JMLR: Workshop and Conference Proceedings 1: 1-10 NIPS 2013 workshop on causality

Learning Causal Structure from Undersampled Time Series

David Danks DDANKS@CMU.EDU
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA, 15213, USA

Sergey Plis SPLIS@MRN.ORG
The MIND Research Network
1101 Yale Blvd. NE
Albuquerque, NM, 87106, USA

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Even if one can experiment on relevant factors, learning the causal structure of a dynamical system can be quite difficult if the relevant measurement processes occur at a much slower sampling rate than the "true" underlying dynamics. This problem is exacerbated if the degree of mismatch is unknown. This paper gives a formal characterization of this learning problem, and then provides two sets of results. First, we prove a set of theorems characterizing how causal structures change under undersampling. Second, we develop an algorithm for inferring aspects of the causal structure at the "true" timescale from the causal structure learned from the undersampled data. Research on *causal* learning in dynamical contexts has largely ignored the challenges of undersampling, but this paper provides a framework and foundation for learning causal structure from this type of complex time series data.

Keywords: Causal learning, causal inference, time series, undersampling

1. Causal Inference, Time Series, and Undersampling

When faced with a difficult causal learning challenge, one often turns to experimentation or interventions as a way to reduce the complexity of the problem (e.g., by eliminating the influence of unobserved common causes). This strategy is substantially more complicated, however, when learning the causal structure of a dynamical system. In particular, the "proper" experimental strategy is often to provide some exogenous shock to the dynamical system, measure its evolution, and then apply a causal learning algorithm to the resulting time series data, such as Demiralp and Hoover (2003); Moneta et al. (2011); Friedman et al. (1999); Voortman et al. (2010). All such algorithms, however, assume that the "measurement timescale" matches the "causal timescale." That is, they assume that we measure fast enough to capture the direct causal relations.[1]

In contrast, we focus in this paper on the quite common situation in which the measurement timescale is slower than the causal timescale; that is, our data are undersampled relative to the "true" speed of causal connections. To make the causal learning prob-

1. Algorithms that learn from independent samples from an equilibrium distribution (of a dynamical system) essentially assume arbitrarily slow measurement timescales, and so are not applicable here.

©2013 David Danks and Sergey Plis.

lem even more realistic, we assume that the degree of undersampling is unknown. These two assumptions—possibility of undersampling plus lack of knowledge about its extent—characterize many application domains, including causal learning from fMRI or learning climatological relations. Moreover, causal learning is truly hard under these conditions, since undersampling can lead to true causal connections disappearing or changing (apparent) direction while spurious causal connections appear at the measurement timescale.

This paper focuses on the theoretical aspects of causal learning given unknown undersampling. We first provide a precise formal characterization of the problem (Section 2). We then provide theorems about the "forward" inference problem (Section 3): how does the true causal structure change for different undersampling rates? We finish by using those theorems to provide the first algorithm (in Section 4) for learning features of the true causal structure given only the causal structure learned at the measurement timescale. The algorithm is almost certainly not complete, but does demonstrate that the learning problem is not hopeless: we can learn about the causal structure of a dynamical system even when we have data undersampled at an unknown rate. The results reported here are part of a larger research project, and so represent only "first steps" towards a full theory of causal prediction and learning given undersampling. In particular, open questions remain about (at least) learning given small datasets, weak causal connections, and small amounts of undersampling. Nonetheless, this paper provides the framework for asking such questions.

2. Representing Temporal Causal Structures

Dynamical causal systems can be modeled using dynamic versions (Murphy, 2002; Eichler, 2006) of causal graphical models (Spirtes et al., 2001; Pearl, 1988; Lakshmikantham et al., 2009); for convenience, we will refer to these as dynamic Bayesian networks or DBNs. Time is modeled as proceeding in discrete steps.[2] We assume that there are no *isochronal* causal connections: that is, there are no direct causal relations between variables in the same timestep. This assumption is relatively innocuous given that we do not assume that we measure at, or even know, the precise causal timescale.

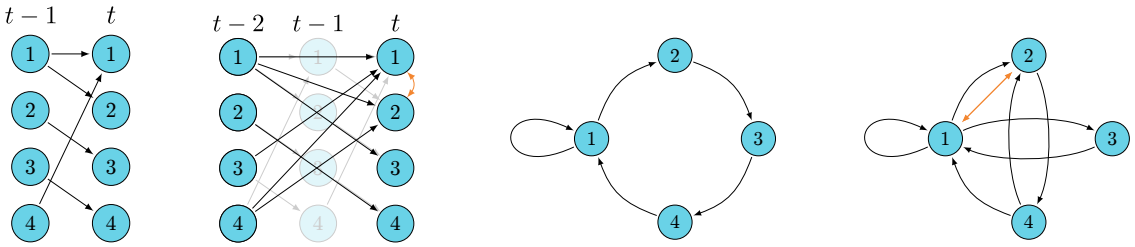
A DBN contains a graph $\mathbf{G}$ over random variables $\mathbf{V}$ at the current timestep t, as well as nodes for $\mathbf{V}$ at each previous timestep in which there is a direct cause of the current values of $\mathbf{V}$ (see Figure 1a for an example). The DBN also has a probability distribution or density $P(n\mathbf{V})$ over all of the nodes in $\mathbf{G}$. $\mathbf{G}$ and $P(n\mathbf{V})$ are connected through the standard *causal Markov*[3] and *causal Faithfulness*[4] (or Stability or Minimality) assumptions. We assume that the "true" causal structure is first-order Markov,[5] and so the only direct causes of variables at t are those at $t-1$. In our setting, $\mathbf{G}$ is thus over $2\mathbf{V}$, and the only permissible edges are $V_i^{t-1} \rightarrow V_j^t$, where possibly $i = j$ (i.e., a variable can cause its own value in the next timestep).

Let $\{t^0, t^1, t^2, \ldots, t^k, \ldots\}$ denote the timesteps at the causal timescale. Say that a data sequence is *(under)sampled at rate u* if the measured timesteps are $\{t^0, t^u, \ldots, t^{ku}, \ldots\}$ (and

2. Our focus throughout will thus be on difference equations. The results in this paper will be applicable to systems governed by differential equations just to the extent that those differential equations can be approximated sufficiently closely by a corresponding set of difference equations.
3. Variables are independent of their non-effects given their direct causes.
4. Variables that are independent conditional on some set are non-adjacent.
5. The Markov order is the largest k such that some $V_i \in \mathbf{V}$ at $t-k$ is a *direct* cause of some $V_j \in \mathbf{V}$ at t.

so the causal timescale is "sampled at rate 1"). Let $\mathbf{G}^1$ denote the true underlying causal structure (i.e., the structure at the causal timescale). To determine the implied $\mathbf{G}$ at other timescales, we first "unroll" $\mathbf{G}^1$ by adding instantiations of $\mathbf{G}^1$ at previous timesteps, where $\mathbf{V}^{t-2}$ bear the same causal relationships to $\mathbf{V}^{t-1}$ that $\mathbf{V}^{t-1}$ bear to $\mathbf{V}^t$, and so forth. In this unrolled (though still time-indexed by t, not absolute times) graph, all $\mathbf{V}$ at intermediate timesteps are not measured, so we marginalize them out to yield a graph $\mathbf{G}^u$.

There are rules for marginalizing out variables (Richardson and Spirtes, 2002), where the result is an *A*cyclic *D*irected *M*ixed *G*raph (ADMG) (Richardson and Spirtes, 2002; Richardson, 2003). In particular, $V_i^{t-u} \to V_j^t$ in $\mathbf{G}^u$ iff there is a directed path from V_i^{t-u} to V_j^t in the unrolled $\mathbf{G}^1$, and $V_i^t \leftrightarrow V_j^t$ in $\mathbf{G}^u$ iff there is a trek between V_i^t and V_j^t whose paths have length $k < u$ in the unrolled $\mathbf{G}^1$ (i.e., a common cause fewer than u timesteps back is being marginalized out). It is straightforward to see that these bidirected edges are the only isochronal edges in $\mathbf{G}^u$, and if such an edge occurs in $\mathbf{G}^u$, then it occurs in $\mathbf{G}^m$ for all $m > u$. The result of "unrolling-and-marginalizing" Figure 1a is shown in Figure 1b. For example, $1^{t-2} \to 3^t$ in $\mathbf{G}^2$ because unrolling $\mathbf{G}^1$ yields a graph in which $1^{t-2} \to 2^{t-1} \to 3^t$. Similarly, $1^t \leftrightarrow 2^t$ in $\mathbf{G}^2$ because $1^t \leftarrow 1^{t-1} \to 2^t$ in the unrolled $\mathbf{G}^1$.



(a) DBN $\mathbf{G}^1$ (b) DBN $\mathbf{G}^2$ (c) $\mathcal{G}^1$ for (1a) (d) $\mathcal{G}^2$ for (1b)

Figure 1: Undersampling example

Unrolling-and-marginalizing a DBN ADMG can be quite computationally complex and unilluminating, so we instead use *compressed graphs*, which encode temporal information in the edges, rather than the variables. Specifically, for arbitrary DBN graph $\mathbf{H}$, $\mathcal{H}$ is its compressed graph representation such that: (i) $\mathcal{H}$ is over *non*-time-indexed versions of $\mathbf{V}$; (ii) $V_i \to V_j$ in $\mathcal{H}$ iff $V_i^{t-1} \to V_j^t$ in $\mathbf{H}$; and (iii) $V_i \leftrightarrow V_j$ in $\mathcal{H}$ iff $V_i^t \leftrightarrow V_j^t$ in $\mathbf{H}$. Compressed graphs can have cycles ($V_i \to V_j \to V_i$ in $\mathcal{H}$ means $V_i^{t-1} \to V_j^t$ and $V_j^{t-1} \to V_i^t$ in $\mathbf{H}$) and even self-loops ($V_i^{t-1} \to V_i^t$ in $\mathbf{H}$ implies $V_i \to V_i$ in $\mathcal{H}$). Note that there is clearly a 1-1 mapping between DBN ADMGs and compressed graphs. The compressed graph representations of $\mathbf{G}^1$ and $\mathbf{G}^2$ are shown in Figures 1c and 1d, respectively.

Instead of unrolling-and-marginalizing $\mathbf{G}^1$, we can determine $\mathcal{G}^u$ (and so $\mathbf{G}^u$) by finding directed sequences[6] in $\mathcal{G}^1$. Specifically, $V_i \to V_j$ in $\mathcal{G}^u$ iff there is a directed sequence from V_i to V_j of length u in $\mathcal{G}^1$. Similarly, $V_i \leftrightarrow V_j$ in $\mathcal{G}^u$ iff there is a balanced trek[7] of length $k < u$ between V_i and V_j in $\mathcal{G}^1$. Given that compressed graphs are computationally simpler to study in the context of undersampling, we focus on $\mathcal{G}^1$ and $\mathcal{G}^u$ in the following sections.

6. A directed path in which nodes can be visited multiple times
7. In this context, a balanced trek of length k between V_i and V_j is a pair of directed sequences $\pi_1 : V_c \to \ldots \to V_i$ and $\pi_2 : V_c \to \ldots \to V_j$ in $\mathcal{G}^*$ such that $length(\pi_1) = length(\pi_2) = k$.

3. "Forward" Inference

We first consider the impact of undersampling on the apparent causal structure at the (possibly undersampled) measurement timescale: given $\mathcal{G}^1$ and some u, what is implied about $\mathcal{G}^u$? This section provides theorems (proofs in the Appendix) characterizing the impact of undersampling, many of which depend on the notion of a *strongly connected component* (SCC). An SCC in a compressed graph $\mathcal{H}$ is a maximal set of nodes $\mathbf{S} \subseteq \mathbf{V}$ such that: for every $X, Y \in \mathbf{S}$ there is a directed path from X to Y. SCCs always pick out cyclic (sub)graphs, since there must be a directed path from X to Y and also from Y to X. The one exception is that a singleton node is always a (perhaps non-maximal) SCC, even if it does not have a self-loop. If $\mathcal{G}^1$ has no (non-singleton) SCCs (i.e., if the compressed graph representation of $\mathbf{G}^1$ is a DAG), then undersampling destroys the directed edges:

Theorem 1 *If $\mathcal{G}^1$ is a DAG (with no self-loops) and l is the length of the longest directed path in $\mathcal{G}^1$, then $\forall u > l$, all edges of $\mathcal{G}^u$ (possibly none) are bidirected.*

Next, suppose $\mathcal{G}^1$ contains nonsingleton SCCs. Any SCC is composed of a set of *simple* loops: directed cycles in which no node is repeated. For example, the graph in Figure 1c is a single SCC with two simple loops (the four-variable loop, and the self-loop at 1). Let $\mathcal{L}_{\mathbf{S}}$ be the set of lengths of the simple loops in SCC $\mathbf{S}$, and let $\gcd(\mathcal{L}_{\mathbf{S}})$ be the greatest common divisor of those simple loop lengths. Interestingly, the gcd is the critical factor in determining asymptotic (as undersampling increases) stability (Theorem 2 below), as well as stable SCC membership regardless of undersampling (Theorem 3). Moreover, if $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$, then the structure converges (as u increases) to a graph we call a *superclique* (Theorem 4).

Theorem 2 *$\exists m \forall i, j > m$[$\mathcal{G}^i$ & $\mathcal{G}^j$ are the same graph] iff $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ for SCC $\mathbf{S} \in \mathcal{G}^1$*

Theorem 3 *$\mathbf{S}$ is an SCC in $\mathcal{G}^u$ for all u iff $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ for SCC $\mathbf{S} \in \mathcal{G}^1$*

Theorem 4 *If $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ for SCC $\mathbf{S} \in \mathcal{G}^1$, then $\exists u$ such that, for all undersampling rates $k > u$, $\mathbf{S}$ forms a superclique: each $X, Y \in \mathbf{S}$ are connected in $\mathcal{G}^k$ via $\leftarrow$, $\rightarrow$, and $\leftrightarrow$ (and so $\forall X \in \mathbf{S}$, X has a self loop).*

Note that the condition of $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ is quite weak, as it only requires that each SCC either have at least one node with a self-loop (i.e., one variable that is a cause of its own value in the next timestep), or be sufficiently "dense" with loops of different lengths. Moreover, if there is a self-loop node, we can bound the time-to-convergence to a superclique in terms of two quantities: for a given node X, define $len_{X\rightarrow} = \max_{B \in \mathbf{S}} \left[\min_{\sigma: X \rightarrow \ldots \rightarrow B} length(\sigma) \right]$ and $len_{\rightarrow X} = \max_{A \in \mathbf{S}} \left[\min_{\pi: A \rightarrow \ldots \rightarrow X} length(\pi) \right]$. Theorem 5 implies that an SCC with a self-loop can become a superclique even with relatively little undersampling.

Theorem 5 *If $\exists Q \in \mathbf{S}$ with a self-loop in $\mathcal{G}^1$, then $\mathbf{S}$ is a superclique in $\mathcal{G}^{len_{Q\rightarrow} + len_{\rightarrow Q}}$, where $(len_{Q\rightarrow} + len_{\rightarrow Q}) \leq 2(|\mathbf{S}| - 1)$ is a tight bound.*

4. Causal Search

The results in the previous section can be used as a basis for causal search: given $\mathcal{G}^u$ for some unknown u, what features of the true underlying $\mathcal{G}^1$ can be inferred? There is clearly a significant problem of underdetermination here. For example, the details of the proof of Theorem 1 imply that *any* compressed graph DAG in which variables have at most one child will, for sufficiently large n, be an empty graph after enough undersampling. So the set of $\mathcal{G}^1$ consistent with $\mathcal{G}^u$ being an empty graph is incredibly large, given that we do not know the actual u. At the same, the problem is not completely intractable, as we may be able to infer partial information about the structure of $\mathcal{G}^1$.

Theorem 3 showed that the node-membership of an SCC does not change when we undersample, regardless of the undersampling rate. There are polynomial-time algorithms for finding the SCCs in a graph (Cormen et al., 2001), and so the first step in inferring $\mathcal{G}^1$ structure is to identify the SCCs in $\mathcal{G}^u$. For some compressed graph $\mathcal{H}$, we can construct the "SCC-graph" $\mathcal{H}_\mathbf{S}$. Each node $S_i \in \mathcal{H}_\mathbf{S}$ corresponds to an SCC $\mathbf{S}_i$ in $\mathcal{H}$, and $S_i \to S_j$ in $\mathcal{H}_\mathbf{S}$ iff there exists at least one $X_i \in \mathbf{S}_i$ and at least one $X_j \in \mathbf{S}_j$ such that $X_i \to X_j$ in $\mathcal{H}$. $\mathcal{H}_\mathbf{S}$ encodes the "SCC parent" relations: $S_i \to S_j$ in $\mathcal{H}_\mathbf{S}$ iff some node in $\mathbf{S}_i$ is a parent in $\mathcal{H}$ of some node in $\mathbf{S}_j$. $\mathcal{H}_\mathbf{S}$ provably has a useful structure (Theorem 6) whose ancestral relations (though not necessarily the direct parent relations) are unchanged by undersampling when $\gcd(\mathcal{L}_\mathbf{S}) = 1$ for all SCCs $\mathbf{S}$ (Corollary 7).

Theorem 6 *For all $\mathcal{H}$, $\mathcal{H}_\mathbf{S}$ is a DAG*

Corollary 7 *If $\gcd(\mathcal{L}_\mathbf{S}) = 1$ for all $\mathbf{S} \in \mathcal{G}^1$, then $\forall i, j$[same ancestral relations in ${\mathcal{G}_\mathbf{S}}^i$ & ${\mathcal{G}_\mathbf{S}}^j$].*

Theorem 4 showed that an SCC $\mathbf{S}$ with $\gcd(\mathcal{L}_\mathbf{S}) = 1$ becomes a superclique with enough undersampling. This convergence destroys all information about the internal structure of $\mathbf{S}$. However, if $\mathbf{S}$ has not yet converged to a superclique, then we can sometimes extract some information from $\mathcal{G}^u$, such as the definite existence of a self-loop in $\mathcal{G}^1$ (Theorem 8).

Theorem 8 *If $\gcd(\mathcal{L}_\mathbf{S}) = 1$ for SCC $\mathbf{S} \in \mathcal{G}^1$ and X is the only self-loop node in $\mathbf{S}$ in $\mathcal{G}^u$, then X has a self-loop in $\mathcal{G}^1$.*

There are currently no theorems demonstrating the full information set that can be extracted, but we can nonetheless assemble the pieces that we do have into Algorithm 1 that extracts constraints on $\mathcal{G}^1$ from some given $\mathcal{G}^u$.

5. Conclusion

This paper provides a formal foundation for causal learning given unknown undersampling, and presents a preliminary algorithm for learning from such data. Rather than focusing on the theory, one could instead approach the problem by trying to develop a score-based procedure that searches more directly through $\langle \mathcal{G}, u \rangle$ space to find possible structures at the causal timescale. Plis and Danks (in prep) implement an MCMC search over this space and show that the $\mathcal{G}^1$ structure can often be partly recovered, though there are significant statistical challenges. We thus have multiple algorithms that can extract some (though obviously not all) features of the "true" causal structure from undersampled time

Algorithm 1: Recovering definite features of $\mathcal{G}^1$

Input: undersampled ADMG $\mathcal{G}^u$

begin construct $\mathcal{G}_{\mathbf{S}}$
  identify SCCs $\mathbf{S}_1, \ldots, \mathbf{S}_k$ in $\mathcal{G}^u$;
  $\mathbf{V_S} \leftarrow \{S_1, \ldots, S_k\}$ & $\mathcal{G}_{\mathbf{S}} \leftarrow$ empty graph over $\mathbf{V_S}$;
  forall the $\langle \mathbf{S}_i, \mathbf{S}_j \rangle$ **do**
    if $\exists X_i \in \mathbf{S}_i, X_j \in \mathbf{S}_j (X_i \rightarrow X_j)$ **then**
      Add $S_i \rightarrow S_j$ to $\mathcal{G}_{\mathbf{S}}$;
begin identify definite self-loops
  foreach $\mathbf{S}_i$ **do**
    if $\exists$ *unique* $X \in \mathbf{S}_i (X \rightarrow X)$ **then** label $X \rightarrow X$ as definite in $\mathcal{G}^1$;

series data. The obvious gap in the present paper is the lack of an algorithm for learning within-SCC structure when the undersampling rate is low enough that the SCC is not yet a superclique. That problem is focus of ongoing research, but the results presented here suggest that the overall problem should be soluble, despite its obvious complexity.

Acknowledgments

Both authors contributed equally to this paper. This research was supported by NSF awards IIS-1318815 (DD) and IIS-1318759 (SP), as well as a James S. McDonnell Foundation Scholar Award (DD).

References

T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2001.

S. Demiralp and K.D. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and statistics*, 65(s1):745–767, 2003.

M. Eichler. Graphical modeling of dynamic relationships in multivariate time series. In B. Schelter, M. Winterhalder, and J. Timmer, editors, *Handbook of time series analysis*, pages 335–372. Wiley-VCH, 2006.

N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *15th Annual Conference on Uncertainty in Artificial Intelligence*, pages 139–147, San Francisco, 1999. Morgan Kaufmann.

V. Lakshmikantham, S. Leela, Z. Drici, and FA McRae. *Theory of causal differential equations*. World Scientific, 2009.

A. Moneta, N. Chlaß, D. Entner, and P. Hoyer. Causal search in structural vector autoregressive models. In *JMLR W&CP: Causality in Time Series*, volume 12, pages 95–114, 2011.

K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, 2002.

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

S. Plis and D. Danks. MCMC search for undersampled causal structure. Technical report, Carnegie Mellon University and Mind Research Network, in prep.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

T. S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2001.

M. Voortman, D. Dash, and M. Druzdzel. Learning why things change: The difference-based causality learner. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 641–650. AUAI Press, 2010.

Supplementary Information: Appendix A. Proofs of theorems

Multiple theorems below use the notion of a *numerical semigroup*: a set $\mathbf{N} \subseteq \mathbb{N}$ such that:

- $0 \in \mathbf{N}$
- $\mathbb{N} \setminus \mathbf{N}$ is finite
- $\forall x, y \in \mathbf{N}, (x + y) \in \mathbf{N}$

Numerical semigroups can always be defined by a basis set $\mathbf{B} = \{x_1, x_2, \ldots, x_r\}$, where $n \in \mathbf{N}$ iff $n = \sum_{i=1}^{r} \alpha_i x_i$ for $\alpha_i \in \mathbb{N}$. A well-known fact about numerical semigroups is:

Observation. $\mathbf{N}$ is a numerical semigroup iff $\gcd(\mathbf{B}) = 1$.

Note also that there is a number $g_{\mathbf{N}}$ for each $\mathbf{N}$ (its so-called Frobenius number) such that $\forall n > g_{\mathbf{N}}, n \in \mathbf{N}$. We can then provide the following Lemma:

Lemma 9 $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ *iff* $\exists m \forall n > m$ *such that, for every* $A, B \in \mathbf{S}$*, there is a directed sequence* $\pi : A \to \ldots \to B$ *&* $length(\pi) = n$*.*

Proof $\Rightarrow$) Let $\mathcal{L}$ be the set of simple loops in $\mathbf{S}$. For any $A, B \in \mathbf{S}$, there is a directed sequence σ_{AB} from A to B that passes through every simple loop at least once. By adding complete loops in the middle of σ_{AB}, we can construct directed sequences from A to B of length: $length(\sigma_{AB}) + \sum_{L \in \mathcal{L}} n_L length(L)$, where $n_i \in \mathbb{N}$. Since $\gcd(\mathcal{L}) = 1$, the set of possible values for the sum forms a numerical semigroup, and so all $l > g_{\mathcal{L}}$ (i.e., larger than the Frobenius number for $\mathcal{L}$) can be formed by some such sum. This implies that, for all $l > (g_{\mathcal{L}} + length(\sigma_{AB}))$, there is a sequence of length l from A to B. The right-hand side of the lemma thus holds when $m = \max_{A,B \in \mathbf{S}} (g_{\mathcal{L}} + length(\sigma_{AB}))$.

$\Leftarrow$) Assume there is some m such that, for all $A, B \in \mathbf{S}$, there are always length $n > m$ sequences from A to B. The length of any directed sequence from A to B can always be expressed as $\sum length(\delta_i) + \sum_{L \in \mathcal{L}} n_L length(L)$, where the δ_i's are directed paths (not loops/sequences). Since $\mathbf{S}$ is finite, there are only finitely many directed paths and all have finite length. Hence, the only way for the assumption to hold is for there to be some k such that every $l > k$ can be formed from $\sum_{L \in \mathcal{L}} n_L length(L)$. Therefore, $\sum_{L \in \mathcal{L}} n_L length(L)$ must be a numerical semigroup, which implies $\gcd(\mathcal{L}) = 1$. ■

Theorem 1 *If* $\mathcal{G}^1$ *is a DAG (with no self-loops) and* l *is the length of the longest directed path in* $\mathcal{G}^1$*, then* $\forall u > l$*, all edges of* $\mathcal{G}^u$ *(possibly none) are bidirected.*

Proof For any $A, B \in \mathbf{V}, A \to B \in \mathcal{G}^u$ iff $\exists \pi \in \mathcal{G}^1 : A \to \ldots \to B$ and $length(\pi) = u$. By assumption, u is greater than the longest directed path in $\mathcal{G}^1$, and so there can be no such π. Hence, the only possible edges in $\mathcal{G}^u$ are bidirected ones. ■

Theorem 2 $\exists m \forall i, j > m$*[*$\mathcal{G}^i$ *&* $\mathcal{G}^j$ *are the same graph]* *iff* $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ *for SCC* $\mathbf{S} \in \mathcal{G}^1$

Proof $\Rightarrow$) Prove the contrapositive: assume $\exists \mathbf{S} \gcd(\mathcal{L}_{\mathbf{S}}) \neq 1$. By Lemma 9, this implies that $\exists A, B \in \mathbf{S}$ such that there is no m for which $\forall n > m, \exists \pi : A \to \ldots \to B$ & $length(\pi) = n$. But that means that, for all u_1 such that $A \to B$ is in $\mathcal{G}^{u_1}$, there is some $u_2 > u_1$ such that $A \to B$ is *not* in $\mathcal{G}^{u_2}$, and vice versa. Hence, $\mathcal{G}^u$ does not stabilize as $u \to \infty$.

$\Leftarrow$) Assume $\forall \mathbf{S} \gcd(\mathcal{L}_{\mathbf{S}}) = 1$. The proof of Theorem 4 (below) shows that every $\mathbf{S}$ converges to a superclique. Now consider $A \in \mathbf{S}_1$ and $B \in \mathbf{S}_2$. If A is an ancestor of B, then there is a directed path π from A to B. By Lemma 9, there is some m such that there are directed sequences from B to B for all lengths $l > m$, and so directed sequences from A to B for all lengths $l^* > (length(\pi) + m)$, and so $A \to B$ in all $\mathcal{G}^u$ for $u > (length(\pi) + m)$. Alternately (and not exclusively), suppose that A and B share an ancestor C. Similar reasoning shows that there must some q such that there is a balanced trek between A and B with C as the head, and with arms of length q. Thus, $A \leftrightarrow B$ in all $\mathcal{G}^u$ for $u > q$. Finally, suppose that A and B share no common ancestor (and so neither is an ancestor of the other). In this case, A and B will be non-adjacent in $\mathcal{G}^u$ for all u, since there are (by assumption) no directed sequences or balanced treks between them. Hence, $\mathcal{G}^u$ eventually converges. ■

Theorem 3 $\mathbf{S}$ *is an SCC in* $\mathcal{G}^u$ *for all* u *iff* $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ *for SCC* $\mathbf{S} \in \mathcal{G}^1$

Proof $\Rightarrow$) Prove the contrapositive: assume $\gcd(\mathcal{L}_{\mathbf{S}}) \neq 1$. Any directed sequence σ from A back to itself will be composed entirely of complete simple loops, and so $length(\sigma) = r \times \gcd(\mathcal{L}_{\mathbf{S}})$ for some r. Now consider some arbitrary $A, B \in \mathbf{S}$ such that $A \to B$. For a given u, there will be a path from A to B in $\mathcal{G}^u$ iff there is a directed sequence from A to B of length ku (for integer k) in $\mathcal{G}^1$. Let $u = \gcd(\mathcal{L}_{\mathbf{S}})$. First, suppose $A \to B$ is the only directed path from A to B in $\mathcal{G}^1$, so any sequence ν from A to B in $\mathcal{G}^1$ must be a sequence from A to itself, followed by $A \to B$. So, any such ν must have $length(\nu) = ru + 1$ for some $r \geq 0$, and so there is no k such that $length(\nu) = ku$ since $u > 1$. Hence, there is no path from A to B in $\mathcal{G}^u$. Now suppose that there are other directed paths $\pi_1, \ldots, \pi_c$ from A to B in $\mathcal{G}^1$. Let $\sigma_1, \ldots, \sigma_d$ be all simple loops in $\mathbf{S}$ that involve $A \to B$. Since σ_j is a simple loop in $\mathbf{S}$, $length(\sigma_j) = r_j \times \gcd(\mathcal{L}_{\mathbf{S}})$ for some r_j. Define μ_{ij} to be the sequence from A to itself formed by π_i followed by σ_j without $A \to B$. By definition, $length(\mu_{ij}) = length(\pi_i) + length(\sigma_j) - 1$. Also, $length(\mu_{ij}) = s_{ij} \times \gcd(\mathcal{L}_{\mathbf{S}})$ for some s_{ij}. Therefore, $length(\pi_i) = s_i \times \gcd(\mathcal{L}_{\mathbf{S}}) + 1$ for some s_i. So for any sequence ν from A to B in $\mathcal{G}^1$ formed by a sequence from A to itself followed by one of the π_i, $length(\nu) = ru + 1$ for some $r \geq 0$, and so there is no k such that $length(\nu) = ku$. Hence, there is no path from A to B in $\mathcal{G}^u$. The node-membership of the SCC thus changes, since $A, B \in \mathbf{S}$ for $\mathcal{G}^1$, but $A, B \notin \mathbf{S}$ for $\mathcal{G}^u$ when $u = \gcd(\mathcal{L}_{\mathbf{S}})$.

$\Leftarrow$) Consider $A, B \in \mathbf{S}$ and some arbitrary u. By Lemma 9, $\exists m \forall n > m \forall A, B \in \mathbf{S} \exists \pi : A \to \ldots \to B$ & $length(\pi) = n$. Let k be the smallest integer such that $ku > m$. By the Lemma, there is a directed sequence σ from A to B in $\mathcal{G}^1$ with $length(\sigma) = ku$, so there is a directed sequence from A to B in $\mathcal{G}^u$ of length k. Now consider $A \in \mathbf{S}$ and $B \notin \mathbf{S}$, and suppose that they in the same SCC in $\mathcal{G}^u$. In that case, there must be directed sequences $\pi_1 : A \to \ldots \to B$ and $\pi_2 : B \to \ldots \to A$ in $\mathcal{G}^u$. By definition of edges in $\mathcal{G}^u$, that means that there are directed sequences $\pi_1^* : A \to \ldots \to B$ and $\pi_2^* : B \to \ldots \to A$ in $\mathcal{G}^1$. Thus, A and B were members of the same SCC in $\mathcal{G}^1$, contrary to assumption. Hence, the membership of $\mathbf{S}$ does not change as u varies. ■

Theorem 4 *If* $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ *for SCC* $\mathbf{S} \in \mathcal{G}^1$*, then* $\exists u$ *such that, for all undersampling rates* $k > u$*,* $\mathbf{S}$ *forms a superclique: each* $X, Y \in \mathbf{S}$ *are connected in* $\mathcal{G}^k$ *via* $\leftarrow$*,* $\rightarrow$*, and* $\leftrightarrow$ *(and so* $\forall X \in \mathbf{S}$*,* X *has a self loop).*

Proof By Lemma 9, there is some m such that every pair of variables A, B is connected by directed sequences of length n for all $n > m$. Let $u = m+1$. This implies that, for all $k > u$, there is a directed sequence of length k from A to B, and so $A \to B$ is part of $\mathcal{G}^k$. Moreover, for any third variable $C \in \mathbf{S}$, there must be sequences with length $\leq m+1$ from C to A, and from C to B. That is, there is a balanced trek with arm-length $\leq m+1$ between A and B, and so $A \leftrightarrow B$ will be part of $\mathcal{G}^k$ for all $k > u$. Hence, $\mathbf{S}$ forms a superclique in $\mathcal{G}^k$ for all $k > u$. ■

Theorem 5 *If $\exists Q \in \mathbf{S}$ with a self-loop in $\mathcal{G}^1$, then $\mathbf{S}$ is a superclique in $\mathcal{G}^{len_{Q\to}+len_{\to Q}}$, where $(len_{Q\to} + len_{\to Q}) \leq 2(|\mathbf{S}| - 1)$ is a tight bound.*

Proof For convergence to a superclique, note that the definitions of $len_{Q\to}$ and $len_{\to Q}$ imply: for any $A, B \in \mathbf{S}$, there is a sequence $\pi : A \to \ldots \to Q \to \ldots \to B$ with $length(\pi) \leq (len_{Q\to} + len_{\to Q})$. Thus, $A \to B$ in $\mathcal{G}^{len_{Q\to}+len_{\to Q}}$ (possibly by adding self-loops of Q to the middle of π). Furthermore, all paths σ from Q to any A and B each have $length(\sigma) \leq len_{Q\to}$. Therefore, there is a balanced (possibly by adding self-loops of Q to one of the arms) trek of length $\leq len_{Q\to} < (len_{Q\to} + len_{\to Q})$ between any A, B. Thus, $A \leftrightarrow B$ for all $u \geq (len_{Q\to} + len_{\to Q})$. Hence, $\mathbf{S}$ in $\mathcal{G}^{len_{Q\to}+len_{\to Q}}$ is a superclique. For the bound, note that, by definition of an SCC, there is a directed sequence from any X to any Y. The shortest such sequence must be a directed path, and so $len_{Q\to}$ and $len_{\to Q}$ both must be $\leq |\mathbf{S}| - 1$ since a directed path uses every node in $\mathbf{S}$ at most once. To see that the bound is tight, let $\mathbf{S}$ be a single simple loop (plus the self-loop at Q). In this case, $len_{\to Q} = |\mathbf{S}| - 1$ when A is Q's child and $len_{Q\to} = |\mathbf{S}| - 1$ when B is Q's parent. ■

Theorem 6 *For all $\mathcal{H}$, $\mathcal{H}_{\mathbf{S}}$ is a DAG*

Proof Suppose $\mathcal{H}_{\mathbf{S}}$ is not a DAG for some $\mathcal{H}$. Let π be a cyclic path in $\mathcal{H}_{\mathbf{S}} : S_{\pi(1)} \to S_{\pi(2)} \to \ldots \to S_{\pi(1)}$. By construction of $\mathcal{H}_{\mathbf{S}}$, there must be, for all $j < length(\pi)$, $X_{o(j)} \in \mathbf{S}_{\pi(j)}, X_{i(j+1)} \in \mathbf{S}_{\pi(j+1)}$ such that $X_{o(j)} \to X_{i(j+1)}$ in $\mathcal{H}$. Because each $\mathbf{S}_j$ is an SCC in $\mathcal{H}$, there is a within-SCC directed path between any two nodes in the SCC. By joining these within-SCC paths with the between-SCC edges, we can construct a path π^* in $\mathcal{H}$ from any node in some $\mathbf{S}_{\pi(i)}$ to any node in some other $\mathbf{S}_{\pi(j)}$, perhaps by going "around the cycle" through $\mathbf{S}_{\pi(1)}$. But that implies that we have only one SCC that includes all nodes from all $\mathbf{S}_{\pi(i)}$, contrary to construction of $\mathcal{H}_{\mathbf{S}}$. ■

Theorem 8 *If $\gcd(\mathcal{L}_{\mathbf{S}}) = 1$ for SCC $\mathbf{S} \in \mathcal{G}^1$ and X is the only self-loop node in $\mathbf{S}$ in $\mathcal{G}^u$, then X has a self-loop in $\mathcal{G}^1$.*

Proof Suppose X does not have a self-loop in $\mathcal{G}^1$, and so $u > 1$. In that case, the self-loop in $\mathcal{G}^u$ must be because of a directed sequence $\pi = X \to Y \to \ldots \to X$ of length u in $\mathcal{G}^1$. Moving the $X \to Y$ edge from being first in π to the end of π yields the directed sequence $\pi^* = Y \to \ldots \to X \to Y$. π^* has length u in $\mathcal{G}^1$, so Y should also have a self-loop in $\mathcal{G}^u$, contrary to the assumption that X is the only self-loop in $\mathbf{S}$. ■