaches. For example, a heuristic coarse-graining scheme computes the pairwise correlation matrix of all state variables, sorts pairs according to their correlation and then coarse-grains these pairs by summing their states (Fig. 4c). Iterative application of this scheme yields distributions of coarse-grained variables that approach a fixed non-Gaussian form, indicating evidence of scaling in both static and dynamic quantities (Fig. 4d). In particular, when data are coarse-grained into clusters of size \( k \), the distribution of non-zero activity approaches a fixed non-Gaussian form and the eigenvalues \( \lambda \) of the covariance matrix can be described by the scaling function \( \lambda = (k/r_{\text{rank}})^\mu \) with \( \mu = 0.7 \) (ref. 34). However, based on this analysis, it is so far not possible to identify a specific universality class.

Coarse-graining data from model simulations, in which the distance to the critical point can be controlled, indicates that such phenomenological coarse-graining procedures indeed have the potential to distinguish critical from non-critical systems. However, these results remain controversial and might not be specific to critical systems only. First, some scaling features apparently persist also in the supercritical regime for finite systems, rendering a distinction difficult. Second, similar scaling features as well as the flow towards a non-Gaussian fixed point can be explained by the presence of several unknown, time-varying latent fields to otherwise independent units. Third, using real-space coarse-graining schemes with continuous sampling units (for instance, in measurements with local field potentials) further runs the risk of introducing spurious correlations into the measurement that can strongly affect estimates of collective phenomena such as avalanche size distributions in neural networks or total magnetization in the Ising model. Coarse-graining procedures thus have strong potential to study scale invariance of systems only accessible through finite observation windows, but great care has to be taken, as hidden latent fields or even the sampling process proper can make a system appear scale-free.

**Solutions beyond scaling theory**

In this section, we describe suitable subsampling-invariant observables and inference approaches that make it possible to overcome the effects of unobserved units.

**Inferring collective dynamics from stationary subsampled activity**

The population dynamics of many complex systems, such as the spread of a virus in a population or the spread of neural activity in a network, can be described by a branching process, which is a stochastic process with an autoregressive representation. In a branching process, the average activity (number of infected individuals or active neurons) at time step \( t \) depends only on the activity in the previous time step or generation:

\[
\langle A_{t+1} | A_t \rangle = m A_t + H,
\]

in which \( A_t \) is a discrete number of active units (infected people or spiking neurons at generation \( t \)), \( m \) is the branching parameter (equivalent to the reproductive number in disease spread) and \( H \) is the mean external input. The process is inherently stochastic, because at every time step, the number of offspring (for instance, the number of new infections \( Y \) caused by each infectious person \( i \) in the previous time step) is a random variable. The mean number of offspring \( m = \langle Y \rangle \) determines whether one observes a growth [\( m > 1 \)], a decline [\( m < 1 \)], a steady state [\( m = 1 \)] or critical phenomena [\( m = 1, H > 0 \)]. Inferring the branching parameter or reproductive number from observed activity is important to understand the underlying spreading dynamics.

In many complex systems, one needs to infer \( m \) from a possibly heavily subsampled observation \( \tilde{A}_t \) of the total activity \( A_t \). A classical
approach to estimate the branching parameter would be solving Eq. (4) for observed data using linear regression, that is, \( n_t = \text{Cov}(\tilde{A}_{t+1}, \tilde{A}_t) / \text{Var}(\tilde{A}_t) \) (refs.85,86). However, this classical estimator is severely biased when applied to a subsampled population, owing to the intrinsic noise of the units or the sampling process (Fig. 5). Instead, it has been proposed to use a multistep regression (MR) estimator\(^{37,71}\), which allows inference of the correct branching parameter \( m \) even from heavily subsampled data. The mathematical basis for this method is to infer an autocorrelation timescale \( \tilde{\tau} = -\Delta t / \ln(m) \) of the process \( \tilde{A}_t \), that, for random spatial sampling, can be shown to remain invariant under subsampling. Thus, although the absolute values of the autocorrelation strength are biased by subsampling, the autocorrelation time is invariant. This estimator was later generalized\(^{44}\) to cyclostationary input, which is common for processes subject to seasonality, circadian rhythms or repeated, trial-based experiments, among others. (A toolbox implementation is also available\(^{88}\).) In addition, a recently
developed Bayesian method increases the accuracy of timescales estimation and enables inference and model selection for cases with various combinations of timescales and oscillation.

Even before the intrinsic timescale was shown to be invariant under subsampling, the idea to use it to characterize the state of a complex system was explored. In particular, strongly subsampled spike recordings from different brain areas revealed that intrinsic timescales reflect the functional hierarchy in the neocortex and areal differences in temporal integration of information. Multistep regression was successfully applied to investigate memory and learning in neuronal networks. In models, it was shown that longer intrinsic timescales (and $m$ close to unity) improve performance on tasks that require temporal integration; however, for tasks that are less complex, a smaller branching parameter $m$ leads to better performance. In that light, the longer intrinsic timescales observed in experiments for higher brain areas may indeed point to longer and more complex integration of information.

A precise estimate of the strength of recurrent interactions helps shed light on neural processing. If recurrent interactions are classified using the branching parameter, for $m = 1$, small changes in the recurrent (synaptic) coupling strength between neurons (through neuromodulation, for example) can considerably alter the intrinsic timescale and thereby the processing properties of the network. A complementary approach to estimate the strength of recurrent interaction in terms of the recurrency $R$ is based on the cross-correlation distribution across neurons and argues along the same line: for $R = 1$, networks are in a sensitive regime in which small changes in the coupling statistics lead to considerable changes of network properties, including the spatial pattern of neuronal coordination and the effective dimensionality of network activity. The sensitivity to small changes in the coupling arises because, in the vicinity of non-equilibrium phase transitions (for example, at $m = 1$ for branching networks), a number of properties, such as the intrinsic timescale, diverge. Therefore, we improve our comprehension of self-organization of neural processing in vivo by understanding how $m$ (or, complementarily, the recurrency or dimensionality) changes across brain areas and with task conditions.

Several experiments examine how the branching parameter $m$, the recurrency and the intrinsic timescale change with task conditions. First, the intrinsic timescales of local activity in a brain area increase with attention, indicating a shift of the neural network dynamics towards the critical point. Second, after a perturbation by stopping sensory input, neural networks transiently decrease their branching parameter but then recover a value close to unity, and thus longer intrinsic timescales, within a few days, presumably by means of homeostasis. Third, cortical stimulation is detected better when the network shows lower instantaneous recurrency before the stimulus is applied, indicating that the larger signal-to-noise ratio may facilitate detection of such stimuli. These results in neuroscience explicitly or implicitly rely on the fact that...
autocorrelation timescales or estimated recurrency can be constructed to be invariant under observation of a small part of the system.

Quantifying dimensionality of population activity

Even when sampling only a small fraction of a complex system, the signals of individual units can be strongly correlated, owing to coupling with a system that imposes coordinated population dynamics. This property has been characterized for brain activity, a case in which even the most advanced experiments sample only hundreds or thousands out of the many millions of behaviourally relevant neurons. For such sampled activity, dimensionality-reduction methods reveal a striking simplicity underlying multivariate neuronal data recorded during execution of different tasks, by various species from different brain areas

high-dimensional neural trajectories for specific computations reside in manifolds of much lower dimension.

This result naturally raises the question of whether the inferred representation would change if more neurons were recorded and, consequently, how reliable the conclusions are. A sufficient condition to guarantee a representative manifold is that the number of recorded neurons exceeds a so-called neural task complexity (NTC). In particular, for a task with $K$ parameters (such as time), the NTC is proportional to the product of parameter-specific ranges $L_k$, $k = 1, \ldots, K$ (such as task duration) and inversely proportional to the product of parameter-specific correlation lengths $\lambda_k$ (such as population autocorrelation time, as discussed above), that is,

$$\text{NTC} \propto \prod_k L_k / \prod_k \lambda_k,$$

provided that the recorded neurons are statistically similar to the full set of neurons (Fig. 5c,d). This theory implies a task-dependent lower bound on the necessary number of random projections (the dimensionality) to accurately recover the geometry of the underlying neural manifold. Indeed, for neural recordings of macaque, it was shown that random projections can be replaced with random subsamples of neurons. Importantly, the geometric distortion (the largest relative difference between the distances of neural activity patterns in projected and full spaces) grows with the logarithm of the number of neurons in a circuit. Hence, a modest number of neurons should suffice for reliable statements about the geometry of neural manifolds.

Inverse problems on subsampled data

The recent explosion in data availability has fostered interest in so-called inverse problems, in which the goal is to calculate from a set of observations the causal relations that produce them. These causal relationships can be implicit, such as in the form of hidden latent variables, or explicit, such as when one reconstructs networks from incomplete data using statistical inference. Even if there is direct information about the system dynamics, network reconstruction is still a non-trivial problem, especially in the presence of higher-order interactions. To reduce inference complexity, one can consider generative network models or focus on properties of interest, such as community detection. These methods can be efficient in inferring the correct network from noisy network measurements. However, if one has no direct information about the edges (as is typical in biological networks), alternative approaches based purely on node dynamics are needed.

Often, the network has to be reconstructed from observed node dynamics without knowing the mechanistic model that gives rise to it (Fig. 6a). Network reconstruction, however, belongs to the class of NP-hard problems and thus poses a major computational challenge. Moreover, correlation and causation are, in general, hard to disentangle without causal interventions. Practically, the simplest approach is to consider nodes in pairs (or sometimes triplets) and estimate dependencies between them with model-free statistical inference, by using methods such as correlation-based measures or various information-theoretical measures. For these methods, hidden (unobserved) nodes can be considered as latent confounders that induce spurious correlations—and hence links—where there are none. (Fig. 6b). To treat confounders, and thereby also unobserved nodes, there have been efforts to include causality in network reconstruction. In addition, one can include higher-order correlations between nodes by capturing the statistics of observed patterns in maximum entropy models. Binary observations (such as presence or absence of spike), this approach results in the class of inverse Ising problems, which have been applied to many systems, including neural networks and gene regulatory networks. In fact, the inverse Ising model can be formulated with higher-order interactions. Moreover, it can be extended with effective non-equilibrium dynamics. Combining such kinetic inverse Ising models with approximate Bayesian inference makes it possible to capture low-order statistics despite sampling only a fraction of network nodes. We note the breadth of approaches for network reconstruction from node dynamics when the underlying dynamics are unknown. How reliable these reconstructions are and how well they capture the effect of hidden nodes depends on how challenging the problem is and how good the data are.

If one can approximate the mechanistic model giving rise to node dynamics, several opportunities exist to capture the effect of hidden nodes. For example, if only a few nodes are absent, then this absence likely affects only the reconstruction of the local neighbourhood of the nodes, causing unstable results with abnormally dense connections, which can be exploited to identify hidden nodes. A more systematic method is to incorporate the effect of unobserved nodes in an effective description of the observed dynamics. In an illustrative manner, the effective dynamics of a subsampled node $x_i$ can be written in the form $\chi_i(t) = \sum W_{ij} x_j(t) + \sum_j \int_0^t dt' M_{ij}(t-t') \chi_j(t') + \chi_i(t)$, in which the first sum describes the local subnetwork dynamics with interactions $W_{ij}$ and the other terms incorporate the effect of unobserved nodes as (non-trivial) memory $M_{ij}(t)$ and noise $\chi_i(t)$. This description is similar in spirit to delay embedding or the Takens theorem. Including these memory and noise terms when analysing subsampled data can improve quantitative assessments of collective properties, such as when applied to identifying dominant memory channels in the signalling network of epidermal growth factor receptors or to distinguishing sources of noise in protein interaction and gene regulation networks.

In addition to indirect approaches, one can explicitly include the hidden states in expectation-maximization-like algorithms. The basic idea of these algorithms is to split the inverse problem into an iterative two-step procedure, in which one step is to infer the weight matrix based on all measured (observed) and estimated (hidden) states, and another step is to update the hidden states based on the observed states and the weight matrix. Such an iterative scheme can, in principle, reconstruct network structure and dynamics from subsampled systems. Example applications include subsampled data of neural dynamics in salamander retina; reconstruction from selected input neurons the activities of the remaining ones; and financial dynamics in the S&P 500 stock market, improving predictions of price changes relevant for profitable trades.
We have presented an overview of promising techniques to study collective properties of complex systems from partial observations. The development of these new approaches is largely driven by the rapid development of experimental techniques that allow better — but far from complete — access to natural complex systems. Does this development mean that one can simply wait until sampling techniques are good enough for a given analysis methods? Certainly not! Despite ongoing progress in experimental techniques, there will probably always be systems that cannot be sampled fully for technical reasons. For large neural systems, such as mammalian brains, it is unclear how one would ever record the activity of all neurons in the deeper brain areas in parallel without harming brain tissue. It thus remains of ongoing interest to complement existing analysis methods to learn as much as possible from subsampled data.

Fig. 6 | Reconstruction of a network under subsampling. a. Even when observing the whole dynamics of each node, a model-free reconstruction of the dependencies network is a challenging problem. However, it can be, in most cases, solved with various methods of statistical inference. b. Missing observation of the dynamics of nodes without additional considerations and modelling of confounders often leads to the wrong reconstruction of the dependencies between observed variables. Blue indicates observed nodes and dynamical traces, grey indicates unobserved and red indicates a spurious inferred connection. c. When assuming a particular dynamical model, reconstruction of the coupling matrix becomes possible by explicitly including an a priori unknown number of hidden nodes and their times series into the model-based reconstruction. Panel c adapted with permission from ref.157, APS.

**Outlook**

We have presented an overview of promising techniques to study collective properties of complex systems from partial observations. The development of these new approaches is largely driven by the rapid development of experimental techniques that allow better — but far from complete — access to natural complex systems. Does this development mean that one can simply wait until sampling techniques are good enough for a given analysis methods? Certainly not! Despite ongoing progress in experimental techniques, there will probably always be systems that cannot be sampled fully for technical reasons. For large neural systems, such as mammalian brains, it is unclear how one would ever record the activity of all neurons in the deeper brain areas in parallel without harming brain tissue. It thus remains of ongoing interest to complement existing analysis methods to learn as much as possible from subsampled data.

Understanding complex systems can strongly profit from combining observations on different scales. The integrated evaluation of multiscale measurements—with different limitations and distortions at each scale—does present a formidable challenge, but it promises to overcome the limitations of sampling at each scale in a synergistic manner, thereby strongly improving inference about the multiscale activity of the full system. To exemplify the recording technique, data from...
Perspective

neuronal cultures or slices, or even in vivo recordings, can include simultaneous recordings using imaging techniques of the calcium fluorescence or electrophysiological recordings from a multi-electrode array. Each recording technique has its weaknesses and advantages. Calcium imaging lacks temporal resolution and requires filtering to extract the activity of individual cells, but has a large spatial yield; complementary multi-electrode recordings only access a tiny fraction of the neurons, but record the activity at full temporal resolution. A promising avenue of research is thus to improve the mathematical toolset that combines data from such imperfect but complementary recording techniques. On a theoretical level, developing a unified theory for windowed and random subsampling might require new scaling theory but, if successful, could greatly forward an accurate assessment of the behaviour of a system. Therefore, considering the combination of different sampling techniques, also across scales, appears to be highly promising, as the interesting emergent dynamics typically evolves across multiple scales.

An important but open question is how to combine samples across temporal scales. Especially for living systems, which are continuously changing, doing so by is by no means a trivial problem. Sticking with the example of the brain, it was shown by monitoring the same neurons over time for a repeated task that neural representations may change spontaneously over time. However, in many systems, experimental constraints might not allow to access multiple spatial scales at the same time but only successively. That limitation might hinder an understanding of how interactions at the microscopic scale affect emergent behaviour at the macroscopic scale and vice versa. Integrating information accessed at different spatial scales, even if not recorded in parallel, is a further open challenge. However, we expect important contributions in the field of non-equilibrium physics to characterize the effect of subsamples across temporal and spatial scales in well-controlled but non-stationary systems.

A gold standard in studying experimental systems is the application of causal interventions. Combining causal interventions at the microscopic scale with recordings at multiple scales may shed light on their interaction mechanisms. Designing optimal protocols for causal interventions under a given subsampling scheme might further accelerate the understanding of emergent phenomena in complex systems across scales.

Overall, the recent development of sampling techniques together with systematic analysis approaches promises that spatial subsampling theories will become more extensive and, with more extensive data availability, also more powerful. This invites future research to tackle the next set of questions, in particular how to design innovative sampling strategies (that is, how to select the sampled components), how to deal with non-representative samples in general, how to combine spatial subsamples of a different kind (by multiscale sampling, for example) and how to design causal interventions. Developing the mathematical tools to tackle these questions promises to open the path to new, unbiased insights into complex systems and collective dynamics.

Published online: 15 November 2022

References

1. Muñoz, M. A. Collapsing universality: Criticality and dynamical scaling in living systems. Rev. Mod. Phys. 90, 031001 (2018).
2. Cross, M. C. & Hohenberg, P. C. Pattern formation outside of equilibrium. Rev. Mod. Phys. 65, 851–1112 (1993).
3. Goldsbrough, J. P. & Langer, J. S. Pattern formation in nonequilibrium physics. Rev. Mod. Phys. 71, S386–S403 (1999).
4. Cross, M. & Greenside, H. Pattern Formation and Dynamics in Nonequilibrium Systems 1st edn (Cambridge Univ. Press, 2009).
5. Desai, R. C. & Kapral, R. Dynamics of Self-Organized and Self-Assembled Structures (Cambridge Univ. Press, 2009).
6. Halatek, J. & Frey, E. Rethinking pattern formation in reaction–diffusion systems. Nat. Phys. 14, 507–514 (2018).
7. Acosta, J. A., Bonilla, L. L., Pérez Vicente, C. J., Ritort, F. & Spigler, R. The Kuramoto model: a simple paradigm for synchronization phenomena. Rev. Mod. Phys. 77, 137–185 (2005).
8. Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. & Zhou, C. Synchronization in complex networks. Phys. Rep. 469, 93–153 (2008).
9. Cavagna, A., Giardina, I. & Grilli, S. T. The physics of flocking: correlation as a compass from experiments to theory. Phys. Rep. 728, 1–62 (2018).
10. Rahmann, P., Penumai, F. & Romanczuk, P. Topological flocking models in spatially heterogeneous environments. Commun. Phys. 4, 206 (2021).
11. Christensen, K. & Moloney, N. R. Complexity and Criticality (World Scientific, 2005).
12. Araújo, N., Grassberger, P., Kahng, B., Schrenk, K. & Ziff, R. Recent advances and open challenges in percolation. Eur. J. Phys. Spec. Top. 223, 2307–2321 (2014).
13. Korchniak, D. J., Orlandi, J. G., Son, S. W. & Davidson, J. Criticality in spreading processes without timescale separation and the critical brain hypothesis. Phys. Rev. X 11, 021059 (2021).
14. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality: an explanation of the 1/1 noise. Phys. Rev. Lett. 59, 381–384 (1987).
15. Jensen, H. J. Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems (Cambridge Univ. Press, 1998).
16. Dickman, R., Munoz, M. A., Vespignani, A. & Zapperi, S. Paths to self-organized criticality. Phys. J. 30, 27–41 (2000).
17. Aschwanden, M. Criticality in Astrophysics: The Statistics of Nonlinear Processes in the Universe (Springer, 2011).
18. Priesemann, G. Self-Organised Criticality: Theory, Models and Characterisation (Cambridge Univ. Press, 2019).
19. Hesse, J. & Gross, T. Self-organized criticality as a fundamental property of neural systems. Front. Syst. Neurosci. 8, 166 (2014).
20. Zierenberg, J., Wilting, J., Priesemann, V. & Levina, A. Description of spreading dynamics by microscopic network models and macroscopic branching processes can differ due to sampling properties of networks. Proc. Natl Acad. Sci. USA 102, 4221–4224 (2005).
21. Serafin, M. et al. True scale-free networks hidden by finite size effects. Proc. Natl Acad. Sci. USA 118, e201352118 (2021).
22. Morstetter, F., Pfeiffer, J., Liu, H. & Carley, K. M. Is the sample good enough? Comparing data from Twitter’s Streaming API with Twitter’s Firehose. Proc. Int. AAAI Conf. Weblogs Soc. Media 7, 400–408 (2013).
23. Priesemann, V., Munk, M. H. & Wibral, M. Subsampling effects in neuronal avalanche distributions recorded in vivo. BMC Neurosci. 10, 40 (2009).
24. Magni, A., Durin, G., Zapperi, S. & Sethna, J. P. Visualisation of avalanches in magnetic thin films: temporal processing. J. Stat. Mech. Theory Exp. 2009, P01020 (2009).
25. Chen, Y.-J., Papankaloao, S., Sethna, J. P., Zapperi, S. & Durin, G. Avalanche spatial structure and multivariable scaling functions: sizes, heights, widths, and views through windows. Phys. Rev. E 84, 061103 (2011).
26. Zierenberg, J., Witling, J., Priesemann, V. & Levina, A. Description of spreading dynamics by microscopic network models and macroscopic branching processes can differ due to coalescence. Phys. Rev. E 81, 023101 (2010).
27. Cabezas, M. A. Criticality in network models. In Nonequilibrium Dynamics of Self-Organized and Self-Assembled Structures (Cambridge Univ. Press, 2021).
