Translated from *Problemy Peredachi Informatsii*, Vol. 32, No. 1, pp. 131–136, January–March, 1996.

J. L. Massey621.391.1 : 519.28

Abstract

A causal interpretation of random variables corresponds to the successive generation of these random variables by a sequence of random experiments, each of which uses the results of previous experiments only. Causality graphs are introduced to describe such a causal representation. It is shown that although the order of the random variables in a causal interpretation is completely arbitrary, causality graphs are nonetheless useful in deducing independencies among random variables.

1. Introduction

This paper is written in honor of the great Russian information theorist, Mark S. Pinsker, on the occasion of his seventieth birthday and in recognition of his own long interest in the information–theoretic aspects of probabilistic dependence (cf. [1, Section I.3]).

In the next section, we introduce the notion of a causal interpretation of random variables as corresponding to the successive generation of these random variables by a sequence of random experiments, each of which uses the results of previous experiments only. Causality graphs are introduced to describe such causal interpretations. We stress that the order of the random variables in a causal interpretation is completely arbitrary. Nonetheless, we show in Section

3 that causality diagrams are useful tools for deducing independencies among random variables. We close in Section 4 showing that not all independencies among random variables can be represented in a single causality diagram and by relating our results to prior work by Pearl [2].

2. Definitions and preliminaries

Here and hereafter, let $X_1, X_2, \ldots, X_N$ be discrete random variables with finite joint entropy, which implies that any joint entropy [e.g., $H(X_2X_4X_6)$] or any conditional entropy [e.g., $H(X_2X_4X_6 \mid X_1X_3)$] involving only these random variables is also finite. By a *causal interpretation* of $X_1, X_2, \ldots, X_N$, we mean an ordered list $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ of these N random variables such that $X_1, X_2, \ldots, X_N$ can be produced by performing a sequence of N random experiments, the n-th of which produces X_{i_n} as its output and may make use of the outputs of the previous $n-1$ random experiments but not of the outputs of the following random experiments. Strictly speaking, we should say only that this sequence of random experiments produces random variables with the same joint probability distribution as the given random variables $X_1, X_2, \ldots, X_N$. But because $X_1, X_2, \ldots, X_N$ are completely described by their joint probability distribution in the sense that, even with repeated trials, an observer who sees only the values of $X_1, X_2, \ldots, X_N$ cannot distinguish whether these random variables were produced by the actual random experiment on which $X_1, X_2, \ldots, X_N$ are defined or by the sequence of random experiments corresponding to the causal

interpretation $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$, we will continue to speak of these latter random experiments as producing the random variables $X_1, X_2, \ldots, X_N$ themselves.

It is important to note that *every one of the $N!$ ordered lists of the random variables $X_1, X_2, \ldots, X_N$ is a valid causal interpretation.* The random experiment producing X_{i_n} when given the outputs $(x_{i_1}, x_{i_2}, \ldots, x_{i_{n-1}})$ of the preceeding random experiments needs simply be designed so as to produce a random variable whose probability distribution is the conditional probability distribution for X_{i_n} given the joint event $X_{i_1} = x_{i_1}$, $X_{i_2} = x_{i_2}$, ...and $X_{i_{n-1}} = x_{i_{n-1}}$. *The order of the random variables in a causal interpretation is completely arbitrary*, but we will see that causal interpretations are nonetheless useful in identifying independencies among random variables.

With the causal interpretation $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ of $X_1, X_2, \ldots, X_N$, it is natural to consider the expansion of the joint entropy of these random variables in the manner

$$H(X_{i_1} X_{i_2} \ldots X_{i_N}) = H(X_{i_1}) + H(X_{i_2} \mid X_{i_1}) + \ldots + H(X_{i_N} \mid X_{i_1} \ldots X_{i_{N-1}}),$$

which we will call the *causal–order expansion* of $H(X_{i_1} X_{i_2} \ldots X_{i_N})$. If the term $H(X_{i_n} \mid X_{i_1} \ldots X_{i_{n-1}})$ is unchanged when certain of the conditioning random variables are removed, i.e., if X_{i_n} is independent of these removed random variables when conditioned on the remaining conditioning random variables, then the random experiment producing X_{i_n} can be performed using only the outputs of the random experiments producing these remaining conditioning random variables. Removing the other conditioning random variables gives what we

will call the *reduced–conditioning* expression for $H(X_{i_n} \mid X_{i_1} \ldots X_{i_{n-1}})$. We can show such known independencies in the *causality graph* , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$, which we define as the directed graph with N vertices, which are labelled by $X_{i_1}, X_{i_2}, \ldots, X_{i_N}$, such that there is an edge from vertex X_{i_k} to vertex X_{i_n} if and only if X_{i_k} is one of the conditioning random variables in the reduced–conditioning expression for $H(X_{i_n} \mid X_{i_1} \ldots X_{i_{n-1}})$. There is sometimes a choice as to which random variables can removed from the conditioning. To obtain uniqueness of the reduced–conditioning expression for $H(X_{i_n} \mid X_{i_1} \ldots X_{i_{n-1}})$, we assume that X_1 is first removed from the conditioning if possible, then X_2 is removed from the conditioning if possible, etc., but the reader will see easily that, although the uniqueness of the reduced conditioning is required in this section, none of the results of Section 3 depend on the rule used for reducing conditioning or, indeed, on whether as many random variables as possible are removed from the conditioning.

Example. Consider the causal interpretation $(X_1, X_2, X_3, X_4, X_5, X_6)$ of the random variables $X_1, X_2,$

X_3, X_4, X_5, X_6 and suppose that $H(X_2 \mid X_1) = H(X_2)$, $H(X_3 \mid X_1X_2) = H(X_3)$, $H(X_4 \mid X_1X_2X_3) = H(X_4 \mid X_1X_2)$, $H(X_5 \mid X_1X_2X_3X_4) = H(X_5 \mid X_2X_3)$ and $H(X_6 \mid X_1X_2X_3X_4X_5) = H(X_6 \mid X_4X_5)$ are the reduced-conditioning expressions for the entropies in the causal–order expansion of $H(X_1X_2X_3X_4X_5X_6)$. The causality graph , $(X_1, X_2, X_3, X_4, X_5, X_6)$ is shown in Fig. 1.

We will say that the random variable X_{i_k} is *causally prior* to the random

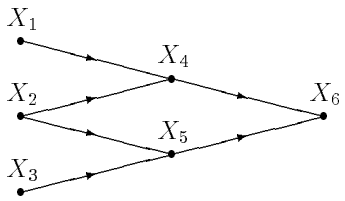

Figure 1: The causality graph , $(X_1, X_2, X_3, X_4, X_5, X_6)$ for the example.

variable X_{i_n} with respect to the causal interpretation $(X_{i_1}, X_{i_2}, \ldots, , X_{i_N})$ if, in the causality graph , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$, there is a directed path from vertex X_{i_k} to vertex X_{i_n}. If there is an edge from vertex X_{i_k} to vertex X_{i_n}, then we will say further that X_{i_k} is *causally directly prior* to X_{i_n}. [For instance, in Fig. 1, X_2 is causally prior to X_4, to X_5 and to X_6 and is causally directly prior to X_4 and to X_5.] Note that X_{i_k} is causally directly prior to X_{i_n} if and only if X_{i_k} is one of the conditioning random variables in the reduced–conditioning expression for $H(X_{i_n} \mid X_{i_1} \ldots X_{i_{n-1}})$.

We will say that the causal interpretation $(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$ is *equivalent* to the causal interpretation $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ if, whenever X_k is causally directly prior to X_n with respect to the causal interpretation $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$, then X_k appears before X_n in the list $(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$. Note that if the causal interpretation $(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$ is equivalent to $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$, then, for every n, *every* random variable X_k causally prior to X_n with respect to $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ must appear before X_n in the list $(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$, which is a fact that we will soon use. The following proposition shows that this notion of equivalence of causal interpretations is indeed an equivalence relation.

Proposition 1. *Two causal interpretations of $X_1, X_2, \ldots, X_N$ have the same causality graph if and only if they are equivalent.*

Example (continued). The causal interpretation $(X_2, X_3, X_5, X_1, X_4, X_6)$ is equivalent to the causal interpretation $(X_1, X_2, X_3, X_4, X_5, X_6)$, whose causality graph was given in Fig. 1, because in the former list both X_1 and X_2 appear before X_4, both X_2 and X_3 appear before X_5, and both X_4 and X_5 appear before X_6. The reader can verify that there are in fact twelve causal interpretations of $X_1, X_2, X_3, X_4, X_5, X_6$ in this equivalence class.

Proof. Two causal interpretations of $X_1, X_2, \ldots, X_N$ having the same causality graph are trivially equivalent. Suppose conversely that the causal interpretation $(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$ is *equivalent* to another causal interpretation that, with no loss of essential generality, we take to be $(X_1, X_2, \ldots, X_N)$. Suppose further that $X_{j_n} = X_k$. Then, before reducing of the conditioning, the n-th term in the causal–order expansion of $H(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$ has the form $H(X_{j_n} \mid X_{j_1} \ldots X_{j_{n-1}}) = H(X_k \mid X_A X_B X_C)$ where X_C is the random vector whose components are the random variables causally directly prior to X_k [in the causal interpretation $(X_1, X_2, \ldots, X_N)$], where X_B is a random vector whose components are the remaining random variables among $X_{j_1} \ldots X_{j_{n-1}}$ that are equal to some X_i with $i < k$, and where X_A is a random vector having components X_i with $i > k$ such that every random variable causally prior to X_k and not in X_A itself is a component of either X_B or X_C. Because all the compo-

nents of X_B and X_C are among $X_1, X_2 \ldots, X_{k-1}$ and $H(X_k \mid X_1 \ldots X_{k-1}) = H(X_k \mid X_C)$, it follows from the causal interpretation $(X_1, X_2, \ldots, X_N)$ that $H(X_k \mid X_B X_C) = H(X_k \mid X_C)$. Similarly, because all the random variables causally prior to each component of X_A are components of X_A or X_B or X_C and $H(X_A \mid X_1 \ldots X_k) =$

$H(X_A \mid X_B X_C)$, it follows that $H(X_A \mid X_B X_C X_k) = H(X_A \mid X_B X_C)$. Hence,

$$\begin{aligned} H(X_k X_A X_B X_C) &= H(X_B X_C) + H(X_k \mid X_C) + H(X_A \mid X_B X_C) \\ &= H(X_A X_B X_C) + H(X_k \mid X_C). \end{aligned}$$

Thus $H(X_k \mid X_A X_B X_C) = H(X_k X_A X_B X_C) - H(X_A X_B X_C) = H(X_k \mid X_C)$. We have now shown that $H(X_{j_n} \mid X_{j_1} \ldots X_{j_{n-1}}) = H(X_k \mid X_C)$. But $H(X_k \mid X_C)$ is precisely the reduced–conditioning expression for $H(X_k \mid X_1 \ldots X_{k-1})$ with respect to the causal interpretation $(X_1, X_2, \ldots, X_N)$. Thus, the causality graph for the causal interpretation $(X_{j_1}, X_{j_2}, \ldots, X_{j_N})$ is the same as that for the causal interpretation $(X_1, X_2, \ldots, X_N)$. $\triangle$

3. Deducing independence

We now show how a causality graph can be used to deduce independence and conditional independence of the random variables labelling its vertices. To this end, let X_A be any random vector with components in $(X_1, X_2, \ldots, X_N)$. Then by the *subgraph* of , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ *causally relevant* to X_A, which we will denote simply by $\Sigma_r(X_A)$, we mean the subgraph of , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$

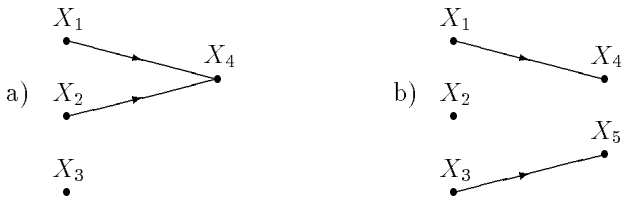

Figure 2: The causally relevant subgraphs (a) $\Sigma_r(X_3, X_4)$ and (b) $\Sigma_r(X_4, X_5 \mid X_2)$ of , $(X_1, X_2, X_3, X_4, X_5, X_6)$ for the example.

consisting of only those vertices that are either components of X_A or causally prior to components of X_A, together with the edges connecting these vertices. The causally relevant subgraph $\Sigma_r(X_3, X_4)$ of , $(X_1, X_2, X_3, X_4, X_5, X_6)$ of Fig. 1 is shown in Fig. 2(a).

Proposition 2. *If all components of the random vector X_A lie in a part of $\Sigma_r(X_A, X_B)$, the subgraph of , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ causality relevant to (X_A, X_B), unconnected (when edges are considered without direction) with the part in which the components of X_B lie, then X_A and X_B are independent.*

Example (continued). From the causally relevant subgraph $\Sigma_r(X_3, X_4)$ of Fig. 2(a), we can conclude that (X_1, X_2, X_4) is independent of X_3; in particular, X_4 is independent of X_3. This deduction is "obvious" from "causal reasoning" applied to the causality graph , $(X_1, X_2, X_3, X_4, X_5, X_6)$ of Fig. 1–but the main purpose of this paper is to justify rigorously such intuitive reasoning.

Proof. Suppose that the components of X_A do indeed lie in a part of $\Sigma_r(X_A, X_B)$ unconnected with the part in which X_B lies. Let X_{A+} and X_{B+} be random vectors whose components are all the random variables in the former

part and latter part, respectively, of $\Sigma_r(X_A, X_B)$, and each with components ordered so that no component precedes a component to which it is causally prior. Then there is an equivalent causal interpretation of $X_1, X_2, \ldots, X_N$ that begins with (X_{A+}, X_{B+}). But no component of X_{A+} is causally prior to a component of X_{B+} so that $H(X_{B+} \mid X_{A+}) = H(X_{B+})$. Thus X_{A+} and X_{B+} are independent and, in particular, the subvectors X_A and X_B are independent. $\triangle$

Let X_A and X_C be any random vectors with components in $(X_1, X_2, \ldots, X_N)$. Then by the *subgraph* of , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ *causally relevant to* X_A *when conditioned on* X_C, which we denote by $\Sigma_r(X_A \mid X_C)$, we mean the graph obtained from $\Sigma_r(X_A, X_C)$ by deleting all edges leaving vertices that are components of X_C. The subgraph $\Sigma_r(X_4, X_5 \mid X_2)$ of , $(X_1, X_2, X_3, X_4, X_5, X_6)$ in Fig. 1 causally relevant to (X_4, X_5) when conditioned on X_2 is shown in Fig. 2(b).

Proposition 3. *If neither* X_A *nor* X_B *has components in common with* X_C *and if all components of the random vector* X_A *lie in a part of* $\Sigma_r(X_A, X_B \mid X_C)$, *the subgraph of* , $(X_{i_1}, X_{i_2}, \ldots, X_{i_N})$ *causally relevant to* (X_A, X_B) *when conditioned on* X_C, *unconnected (when edges are considered without direction) with the part in which the components of* X_B *lie, then* X_A *and* X_B *are independent when conditioned on* X_C .

Example (concluded). From the subgraph $\Sigma_r(X_4, X_5 \mid X_2)$ in Fig. 2(b) of the causality graph

, $(X_1, X_2, X_3, X_4, X_5, X_6)$ of Fig. 1, we can conclude that (X_1, X_4) is indepen-

dent of (X_3, X_5) when conditioned on X_2; in particular, X_4 is independent of X_5 when conditioned on X_2. But we cannot conclude that X_4 is independent of X_5 when conditioned on X_6 because $\Sigma_r(X_4, X_5 \mid X_6) =$, $(X_1, X_2, X_3, X_4, X_5, X_6)$, which has no unconnected parts.

PROOF. Suppose that the components of X_A lie in a part of $\Sigma_r(X_A, X_B \mid X_C)$ unconnected with the part in which X_B lies. This implies, in the causally relevant subgraph $\Sigma_r(X_A, X_B, X_C)$, that both X_A and X_B cannot have components causally prior to X_C. We suppose then that there may be components of X_A, but not of X_B, causally prior to X_C. It follows that there is an equivalent causal interpretation of $(X_1, X_2, \ldots, X_N)$ that begins with (X_A, X_C, X_B). Moreover, because all the random variables causally directly prior to components of X_B must be components of either X_C or X_B itself, $H(X_B \mid X_A X_C) = H(X_B \mid X_C)$. Thus $H(X_A X_C X_B) = H(X_A X_C) + H(X_B \mid X_C) = H(X_C) + H(X_A \mid X_C) + H(X_B \mid X_C)$. Hence, $H(X_A X_B \mid X_C) = H(X_A X_C X_B) - H(X_C) = H(X_A \mid X_C) + H(X_B \mid X_C)$, which shows that X_A and X_B are indeed independent when conditioned on X_C. $\triangle$

4. Concluding remarks

It is not difficult to see that the deductions of independence given in Propositions 2 and 3 are the strongest that can be made from the causality graph alone in the sence that one can always define random variables $X_1, X_2, \ldots, X_N$ that would have the same causality diagram but for which there would no inde-

Figure 3: The causality graph of a secret–key cipher (a) in general and (b) when the cipher provides perfect secrecy where X_1, X_2 and X_3 are the plaintext, secret key and cryptogram, respectively.

pendencies or conditional independencies among these random variables beyond those specified by Propositions 2 and 3.

It must be stressed, however, that not every known independence of random variables can be incorporated into a single causality graph. To see this, suppose that X_1, X_2 and X_3 are the plaintext, secret key and cryptogram, respectively, for some cipher. The secret key is always chosen independently of the plaintext so that the "natural" causality diagram for these random variables is that shown in Fig. 3(a). But if the cipher provides *perfect secrecy* in the sense of Shannon [3], then the plaintext X_1 and the cryptogram X_3 are also independent so that the causality diagram in Fig. 3(b) is also valid. Each of these two causality diagrams provides independence information about X_1, X_2 and X_3 that cannot be garnered from the other. We remark further that all the results of this paper obviously hold also for the case where $X_1, X_2, \ldots, X_N$ are continuous random variables with finite joint differential entropy.

We cannot claim much novelty for this paper. What we have called a

causality graph coincides, when no additional random variables can be removed from the reduced–conditioning expressions for $H(X_{i_n} \mid X_{i_1} \ldots X_{i_{n-1}})$, $n = 2, 3, \ldots, N$, that were used to generate this graph, with what Pearl calls a *Bayesian network*, and our Propositions 2 and 3 characterize, perhaps more clearly, what Pearl calls *d-separation* in such networks (cf. [2, p. 117]). Moreover, Pearl explicitly points out the dependence of the Bayesian network on the ordering of the random variables in question (cf. [2, p. 117]), the inability of such a network to model all independencies that may be known (cf. [2, p. 126]), and the general absence of independencies not deducible from this network, (cf. [2, p. 122]). At most we have formulated a sharper notion of "causality" and given a more transparent mathematical formulation that facilitates proofs. The reader is also referred to [2] for an historical account of the development of methods for deducing independence, both in probabilistic reasoning and in more general models of reasoning.

REFERENCES

1. M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden Day, San Fransisco (1964). 2. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, Calif. (1988). 3. C. E. Shannon, "Communication theory of secrecy systems", *Bell System Tech. J.*, **28**, No. 4, 656–715 (1949).