Discrete Mathematics and Theoretical Computer Science **AB(DMCS)**, 2003, 11–30

Optimal Nonlinear Prediction of Random Fields on Networks

Cosma Rohilla Shalizi†

Center for the Study of Complex Systems, 4485 Randall Laboratory, University of Michigan, Ann Arbor, MI 48109 USA

It is increasingly common to encounter time-varying random fields on networks (metabolic networks, sensor arrays, distributed computing, etc.). This paper considers the problem of optimal, nonlinear prediction of these fields, showing from an information-theoretic perspective that it is formally identical to the problem of finding minimal local sufficient statistics. I derive general properties of these statistics, show that they can be composed into global predictors, and explore their recursive estimation properties. For the special case of discrete-valued fields, I describe a convergent algorithm to identify the local predictors from empirical data, with minimal prior information about the field, and no distributional assumptions.

Keywords: Networks, random fields, sufficient statistics, nonlinear prediction, information theory, recursive estimation

1 Introduction

Within the field of complex systems, most interest in networks has focused either on their structural properties (Newman, 2003), or on the behavior of known dynamical systems coupled through a networks, especially the question of their synchronization (Pikovsky *et al.*, 2001). Statistical work, ably summarized in the book by Guyon (1995), has largely (but not exclusively) focused on characterization and inference for *static* random fields on networks. In this paper, I consider the problem of predicting the behavior of a random field, with unknown dynamics, on a network of fixed structure. Adequate prediction involves reconstructing those unknown dynamics, which are in general nonlinear, stochastic, imperfectly measured, and significantly affected by the coupling between nodes in the network. Such systems are of interest in many areas, including biochemistry (Bower and Bolouri, 2001), sociology (Young, 1998), neuroscience (Dayan and Abbott, 2001), decentralized control (Šijlak, 1990; Mutambara, 1998) and distributed sensor systems (Varshney, 1997). Cellular automata (Ilachinski, 2001) are a special case of such systems, where the graph is a regular lattice.

There are two obvious approaches to this problem of predicting network dynamics, which is essentially a problem of system identification. One is to infer a global predictor, treating entire field configurations as measurements from a time series; this is hugely impractical, if only because on a large enough network,

†Supported by a grant from the James S. McDonnell Foundation

no configuration ever repeats in a reasonable-sized data sample. The other straightforward approach is to construct a distinct predictor for each node in the network, treating them as so many isolated time-series. While more feasible, this misses any effects due to the coupling between nodes, which is a serious drawback. In these contexts, we often know very little about the causal architecture of the systems we are studying, but one of the things we do know is that the links in the network are causally relevant. In fact, it is often *precisely* the couplings which interest us. However, node-by-node modeling ignores the effects of the couplings, which then show up as increased forecast uncertainty at best, and systematic biases at worst. We cannot guarantee optimal prediction unless these couplings are taken into account. (See the end of Section 3.1.) I will construct a distinct predictor for each node in the network, but these predictors will explicitly take into account the couplings between nodes, and use them as elements in their forecasts.

By adapting tools from information theory, I construct optimal, nonlinear *local* statistical predictors for random fields on networks; these take the form of minimal sufficient statistics. I describe some of the optimality properties of these predictors, and show that the local predictors can be composed into global predictors of the field's evolution. Reconstructing the dynamics inevitably involves reconstructing the underlying state space, and raises the question of determining the state from observation, i.e., the problem of filtering or state estimation. There is a natural translation from the local predictors to a filter which estimates the associated states, and is often able to do so with *no* error. I establish that it is possible to make this filter recursive without loss of accuracy, and show that the filtered field has some nice Markov properties. In the special case of a discrete field, I give an algorithm for approximating the optimal predictor from empirical data, and prove its convergence.

Throughout, my presentation will be theoretical and abstract; for the most part, I will not deal with practical issues of implementing the method. In particular, I will slight the the important question of how much data is needed for reliable prediction. However, the method *has* been successfully applied to fields on regular lattices (see Section 5 below).

I will make extensive use of (fairly basic) information theory and properties of conditional independence relations, and accordingly will assume some familiarity with conditional measures. I will not pay attention to measure-theoretic niceties, and shall assume that the random field is sufficiently regular that all the conditional measures I invoke actually exist and are conditional probability distributions. Readers, for their part, may assume all functions to be measurable functions.

The next section of the paper establishes the basic setting, notation, and preliminary results, the latter mostly taken without proof from standard references on information theory and conditional independence. Section 3 constructs the local predictors and establishes their main properties. Section 4 establishes the results about transitions between states which are related to recursive filtering. Section 5 discusses an algorithm for identifying the states from empirical data on discrete fields.

2 Notation and Preliminaries

2.1 The Graph and the Random Field $X(\langle v,t\rangle)$

Consider a fixed graph, consisting of a set of nodes or vertices V, and undirected edges between the nodes E. An edge is an ordered pair (v_1, v_2), indicating that it is possible to go directly from v_1 to v_2; the set of edges is thus a binary relation on the nodes. We indicate the presence of this relation, i.e. of a direct path, by $v_1 E v_2$; since the edges are undirected, this implies that $v_2 E v_1$. There is a path of length k between two nodes if $v_1 E^k v_2$. In addition to the graph, we have a time coordinate which proceeds in integer ticks: $t \in \mathbb{N}$

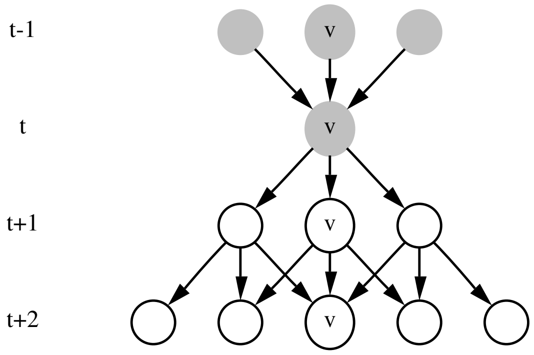

Fig. 1: Schematic of the light-cones of node $\langle v,t\rangle$. Time runs vertically downward. For visual simplicity, vertices are drawn as though the graph was a one-dimensional lattice, but no such assumption is necessary. The gray circles denote the past light-cone of $\langle v,t\rangle$, and the white ones its future light-cone. Note that $\langle v,t\rangle$ is included in its past light-cone, resulting in a slight asymmetry between the two cones.

or $\in \mathbb{Z}$. We need a way to refer to the combination of a vertex and a time; I will call this a *point*, and write it using the ordered pair $\langle v,t\rangle$.

At each point, we have a random variable $X(\langle v,t\rangle)$, taking values in the set $\mathcal{A}$, a "standard alphabet" (Gray, 1990), such as a set of discrete symbols or a finite-dimensional Euclidean vector space; this is the random field. Let c be the maximum speed of propagation of disturbances in the field. Now define the *past light-cone* of the point $\langle v,t\rangle$ as the set of all points where the field could influence the field at $\langle v,t\rangle$ (including $\langle v,t\rangle$); likewise the *future light-cone* is all the points whose fields could be influenced by the field at $\langle v,t\rangle$ (excluding $\langle v,t\rangle$). We will mostly be concerned with the configurations in these light-cones, rather than the sets of points themselves; the configurations in the past and future light-cone are respectively

$$L^-(\langle v,t\rangle) \equiv \bigcup_{\tau\geq 0}\left\{X(\langle u,t-\tau\rangle)\middle|\bigvee_{k=0}^{k=c\tau} uE^k v\right\} \text{ and} \tag{1}$$

$$L^+(\langle v,t\rangle) \equiv \bigcup_{\tau\geq 1}\left\{X(\langle u,t+\tau\rangle)\middle|\bigvee_{k=0}^{k=c\tau} vE^k u\right\}. \tag{2}$$

Figure 1 is a schematic illustration of the past and future light-cones[‡] There is a certain distribution over future light-cone configurations, conditional on the configuration in the past; this is $\mathrm{P}(L^+(\langle v,t\rangle)|L^-(\langle v,t\rangle) = l^-)$, which I shall abbreviate $\mathrm{P}(L^+|l^-)$. Note that, in general, the light-cones of distinct vertices have completely different shapes, and so neither their configurations nor the distribution over those configurations are comparable.

[‡] The term "light-cone" is used here in analogy with relativistic physics, but it's just an analogy.

2.2 Mutual Information

The mutual information between two random variables X and Y is

$$I[X;Y] \equiv \mathbf{E}_{X,Y}\left[\log_2 \frac{\mathrm{P}(X=x,Y=y)}{\mathrm{P}(X=x)\mathrm{P}(Y=y)}\right] , \tag{3}$$

where $\mathbf{E}$ is mathematical expectation, and P is the probability mass function for discrete variables, and the probability density function for continuous ones. That is, it is the expected logarithm of the ratio between the actual joint distribution of X and Y, and the product of their marginal distributions. $I[X;Y] \geq 0$, and $I[X;Y] = 0$ if and only if X and Y are independent.

The conditional mutual information $I[X;Y|Z]$, is

$$I[X;Y|Z] \equiv \mathbf{E}_{X,Y,Z}\left[\log_2 \frac{\mathrm{P}(X=x,Y=y|Z=z)}{\mathrm{P}(X=x|Z=z)\mathrm{P}(Y=y|Z=z)}\right] \tag{4}$$

Conditional mutual information is also non-negative.

We will need only two other properties of mutual information. The first is called the "data-processing inequality". For any function f,

$$I[f(X);Y] \leq I[X;Y] \text{ and} \tag{5}$$

$$I[f(X);Y|Z] \leq I[X;Y|Z] . \tag{6}$$

The other is called the "chain rule".

$$I[X,Z;Y] = I[X;Y] + I[Z;Y|X] \tag{7}$$

For more on the nature and uses of mutual information, and proofs of Equations 5–7, see any text on information theory, e.g. Gray (1990).

2.3 Conditional Independence

Two random variables, X and Y, are conditionally independent given a third, T, in symbols $X \perp\!\!\!\perp Y|T$, when

$$\mathrm{P}(X,Y|T) = \mathrm{P}(X|T)\mathrm{P}(Y|T) . \tag{8}$$

$X \perp\!\!\!\perp Y|T$ if and only if

$$\mathrm{P}(Y|X,T) = \mathrm{P}(Y|T) \text{ and} \tag{9}$$

$$I[X;Y|T] = 0 . \tag{10}$$

Conditional independence has many implicative properties; we will use some of the standard ones, proofs of which may be found in any book on graphical models (Lauritzen, 1996; Pearl, 2000; Spirtes *et al.*, 2001).

$$(A \perp\!\!\!\perp B|CD) \wedge (A \perp\!\!\!\perp D|CB) \Rightarrow (A \perp\!\!\!\perp BD|C) \tag{11}$$

$$(A \perp\!\!\!\perp BC|D) \Rightarrow (A \perp\!\!\!\perp B|CD) \tag{12}$$

$$(A \perp\!\!\!\perp B|C) \wedge (A \perp\!\!\!\perp D|CB) \Rightarrow (A \perp\!\!\!\perp BD|C) \tag{13}$$

Sadly,

$$(A \perp\!\!\!\perp B) \quad \not\Rightarrow \quad (A \perp\!\!\!\perp B|C) \tag{14}$$
$$(A \perp\!\!\!\perp B|C) \quad \not\Rightarrow \quad (A \perp\!\!\!\perp B|CD) \tag{15}$$

The following property, while not included in most lists of conditional independence properties, is of some use to us:

$$(A \perp\!\!\!\perp B|C) \quad \Rightarrow \quad (A \perp\!\!\!\perp f(B)|C) \tag{16}$$

for any nonrandom function f. It is a direct consequence of the non-negativity of conditional mutual information, and Equations 6 and 10.

2.4 Predictive Sufficient Statistics

Any function on the past light-cone defines a *local statistic*, offering some summary of the influence of all the points which could affect what happens at $\langle v,t \rangle$. A good local statistic conveys something about "what comes next", i.e., L^+. We can quantify this using information theory.

To be slightly more general, suppose we know the random variable X and wish to predict Y. Any function $f(X)$ defines a new random variable $F = f(X)$ which is a *statistic*. Because F is a function of X, the data processing inequality (Equation 5) says $I[Y;F] \leq I[Y;X]$.

Definition 1 (Sufficient Statistic) *A statistic $F = f(X)$ is* sufficient *for Y when $I[Y;F] = I[X;F]$.*

A sufficient statistic, that is to say, is one which is as informative as the original data. While I will not elaborate on this here, it is very important to remember that any prediction method which uses a non-sufficient statistic can be bettered by one which does. No matter what the loss function for prediction, the optimal predictor can always be implemented by an algorithm which depends solely on a sufficient statistic (Blackwell and Girshick, 1954).

Here are two important criteria, and consequences, of sufficiency.

Proposition 1 *F is a sufficient statistic, as in Definition 1, if and only if, $\forall x$,*

$$\mathrm{P}(Y|X=x) \quad = \quad \mathrm{P}(Y|F=f(x)) \; . \tag{17}$$

See Kullback (1968, Sections 2.4 and 2.5). (Some authors prefer to use this as the definition of sufficiency.)

Lemma 1 *F is a sufficient statistic if and only if X and Y are conditionally independent given F, i.e., $X \perp\!\!\!\perp Y|F$.*

Proof. "Only if": begin with the observation that

$$\mathrm{P}(Y|X, f(X)) = \mathrm{P}(Y|X) \; , \tag{18}$$

no matter what the function[§]. Hence $\mathrm{P}(Y|X,F) = \mathrm{P}(Y|X)$. But if $Y \perp\!\!\!\perp X|F$, then $\mathrm{P}(Y|X,F) = \mathrm{P}(Y|F)$. Hence $\mathrm{P}(Y|X) = \mathrm{P}(Y|F)$. "If": start with $\mathrm{P}(Y|F) = \mathrm{P}(Y|X)$. As before, by Equation 18, since $F = f(X)$,

[§] Let $\mathcal{A} = \sigma(X)$ be the sigma-algebra generated by X, and $\mathcal{B} = \sigma(f(X))$. Clearly $\mathcal{B} \subseteq \mathcal{A}$, so $\sigma(X, f(X)) = \mathcal{A}B = \mathcal{A} = \sigma(X)$. Thus, the sigma-algebra involved when we condition on X is the same as when we condition on X and $f(X)$ at once.

$P(Y|X) = P(Y|X,F)$. Hence $P(Y|F) = P(Y|X,F)$, so (Eq. 9), $X \perp\!\!\!\perp Y|F$. Since (2) and (3) are equivalent, as are (1) and (2), (1) and (3) are also equivalent. □

While all sufficient statistics have the same predictive power, they are not, in general, equal in other respects. In particular, some of them make finer distinctions among past light-cones than others. Since these are distinctions without a difference, we might as well, in the interests of economy, eliminate them when we find them. The concept of a *minimal* sufficient statistic captures the idea of eliminating all the distinctions we can get rid off, without loss of predictive power.

Definition 2 (Minimal Sufficiency) *F is a* minimal sufficient statistic *for predicting Y from X if and only if it is sufficient, and it is a function of every other sufficient statistic.*

3 Construction and Properties of Optimal Local Predictors

3.1 Minimal Local Sufficient Statistics

We have observed that for each past light-cone configuration l^- at a point, there is a certain conditional distribution over future light-cone configurations, $P(L^+|l^-)$. Let us say that two past configurations, or "pasts", are equivalent if they have the same conditional distribution,

$$l_1^- \sim l_2^- \quad \Leftrightarrow \quad P(L^+|l_1^-) = P(L^+|l_2^-) \tag{19}$$

Let us write the equivalence class of l^-, that is, the set of all pasts it is equivalent to, as $[l^-]$. We now define a local statistic, which is simply the equivalence class of the past light-cone configuration.

Definition 3 (Local Causal State) *The local causal state at $\langle v,t\rangle$, written $S(\langle v,t\rangle)$, is the set of all past light-cones whose conditional distribution of future light-cones is the same as that of the past light-cone at $\langle v,t\rangle$. That is,*

$$\begin{aligned} S(\langle v,t\rangle) = \varepsilon(l^-) &\equiv [l^-] &(20)\\ &= \left\{\lambda \middle| P(L^+|\lambda) = P(L^+|l^-)\right\} &(21)\end{aligned}$$

The name "causal state" was introduced for the analogous construction for time series by Crutchfield and Young (1989). I will continue to use this term here, for two reasons. First, without getting into the vexed questions surrounding the nature of causality and the properties of causal relations, it is clear that the causal states have at least the "screening" properties (Salmon, 1984) all authorities on causality require, and may well have the full set of counter-factual or "intervention" properties demanded by Pearl (2000) and Spirtes *et al.* (2001). (When and whether they meet the stricter criteria is currently an open question.) Second, and decisively, these objects need a name, and the previous terms in the literature, like "action-test pairs in a predictive state representation" (Littman *et al.*, 2002), "elements of the statistical relevance basis" (Salmon, 1984), or "states of the prediction process" (Knight, 1975), are just too awkward to use.

Theorem 1 (Sufficiency of Local Causal States) *The local causal states are sufficient.*

Proof: First, we will find the distribution of future light-cone configurations conditional on the causal state. Clearly, this is the average over the distributions conditional on the past light-cones contained in the state. (See Loéve (1955, Section 25.2) for details.) Thus,

$$P(L^+|S=s) = \int_{\lambda \in \varepsilon^{-1}(s)} P(L^+|L^- = \lambda) P(L^- = \lambda|S=s) d\lambda\,. \tag{22}$$

By construction, $\mathrm{P}(L^+|L^- = \lambda)$ is the same for all λ in the domain of integration, so, picking out an arbitrary representative element l^-, we pull that factor out of the integral,

$$\mathrm{P}(L^+|S = s) \quad = \quad \mathrm{P}(L^+|L^- = l^-) \int_{\lambda \in \varepsilon^{-1}(s)} \mathrm{P}(L^- = \lambda|S = s) d\lambda \; . \tag{23}$$

But now the integral is clearly 1, so, for all l^-,

$$\mathrm{P}(L^+|S = \varepsilon(l^-)) \quad = \quad \mathrm{P}(L^+|L^- = l^-) \; . \tag{24}$$

And so, by Proposition 1, S is sufficient. □

Corollary 1 *The past and future light-cones are independent given the local causal state.*

Proof: Follows immediately from the conclusion of the theorem and Lemma 1. □

Corollary 2 *Let K be the configuration in any part of the future light-cone. Then*

$$I[K;R] \quad \leq I[K;S] \; . \tag{25}$$

This corollary is sometimes useful because $I[L^+;R]$ can be infinite for all non-trivial statistics.

Theorem 2 (Minimality of Local Causal States) *The local causal state is a minimal sufficient statistic.*

Proof: We need to show that, for any other sufficient statistic $R = \eta(L^-)$ there is a function h such that $S = h(R)$. From Proposition 1 $\mathrm{P}(L^+|R = \eta(l^-)) = \mathrm{P}(L^+|l^-)$. It follows that $\eta(l_1^-) = \eta(l_2^-)$ only if $\mathrm{P}(L^+|l_1^-) = \mathrm{P}(L^+|l_2^-)$. This in turn implies that $l_1^- \sim l_2^-$, and so $\varepsilon(l_1^-) = \varepsilon(l_2^-)$. Thus, all histories with a common value of $\eta(l^-)$ also have the same value of $\varepsilon(l^-)$, and one can determine S from R. Hence the required function exists. □

Corollary 3 (Uniqueness of the Local Causal States) *If $R = \eta(L^-)$ is a minimal local sufficient statistic, then η and ε define the same partition of the data.*

Proof: Since S is a minimal statistic, it is a function of R, and for some function h, $S = h(R)$. But since R is also minimal, there is a function g such that $R = g(S)$. Hence $\varepsilon(l_1^-) = \varepsilon(l_2^-)$, if and only if $\eta(l_1^-) = \eta(l_2^-)$. □

In the introduction, I claimed that if we tried to predict the future of the network by building a separate model for each vertex, taken in isolation, the results would generally be sub-optimal. Theorem 2 and Corollary 3 vindicate this claim. The only way predictions based on the vertex-by-vertex procedure could be as good as ones based on the full light-cone is if *all* of the rest of the light-cone was *always* irrelevant. This would mean that there was, effectively, no coupling whatsoever between the vertices.

3.2 Statistical Complexity of the Field

Much thought has gone into the problem of defining a measure of complexity that is not simply a measure of randomness, as are Shannon entropy and the algorithmic information (Badii and Politi, 1997). Perhaps the best suggestion is the one which seems to have originated with Grassberger (1986), that the complexity of a system is the minimal amount of information about its state needed for optimal prediction. This suggests, following Crutchfield and Young (1989), that we identify the complexity of the system with the amount of information needed to specify its causal state. Crutchfield and Young called this quantity the *statistical complexity*.

In the case of random fields, it is more appropriate to look at a local, point-by-point version of this quantity. That is, the *local statistical complexity* is

$$C(\langle v,t\rangle) \equiv I[S(\langle v,t\rangle);L^-(\langle v,t\rangle)] \ . \tag{26}$$

Dropping the argument, we see that because of Equation 5 and the fact that S is a minimal statistic,

$$I[R;L^-] \geq C \tag{27}$$

for any other sufficient statistic R. In fact, one can start with Equations 25 and 27, maximal predictive power and minimal complexity, as axioms, and from them derive all the properties of the local causal states (Shalizi and Crutchfield, 2001).

A useful property of the statistical complexity is that it provides an upper bound on the predictive information.

$$I[L^+;L^-] \leq I[L^-;S] \tag{28}$$

Proof: Use the chain rule for information, Equation 7.

$$I[S,L^+;L^-] = I[S;L^-]+I[L^+;L^-|S] \tag{29}$$

$$= I[S;L^-] \ , \tag{30}$$

since Corollary 1 and Equation 10 imply $I[L^+;L^-|S]=0$. Using the chain rule the other way,

$$I[S,L^+;L^-] = I[L^+;L^-]+I[S;L^-|L^+] \tag{31}$$

$$\geq I[L^+;L^-] \ , \tag{32}$$

since $I[S;L^-|L^+] \geq 0$.

Shalizi and Crutchfield (2001) give detailed arguments for why C is the right way to measure complexity; here I will just mention two more recent applications. First, when our random field can be interpreted as a macroscopic variable which is a coarse-graining of an underlying microscopic system, as in statistical mechanics, then C is the amount of information the macro-variable contains about the micro-state (Shalizi and Moore, 2003). Second, the rate of change of C over time provides a quantitative measurement of self-organization in cellular automata (Shalizi and Shalizi, 2003).

3.3 Composition of Global Predictors from Local Predictors

We have just seen that, if we are interested in the future of an individual point, we can do optimal prediction with only a knowledge of its local causal state. One might worry, however, that in compressing the past light-cone down to the causal state, we have thrown away information that would be valuable on a larger scale, that would help us if we wanted to predict the behavior of multiple vertices. In the limit, if we wanted to predict the behavior of the entire network, and had only the local causal states available to us, how badly would we be hampered?

To address these issues, let us turn our thoughts from a single point to a connected set of vertices taken at a common time t, or a *patch*. The patch has a past and future light-cone (P^- and P^+, respectively),

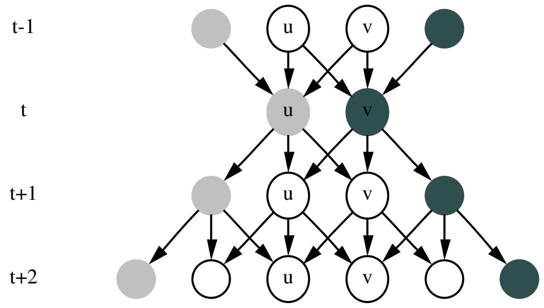

Fig. 2: The space-time regions for a two-point patch. Points which belong exclusively to the light-cones of the point on the left ($\langle u,t\rangle$) are shaded light grey; those which belong exclusively to the light-cones of the other point ($\langle v,t\rangle$) are shaded dark grey. The areas of overlap (C^- and C^+) is white. Note that, by the definition of light-cones, the configuration in U^- can have no effect on that in V^+ or vice versa.

which are just the unions of the cones of its constituent points. Just as we did for points, we can construct a causal state for the patch, which we'll call the *patch causal state*, which has all the properties of the local causal states. We now ask, what is the relationship between the local causal states of the points in the patch, and the patch causal state? If we try to predict the future of the patch using just the local states, are our predictions necessarily impaired in any way?

The answer, it turns out, is *no.*

Lemma 2 (Patch Composition) *The causal state of a patch at one time is uniquely determined by the composition of all the local causal states within the patch at that time.*

Proof: We will show that the composition of local causal states within the patch is a sufficient statistic *of the patch*, and then apply minimality.

Consider first a patch consisting of two spatially adjacent points, $\langle u,t\rangle$ and $\langle v,t\rangle$. Define the following variables:

$$
\begin{aligned}
C^- &= L^-(\langle u,t\rangle) \cap L^-(\langle v,t\rangle) \\
U^- &= L^-(\langle u,t\rangle) \setminus C^- \\
V^- &= L^-(\langle v,t\rangle) \setminus C^-
\end{aligned}
$$

Thus $L^-(\langle u,t\rangle) = U^- \cup C^-$, and likewise for $L^-(\langle v,t\rangle)$. Define U^+, C^+ and V^+ similarly. (Figure 2 gives a picture of these regions.) Now consider the configurations in these regions. We may draw a diagram of effects or influence, Figure 3, which should not be confused with the graph of the network.

Corollary 1 tells us that every path from U^- or C^- to U^+ must go through $S(\langle u,t\rangle)$. By the very definition of light-cones, there cannot be a path linking V^- to U^+. Therefore there cannot be a link from $S(\langle v,t\rangle)$ to U^+. (Such a link would in any case indicate that U^+ had a dependence on C^- which was not mediated by $S(\langle u,t\rangle)$, which is false.) All of this is true, *mutatis mutandis*, for V^+ as well.

Now notice that every path from variables in the top row — the variables which collectively constitute P^- — to the variables in the bottom row — which collectively are P^+ — must pass through either

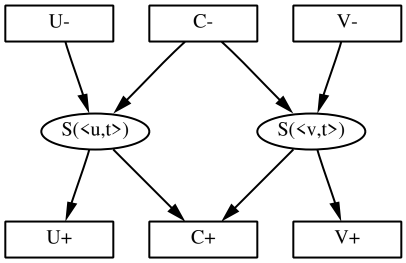

Fig. 3: Diagram of effects for the two-node patch. Arrows indicate the direction of influence; causes to effects; the absence of an arrow between two nodes indicates an absence of direct influence.

$S(\langle u,t\rangle)$ or $S(\langle v,t\rangle)$. The set $Z = \{S(\langle u,t\rangle), S(\langle v,t\rangle)\}$ thus "blocks" those paths. In the terminology of graphical models (Pearl, 2000, p. 18), Z *d-separates* P^- and P^+. But d-separation implies conditional independence (*ibid.*). Thus P^- and P^+ are independent given the composition of $S(\langle u,t\rangle)$ and $S(\langle v,t\rangle)$. But that combination is a function of P^-, so Lemma 1 applies, telling us that the composition of local states is sufficient. Then Theorem 2 tells us that there is a function from the composition of local states to the patch causal state.

Now, the reader may verify that this argument would work if one of the two "nodes" above was really itself a patch. That is, if we break a patch down into a single node and a sub-patch, and we know their causal states, the causal state of the larger patch is fixed. Hence, by mathematical induction, if we know all the local causal states of the nodes within a patch, we have fixed the patch causal state uniquely. □

Theorem 3 (Global Prediction) *The future of the entire network can be optimally predicted from a knowledge of all the local causal states at one time.*

Proof: Sufficiency, as we've said, implies optimal prediction. Let us check that the combination of all the local causal states is a sufficient statistic for the global configuration. Apply Lemma 2 to the patch of the entire lattice. The proof of the lemma goes through, because it in no way depends on the size or the shape of the past, or even on the patch being finite in extent. Since the patch causal state for this patch is identical with the global causal state, it follows that the latter is uniquely fixed by the composition of the local causal states at all points on the lattice. Since the global causal state is a sufficient statistic, the combination of all local causal states is too. □

Thus, remarkably enough, the local optimal predictors contain all the information necessary for global optimal prediction as well. There doesn't seem to be any way to make this proof work if the local predictors are not based on the full light-cones.

4 Structure of the Field of States $S(\langle v,t\rangle)$

Let us take a moment to recap. We began with a dynamic random field X on the network. From it, we have constructed another random field on the the same network, the field of the local causal states S. If we want to predict X, whether locally (Theorem 1) or globally (Theorem 3), it is sufficient to know S. This situation resembles attractor reconstruction in nonlinear dynamics (Kantz and Schreiber, 1997), state-space models in time series analysis (Durbin and Koopman, 2001), and hidden Markov models in signal processing (Elliot *et al.*, 1995). In each case, it helps, in analyzing and predicting our observations, to regard them as distorted measurements of another, unseen set of state variables, which have their own dynamics. Let us, therefore, call all these things "hidden-state models".

The hidden state space always has a more tractable structure than the observations, e.g., the former is Markovian, or even deterministic, but the latter is not. Now, usually the nice structure is simply *demanded* by us a priori, and it is an empirical question whether the demand can be met, whether any hidden-state model with that structure can adequately account for our observations. However, in the case of time series, one can *construct* hidden states, analogous to those built in Section 3, and show that these *always* have nice structural properties: they are always homogeneous Markov processes, and so forth (Knight, 1975; Shalizi and Crutchfield, 2001). This analogy gives us reason to hope that the causal states we have constructed always have nice properties, which is indeed the case.

In this section, I am going to establish some of the structural properties of the field of causal states, which involve the relations between states at different points. Section 4.1 shows how the state at one point may be used to partially determine the state at other points. Section 4.2 applies those results to the problem of designing a filter to estimate the state field S from the observation field X. Finally, Section 4.3 considers the Markov properties of the S field. Note that, if we tried to build a hidden-state model for each vertex separately, the result would lack these *spatial* properties, as well as being a sub-optimal predictor (see the end of Section 3.1).

4.1 Transition Properties

The basic idea motivating this section is that the local causal state at $\langle v,t\rangle$ should be an adequate summary of $L^-(\langle v,t\rangle)$ for *all* purposes. We have seen that this is true for both local and non-local prediction. Here we consider the problem of determining the state at another point $\langle u,s\rangle$, $s \geq t$. In general, the past light-cones of $\langle v,t\rangle$ and $\langle u,s\rangle$ will overlap. If we have $S(\langle v,t\rangle)$, and want $S(\langle u,s\rangle)$, do we need to know the actual contents of the overlapping region, or can we get away with just knowing $S(\langle v,t\rangle)$?

It seems clear (and we will see that it is true) that determining $S(\langle u,s\rangle)$ will require knowledge of the "new" observations, the ones which are in the past light-cone of $\langle u,s\rangle$ but not in $L^-(\langle v,t\rangle)$. This data is relevant to $\langle u,s\rangle$, but inaccessible at $\langle v,t\rangle$, and couldn't be summarized in $S(\langle v,t\rangle)$. This data also needs a name; let us call it the *fringe* seen when moving from $\langle v,t\rangle$ to $\langle u,s\rangle$. (See Figures 2 and 4.)

The goal of this section is to establish that $S(\langle u,s\rangle)$ is completely determined by $S(\langle v,t\rangle)$ and the fringe. The way I do this is to show that $L^+(\langle u,s\rangle)$ and $L^-(\langle u,s\rangle)$ are independent, conditional on $S(\langle v,t\rangle)$ and the fringe. This means that the latter two, taken together, are a sufficient statistic, and then I invoke Theorem 2. An alternate strategy would be to consider a transducer whose inputs are fringes, read as the transducer moves across the graph, and whose outputs are local states. Our problem then would be to show that the transducer's transition rules can always be designed to ensure that the states returned track the actual states. This way of framing the problem opens up valuable connections to the theory of spatial automata and spatial languages (Lindgren *et al.*, 1998), but I prefer a more direct approach by way of conditional independence.

Even so, the proof is frankly tedious. First, I show that the desired property holds if we consider successive states at the same vertex. Next, I show that it holds for simultaneous states at neighboring vertices. Then I extend those results to relate states at points connected by an arbitrary path in space and time. Finally, I show that the result is independent of the precise path chosen to link the points. Sadly, I have not been able to find a simpler argument.

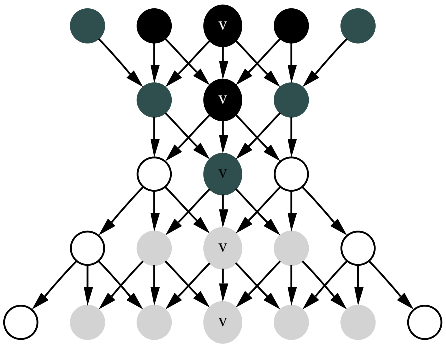

Fig. 4: Space-time regions for the time-forward transition from $\langle v,t\rangle$ to $\langle v,t+1\rangle$. Black points: $L^-(\langle v,t\rangle)$, the past light-cone of $\langle v,t\rangle$. Dark grey: N^-, the part of $L^-(\langle v,t+1\rangle)$, the past light-cone of $\langle v,t+1\rangle$, outside the past light-cone of $\langle v,t\rangle$. Light grey: $L^+(\langle v,t+1\rangle)$, the future light-cone of $\langle v,t+1\rangle$. White: M^+, the part of $L^+(\langle v,t\rangle)$ outside $L^+(\langle v,t+1\rangle)$.

4.1.1 Temporal Transitions

We want to move forward in time one step, while staying at the same vertex. Call the point we start at $\langle v,t\rangle$, and its successor $\langle v,t+1\rangle$. The whole of the new future light cone is contained inside the old future light cone, and vice versa for the past cones. So let's define the following variables:

$$\begin{aligned} N^- &= L^-(\langle v,t+1\rangle) \setminus L^-(\langle v,t\rangle) \\ M^+ &= L^+(\langle v,t\rangle) \setminus L^+(\langle v,t+1\rangle) \; ; \end{aligned}$$

N^- is the fringe. (Figure 4 pictures these regions.)

Lemma 3 *The local causal state at $\langle v,t+1\rangle$ is a function of the local causal state at $\langle v,t\rangle$ and the time-forward fringe N^-.*

Proof. Start by drawing the diagram of effects (Figure 5).

M^+ and $L^+(\langle v,t+1\rangle)$ jointly constitute $L^+(\langle v,t\rangle)$, so there must be paths from $S(\langle v,t\rangle)$ to both of them. Now, $S(\langle v,t+1\rangle)$ renders $L^-(\langle v,t+1\rangle)$ and $L^+(\langle v,t+1\rangle)$ conditionally independent. Hence it should d-separate them in the graph of effects. But $L^-(\langle v,t\rangle)$ is part of $L^-(\langle v,t+1\rangle)$ and has a direct path to $S(\langle v,t\rangle)$. This means that there cannot be a direct path from $S(\langle v,t\rangle)$ to $L^+(\langle v,t+1\rangle)$; rather, the path must go through $S(\langle v,t+1\rangle)$. (We indicate this in the graph by a dotted arrow from $S(\langle v,t\rangle)$ to $L^+(\langle v,t+1\rangle)$. Similarly, $L^-(\langle v,t\rangle)$ certainly helps determine $S(\langle v,t+1\rangle)$, but it need not do so directly. In fact, it cannot: $S(\langle v,t\rangle)$ must d-separate $L^-(\langle v,t\rangle)$ and $L^+(\langle v,t\rangle)$, i.e., must d-separate $L^-(\langle v,t\rangle)$ from $L^+(\langle v,t+1\rangle)$ and M^+. Hence the influence of $L^-(\langle v,t\rangle)$ on $S(\langle v,t+1\rangle)$ must run through $S(\langle v,t\rangle)$. (We indicate this, too, by a dotted arrow from $L^-(\langle v,t\rangle)$ to $S(\langle v,t+1\rangle)$.)

Now it is clear that the combination of $S(\langle v,t\rangle)$ and N^- d-separates $L^-(\langle v,t+1\rangle)$ from $L^+(\langle v,t+1\rangle)$, and hence makes them conditionally independent. But now the usual combination of Lemma 1 and Theorem 2 tell us that there's a function from $S(\langle v,t\rangle), N^-$ to $S(\langle v,t+1\rangle)$. $\square$

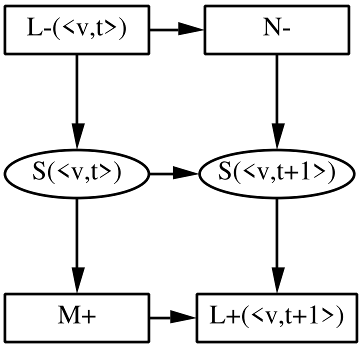

Fig. 5: Influence diagram for the variables involved in a time-forward transition.

4.1.2 Spatial Transitions

Lemma 4 *Let $\langle u,t \rangle$ and $\langle v,t \rangle$ be simultaneous, neighboring points. Then $S(\langle v,t \rangle)$ is a function of $S(\langle u,t \rangle)$ and the fringe in the direction from $\langle u,t \rangle$ to $\langle v,t \rangle$, V^-.*

Here the breakdown of the past and future light-cone regions is the same as when we saw how to compose patch causal states out of local causal states in Section 3.3, as is the influence diagram; we'll use the corresponding terminology, too. (See Figures 2 and 3, respectively.) What we hope to show here is that conditioning on the combination of $S(\langle u,t \rangle)$ and V^- makes $L^+(\langle v,t \rangle)$ independent of V^- and C^-. Unfortunately, as the reader may verify by inspecting the diagram, our conditional variables no longer d-separate the other variables (since they have an unblocked connection through $S(\langle v,t \rangle)$). All is not lost, however: d-separation implies conditional independence, but not conversely.

Abbreviate the pair of variables $\{S(\langle u,t \rangle), V^-\}$ by Z. Now, $S(\langle v,t \rangle)$ is a (deterministic) function of C^- and V^-. Hence it is also a function of Z and C^-. Thus $\mathrm{P}(V^+|S(\langle v,t \rangle),Z,C^-) = \mathrm{P}(V^+|Z,C^-)$. But this tells us that

$$V^+ \perp\!\!\!\perp S(\langle v,t \rangle)|Z,C^- \tag{33}$$

From d-separation, we also have

$$V^+ \perp\!\!\!\perp C^-|Z,S(\langle v,t \rangle) \tag{34}$$

Applying Eq. 11,

$$V^+ \perp\!\!\!\perp S(\langle v,t \rangle),C^-|Z \tag{35}$$

Applying Eq. 12,

$$V^+ \perp\!\!\!\perp C^-|Z \tag{36}$$

Since $Z = Z,V^-$,

$$V^+ \perp\!\!\!\perp C^-|Z,V^- \tag{37}$$

The following conditional independence is odd-looking, but trivially true:

$$V^+ \perp\!\!\!\perp V^-|Z \tag{38}$$

And it, along with Eq. 13, gives us

$$V^+ \perp\!\!\!\perp C^-, V^- | Z \tag{39}$$

A similar train of reasoning holds for C^+. Thus, the entire future light cone of $\langle v,t\rangle$ is independent of that point's past light cone, given $S(\langle u,t\rangle)$ and V^-. This tells us that $\{S(\langle u,t\rangle), V^-\}$ is sufficient for $L^+(\langle v,t\rangle)$, hence $S(\langle v,t\rangle)$ is a function of it. □

4.1.3 Arbitrary Transitions

Lemma 5 *Let $\langle u,t\rangle$ and $\langle v,s\rangle$ be two points, $s \geq t$. Let Γ be a spatio-temporal path connecting the two points, arbitrary except that it can never go backwards in time. Let F_Γ be the succession of fringes encountered along Γ. Then $S(\langle v,s\rangle)$ is a function of $S(\langle u,t\rangle)$, Γ and F_Γ,*

$$S(\langle v,s\rangle) = g(S(\langle u,t\rangle), \Gamma, F_\Gamma) \tag{40}$$

for some function g.

Proof. Apply Lemma 3 or 4 at each step of Γ. □

Theorem 4 *Let $\langle u,t\rangle$ and $\langle v,s\rangle$ be two points as in the previous lemma, and let Γ_1, Γ_2 be two paths connecting them, and F_{Γ_1} and F_{Γ_2} their fringes, all as in the previous lemma. Then the state at $\langle v,s\rangle$ is independent of which path was taken to reach it,*

$$g(S(\langle u,t\rangle), \Gamma_1, F_{\Gamma_1}) = g(S(\langle u,t\rangle), \Gamma_2, F_{\Gamma_2}) . \tag{41}$$

Proof. Suppose otherwise. Then either the state we get by going along Γ_1 is wrong, i.e., isn't $S(\langle v,s\rangle)$, or the state we get by going along Γ_2 is wrong, or both are.

$$S(\langle v,s\rangle) \neq g(S(\langle u,t\rangle), \Gamma_1, F_{\Gamma_1}) \quad \vee \quad S(\langle v,s\rangle) \neq g(S(\langle u,t\rangle), \Gamma_2, F_{\Gamma_2}) \tag{42}$$

$$L^+(\langle v,s\rangle) \not\perp\!\!\!\perp L^-(\langle v,s\rangle) | S(\langle u,t\rangle), \Gamma_1, F_{\Gamma_1} \quad \vee \quad L^+(\langle v,s\rangle) \not\perp\!\!\!\perp L^-(\langle v,s\rangle) | S(\langle u,t\rangle), \Gamma_2, F_{\Gamma_2} \tag{43}$$

$$\neg(L^+(\langle v,s\rangle) \perp\!\!\!\perp L^-(\langle v,s\rangle) | S(\langle u,t\rangle), \Gamma_1, F_{\Gamma_1} \quad \wedge \quad L^+(\langle v,s\rangle) \perp\!\!\!\perp L^-(\langle v,s\rangle) | S(\langle u,t\rangle), \Gamma_2, F_{\Gamma_2}) \tag{44}$$

But, by Lemma 5, $L^+(\langle v,s\rangle) \perp\!\!\!\perp L^-(\langle v,s\rangle) | S(\langle u,t\rangle), \Gamma_1, F_{\Gamma_1}$ and $L^+(\langle v,t\rangle) \perp\!\!\!\perp L^-(\langle v,s\rangle) | S(\langle u,t\rangle), \Gamma_2, F_{\Gamma_2}$. Hence transitions must be path-independent. □

4.2 Recursive Filtering

Because the causal states are logical constructions out of observational data, in principle the state at any point can be determined exactly from looking at the point's past light-cone. It may be, however, that for some fields one needs to look infinitely far back into the past to fix the state, or at least further back than the available data reaches. In such cases, one will generally have not exact knowledge of the state, but either a set-valued type estimate (i.e., $S \in \mathcal{S}$, for some set of states $\mathcal{S}$), or a distribution over states, supposing you have a meaningful prior distribution.

A naive state-estimation scheme, under these circumstances, would produce an estimate for each point separately. The transition properties we have just proved, however, put *deterministic* constraints on which states are allowed to co-exist at different points. In particular, we can narrow our estimates of the state at each point by requiring that it be consistent with our estimates at neighboring points. Applied iteratively, this will lead to the tightest estimates which we can extract from our data. If we can fix the state at even one point exactly, then this will propagate to all points at that time or later.

More generally, rather than first doing a naive estimate and then tightening it, we can incorporate the deterministic constraints directly into a recursive filter. Under quite general circumstances, as time goes on, the probability that such a filter will *not* have fixed the state of at least one point goes to zero, and once it fixes on a state, it stays fixed. Such filters can be implemented, at least for discrete fields, by means of finite-state transducers; see Shalizi (2001, Chapter 10) for examples (further specialized to discrete fields on regular lattices).

4.3 Markov Properties

Recursive estimation is a kind of Markov property (Nevel'son and Has'minskiĭ, 1972/1976), and the fact that it can work exactly here is very suggestive. So, too, is the fact that the past and the future light-cones are independent, conditional on the causal state. It would be very nice if the causal states formed a Markov random field; for one thing, we could then exploit the well-known machinery for such fields. For instance, we would know, from the equivalence between Gibbs distributions and Markov fields (Guyon, 1995), that there was an effective potential for the interactions between the states across the network.

Definition 4 (Parents of a Local Causal State) *The parents of the local causal state at $\langle v,t\rangle$ are the causal states at all points which are one time-step before $\langle v,t\rangle$ and inside its past light-cone:*

$$A(\langle v,t\rangle) \quad \equiv \quad \{S(\langle u,t-1\rangle)\,|\,(u=v)\vee(uEv)\} \tag{45}$$

Lemma 6 *The local causal state at a point, $S(\langle v,t\rangle)$, is independent of the configuration in its past light cone, given its parents.*

$$S(\langle v,t\rangle) \perp\!\!\!\perp L^{-}(\langle v,t\rangle)|A(\langle v,t\rangle) \tag{46}$$

Proof. $\langle v,t\rangle$ is in the intersection of the future light cones of all the node in the patch at $t-1$. Hence, by the arguments given in the proof of the composition theorem, it is affected by the local states of all those nodes, *and by no others*. In particular, previous values of the configuration in $L^{-}(\langle v,t\rangle)$ have no direct effect; any influence must go through those nodes. Hence, by d-separation, $S(\langle v,t\rangle)$ is independent of $L^{-}(\langle v,t\rangle)$. □

Theorem 5 (Temporal Markov Property) *The local causal state at a point, $S(\langle v,t\rangle)$, is independent of the local causal states of points in its past light cone, given its parents.*

Proof. By the previous lemma, $S(\langle v,t\rangle)$ is conditionally independent of $L^{-}(\langle v,t\rangle)$ given its parents. But the local causal states in its past light cone are a function of $L^{-}(\langle v,t\rangle)$. Hence by Equation 16, $S(\langle v,t\rangle)$ is also independent of those local states. □

Comforting though that is, we would really like a stronger Markov property, namely the following.

Conjecture 1 (Markov Field Property) *The local causal states form a Markov field in space-time.*

Argument for why this is plausible. We've seen that, temporally speaking, a Markov property holds: given a patch of nodes at one time, they are independent of their past light cone, given their causal parents. What we need to add for the Markov field property is that, if we condition on *present* neighbors of the patch, as well as the parents of the patch, then we get independence of the states of all points at time t or earlier. It's plausible that the simultaneous neighbors are informative, since they are also causal descendants of causal parents of the patch. But if we consider any more remote node at time t, its last common causal ancestor with any node in the patch must have been before the immediate parents of the patch, and the effects of any such local causal state are screened off by the parents.

Unfortunately, this is not really rigorous. In the somewhat specialized case of discrete fields on regular lattices, a number of systems have been checked, and in all cases the states have turned out to be Markov random fields.

5 State Reconstruction for Discrete Fields

The local causal states are a particular kind of hidden-state model; they combine optimal prediction properties (Section 3) with the nice structural properties of hidden-state models (Section 4). It is only human to be tempted to use them on actual systems. In the special case of discrete-valued fields, one can reliably identify the causal states from empirical data, with minimal assumptions. This section describes an algorithm for doing so, building on earlier procedures for cellular automata (Shalizi and Shalizi, 2003). As always, I take structure of the network to be known and constant.

Assume that for each point we have estimates of the conditional probability distribution $\mathrm{P}(L^+,L^-)$ over light-cones for each point. (These could come, for instance, from the empirical distribution in an ensemble of networks with the same structure, or from the same network observed over time, if certain ergodicity assumptions can be made.) The resulting conditional distributions, $\mathrm{P}(L^+|l^-)$, can be treated as multinomials. We then cluster past configurations, point by point, based on the similarity of their conditional distributions. We cannot expect that the estimated distributions will match exactly, so we employ a standard test, e.g. χ^2, to see whether the discrepancy between estimated distributions is significant. These clusters are then the estimated local causal states. We consider each cluster to have a conditional distribution of its own, equal to the weighted mean of the distributions of the pasts it contains.

As a practical matter, we need to impose a limit on how far back into the past, or forward into the future, the light-cones are allowed to extend — their depth. Also, clustering cannot be done on the basis of a true equivalence relation. Instead, we list the past configurations $\{l_i^-\}$ in some arbitrary order. We then create a cluster which contains the first past, l_1^-. For each later past, say l_i^-, we go down the list of existing clusters and check whether $\mathrm{P}_t(L^+|l_i^-)$ differs significantly from each cluster's distribution. If there is no difference, we add l_i^- to the first matching cluster and update the latter's distribution. If l_i^- does not match any existing cluster, we make a new cluster for l_i^-. (See Figure 6 for pseudo-code.) As we give this procedure more and more data, it converges in probability on the correct set of causal states, independent of the order in which we list past light-cones (see below). For finite data, the order of presentation matters, but we finesse this by randomizing the order.

Suppose that the past and the future light-cones are sufficiently deep that they suffice to distinguish the causal states, i.e., that if we had the exact distribution over light-cones, the true causal states would coincide exactly with those obtained from the distribution over limited-depth light-cones. Then, conditioning on the limited-depth past cones makes futures independent of the more remote past. Indeed, every time we examine the future of a given past, we take an *independent* sample from an unchanging distribution over futures. Thus, the strong law of large numbers tells us that the empirical probability of any future configuration will converge on the true probability almost surely. Since, with finite future depth, there are only finitely many possible future configurations, the conditional distribution as a whole also converges almost surely. And, if there are only finitely many vertices, we get over-all convergence of all the necessary conditional distributions.

```
U ← list of all pasts in random order
Move the first past in U to a new state
for each past in U
    noMatch ← TRUE
    state ← first state on the list of states
    while (noMatch and more states to check)
        noMatch ← (Significant difference between past and state?)
        if (noMatch)
            state ← next state on the list
        else
            Move past from U to state
            noMatch ← FALSE
    if (noMatch)
        make a new state and move past into it from U
```

Fig. 6: Algorithm for grouping past light-cones into estimated states

For this analysis to hold, we need to know three things.

1. The structure of the graph.
2. The constant c which is the maximum speed at which information can propagate.
3. The minimum necessary depth of past and future cones.

Note that if we over-estimate (2) and (3), we will not really harm ourselves, but under-estimates *are* harmful, since in general they will miss useful bits of predictive information.

Note that this analysis, like the algorithm, is somewhat crude, in that it doesn't make use of any of the properties of the causal states we established earlier. In the analogous case of time series, using those properties can greatly speed up the reconstruction process, and allows us to not merely prove convergence, but to estimate the rate of convergence (Shalizi *et al.*, 2002). Probably something like that could be done here, with the additional complication that each vertex must get its own set of states.

6 Conclusion

The purpose of this paper has been to define, mathematically, optimal nonlinear predictors for a class of complex systems, namely dynamic random fields on fixed, undirected graphs. Starting with the basic idea of maximizing the predictive information, we constructed the *local causal states* of the field, which are minimal sufficient statistics, and so the simplest possible basis for optimal prediction. We then examined how to combine these states for non-local prediction, and the structure of the causal-state field. The last section described an algorithm for reconstructing the states of discrete fields.

Acknowledgements

This paper grew out of earlier work on local prediction for fields on regular lattices with Rob Haslinger, Kristina Shalizi and Jacob Usinowicz; it is a pleasure to acknowledge their support. I have also benefited from discussions with Jim Crutchfield, Dave Feldman, Christian Lindgren, Cris Moore, Mats Nordahl, Mitch Porter and Karl Young. I am grateful to Scott Page for the generous construction he puts on the phrase "complexity and diversity", and to the organizers of DMCS 2003 for the opportunity to participate. Finally, I wish to thank Kara Kedi and Kristina Shalizi for being perfect.

References

Badii, Remo and Antonio Politi (1997). *Complexity: Hierarchical Structures and Scaling in Physics*, vol. 6 of *Cambridge Nonlinear Science Series*. Cambridge: Cambridge University Press.

Blackwell, David and M. A. Girshick (1954). *Theory of Games and Statistical Decisions*. New York: Wiley. Reprinted New York: Dover Books, 1979.

Bower, James M. and Hamid Bolouri (eds.) (2001). *Computational Modeling of Genetic and Biochemical Networks*, Computational Molecular Biology, Cambridge, Massachusetts. MIT Press.

Crutchfield, James P. and Karl Young (1989). "Inferring Statistical Complexity." *Physical Review Letters*, **63**: 105–108.

Dayan, Peter and Laurence F. Abbott (2001). *Theoretical Neuroscience*. Cambridge, Massachusetts: MIT Press. URL `play.ccs.brandeis.edu/abbott/book/`.

Durbin, James and Siem Jam Koopman (2001). *Time Series Analysis by State Space Methods*, vol. 24 of *Oxford Statistical Science Series*. Oxford: Oxford University Press.

Elliot, Robert J., Lakhdar Aggoun and John B. Moore (1995). *Hidden Markov Models: Estimation and Control*, vol. 29 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. New York: Springer-Verlag.

Grassberger, Peter (1986). "Toward a Quantitative Theory of Self-Generated Complexity." *International Journal of Theoretical Physics*, **25**: 907–938.

Gray, Robert M. (1990). *Entropy and Information Theory*. New York: Springer-Verlag. URL `www-ee.stanford.edu/~gray/it.html`.

Guyon, Xavier (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer Series in Statistics: Probability and Its Applications. Berlin: Springer-Verlag.

Ilachinski, Andrew (2001). *Cellular Automata: A Discrete Universe*. Singapore: World Scientific.

Kantz, Holger and Thomas Schreiber (1997). *Nonlinear Time Series Analysis*, vol. 7 of *Cambridge Nonlinear Science Series*. Cambridge, England: Cambridge University Press.

Knight, Frank B. (1975). "A Predictive View of Continuous Time Processes." *The Annals of Probability*, **3**: 573–596.

Kullback, Solomon (1968). *Information Theory and Statistics*. New York: Dover Books, 2nd edn. First edition New York: Wiley, 1959.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford University Press.

Lindgren, Kristian, Cristopher Moore and Mats Nordahl (1998). "Complexity of Two-Dimensional Patterns." *Journal of Statistical Physics*, **91**: 909–951. URL `arxiv.org/abs/cond-mat/9804071`.

Littman, Michael L., Richard S. Sutton and Satinder Singh (2002). "Predictive Representations of State." In *Advances in Neural Information Processing Systems 14* (Thomas G. Dietterich and Suzanna Becker and Zoubin Ghahramani, eds.), pp. 1555–1561. Cambridge, Massachusetts: MIT Press. URL `www.eecs.umich.edu/~baveja/Papers/psr.pdf`.

Loéve, Michel (1955). *Probability Theory*. New York: D. Van Nostrand Company, 1st edn.

Mutambara, Arthur G. O. (1998). *Decentralized Estimation and Control for Multisensor Systems*. Boca Raton, Florida: CRC Press.

Nevel'son, M. B. and R. Z. Has'minskiĭ (1972/1976). *Stochastic Approximation and Recursive Estimation*, vol. 47 of *Translations of Mathematical Monographs*. Providence, Rhode Island: American Mathematical Society. Translated by the Israel Program for Scientific Translations and B. Silver from *Stokhasticheskaia approksimatsia i rekurrentnoe otsenivanie*, Moscow: Nauka.

Newman, Mark E. J. (2003). "The Structure and Function of Complex Networks." *SIAM Review*, **45**: 167–256. URL `arxiv.org/abs/cond-mat/0303516`.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Pikovsky, Arkady, Michael Rosenblum and Jürgen Kurths (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*, vol. 12 of *Cambridge Nonlinear Science Series*. Cambridge, England: Cambridge University Press.

Salmon, Wesley C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Shalizi, Cosma Rohilla (2001). *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. Ph.D. thesis, University of Wisconsin-Madison. URL `bactra.org/thesis/`.

Shalizi, Cosma Rohilla and James P. Crutchfield (2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics*, **104**: 817–879. URL `arxiv.org/abs/cond-mat/9907176`.

Shalizi, Cosma Rohilla and Cristopher Moore (2003). "What Is a Macrostate? From Subjective Measurements to Objective Dynamics." *Studies in the History and Philosophy of Modern Physics*, **submitted**. URL `arxiv.org/abs/cond-mat/0303625`.

Shalizi, Cosma Rohilla and Kristina Lisa Shalizi (2003). "Quantifying Self-Organization in Cyclic Cellular Automata." In *Noise in Complex Systems and Stochastic Dynamics* (Lutz Schimansky-Geier, Derek Abbott, Alexander Neiman and Christian Van den Broeck, eds.), vol. 5114 of *Proceedings of SPIE*. Bellingham, Washington: SPIE. URL `bactra.org/research/`.

Shalizi, Cosma Rohilla, Kristina Lisa Shalizi and James P. Crutchfield (2002). "An Algorithm for Pattern Discovery in Time Series." Electronic pre-print. URL `arxiv.org/abs/cs.LG/0210025`.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press, 2nd edn.

Šijlak, Dragoslav S. (1990). *Decentralized Control of Complex Systems*, vol. 184 of *Mathematics in Science and Engineering*. Boston: Academic Press.

Varshney, Pramod K. (1997). *Distributed Detection and Data Fusion*. Signal Processing and Data Fusion. Berlin: Springer-Verlag.

Young, H. Peyton (1998). *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton: Princeton University Press.