Proceedings of Machine Learning Research vol 213:1–13, 2023 2nd Conference on Causal Learning and Reasoning

Causal Learning through Deliberate Undersampling

Kseniya Solovyeva KS.P.SOLO@GMAIL.COM
TReNDS center, Georgia State University, Atlanta

David Danks DDANKS@UCSD.EDU
University of California, San Diego

Mohammadsajad Abavisani S.ABAVISANI@GATECH.EDU
TReNDS center, Georgia Institute of Technology

Sergey Plis S.M.PLIS@GMAIL.COM
TReNDS center, Georgia State University, Atlanta

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Domain scientists interested in causal mechanisms are usually limited by the frequency at which they can collect the measurements of social, physical, or biological systems. A common and plausible assumption is that higher measurement frequencies are the only way to gain more informative data about the underlying dynamical causal structure. This assumption is a strong driver for designing new, faster instruments, but such instruments might not be feasible or even possible. In this paper, we show that this assumption is incorrect: there are situations in which we can gain additional information about the causal structure by measuring more *slowly* than our current instruments. We present an algorithm that uses graphs at multiple measurement timescales to infer underlying causal structure, and show that inclusion of structures at slower timescales can nonetheless reduce the size of the equivalence class of possible causal structures. We provide simulation data about the probability of cases in which deliberate undersampling yields a gain, as well as the size of this gain.

1. Introduction

Domain scientists often study systems of interest through the lens of complex multivariate time series collected at fixed intervals as measurements of the interacting parts of the system. The search for causal structures, namely the search for causal relationships between several random variables, has been widely studied at least since Spirtes et al. (1993) with applications in a range of sciences, including social sciences (Blalock, 2017), climatology (Runge et al., 2019), and neuroscience (Zhang et al., 2020). Often, the goal is identification of a form of causal relationship among the parts of the system. The subject of our work is dynamic causal structures that arise as a result of these studies. It has been noted in a number of works that the initial causal structures derived from the data may differ depending on the frequency with which measurements are made (Danks and Plis, 2013; Rhea et al., 2015; Gong et al., 2017).

In multivariate dynamical systems there is a minimal characteristic time scale at which variables of interest exchange information. We call it the causal timescale as opposed to the measurement timescale at which the data is collected. Typically, the measurement timescale is significantly slower than the causal time scale—that is, we have undersampling—which can result in unmeasured intermediate states. For example, the typical resolution of an fMRI scanner is around 2000 ms (Constable, 2006), but the causal interactions of groups of neurons are likely at least 10-20 orders of magnitude faster (Perri et al., 2014; Freiwald et al., 1999). Moreover, in many cases, we do not know the extent

© 2023 K. Solovyeva, D. Danks, M. Abavisani & S. Plis.

of undersampling, and so we do not necessarily know how to infer causal structure from apparent, but perhaps misleading, causal connections (Breitung and Swanson, 2002; Danks and Plis, 2013; Gong et al., 2015). That is, undersampling can lead to causal "illusions," much as a restriction to only one perspective or vantage point can lead to optical illusions.

There are algorithms for searching for a causal structure during undersampling (when the measurement is performed k times less than the time of the causal interaction) with known k (Silvestrini and Veredas, 2008; Gong et al., 2017) and unknown k (Hyttinen et al., 2017; Plis et al., 2015; Abavisani et al., 2023). In each of these cases, there remains an error in the estimation of the causal structure, which decreases with increasing frequency of measurement.

A natural response to the challenge of undersampling is to conclude that we simply should measure faster. If we could somehow measure at the causal time scale, then we could avoid the challenges of undersampling. More generally, a natural intuition is that faster data collection always provides more information about causal interactions in a system, and so we should invest in developing better instrumentation (Fan et al., 2022; Kalhan et al., 2022). But while faster measurements almost always reduce equivalence class size (or at least, do not increase it), faster measurements are typically more costly, whether in time, resources, or personnel. In some cases, it may not even by possible to measure sufficiently quickly without significant complications. For example, increasing the frequency of blood sample collections could enable better understanding of a patient's health, but also impose a significant negative impact on their quality of life. Moreover, measuring *too* fast can introduce additional noise into our system measurements (or perhaps impact the system itself), thereby increasing the data collection requirements. Finally, a single-minded focus on increasing measurement speed can lead researchers to disregard previously-collected, though slower, data when a faster instrument is available.

The central idea in this paper is that this focus on faster measurement is, in some cases, unwarranted. In particular, one can often achieve equal improvements in learning by using *slower* measurements, not only faster ones. The basic intuition is that the slower measurements can provide a second "perspective" on the causal system that potentially reveals insights that were hidden from the first perspective. The present paper provides the first algorithm for reducing the causal structure error using additional measurements at a lower (i.e., slower) measurement frequency.

In general, if we only have measurements at a single timescale, then a slower measurement frequency will typically lead to an increased size for the equivalence class of causal structures that could have produced the observed data. In this sense, the "natural intuition" is correct: if we can only have one dataset, then we should usually aim to have a faster timescale (up to the actual causal timescale). Of course, we are rarely limited in this way; in particular, we can often obtain a second dataset without substantial effort using the same equipment, at least if we are willing to use a slower timescale. The results in this paper show that we can gain significant information by taking this easier path, though the slower timescale must be "coprime" with respect to the original timescale.[1]

This result is arguably surprising—we can learn more by measuring more slowly—but analogical phenomena exist in other domains. Most notably, consider the case of image super-resolution. In this phenomenon, one can obtain substantial increases in image resolution not by using a higher-resolution imaging device, but rather by using a same-resolution device that is offset by a distance that is not an integer multiple of the pixel size. Alternately, one can simply use a *lower*-resolution imaging device;

1. The exact conditions are explained below.

if the higher resolution is not an integer multiple of the lower resolution, then there will necessarily be many "offset" pixels, and so some super-resolution will be possible.

We emphasize that we are not proposing to use measurement at a slower timescale to enable super-resolution of the data itself, not least because the two data streams need not be collected simultaneously. However, the same intuition holds with regards to causal structure: by capturing an undersampled time series with a different "offset," we can learn a more detailed picture of the underlying causal structure. For example, if we use a slower measurement speed in fMRI, then although we do not improve the temporal resolution of the voxel-level data stream, we do improve the temporal resolution of the causal structures in the brain that generate those data streams. That is, we get a "resolution gain" in the causal system, rather than the independent variables.

The present paper thus (i) shows the existence of the counter-intuitive phenomenon that using slower measurements can help as much as faster ones; which thereby (ii) enables scientists to learn more about causal structure from undersampled data using existing measurement instruments, rather than waiting for the development of novel measurement methodologies.

2. Notation and definitions

2.1. Theoretical model

A commonly used modeling framework for dynamical causal systems is based on dynamic (causal) graphical models (Shafer, 1995; Pearl, 1988; Lakshmikantham et al., 2010), often referred to as dynamic Bayesian networks (DBNs). A specific DBN includes random variables $\mathbf{V}$ at the current time step t, as well as nodes for $\mathbf{V}$ at each previous time step in which there is a direct cause of the current values of $\mathbf{V}$. We assume that the underlying causal structure can be represented by a Markov order one graphical model in which no causal interactions happen instantaneously. In this work, we additionally assume absence of instantaneous associations, i.e. causal sufficiency. This assumption is relatively innocuous since we are not assuming that we *measure* at this speed. For convenience, we present the DBN in the form of a compressed graph G in which time information is encoded implicitly in the edges (i.e., $X \to Y$ in G means $X^{t-1} \to Y^t$ in the underlying DBN).

Let $t^0, t^1, t^2, \ldots, t^k, \ldots$ denote the time steps of the underlying causal structure and call the graph G or G^1 corresponding to this structure a *causal graph*. The graph G can contain self-loops, or edges leading from a vertex X to itself (i.e., $X^{t-1} \to X^t$). The data are undersampled with rate u if the measured time steps are $t^0, t^u, t^{2u}, \ldots, t^{ku}, \ldots$. We denote the corresponding measurement timescale graph as G^u, and say that it is *measured* or *undersampled* with rate u. For example, in Figure 1, the first graph is the *causal graph*, the second one is *measured* with $u = 2$, the third one is *measured* with $u = 3$, etc. Undersampling can produce unobserved common causes of variables at the current timestep; we denote those same-timestep associations using bidirectional (red) edges, in contrast with the between-timestep causal relationships in directed (black) edges.

Let $[\![H]\!]$ denote the equivalence class of measured graph H: the set of all graphs G such that $\exists u$ such that $G^u = H$. And let $[\![H_1, H_2, \ldots, H_k]\!]$ denote the set of all G^1 such that $\exists u_1, u_2, \ldots, u_k$ such that $G^{u_1} = H_1, G^{u_2} = H_2, \ldots, G^{u_k} = H_k$. Note, we may sometimes abuse this notation and with $[\![H]\!]$ denote the cardinality of the equivalence class set of H.

We say that a graph G^1 exhibits *non-monotonic undersampling* if and only if $\exists u < v$ such that $[\![G^u, G^v]\!] \neq [\![G^u]\!]$. That is, these are the graphs for which, given some u, v, our uncertainty about the possible G^1 does not monotonically increase as the undersampling increases. More practically, this property implies that additionally measuring G^v enables us to further reduce the space of possible

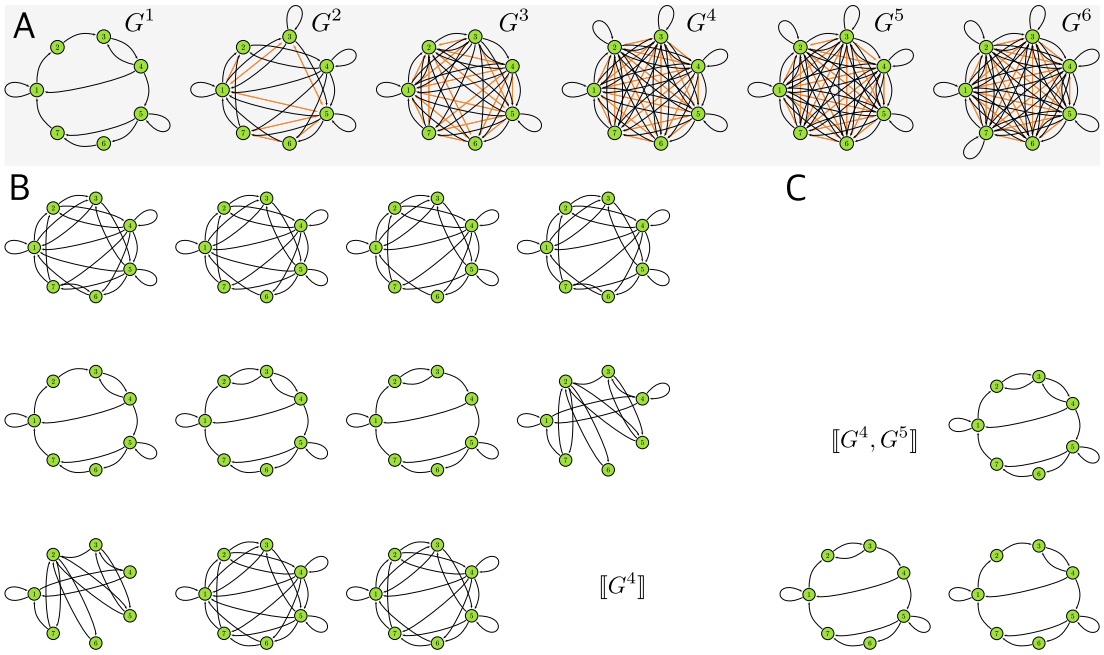


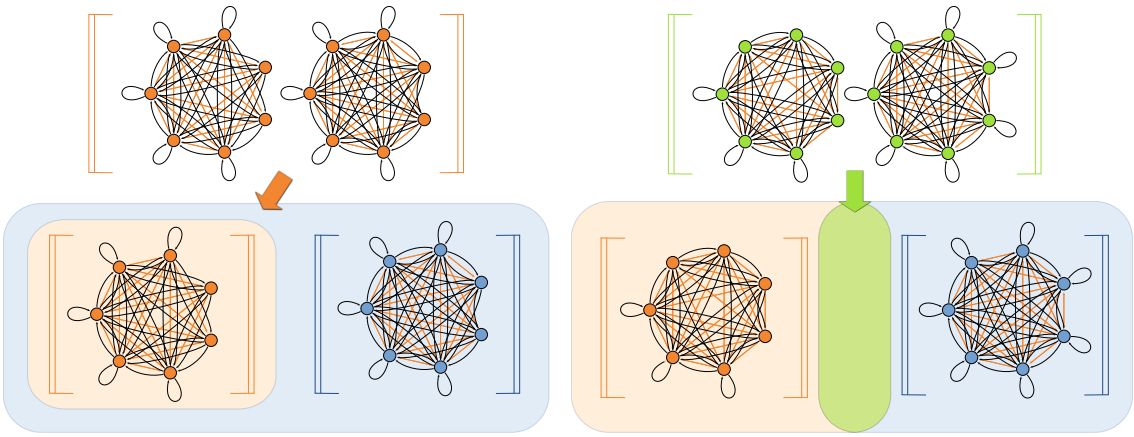
Figure 1: Example of a non-monotonic graph (A). G^1 is the causal graph, while others are its undersampled versions. Black arrows indicate (apparent) direct causal connections, orange bidirectional arrows indicate correlations. The equivalence class of the graph G^4: $[\![G^4]\!]$ (B). The equivalence class of graphs jointly satisfying G^4 and G^5: $[\![G^4, G^5]\!]$ (C). The gain for this case is calculated as $gain = \log_{10} \frac{[\![G^4]\!]}{[\![G^4,G^5]\!]} = \log_{10} \frac{11}{3} \approx 0.56$

G^1, even though G^v is even more undersampled than G^u (see 2). In contrast, graphs that exhibit *monotonic undersampling* are those where $\forall u < v : [\![G^u]\!] \subseteq [\![G^v]\!]$. For clarity, a specific example of a graph with non-monotonic undersampling is shown in 1. For a graph with monotonic undersampling, the classes presented on the left and bottom right would be equal.

2.2. Computational model

We now aim to show that some graphs exhibit non-monotonic undersampling; that is, we can reduce our causal uncertainty by deliberately undersampling even more. For illustrative purposes, we focus on computational experiments using graphs that are strongly connected components (SCCs) with gcd $= 1$ (gcd - the greatest common divisor of the length of all elementary cycles (or simple loops) of an SCC) according to Abavisani et al. (2023). An SCC is a maximal set of nodes $\mathbf{N}$ such that $\forall A, B \in \mathbf{N}$ there is a path $\pi : A \to \ldots \to B$. While SCCs are a subset of all possible graphs, in many scientific contexts, such as biological systems (e.g., homeostatic feedback loops), they are a particularly important subset. More generally, many complicated systems exhibit a phenomenon coined "near decomposability" by Herb Simon (Simon (1982)): they consist of "components" with dense within-module causal connections (including feedback loops) and sparse between-module causal connections. That is, these components are typically SCCs.

The whole set of graphs with directed and bidirected edges can be divided into those that are reachable versus unreachable by undersampling. That is, a graph H is reachable by undersampling if $\exists u > 1, G^1$ such that $G^u = H$ and unreachable otherwise. The proportion of reachable graphs among



(*a*) A monotonic undersampling step (*b*) A non-monotonic undersampling step

Figure 2: Venn diagrams illustrating relations between equivalence classes for two different measured graphs G and H. The blue, orange, and green colors denote measurement timescale graphs and their equivalence classes at different undersampling rates. $[\![G]\!]$ denotes the equivalence class of the graph G. Figure 2(*a*): monotonically increasing undersampling. The left graph is G^5 and the right is G^6. The equivalence class $[\![G^5]\!]$ is a subset of the equivalence class of $[\![G^6]\!]$. In this case, $[\![G^5, G^6]\!] = [\![G^5]\!]$ Figure 2(*b*): non-monotonically increasing undersampling. The left graph is H^5 and the right is H^6. Equivalence classes overlap but $[\![H^5]\!]$ is not a subset of $[\![H^6]\!]$. Although the class sizes still relate as $[\![H^5]\!] \leq [\![H^6]\!]$, this leads to $[\![H^5, H^6]\!] < [\![H^5]\!]$.

random graphs is small,[2] so it makes sense to generate G^u from a known class of graphs G so we can better discern patterns in the value of deliberate undersampling. In particular, if G^1 consists of a single SCC with gcd = 1, then all G^u are also a single SCC (Danks and Plis, 2013). In practical terms, we focus on n-node graphs composed of a simple cycle of length n plus k random edges (possibly self-loops) such that gcd = 1.

3. A Primer on `clingo` and Answer Set Programming (ASP)

`clingo` (Gebser et al., 2011) combines a grounder `gringo` and a solver `clasp`. `clingo` is a declarative programming system based on logic programs and their answer sets, used to accelerate solutions of computationally involved combinatorial problems. The grounder converts all parts of a `clingo` program to "atoms," (grounds the statements) and the solver finds "stable models." In ASP, the answer set is a model in which all the atoms are derived from the program and each "answer" is a stable model where all the atoms are simultaneously true.

A general `clingo` program includes three main sections, which we show below using our algorithm as an example:

1. **Facts:** these are the known elements of the problem. For example, the input to Listing 1 is a graph for which we know the edges. A directed edge from node 1 to node 5 is in $\mathcal{H}$ translates to *`hdirected(1,5)`* (line 1) or if node 1 is part of the SCC number 2, we state this fact in `clingo` by *`scc(1,2)`* (line 2).

2. Since H can have bidirected edges, the space of H is bigger than the space of G^1 by a factor of $2^{(n \text{ choose } 2)}$. Any particular G^1 can lead to multiple graphs in H-space (as u varies), but H-space is only sparsely covered by various G^u.

2. **Rules:** much like an `if-else` statement, a rule in `clingo` consists of a body and a head, formatted as `head :- body.` If all the literals in the body are true, then the head must also be true. Rules can include variables (starting with capital letters), and they are used to derive new facts after grounding. For example:

```
directed(X, Y, 1) :- edge1(X, Y).
```

means that for any instantiations of the variables X and Y, if we have an edge from X to Y, there is a directed path from X to Y of length 1. Before this line, if the model contained the fact `edge1(2,3)`, this line would generate a new fact: `directed(2,3,1)`.

Another type of rule is the "choice rule" that describes all the possible ways to choose which atoms are included in the model. For example, in line 5 of Listing 1 we used a choice rule to state that the undersampling rate `u` can be anything from 1 to `maxu`. The cardinality constraint:

```
{u(1..20)}.
```
(1)

will generate 2^{20} different models (they will not all actually be generated if they conflict with other predicate in each model, or else it would not be possible). In each of these 2^{20} models, one subset of all possible atoms generated with this choice rule exists (ϕ, `{u(1)}`, `{u(1), u(2)}`, ...). An example of an unconstrained choice rule is line 6 in Listing 1, where we want to generate one model for each possible way edges can be present in a graph between two nodes X and Y. We can also limit the choice rule. In our problem, only one undersampling rate is present at each solution. We limit the cardinality constraint to have only one member in each model:

```
1 {u(1..20)} 1.
```
(2)

the 1 on the left is the minimum instantiations of this atom in the model and the 1 on the right is the maximum. Therefore, we only generate $\binom{20}{1} = 20$ models with this rule, namely one for each undersampling rate. Having several choice rules will multiply the number of generated models by each choice rule.

3. **Integrity Constraints:** if choice rules are to generate new models, integrity constraints are there to remove the wrong models from the answers set. More specifically, an integrity constraint is of the form:

```
:- L0, L1, ...  .
```
(3)

where literals $L_0, L_1, ...$ cannot be simultaneously positive. For example, in line 16 of Listing1, we have:

```
:- edge1(X, Y), scc(X, K), scc(Y, L), K != L,
        sccsize(L, Z), Z > 1, not dag(K,L).
```
(4)

for cases where the graph consists of several SCCs that are connected using a DAG. If the SCCs are connected by a cyclic directed graph, then the whole graph will become one big Strongly Connected Component. Integrity constraint 4 states that if there is not a directed edge from a node in SCC K to a node in SCC L as part of the initial DAG, there cannot be such `edge1(X, Y)` from node X to node Y, if node X is in SCC K and node Y is in SCC L.

```
1  %( * input graph 1 edge specifications here * e.g.: hdirected(1,5,1) ... )
2  %( * input graph 2 edge specifications here * e.g.: hdirected(1,5,2) ... )
3  #const n = 10, maxu = 20
4  node(1..n).
5  1 {u(1..maxu, 1)} 1.
6  1 {u(1..maxu, 2)} 1.
7  {edge1(X,Y)} :- node(X), node(Y).
8  directed(X, Y, 1) :- edge1(X, Y).
9  directed(X, Y, L) :- directed(X, Z, L-1), edge1(Z, Y), L <= U, u(U, _).
10 bidirected(X, Y, U) :- directed(Z, X, L), directed(Z, Y, L), node(X;Y;Z), X < Y, L < U, u(U, _).
11 :- directed(X, Y, L), not hdirected(X, Y, K), node(X;Y), u(L, K).
12 :- bidirected(X, Y, L), not hbidirected(X, Y, K), node(X;Y), u(L, K), X < Y.
13 :- not directed(X, Y, L), hdirected(X, Y, K), node(X;Y), u(L, K).
14 :- not bidirected(X, Y, L), hbidirected(X, Y, K), node(X;Y), u(L, K), X < Y.
15 :- M = N, {u(M, 1..2); u(N, 1..2)} == 2, u(M, _), u(N, _).
```

Listing 1: Clingo code for sRASL

4. dRASL

In this section, following Abavisani et al. (2023), we present and explain the `clingo` answer set programming (ASP) solver implementation of a solution for the rate agnostic causal learning (RASL) problem (Plis et al., 2015).The RASL algorithm finds all graphs from the equivalence class of a single graph H. We follow the conventions of Abavisani et al. (2023) and call the algorithm dRASL, were d stands for "deliberate undersampling." Our modification of the RASL algorithm is a generalization. dRASL finds a general equivalence class for any number of graphs $[\![H_1, H_2, \ldots, H_k]\!]$. Listing 1 presents the essential parts of the `clingo` code for the problem. `clingo` is a declarative programming language that operates on predicates. Declarative compression is not algorithmic, so we describe below the `clingo` code which is identical commonly used to ASP encoding (Hyttinen et al., 2014). Lines $1-2$ are generated automatically by Python given two input graphs estimated at "co-prime" measurement rates (or generated ensuring this property).

These are the graph-specific predicates. The first line lists the directed and bidirected edges of both graphs H^u and H^{u+1}. For example, existence of edge from node 1 to node 5 in H^u is translated to `hdirected(1, 5, 1)`. Line 2 focuses on the same but for H^{u+1}. Line 4 sets the number of nodes in H and sets the maximum undersampling rate to consider. The undersampling rate needs to have a fixed upper bound. However, it is often possible to guess this upper bound from domain-specific considerations. For example, in functional MRI, the maximum undersampling rate can arguably be set to 20 as the difference between hundred milliseconds of the neural causal timescale and 2 second measurement time scale hardly can be more than 20. Also, the upper bound for u can be determined mathematically. For example, for any graph with n nodes and shortest simple loop length $s \geq 2$ is $\leq (n-1)^2 + 1 - \frac{1}{2}(s-2)(2n-s-3)$ according to (Heap and Lynn, 1964). Another bound for the same involving graph diameter is given in (Plis et al., 2015).

Lines $5, 6$ set two different predicates for under sampling rates, one for each H. These undersamplings can be anywhere from 1 to `maxu`. Line 7 assumes the existence of all possible edges in the graph G^1. Lines $8, 9, 10$ generate predicates for directed and bidirected paths of lengths L, practically encoding the forward algorithm of undersampling. Line 9 imposes connection conditions between directed edges in graphs G^L, G^{L-1}, and G^1. Line 10 describes how bidirectional edges are defined in a graph G^L. Lines $11-14$ ensure that the candidate G^1 agrees with both H^u and H^{u+1} in terms of having the same directed and bidirected paths of different lengths, as well as absence of the same directed and bidirected paths of different lengths. In other words, these lines exclude all solutions in

which the generated possible graph G^1 with undersampling rates $u(L, K)$ do not match the initially given graphs H^u and H^{u+1}. Finally, line 15 makes sure that H^u and H^{u+1} are indeed distinct graphs.

The main results of the correctness of this algorithm are based on the correctness of the problem formulation in Listing 1, which follows by construction, and validity and completeness of `clingo`'s ASP solver algorithms. The latter is proven elsewhere, including (Hyttinen et al., 2017; Abavisani et al., 2023)

5. Results

In this section, we present the results of our computational experiments on the potential learning benefits of deliberate undersampling. In particular, we empirically demonstrate that non-monotonic undersampling is quite common in the class of SCCs with gcd = 1, which forms the largest, and arguably the most practically relevant, class of SCCs.

5.1. Estimating prevalence of non-monotonic undersampling graphs

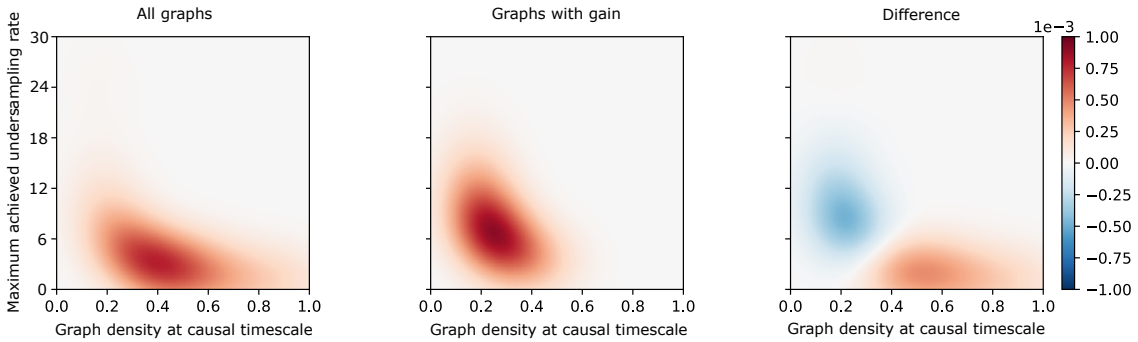


Figure 3: Kernel density estimation (KDE) of graphs with gcd $= 1$. X-axis: The density of the causal graph G^1. Y-axis: The undersampling rate after which the graph G^1 converges to a super-clique or cyclically repeats. (a) Normalized KDE for all SCCs. (b) Normalized KDE for SCCs with *non-monotonic undersampling* mode. (c) Difference between KDE of all SCCs and SCCs with *non-monotonic undersampling* modes.

The purpose of this computational experiment is to find a lower bound on the proportion of *non-monotonic undersampling* causal graphs G. To estimate this bound, we generated 100 SCCs with n vertices connected in a large ring and a random arrangement of k additional edges for each value of $n \in \{5, 6, 7, 8, 9\}$ and $k \in [1, 30]$. Only graphs with gcd $= 1$ were considered as the most frequent and the most practically relevant. Presence of even a single auto-correlation in an SCC leads to gcd $= 1$ and auto-correlations are common in most studied systems.

Given a randomly generated SCC G^1, we produced a list of G^u graphs for all u values available for this graph before convergence. After that incrementing u by 1 we sequentially ran sRASL of Abavisani et al. (2023) to produce $[\![G^u]\!]$ and our dRASL modification on (G^u, G^{u+1}) to obtain $[\![G^u, G^{u+1}]\!]$ The SCC was considered as non-monotonic in the evaluation if there is any u such that $[\![G^u, G^{u+1}]\!] < [\![G^u]\!]$ or, in other words, there is a non-zero gain in using dRASL (see 2b).

As a visualization of the properties of graphs with a gain and without, we present graphs of kernel density estimation (KDE) 3. Notably, the density of graph with gain (3b) is a lower bound on the true density of such graphs, since it follows from the definition of *non-monotonic undersampling*

graphs that v can be anything greater than u as long as they are co-prime, while we restrict ourselves to the case $v = u + 1$. Moreover, due to computational limitations, we used the dRASL algorithm to test u for the random SCCs only up to a maximum u, which is lower than the maximum value of u for graphs with k edges added to the initial ring, when $k < 4$. Therefore, the estimate for graphs with $k < 4$ could be much higher than that presented in the graph.

As we can see from 3, deliberate undersampling is most helpful for the sparse G^1 graphs. In 3 the X-axis represents the density of the original SCC we generated (as described in the Section 2.2), or the graph on the causal time scale. Such graphs (3a) are not uniformly distributed from 0 to 1 in density, but are distributed from the minimum to the maximum values of the function $(n + k)/n^2$. The Y-axis shows how many different graphs can be obtained from the original graph G by the undersampling operation G^u. 3b shows the kernel density estimate for those graphs that, according to our lower bound estimate, turned out to be *non-monotonic undersampling*.

From the difference in the normalized KDE distributions (3c) of all graphs and graphs for which deliberate undersampling leads to a gain, we see that SCCs with gain are most likely to be found among graphs with a lower density and higher maximum u.

5.2. Density of measured *non-monotonic undersampling* graphs

Above, we considered density of SCCs that have non-monotonic undersampling mode relative to total number of graphs. From a practical point of view, the properties of the measured G^u, for which deliberate undersampling gives a gain, seem more interesting. On the histogram in 4, the abscissa shows the density of the causal timescale graph G^1, and the ordinate shows the density of the measurement timescale graph for all *non-monotonic undersampling* graphs in the experiment. Density is computed as N/n^2, where N - is the number of all edges of the current graph and n - the number of nodes. As can be seen from the figure, the largest number of measured graphs is in the density range from 0.5 to 0.8, while the density of causal graphs does not exceed 0.6.

5.3. Statistics of the gain of *non-monotonic undersampling* graphs

The purpose of this experiment is to estimate the gain given by the deliberate undersampling. The gain is calculated as $gain = \log_{10} r_1/r_{12}$. Where $r_1 = [\![G^u]\!]$ and $r_{12} = [\![G^u, G^{u+1}]\!]$. This means, that, if $gain = 2$, the deliberate undersampling reduces the number of solutions 100-fold. For clarity, a specific example of a non-monotone graph is shown in 1. The gain for this case is calculated as $gain = \log_{10} \frac{[\![G_4]\!]}{[\![G_4,G_5]\!]} = \log_{10} \frac{11}{3} \approx 0.56$. It is worth noting that the gain in this section determines how the number of solutions (the size of the equivalence class) changes when an additional, slower dimension is added. It would help to revisit our fMRI example, where we have assumed that the causal time scale is 100 ms, the time resolution is 2000 ms, which led to the undersampling rate of 20.[3] In this case, adding a measurement for $u + 1$ corresponds to a measurement with a time resolution of 2100 ms. In this case, we measured the graph G^u and got r_1 possible graphs, from which we can get the graph G^u by undersampling. When we collect the data at the new, slower, temporal resolution, we get an additional constraint that the graphs in the causal temporal resolution must comply with, and the size of the equivalence class became r_{12}. Two histograms (5) show the same *non-monotonic undersampling* graphs in two different time scales - in the causal (or original

3. It is more common to study interaction between large groups of neurons rather than individual cells. These groups, commonly referred to as intrinsic networks or regions of interest (ROIs), contain multiple layers of neurons so the "causal timescale of an ROI" could easily be as high as 1000 ms (i.e., $u = 2$)

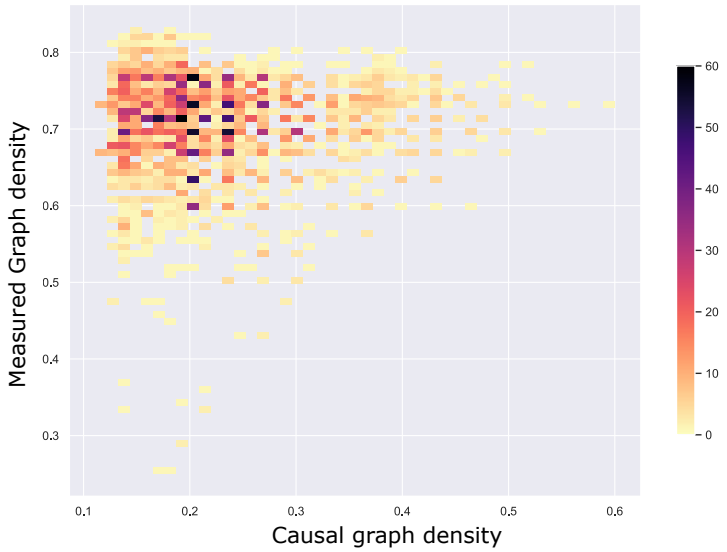


Figure 4: The histogram of *non-monotonic undersampling* graphs G. X-axis: The density of the graph in causal timescale G^1. Y-axis: The density of the graph in measured timescale G^u. Color map: number of graphs. All graphs are n-cyclic graphs with k additional random edges ($k = 1, 2, ..30$; 100 samples for each k). $n = 5, 6, 7, 8, 9$ (total 15000).

.

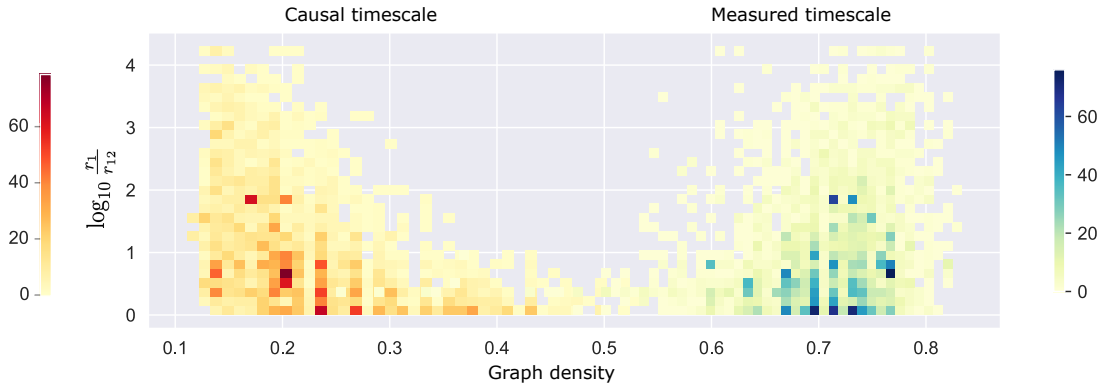


Figure 5: The histogram of the gain of *non-monotonic undersampling* graphs G. X-axis: The density of the graph. Y-axis: Decimal logarithm of r_1/r_{12} where $r_1 = [\![G^u]\!]$ and $r_{12} = [\![G^u, G^{u+1}]\!]$. Z(colour): number of graphs. All graphs are n-cyclic graphs with k additional random edges ($k = 1, 2, ..30$; 100 samples for each k). $n = 5, 6, 7, 8, 9$ (total 15000). (left): The histogram in causal timescale. (right): The histogram in measured timescale.

G^1 graphs) and in the measured (or G^u graphs). Although there are individual cases for which the gain is 3-4 orders of magnitude, but most of them are concentrated in the area of gain from units to hundreds (0-2 orders of magnitude).

It is interesting to relate the densities of our non-monotonic graphs to real world graphs. The density of measurement and causal timescale graphs in the real world depends on the nature of systems at investigation. For example, the density of social graphs can vary depending on whether we are considering the entire social network of hundreds of millions of users or a small part of it. Similarly, for brain networks the density of graphs may differ condition on whether we consider individual neurons or entire regions of the brain as the nodes. Often the density of causal graphs (which characterize interaction dynamics) is unknown, but its order can be estimated from the density (d) of structural graphs. For example: peer-to-peer (web) file sharing $d = 0.2258, N = 6049$ (Iamnitchi et al., 2004), constraint programming $d = 0.33, N = 240$ (Melancon, 2006), web pages $d = 0.2088, N = 589$ (Melancon, 2006), passenger air traffic 2000 $d = 0.112, N = 1148$ (Melancon, 2006), social network (IMDB) $d = 0.1794, N = 419$ (Auber et al., 2003), social graph of US and Chinese entrepreneurs $d = 0.1 - 0.8, N = 2193$ (Burt, 2019)), brain networks $d = 0.26 - 0.38$ (Fallani et al., 2008). As we can see, the orders of graph densities correlate with the densities of the non-monotone graphs we have considered in the causal timescale.

6. Conclusion

This work highlights the problem of distortion of the measured causal structure when measuring with a time resolution lower than the characteristic causal time. We have positively answered the question whether it is possible to use deliberate undersampling to reduce indeterminacy of the problem compared to the algorithms using data at a single sampling rate. Surprisingly, in many cases deliberate undersampling gives a desired effect of reducing the "causal illusion". Having considered a specific class of graphs for which the causal system does not degenerate under slow measurement, we have shown the statistical features of the density of such graphs and the gain that can be obtained using deliberate undersampling. Our work serves as a demonstration that there is much to gain when faster measurements are not an option for any of the possible reasons. One of the larger expected impacts of this result may come from a potential reuse of historical data collected at slower timescales, when datasets with somewhat faster sampling rates are already available. Limitations of our work include the missing theoretical investigation on the conditions when deliberate undersampling is helpful in getting gaining information. Ideally, we would want a test on the measured graph to tell whether deliberate undersampling would help. However, even characterization of the properties conducive to the gain of G^1 graphs in the causal timescale could be insightful. This, however, is still work in progress.

Acknowledgments

This work was supported by NIH R01MH129047 and in part by NSF 2112455, and NIH 2R01EB006841. We are grateful to Antti Hyttinen, Matti Järvisalo, and Frederick Eberhardt for discussions on `clingo`.

References

Mohammadsajad Abavisani, David Danks, and Sergey Plis. Constraint-based causal structure learning from undersampled graphs. *ICLR*, 2023.

David Auber, Yves Chiricota, Fabien Jourdan, and Guy Melançon. Multiscale visualization of small world networks. In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, pages 75–81. IEEE, 2003.

Jr Blalock. *Causal models in the social sciences*. Routledge, 2017.

Jörg Breitung and Norman R Swanson. Temporal aggregation and spurious instantaneous causality in multiple time series models. *Journal of Time Series Analysis*, 23(6):651–665, 2002.

Ronald S Burt. Network disadvantaged entrepreneurs: Density, hierarchy, and success in china and the west. *Entrepreneurship Theory and Practice*, 43(1):19–50, 2019.

R Todd Constable. Challenges in fMRI and its limitations. In *Functional MRI*, pages 75–98. Springer, 2006.

David Danks and Sergey Plis. Learning Causal Structure from Undersampled Time Series. 12 2013. doi: 10.1184/R1/6492101.v1. URL https://kilthub.cmu.edu/articles/journal_contribution/Learning_Causal_Structure_from_Undersampled_Time_Series/6492101.

F De Vico Fallani, L Astolfi, F Cincotti, D Mattia, A Tocci, S Salinari, MG Marciani, H Witte, A Colosimo, and F Babiloni. Brain network analysis from high-resolution eeg recordings by the application of theoretical graph indexes. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(5):442–452, 2008.

Jingjing Fan, Kevin Sitek, Bharath Chandrasekaran, and Abhra Sarkar. Bayesian tensor factorized mixed effects vector autoregressive processes for inferring Granger causality patterns from high-dimensional neuroimage data. *arXiv preprint arXiv:2206.10757*, 2022.

Winrich A Freiwald, Pedro Valdes, Jorge Bosch, Rolando Biscay, Juan Carlos Jimenez, Luis Manuel Rodriguez, Valia Rodriguez, Andreas K Kreiter, and Wolf Singer. Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *Journal of neuroscience methods*, 94(1): 105–119, 1999.

Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider. Potassco: The Potsdam answer set solving collection. *Ai Communications*, 24(2):107–124, 2011.

Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pages 1898–1906. PMLR, 2015.

Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal discovery from temporally aggregated time series. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2017. NIH Public Access, 2017.

BR Heap and MS Lynn. The index of primitivity of a non-negative matrix. *Numerische Mathematik*, 6(1): 120–141, 1964.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, 90:208–225, 2017.

Adriana Iamnitchi, Matei Ripeanu, and Ian Foster. Small-world file-sharing communities. In *IEEE INFOCOM 2004*, volume 2, pages 952–963. IEEE, 2004.

Shivam Kalhan, Jessica McFadyen, Naotsugu Tsuchiya, and Marta I Garrido. Neural and computational processes of accelerated perceptual awareness and decisions: A 7T fMRI study. *Human Brain Mapping*, 2022.

Vangipuram Lakshmikantham, Srinivasa Leela, Zahia Drici, and FA McRae. *Theory of causal differential equations*, volume 5. Springer, 2010.

Guy Melancon. Just how dense are dense graphs in the real world? a methodological note. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7, 2006.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Rinaldo Livio Perri, Marika Berchicci, Donatella Spinelli, and Francesco Di Russo. Individual differences in response speed and accuracy are associated to specific brain activities of two interacting systems. *Frontiers in behavioral neuroscience*, 8:251, 2014.

Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate-agnostic (causal) structure learning. *Advances in neural information processing systems*, 28, 2015.

Christopher K Rhea, Adam W Kiefer, W Geoffrey Wright, Louisa D Raisbeck, and F Jay Haran. Interpretation of postural control may change due to data processing techniques. *Gait & posture*, 41(2):731–735, 2015.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in Earth system sciences. *Nature communications*, 10(1):1–13, 2019.

Scheines Shafer, Glenn. Causation, prediction and search (book review). *Synthese*, 104(1):161–176, 1995.

Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497, 2008.

Herbert Simon. The structure of complex systems. *Models of bounded rationality*, 3, 1982.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer, 1993.

Gemeng Zhang, Aiying Zhang, Biao Cai, Zhuozhuo Tu, Vince D Calhoun, and Yu-Ping Wang. Detecting abnormal connectivity in schizophrenia via a joint directed acyclic graph estimation model. *arXiv preprint arXiv:2010.13029*, 2020.