Synthese manuscript No.
(will be inserted by the editor)

Amalgamating evidence of dynamics

David Danks · Sergey Plis

Received: date / Accepted: date

Abstract Many approaches to evidence amalgamation focus on relatively static information or evidence: the data to be amalgamated involve different variables, contexts, or experiments, but not measurements over extended periods of time. However, much of scientific inquiry focuses on dynamical systems; the system's behavior over time is critical. Moreover, novel problems of evidence amalgamation arise in these contexts. First, data can be collected at different measurement timescales, where potentially none of them correspond to the underlying system's causal timescale. Second, missing variables have a significantly different impact on time series measurements than they do in the traditional static setting; in particular, they make causal and structural inference much more difficult. In this paper, we argue that amalgamation should proceed by integrating *causal knowledge*, rather than at the level of "raw" evidence. We defend this claim by first outlining both of these problems, and then showing that they can be solved only if we operate on causal structures. We therefore must use causal discovery methods that are reliable given these problems. Such methods do exist, but their successful application requires careful consideration of the problems that we highlight.

Keywords Dynamical systems · Timescale · Latent variables · Causal inference · Causal discovery

D. Danks
Departments of Philosophy & Psychology
161 Baker Hall
Carnegie Mellon University
Tel.: +1 412-268-8047
Fax: +1 412-268-1440
E-mail: ddanks@cmu.edu

S. Plis
The Mind Research Network
1101 Yale Blvd NE
E-mail: s.m.plis@gmail.com

1 Learning about dynamics

Present-day science proceeds through a division of cognitive and laboratory labor: no one individual, or even research group, can fully address scientific problems of any significant size. Instead, modern sciences operate through widespread "divide-and-conquer," where the divisions might be based on domain, method, or question. This distributed effort inevitably leads to integration challenges, as we aim to bring together different data and theories into relatively coherent pictures of the world. As a result, philosophers of science and epistemologists have recently taken a renewed interest in these types of integration and amalgamation challenges (e.g., [6, 19, 4, 31]): how can or should we combine similar, overlapping, or related scientific claims and evidence into a unified whole (or small set of possible wholes)?

In this paper, we concentrate on a relatively under-explored setting for integration and amalgamation challenges—namely, discovery and confirmation of models of dynamical systems (broadly understood) using time series data. Many scientific investigations aim to understand a system that dynamically changes over time, such as neural processes in the brain, transactions between firms in an economy, long-range climatic teleconnections involving ocean indices (e.g., El Niño), or students' knowledge development in an online course. In all of these cases, we have observations of the system (or multiple systems of the same type) as it changes over time. Since these systems are complex and important, we often have multiple research groups or measurement methods, resulting in multiple time series datasets that must somehow be amalgamated together to yield a coherent model of the interactions among the underlying system elements.

An additional desideratum for much of this scientific research is to learn *causal* models of the underlying dynamical systems. For many (though not all) of these systems, we want to not only predict their behavior, but also design effective interventions—actions, policies, modifications, and so forth—to control them and achieve relevant goals. Knowledge that A predicts B is not sufficient to know whether an exogenous change in A (e.g., by us) will yield a corresponding change in B; instead, we require causal knowledge. Thus, our amalgamation methods should ideally be capable of identifying, discovering, or working with underlying causal structures, and not only phenomenological or measurement-level characterizations of the various time series datasets.

However, this desire for causal models raises two novel problems for evidence amalgamation in the time series setting. First, the measurements in different time series are often undersampled relative to the "true" timescale of the underlying causal dynamics. Our measurement methods sometimes cannot capture the system state at the speed with which it actually changes, whether for technical, financial, or ethical reasons. As a concrete example, consider efforts to learn causal and communication structures within human brains. While we do not know the exact causal timescale of the brain, it is surely faster than the ≈ 2000 ms time between measurements in (standard) fMRI experiments. Our time series data about any particular brain includes

only some of the causally relevant timepoints and changes, which significantly complicates inference. Moreover, in many cases, we do not know the extent of undersampling, nor do all measurement methods undersample to the same degree. Undersampling—more generally, a mismatch between the underlying causal timescale and the various measurement timescales—is a significant novel challenge to integration and amalgamation that arises only for time series.

Second, our time series measurements frequently include only a subset of the causally significant variables (for the system). For example, there might be some unobserved common causes of measured features. The possibility of latent common causes, and latent variables more generally, is a well-known problem for causal inference. The problem is significantly exacerbated for dynamical causal systems. Latent variables in "static" settings typically introduce only local complications in the observed joint probability distribution or density; their impact is rarely global.[1] In contrast, latent variables in dynamical systems typically create complexities throughout the observational data; the impact is almost always global.[2] In addition, additional types of latent variables present problems for inference, amalgamation, and integration in dynamical settings. For example, suppose we have $X \to L \to Z$, but where L is unobserved. In the static setting, L is unproblematic: it is an unobserved mediator, but does not cause inferential problems. In contrast, in the dynamical setting, L makes inference and amalgamation much harder (as we will see in Section 3).

Overall, both of these challenges pose serious threats to the very possibility of causal structure learning from diverse, observed time series data. Evidence amalgamation is qualitatively different (and harder) in the dynamical, time series domain. The arguments in this paper are guided by the overarching idea that we can best address these amalgamation challenges by integrating "local" causal knowledge, rather than trying to amalgamate the data or evidence directly. Of course, such an approach requires that we have reliable methods for causal structure discovery given these types of data, and the above two problems make such discovery significantly harder. Matters are not hopeless, however, as suitable methods given these challenges have recently emerged. In the remainder of this paper, we examine these problems in more detail: undersampling and timescale mismatch (Section 2) and latent, unobserved variables (Section 3). For clarity, we will largely avoid technical or mathematical details (except when necessary), but we emphasize that everything discussed here can be represented precisely and formally.

[1] We will be more precise about this claim in Section 3.

[2] The two exceptions are (1) a constant-valued (and so not really variable) latent; and (2) a latent that is a cause of only one other variable (including its own future state) at each moment in time (and so essentially functions as an additional source of noise).

2 The challenge of multiple timescales

2.1 Timescales and their mismatch

One set of amalgamation challenges arises from discrepancies in the relevant "timescales" (i.e., rates at which things happen). As a simple example, consider an experiment in which we measure the temperature at the same location at 12:00 noon every day. The "measurement timescale" for this experiment is once per day, as that is the rate at which we measure the state of the world. More generally, almost all experiments have a measurement timescale that corresponds to the rate at which the experimenters (or their devices) make measurements of the world. There are numerous potential complications here, as different devices might measure at different rates, or some measurements might be triggered by external events rather than occurring on a regular schedule. However, if we allow our dataset to have missing measurements, then we can essentially always establish a measurement timescale for an experiment.[3]

We can also sensibly think about a "causal timescale" for a system, as different causal processes can require different lengths of time to occur or progress. As a very rough characterization, the timescale for a causal connection $C \to E$ can be understood as the length of time it takes for an exogenous change in C to (start to) influence E. For example, economic demand for a product does not immediately change the supply of that product; the causal influence takes time to propagate. Moreover, that timescale is different (and significantly faster) than the causal timescale at which increases in atmospheric carbon dioxide influence global climate. Different systems exhibit different causal timescales.[4] Of course, we often do not know the causal timescale of a system; discovery of the causal timescale is frequently part of our scientific investigations.

One might immediately object that 'causal timescale' implies an inappropriate discretization of time or the causal "flows": the worry is that if causal influence occurs continuously in time, then the only possible causal timescale would be "infinitely fast" (which would be scientifically unusable). At the very least, though, all causal influences take some time to propagate, and so there must be some bound on the causal timescale (perhaps derivable from the spatial separation of the relata).[5] This "propagation time" thus provides a non-zero lower bound for the time it takes an exogenous change to have an influence (i.e., how we characterized 'causal timescale'). Moreover, as long as the system variation is not infinitely quick, we can use the Nyquist-Shannon

[3] For example, suppose device D_1 measures once every two seconds, and device D_2 once per three seconds. In this case, the measurement timescale is every one second, but with numerous missing values (e.g., D_1 missing values in every other timestep).

[4] And different causal relations within a system can have different timescales. In general, we assume that the causal timescale of a system is the greatest common divisor of the timescales of the causal relations within the system.

[5] And perhaps time truly is discretized at fine scales [28, 1].

sampling theorem to determine a discrete, effective (causal) sampling rate.[6] More generally, the vast majority of scientifically interesting causal relations take time to occur, and so we can safely understand a system as having a causal timescale, particularly since we make no assumptions about our knowledge of that causal timescale, or our ability to measure at it.

Now consider the problem of evidence amalgamation from multiple experiments on a dynamical system. In practice, these experiments will typically involve different measurement timescales, and so we cannot simply merge the data, even if we ignore all of the challenges that arise from different measurement methods, different populations, and so forth. However, these experiments all (supposedly) measure the same type of causal system, so we might instead attempt to merge the experiments by examining what they tell us about the underlying causal system. That is, rather than trying to merge the data, we could try to merge the results or implications of the data about the target causal system. The measurement timescales might differ between experiments, but the causal timescale presumably does not. Thus, amalgamation of causal knowledge sidesteps some challenges that (naïve) evidence amalgamation faces.

This intuition is a sensible one, but it faces a significant inferential challenge. Most existing algorithms that learn causal structure from time series assume that the measurement timescale is approximately equal to the relevant causal timescale. However, this assumption is violated in many experiments. For example, in fMRI experiments, the measurements are significantly slower than the relevant causal relations, whether because hemodynamics are slow relative to neuronal causes, or because data acquisition is slow relative to artifacts such as cardiac signals.[7] Moreover, the failure of the assumption of "same timescale" matters: these existing inference methods can be completely unreliable even if all other assumptions are satisfied [27]. Every type of error is possible: missing causes (false negatives), incorrect inferences of causation (false positives), and reversed causal direction.

As a practical example of these issues, consider learning brain network structure using magnetoencephalography (MEG) data (timescale: measurement each 1 ms) and fMRI data (timescale: every 2000 ms). In this case, many of the traditional evidence amalgamation concerns can be avoided by collecting these data from the same human subjects in the same experimental paradigm under similar conditions [25]. Nonetheless, the different measurement timescales pose a problem for data integration. One might hope to amalgamate data of this type at the measurement level, but the MEG and fMRI connectivity graphs inferred from their respective types of data are quite different. Even high-level network properties change depending on the sampling rate [25], even though both measurement processes are measuring signals that are arguably caused by the very same underlying neural activity [18]. The intuitive idea that we should amalgamate at the level of inferred

[6] Thanks to an anonymous reviewer for suggesting this idea.

[7] Event-related fMRI designs can potentially get much tighter temporal resolution, though only under significantly stronger assumptions.

causal structure does not work, precisely because the violation of the "same timescale" assumption means that the inference algorithms are unreliable. Of course, there might be many reasons for the different inferred "causal" structures at the respective measurement timescales, as fMRI and MEG do not actually measure exactly the same sources. However, similar (inferred) structure variation as a function of sampling rate and speed has also been found across resting state fMRI experiments that all use the same modality [2, 16] (see [20] for additional concerns about structural inference from resting state fMRI data).[8]

In general, the strategy of amalgamating causal knowledge will avoid many evidence amalgamation problems, but only if the causal inference methods are reliable, which they are not given existing measurement methods. And if the causal outputs "learned" from different experiments are unreliable, then our amalgamation will be "garbage in, garbage out." Instead, we require inferential methods that can learn *causal* timescale information from data obtained at a different measurement timescale. We describe several such methods below, but we first explain (in more precise detail) why causal inference is challenging when the measurement and causal timescales differ.

2.2 From causal timescale to measurement timescale

Formally, we represent causal structures as causal graphical models, which have been used to represent causal systems in a wide variety of domains. Dynamic Bayesian neworks [21, 9] are perhaps the best-known of these dynamical causal graphs, though they are not the only type. For our present purposes, we only need the idea that causation is represented by directed arrows: $X \to Y$ means that X is a direct cause of Y (given the other variables in the causal graph), where 'direct cause' is understood instrumentally as "there are conditions in which an exogenous change to X will probabilistically lead to a change in Y, even when all other variables in the graph are held fixed" (see [35] for a careful explication of this intuition). Causal graphical models are significantly more complicated than this, but the inferential impact of timescale mismatch can be explained with this simple characterization.[9]

We explicitly model time in our causal graph by including nodes for random variables V at both times t and $t+1$ (see left-most graph in Figure 1). For convenience, we assume that our causal system is Markov order one, which means that no *direct* causal influence requires more than one timestep to propagate. We emphasize, though, that everything we say here generalizes to Markov orders greater than one. We also assume that there are no isochronal edges (i.e., no causal connections between variables at the same moment) at the causal timescale, though of course there might be such (apparent) connections

[8] Thanks to an anonymous reviewer for noting that similar phenomena occur with resting state fMRI data.

[9] Readers interested in the full framework of causal graphical models are encouraged to consult one of the many mathematical introductions on the topic [22, 30, 15].

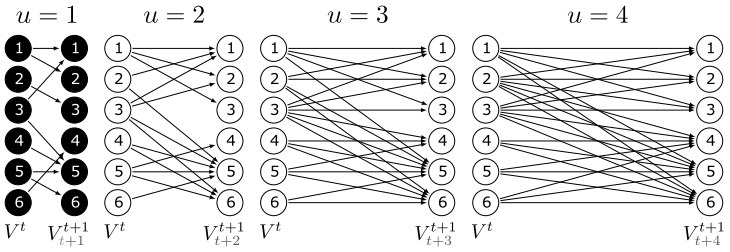

Fig. 1 Undersampling effect on a dynamical causal graph. Superscripts denote measurement time index; subscripts denote causal time index; u is the degree of undersampling.

at the measurement timescale.[10] We denote this dynamical causal graph (at the causal timescale) by $\mathcal{G}^1$; for causal graphs, superscripts will denote the undersampling rate, where 1 means "no undersampling."

We are principally concerned with *undersampling*: situations in which the measurement timescale is significantly slower than the causal timescale. More precisely, we model undersampling as missing values for measurements (for every variable) at particular times. For example, if the measurement timescale is twice as slow as the causal timescale, then we fail to record every other (causal) timestep. Figure 1 shows the apparent structural changes in $\mathcal{G}^1$ as the degree of undersampling (denoted by u) increases.[11] As Figure 1 shows, the apparent structure at the measurement timescale can be quite different from the actual causal structure at $u = 1$. Thus, we cannot amalgamate causal knowledge if we use methods that assume sameness of causal and measurement timescales, as those methods can yield quite different outputs, simply because of undersampling.

In fact, we can be much more precise about the nature and prevalence of errors due to undersampling. For convenience, we shift to representing dynamical causal graphs in terms of compressed graphs that encode time implicitly in the edges: (i) $X \to Y$ means $X^t \to Y^{t+1}$; and (ii) $X \leftrightarrow Y$ means $X^{t+1} \leftarrow L^t \to Y^{t+1}$ for some variable L. Figure 2 shows a dynamical causal graph and its compressed counterpart. For our problem space, these two representations are provably equivalent: there is a $1-1$ mapping between dynamical causal graphs and compressed graphs [7].

The general observation that undersampling can lead to unreliable causal learning has been made by other researchers [27], but there are methods to determine, in an efficient manner, exactly which causal timescale causal relationships will be hidden given undersampling, and which indirect connections will appear to be direct [7]. In general, undersampling leads to causal paths composed of multiple edges appearing to be direct causal connections. For ex-

[10] These two assumptions are often made in scientific practice, e.g., in neuroimaging [36, 26, 3, 14, 17].

[11] For clarity, we omit the additional edges that represent connections due to common causes that are unobserved due to undersampling. For example, if $u = 2$, then 1 and 2 at the second (measurement) timestep have an unobserved common cause—namely, 1 at the previous (unmeasured) timestep.

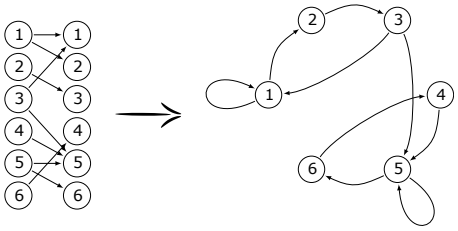

Fig. 2 An example of obtaining a compressed representation of a Markov order one DBN (bidirected edges not shown).

ample, if the undersampling rate is $u = 3$, then $X_1 \to X_2 \to X_3 \to X_4$ in $\mathcal{G}^1$ will manifest as $X_1 \to X_4$ in $\mathcal{G}^3$. This type of reasoning can be generalized to prove results about the impacts of undersampling. As one instance, if $\mathcal{G}^1$ (in its compressed form) is a directed acyclic graph—that is, there are no causal cycles over time (not even self-loops)—then the resulting upper bound on the length of causal paths implies that increases in undersampling lead to monotonic reductions in the number of (apparent) between-time causal connections (see top two rows of Figure 3).

More generally, the determining factor for the effect of undersampling is the graphical structure of the *strongly connected components* (SCCs)—maximal sets of nodes such that each node can be reached from any other. Node membership in SCCs is invariant given undersampling (under a weak additional condition); that is, feedback groups are not broken apart by undersampling, though the particular order or arrangement of variables in the feedback loops can change [7]. In particular, as $u \to \infty$, the SCC will either converge to a stable graph for all $u > u_{converge}$ (see Figure 4), or will settle into a periodic "oscillation" across a set of graphs. This SCC behavior is determined by its internal structure,[12] and is immediately derivable from the structure of $\mathcal{G}^1$. More generally, for a given causal timescale causal graph, one can provably determine its apparent changes given undersampling of particular degrees, or even given only information about the direction of change in undersampling [7].

2.3 From measurement timescale to causal timescale

These results about changes in the causal graph due to undersampling can be used to support the opposite inference: given a measurement timescale graph $\mathcal{H}$, we can ask which causal timescale graphs could have produced it. And in so doing, we can obtain exactly the content required for amalgamation of causal knowledge about the dynamical system. Of course, this inference can be quite difficult due to underdetermination: there will sometimes be multiple $\mathcal{G}^1$ that "look like" $\mathcal{H}$ given some undersampling. Nonetheless, there are algorithms

[12] Specifically, if the greatest common divisor of the lengths of the simple loops in the SCC is 1, then it will converge; otherwise, it will oscillate.

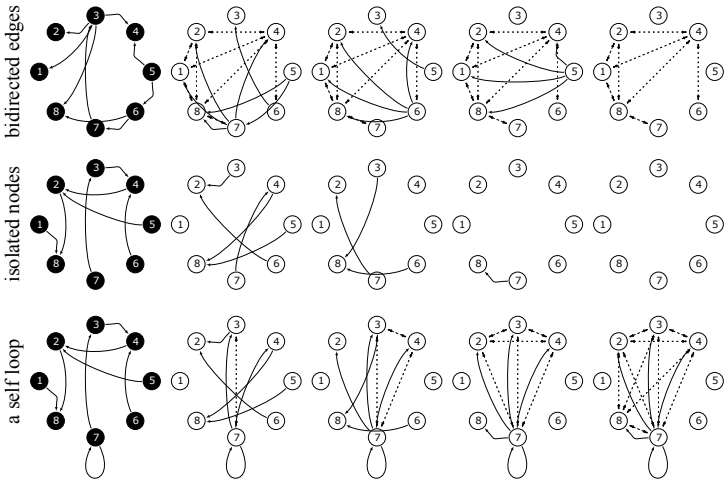

Fig. 3 Limiting behavior of apparent causal structure for causal time scale interactions described by directed acyclic graphs that allow loops of length one.

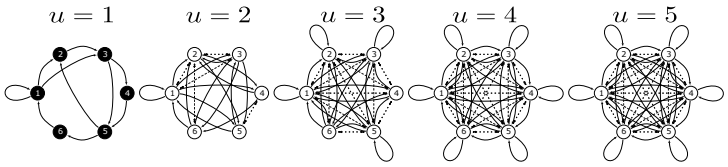

Fig. 4 A strongly connected component at a certain undersampling rate may appear as fully connected with all possible edges present.

that infer causal timescale structure from measurement timescale data, given either known or unknown undersample rate.

Mesochronal Structure Learning (MSL) algorithms assume the undersampling rate u is known, and then learn the set of possible underlying $\mathcal{G}^1$ such that $\mathcal{G}^u = \mathcal{H}$. In particular, if u is known, then we know that each measurement timescale edge corresponds to a path with u edges in the underlying causal structure. For example, if $u = 2$, then each edge in $\mathcal{H}$ corresponds to a two-edge path in $\mathcal{G}^1$. The MSL algorithms sequentially determine the intermediate nodes in each of those u-edge $\mathcal{G}^1$ paths, thereby producing the $\mathcal{G}^1$s that exactly capture the causal connections in $\mathcal{H}$. Computational efficiency is achieved by intelligently pruning the search tree branches during search. Several different versions of MSL have been developed using different search procedures, and all are provably correct, complete, and reliable [24].

Interestingly, simulations with the MSL algorithms demonstrated that, for reasonably sparse $\mathcal{H}$, there were typically few, and perhaps only one, $\mathcal{G}^1$ that could have generated it. This finding suggests that the learning problem is not intractable, as significant causal timescale structure can often be recovered from measurement timescale data. Even when more than one $\mathcal{G}^1$ is possi-

ble, there are usually sufficiently few that domain experts can examine the possibilities directly. There are no known computational complexity results for the full "causal inference from undersampled data" problem, though it is likely to be NP-hard. Causal timescale structure discovery given a measurement timescale structure is known to be NP-complete if one uses only directed edges in $\mathcal{H}$ [12, 13]; it is unknown whether the bidirected edges in $\mathcal{H}$ make discovery easier or harder. In practical terms, the MSL algorithms can (as of 2017) handle graphs with up to 70 variables when u and the graph density are small, and correspondingly smaller numbers of variables as u or the graph density increases [12, 13]. However, the MSL algorithms require the significant assumption that the undersampling rate u is known. In practice, we often do not know the exact timescale mismatch between the measurement and causal timescales. Furthermore, if we are trying to amalgamate different datasets, then the timescale mismatches may be different for each dataset.

Rate Agnostic Structure Learning (RASL) algorithms drop this key assumption of the MSL algorithms; they do not assume that u is known. More precisely, for a given measurement timescale graph $\mathcal{H}$, define $[\![\mathcal{H}]\!]$, the *undersampling equivalence class* of $\mathcal{H}$, to be the set containing every $\mathcal{G}^1$ for which there is *some* u (not necessarily the same for each $\mathcal{G}^1$) such that $\mathcal{G}^u = \mathcal{H}$. That is, $[\![\mathcal{H}]\!]$ contains all of the $\mathcal{G}^1$ that could look like $\mathcal{H}$ given *some* degree of undersampling. The RASL algorithms iteratively construct the elements of $[\![\mathcal{H}]\!]$ by gradually adding possible edges to candidate graphs and checking whether the candidate graph conflicts with $\mathcal{H}$.[13] If they do conflict, then (provably) no super-graph of that candidate can be in $[\![\mathcal{H}]\!]$, so we do not need to search further along that branch. This single test significantly reduces the search space, and makes the RASL algorithms computationally feasible.

Different RASL algorithms result from different schemata for iteratively adding-and-testing candidate edges or loops, where runtimes depend strongly on the method; one version could be more than 1000 times faster than another. All of these RASL algorithms are provably correct, complete, and reliable. They demonstrate that timescale mismatch is not an insurmountable problem: given measurement timescale data, we can frequently learn a substantial amount about the causal timescale system, even when the two timescales diverge to an unknown degree [23].

The RASL algorithms are the first inverse inference algorithms for unknown undersampling, and they are approaching usability for actual scientific problems. RASL is lower-bounded in computational complexity by the MSL algorithms, so is also likely (at least) NP-hard. The fastest RASL algorithms can (as of 2017) be applied to datasets with around 10 variables. While this might seem like a small number, that is approximately the number of regions of interest (ROIs) in many neuroimaging experiments, and multiple other scientific contexts also deal with comparable numbers of variables of interest. Current versions of RASL are non-parametric—the algorithm takes a mea-

[13] Technically, a conflict occurs between candidate $\mathcal{G}^1$ and $\mathcal{H}$ when $\forall u\{\mathcal{G}^u \not\subseteq \mathcal{H}\}$

surement timescale graph as an input, rather than "raw" data—and so one key research challenge is to develop methods that better estimate that input graph, as well as exploit the specific parametric and dynamical properties in a particular domain (e.g., measurement characteristics of different neuroimaging modalities).

At a high level, the MSL and RASL algorithms can be transformed into multi-measurement timescale variants by "post-processing" their outputs: the only possible $\mathcal{G}^1$ are those that are possible for each dataset separately. That is, we can (in theory) straighforwardly amalgamate the causal information produced by these algorithms, as the outputs are all at the same timescale—namely, the causal one. In practice, however, the amalgamation is rarely this easy. The simple intersection of output sets will often be empty, as there may be *no* $\mathcal{G}^1$ that is maximally consistent with each individual dataset. An open research question is how best to perform amalgamation of causal structures in these cases to yield scientifically useful multi-timescale versions of the MSL and RASL algorithms. This work is particularly challenging for RASL, as it generally[14] yields larger—sometimes, much larger—output sets than MSL [24, 23], but is also often the more appropriate algorithm (e.g., if we know only the relative sampling rates of the measurement methods).

Other methods have recently been developed to infer causal structure even in the presence of undersampling or timescale mismatch. For example, non-parametric methods have been extended to use modern constraint satisfaction problem solving tools, as well as answer set programming [12, 13]. Alternatively, earlier parametric approaches [29] have been recently revived to try to directly estimate the (linear) causal strengths as the causal timescale, both for fixed u [11] and mixed undersampling rates [32]. Timescale mismatch is thus a potentially tractable challenge for amalgamation of causal structure. Unfortunately, it is not the only distinct issue for dynamical or time series evidence.

3 The challenge of latent variables

3.1 Latent variables and causal inference

A different type of amalgamation challenge arises when variables are measured in only some experiments or datasets, as we can get seemingly conflicting results depending on exactly what we do (or don't) include. As a concrete example, suppose that dataset $\mathbf{D}_1$ measures $\{X, Y, A\}$, $\mathbf{D}_2$ measures $\{X, Y, B\}$, and $\mathbf{D}_3$ measures $\{X, Y, C\}$. Further suppose we use regression to try to learn from our data, and we find a zero[15] regression coefficient between X and Y in $\mathbf{D}_1$ and $\mathbf{D}_3$, but a significantly non-zero coefficient for $\mathbf{D}_2$.[16] How should

[14] The RASL output always contains at least as many graphs as the MSL output, as the latter considers only a single u. In practice, though, their output sets are frequently equal.

[15] More precisely, not statistically significant from zero.

[16] This would happen if, e.g., the underlying causal structure is: $X \to C \to B \leftarrow A \leftarrow Y$.

we amalgamate this evidence (assuming we do not have unique identifiers for each individual)? It is well-known that different causal inferences are licensed from the same data, depending on whether latent (or unobserved) variables are possible [10]. And for these three datasets, unobserved variables for each $\mathbf{D}$ are not merely possible, but rather are known: for example, we know that B and C are not measured in $\mathbf{D}_1$. Thus, we face a significant challenge in amalgamating the evidence in $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$.

As with the case of mismatched timescales, we can potentially sidestep many of these issues by instead amalgamating at the level of causal or structural knowledge. That is, if we can determine the causal structure underlying each dataset, then we can amalgamate this information by finding the "global" structures that could have produced these local datasets. This strategy has been pursued for static data (i.e., non-time series) in the Integration of Overlapping Networks (ION)[5, 6, 34] and Integration of Overlapping Datasets (IOD)[33] algorithms.

The ION and IOD algorithms reliably, correctly, and efficiently learn the "global" causal structures (i.e., over *all* measured variables) that are consistent with the "local" causal structures inferred from particular datasets. These algorithms can work directly on the local data (IOD), or include prior expert knowledge about the (plausible) local causal structures (ION). As an abstract, toy example, suppose we discover causal structures for two different datasets: from $\mathbf{D}_1$, we find $X \to Y \leftarrow Z$; from $\mathbf{D}_2$, we find $Y \to A \leftarrow B$. Now consider whether there is a causal connection between, say, B and Z. If $B \to Z$ (in the global structure), then B would be an indirect cause of Y (in the global structure) since we would have $B \to Z \to Y$. The *absence* of $B \to Y$ in the local structure for $\mathbf{D}_2$ means that B is not an indirect cause of Y, so we cannot have $B \to Z$ in the global structure. Similarly, if $B \leftarrow Z$ or $B \leftrightarrow Z$ (i.e., a globally unobserved common cause of the two), then there would be a common cause of Y and B (that is unobserved in $\mathbf{D}_2$), which contradicts our previous local learning. Thus, we can conclude that there cannot be any direct causal connection between B and Z, even though we never collected a (local) dataset that contains both. This toy example shows how to infer absence of causal connection, but other local structure patterns can imply definite presence of connections. For example, suppose $\mathbf{D}_1$ implies $X \to A \to Y$; and $\mathbf{D}_2$ implies $X \to B \to Y$. The first graph says that every path from X to Y includes A; the second graph says that every such path flows through B. The only two possible graphs that satisfy both path constraints are: $X \to A \to B \to Y$ and $X \to B \to A \to Y$. Thus, the local structures jointly imply that there must be an A—B causal connection, either $A \to B$ or $A \leftarrow B$. With additional background assumptions, more identification of global causal structure might be possible from the local structures.

The ION and IOD algorithms can be quite powerful, though they still face underdetermination challenges [19]. Moreover, much of their power comes precisely because they do not bother (much[17]) with amalgamating the evidence

[17] IOD actually does perform a type of limited evidence amalgamation.

directly. Instead, they focus on integrating or amalgamating the causal knowledge that we extract from the evidence. That latter content, as well as the algorithms themselves, can represent the possibility of latent or unobserved variables, and so they are not affected in the same way by those latents. They do, however, depend on our ability to extract this causal knowledge in the first place, and that challenge is significantly greater for dynamical systems.

Suppose that we are studying a non-dynamical system where the underlying causal structure is $A \to L \to B$, but we do not observe L. This case presents no particular causal inference challenges: a causal output of the form $A \to B$ always allows for the possibility of an unobserved mediator in that causal connection. Now suppose we have exactly the same structure, but in a dynamical (compressed) graph, and where each variable has a self-loop (i.e., causes itself in the next timestep). Recall that these arrows implicitly encode temporal lag, so this graph represents $A^{t-1} \to L^t$ and $L^{t-1} \to B^t$. In this case, L presents a serious challenge, as the system is no longer Markov order one. Instead, values of A arbitrarily far back in the past can appear to "directly" cause the current state of B, as there are paths of the form $A^{t-k} \to L^{t-k+1} \to ... \to L^{t-1} \to B^t$ in which every mediating variable is unobserved.

The possibility of latent variables in a dynamical system creates novel statistical challenges. For example, an infinite Markov order means that variables arbitrarily far in the past are direct (relative to the measured variables) causes of variables in the current timestep. Our finite sample data obviously do not extend arbitrarily far back in time, though, so we cannot directly test for infinite Markov order. Instead, we must make an inductive leap at some point. A different statistical challenge arises for these "long-distance" connections, as they can become quite weak (though still non-zero). In general, the correlation between A and B induced by paths of the form $A^{t-k} \to L^{t-k+1} \to ... \to L^{t-1} \to B^t$ will decrease as k increases. Thus, for finite sample data, we might not be able to detect the presence of this path, regardless of the particular statistical method that we use, and so might not accurately estimate the Markov order.

The possibility of latent variables also leads to much greater underdetermination worries in the dynamical setting, as there will typically be a huge number of ways to add latent variables in order to save the phenomena. Consider the simple case of a measurement graph in which A^t is directly caused by $A^{t-2}, A^{t-4}, A^{t-6}, ...$ One possibility—in fact, the one with the fewest unobserved variables—is the causal graph $A \rightleftarrows B \rightleftarrows C$, where no variables have self-loops. However, there are many other causal graphs that explain these measurements; for example, perhaps the true causal graph has multiple causal loops involving A that just happen to be of length $2, 4, 6, ...$ Any causal inference method for time series data (with the possibility of latent variables) must address *both* statistical and theoretical/underdetermination challenges, and so amalgamation via causal information must also confront these issues. For reasons of space, we do not consider the statistical issues, but we turn now to the theoretical underdetermination challenges.

3.2 From measured variables to "simplest" causal strutures

The underdetermination challenges for causal inference from time series data are significantly harder than for static data. Suppose that we have a dynamical causal graph $\mathcal{G}_{\mathbf{V}\cup\mathbf{L}}$ over a large set of variables $\mathbf{V} \cup \mathbf{L}$. There are computationally efficient methods [8] for determining the implied dynamic causal graph $\mathcal{G}_{\mathbf{V}}$ that represents the causal relations between variables only in $\mathbf{V}$. We will refer to $\mathcal{G}_{\mathbf{V}}$ as a *reduction* of $\mathcal{G}_{\mathbf{V}\cup\mathbf{L}}$; in the other direction, we call $\mathcal{G}_{\mathbf{V}\cup\mathbf{L}}$ an *extension* of $\mathcal{G}_{\mathbf{V}}$. For example, if we start with the dynamic causal graph $A \to B \to C$ (where we drop temporal superscripts when the edges take one timestep), then we can reduce it to $A^{t-2} \to C$. Alternately, if we start with $X^{t-3} \to Y$, then we can extend it to $X \to L_1 \to L_2 \to Y$. One final notational piece will be helpful: rather than including an edge for every past cause, we simply associate with each $X \to Y$ a set containing the "lags" for which there is an apparent direct causal connection. For example, if $X^{t-3} \to Y$ and $X^{t-2} \to Y$, then we will write $X \to Y$ with the associated set $\{2, 3\}$. These associated sets are the *edge lag sets* for a particular edge.

The underdetermination problem can now be stated more precisely: if $\mathcal{G}_{\mathbf{V}}$ has at least one edge lag set with infinitely many elements, then there are infinitely many extensions of $\mathcal{G}_{\mathbf{V}}$. Moreover, if there are latent variables, then it is quite easy to get an edge lag set with infinitely many elements. For example, if *any* of the latent variables has a self-loop (i.e., $L^{t-1} \to L^t$), then the antecedent will be satisfied. And in all of those cases, there will be infinitely many ways to extending $\mathcal{G}_{\mathbf{V}}$ to yield a dynamical causal graph over $\mathbf{V} \cup \mathbf{L}$ that implies exactly the observed causal connections over $\mathbf{V}$.[18]

One option in the face of this extreme underdetermination is simply to embrace it: we could provide a characterization of those infinitely many extensions, and then say no more. A different option, and the one that we consider here, is to prefer the "simplest" dynamical causal graph extension, understood as the Markov order one graph that postulates the fewest number of unobserved variables. Of course, this preference might be incorrect; perhaps the world really is more complex than is required to explain the data. However, in the absence of additional domain knowledge, we contend that this use of Ockham's Razor—postulate as few additional variables as necessary to explain the observed data—is a reasonable preference.

Even given the Ockham assumption ("prefer extensions with the fewest number of additional variables"), we face a significant challenge to infer that extension from an observed dynamical causal graph $\mathcal{G}_{\mathbf{V}}$, partly because we do not know how many latents will be required. However, we can exploit

[18] Technically minded readers might observe that there are usually infinitely many "extensions" in the static case as well. For example, if $X \leftarrow L \to Y$ for latent L is possible, then the same graph with two latent common causes L_1, L_2 will also fit the data. Thus, the dynamical case does not seem harder in this regard. However, static latents can typically be "collapsed" together (in a technical sense) in a way that these dynamical latents cannot. That is, underdetermination in the static case involves graphs that are the same in many key respects; in the dynamical case, the graphs are quite different.

a significant constraint. Recall that edges are associated with edge lag sets, where the elements in the edge lag set encode the previous timesteps for which the cause C appears to directly cause the effect E. Put differently, the edge lag sets encode the lengths of the paths from C to E that involve only latent variables. This is why we get an infinite edge lag set if there is even a single simple loop involving a latent on a path from C to E.

More generally, suppose that there is a single path, perhaps with multiple loops, from C to E involving only latent variables. The set of possible path lengths can be described in terms of a linear combination of the lengths of the simple loops, and the length of the directed path to which they attach (and so can act as a "backbone"). The algebraic expression for this set of path lengths can get quite complex, though it always can be expressed in terms of a finite series of set sums and set products. If there are multiple paths ("backbones"), then the edge lag set is the union of these sets (each describable with an algebraic representation). Computationally, the marginalization of latents can be performed either sequentially (latent by latent) or in batch mode, where these are provably equivalent [8].

In theory, the inverse operation—obtaining a compressed graph extension that will result in the observed dynamical causal graph when marginalized—is "only" a matter of searching the space of extensions to find one that fits the data with the fewest number of latents. Of course, any brute force approach to this search would be computationally hopeless. Instead, we need to efficiently construct a dynamical causal graph extension so that it implies all and only the observed edge lag sets.

Any algorithm that finds the most parsimonious extension will clearly need to reuse across paths the simple loops (involving latents) whenever possible. Moreover, this desideratum can be readily satisfied because there is an empirical signal for when paths share a latent: for all pairs of edges $X \to Y$ and $A \to B$ in the observed graph $\mathcal{G}_{\mathbf{V}}$, if the underlying paths in $\mathcal{G}_{\mathbf{V} \cup \mathbf{L}}$ share even a single latent, then there will be $X \to B$ and $A \to Y$ in $\mathcal{G}_{\mathbf{V}}$.[19] Thus, if there is *not* $X \to B$, then we know that the paths underlying $X \to Y$ and $A \to B$ do not share any edges (or latents). We can thus reduce the size of the search problem, since we can "modularize" $\mathcal{G}_{\mathbf{V}}$ into maximal bipartite cliques $(\mathbf{V}_1, \mathbf{V}_2)$—the pairs of (largest) sets such that $\forall A \in \mathbf{V}_1, B \in \mathbf{V}_2 \{A \to B\}$. Notably, this modularization will typically be a cover, not a partition; a random variable may belong to multiple maximal bipartite cliques.

Given this modularization of $\mathcal{G}_{\mathbf{V}}$, we can analyze each bipartite clique separately, since we know that there should be maximal sharing of latents within each bipartite clique, and no sharing of latents between cliques. The search for a minimal extension now consists of sequentially constructing simple loops to explain elements of edge lag sets, and then reconciling these loops among the edges. Alternatively, we can first express each edge lag set with simple loops in a bipartite clique and then reconcile the loops. The current implementation

[19] We conjecture that this is actually an "if and only if" when $\mathcal{G}_{\mathbf{V} \cup \mathbf{L}}$ is a minimal extension.

of this process involves two distinct steps; each is provably node-minimal, and we conjecture that the pair of steps is globally node-minimal.

Returning to the overall problem of evidence amalgamation, suppose that we have multiple time series, each with the possiblity of latent variables. By using this algorithm, we can infer minimal extensions (one for each time series) that explain the observed data using the fewest number of latent variables. We can then amalgamate this *causal* information in a relatively straightforward way; for example, the earlier-discussed ION algorithm [34] could theoretically find the set of global graphs (over both measured and inferred latent variables) that imply exactly these causal connections.[20] As with the case of timescale mismatch, we can combine multiple experiments from multiple sources by integrating inferred causal structures.

4 Conclusion

In general, we contend that evidence amalgamation is sometimes best addressed by thinking about the underlying structures that generated the evidence, as we can thereby sidestep some of the standard problems of evidence amalgamation. For example, one challenge in evidence amalgamation is the possibility of different background conditions in different experiments. As a practical example, suppose that X does not cause Y in any individual, but that the base rates of X and Y are both higher in population S_1 than in S_2. If we simply merge data from S_1 and S_2, then we will find an association between X and Y, which suggests a causal connection of some sort. If we instead merge the causal structures inferred from each dataset, then we will correctly learn that there is no causal connection between them (since we will learn "no connection" from each dataset).[21]

Of course, the shift to amalgamating causal information presupposes that we can reliably discover or infer those causal structures. Algorithms for causal discovery from "static" data have been heavily studied over the past 25 years, and their strengths, weaknesses, and assumptions are relatively well-understood. For example, we know the conditions in which a latent variable will be inferrable, or will significantly complicate the discovery of causal connections between measured variables [30]. In contrast, there has been much less focus on conditions in which we can reliably infer dynamical causal structure from time series data.

This paper has focused on two different challenges to reliable discovery of dynamical causal graphs, both of which arise regularly in scientific practice. Timescale mismatch leads to apparently direct causal connections that are

[20] ION would need significant modifications, though, since there is no a priori way to "match" newly introduced latents across graphs. A computationally hopeless algorithm would be to run ION on all possible relabellings of the latents, and then choose the result that minimizes a suitable criterion, such as number of resulting edges. Clearly, a more efficient version would need to be developed.

[21] More generally, merging at the structural level will be superior whenever there is shared structure but varying parameters [33].

actually only indirect in the generating causal structure. Latent variables similarly transform "long-distance" causal connections into direct edges between variables at significantly different times. In both cases, we thus risk having highly misleading inputs to the causal structure amalgamation process.

These inference challenges are distinctive to the time series case, and have gone largely undiscussed in both philosophical and technical literatures. This is particularly surprising given their prevalence in many scientific domains, and given the serious problems that they present for causal inference [27]. Matters are not hopeless, however, as we have identified causal discovery algorithms that are reliable for each challenge. Moreover, the different algorithms can be used in sequence for reliable causal discovery even if both challenges are present.[22] Amalgamation of causal structure (as a way to sidestep some evidence amalgamation challenges) is distinctively harder in the dynamical or time series case, but there are concrete solutions to these challenges.

Acknowledgements Thanks to two anonymous reviewers for their valuable comments and feedback. Thanks to Jianyu Yang for his contributions to the MSL algorithm; Cynthia Freeman for collaboration on the RASL work; and Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo for their work on the constraint-centric formulation of the undersampling problem. Special thanks to Ian Beaver for helping with the code. The latent variable work was conducted in close collaboration with Erich Kummerfeld and Isaac Davis. DD was supported by NSF IIS-1318815 & NIH U54HG008540 (from the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative). SP was supported by NSF IIS-1318759 & NIH R01EB006841. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Ambjorn J, Jurkiewicz J, Loll R (2009) Quantum gravity: the art of building spacetime. Approaches to Quantum Gravity, Cambridge University Press, Cambridge pp 341–359
2. Boubela R, Kalcher K, Nasel C, Moser E (2014) Scanning fast and slow: current limitations of 3 Tesla functional MRI and future potential. Frontiers in Physics 2
3. Burge J, Lane T, Link H, Qiu S, Clark VP (2009) Discrete dynamic bayesian network analysis of fMRI data. Human Brain Mapping 30(1):122–137
4. Cartwright N (1999) The dappled world: A study of the boundaries of science. Cambridge University Press, Cambridge
5. Danks D (2002) Learning the causal structure of overlapping variable sets. In: Discovery Science, Springer, pp 178–191
6. Danks D (2005) Scientific coherence and the fusion of experimental results. British Journal for the Philosophy of Science 56:791–807
7. Danks D, Plis S (2013) Learning causal structure from undersampled time series. In: JMLR: Workshop and Conference Proceedings, vol 1, pp 1–10

[22] We omit details for reasons of space.

8. Davis I, Kummerfeld E, Danks D, Plis S (2015) Inferring observed structure for dynamic graphs with unobserved variables. Tech. Rep. CMU-PHIL-193, Carnegie Mellon University, Department of Philosophy
9. Eichler M (2006) Graphical modeling of dynamic relationships in multivariate time series. In: Schelter B, Winterhalder M, Timmer J (eds) Handbook of time series analysis, Wiley-VCH, chap 14, p 335
10. Glymour C, Madigan D, Pregibon D, Smyth P (1997) Statistical themes and lessons for data mining. Data Mining and Knowledge Discovery 1:11–28
11. Gong M, Zhang K, Schoelkopf B, Tao D, Geiger P (2015) Discovering temporal causal relations from subsampled data. In: Proc. ICML, pp 1898–1906
12. Hyttinen A, Plis SM, Järvisalo M, Eberhardt F, Danks D (2016) Causal discovery from subsampled time series data by constraint optimization. In: Antonucci A, Corani G, de Campos CP (eds) Probabilistic Graphical Models - Eighth International Conference, PGM 2016, Lugano, Switzerland, September 6-9, 2016. Proceedings, JMLR.org, JMLR Workshop and Conference Proceedings, vol 52, pp 216–227, URL http://jmlr.org/proceedings/papers/v52/hyttinen16.html
13. Hyttinen A, Plis SM, Järvisalo M, Eberhardt F, Danks D (in press) A constraint optimization approach to causal discovery from subsampled time series data. International Journal of Approximate Reasoning
14. Kim D, Burge J, Lane T, Pearlson GD, Kiehl KA, Calhoun VD (2008) Hybrid ICA–Bayesian network approach reveals distinct effective connectivity differences in schizophrenia. Neuroimage 42(4):1560–1568
15. Koller D, Friedman N (2009) Probabilistic Graphical Models: Principles and Techniques. The MIT Press
16. Lewis LD, Setsompop K, Rosen BR, Polimeni JR (2016) Fast fMRI can detect oscillatory neural activity in humans. Proceedings of the National Academy of Sciences 113(43):E6679–E6685, DOI 10.1073/pnas.1608117113, http://www.pnas.org/content/113/43/E6679.full.pdf
17. Li J, Wang ZJ, Palmer SJ, McKeown MJ (2008) Dynamic Bayesian network modeling of fMRI: a comparison of group-analysis methods. Neuroimage 41(2):398–407
18. Logothetis NK, Wandell BA (2004) Interpreting the BOLD signal. Annual Review of Physiology 66(1):735–769
19. Mayo-Wilson C (2014) The limits of piecemeal causal inference. British Journal for the Philosophy of Science 65:213–249
20. McCaffrey J, Danks D (in press) Mixtures and psychological inference with resting state fMRI. British Journal for the Philosophy of Science
21. Murphy K (2002) Dynamic Bayesian networks: Representation, inference and learning. PhD thesis, UC Berkeley
22. Pearl J (2000) Causality: models, reasoning, and inference. Cambridge Univ Pr

23. Plis S, Danks D, Freeman C, Calhoun V (2015) Rate agnostic (causal) structure learning. In: Advances in Neural Information Processing Systems 28, Curran Associates, Inc., pp 1–9
24. Plis S, Danks D, Yang J (2015) Mesochronal structure learning. In: Proceedings of the Thirty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-15), AUAI Press, Corvallis, Oregon
25. Plis SM, Weisend MP, Damaraju E, Eichele T, Mayer A, Clark VP, Lane T, Calhoun VD (2011) Effective connectivity analysis of fMRI and MEG data collected under identical paradigms. Computers in Biology and Medicine 41(12):1156–1165
26. Rajapakse JC, Zhou J (2007) Learning effective brain connectivity with dynamic Bayesian networks. Neuroimage 37(3):749–60
27. Seth AK, Chorley P, Barnett LC (2013) Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. Neuroimage 65:540–555
28. Shapere A, Wilczek F (2012) Classical time crystals. Physical review letters 109(16):160,402
29. Silvestrini A, Veredas D (2008) Temporal aggregation of univariate and multivariate time series models: a survey. Journal of Economic Surveys 22(3):458–497
30. Spirtes P, Glymour C, Scheines R (2001) Causation, prediction, and search, vol 81. MIT press
31. Stegenga J (2011) Is meta-analysis the platinum standard of evidence? Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 42(4):497–507
32. Tank A, Fox E, Shojaie A (2016) Identifiability of non-Gaussian structural VAR models for subsampled and mixed frequency time series. In: SIGKDD Workshop on Causal Discovery
33. Tillman RE, Spirtes P (2011) Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In: Proceedings of the International Conference on Artificial Intelligence and Statistics
34. Tillman RE, Danks D, Glymour C (2009) Integrating locally learned causal structures with overlapping variables. Advances in Neural Information Processing Systems 21:1665–72
35. Woodward J (2003) Making things happen: A theory of causal explanation. Oxford University Press, Oxford
36. Zheng X, Rajapakse JC (2006) Learning functional structure from fMRI images. Neuroimage 31(4):1601–13, DOI 10.1016/j.neuroimage.2006.01.031