Directed Information and Conditional Mutual Information

Peter Harremoës
Department of Mathematics
University of Copenhagen
Universitetsparken 5, DK-2100 Copenhagen O
Denmark
Email: moes@math.ku.dk

***Abstract*— We study directed information in Bayesian networks and related structures. Mutual information is split into directed information and residual information. Some basic equations for directed information and residual information are determined.**

I. Introduction

Shannon introduced mutual information and demonstrated how useful this quantity is to analyze a communication setup [1]. For instance the capasity of a simple information channel equals the maximum mutual information between input and output where the maximum is taken over all input distributions. This result is remarkabel in that mutual information is symmetric in the two arguments but an information channel is highly unsymmetric in the sense that the sender has (partly or full) control over the input, but the sender has only indirect control over the distribution of output. Thus, the concept of mutual information does not capture that information flows from the input to the output.

In channels with feedback there is both a flow of information from input to output and a flow from output to input. The concept of directed information has been introduced to analyse channels with feedback, see [2], [3], [4], [5] and . Here we shall go one step further in order to understand what it means that information flows. This question is both of philosophical, theoretical and practical interest. There are many situations where we need more precise concepts in order to understand to what extend information is flowing. We just mention quantum entanglement and secrecy sharing as concepts for which it is not so obvious whether these concepts represent quantities that can flow.

In this paper we shall be more modest and restrict our attention to Bayesian networks, because these networks are important in modelling causality. Thus, for Bayesian networks there is a natural direction of causality and time.

II. Bayesian networks

By a variable we shall understand any measurable mapping from a probability space into a finite space. The finiteness condition is assumed for simplicity and most of our results can easily be extended to variables that can take infinitely many values. Variables will be denoted with capital letters. If X and Y are variables then $X \cup Y$ shall denote the vector valued variable (X, Y). If (X_i) is a (finite or infinite) sequence of variables then X_m^n shall denote the sequence $(X_m, X_{m+1}, ..., X_n)$ where $m \leq n$. Bold face capital letters will indicate infinite sequences.

If the variables X and Y are independent for each value z of the variable Z we say that X and Y are *conditionally independent given* Z and write $(X \perp Y \mid Z)$. The relation conditional independence is a *semi graphoid relation*, which means that

$$(X \perp Y \mid Z) \Longleftrightarrow (Y \perp X \mid Z) \tag{1}$$

and

$$(X \perp Y \cup Z \mid W) \Longleftrightarrow (X \perp Y \mid W) \text{ and } (X \perp Z \mid Y \cup W). \tag{2}$$

We introduce *conditional mutual information* $I(X; Y \mid Z)$ and note that $I(X; Y \mid Z) \geq 0$ with equality if and only if $(X \perp Y \mid Z)$. The relations (1) and (2) for conditional independence have counterparts for conditional independence:

$$I(X; Y \mid Z) = I(Y; X \mid Z) \tag{3a}$$

$$I(X; Y \cup Z \mid W) = I(X; Y \mid W) + I(X; Z \mid Y \cup W). \tag{3b}$$

Note that these equations implies the semi graphoid relations. We also get the important relation

$$I(X; Y \mid Z) = I(X \cup Z; Y) - I(Z; Y) \tag{4}$$

that can be used to define conditional mutual information from mutual information. Now (3a) and (4) together implies (3b).

Semi graphoid relations have proved important in understanding causality [7], but they are difficult to classify. In modelling causality one therefore prefer semi graphoid relations described by Bayesian networks or related graphical models. Here we shall focus on Bayesian networks. A *Bayesian network* is a directed acyclic graph with a variable associated to each node in the graph. If there is a directed path from the variable X to the variable Y in the network we shall write $X \ll Y$ and note that $\ll$ is a partial ordering of the nodes and the associated variables. *The ancestor set* $a(Y)$ is defined as the set of variables X such that there exists a path to a variable in Y.

We shall restrict to *syncronious networks* consisting of a number of layers on top of each other such that any arrow in

the graph points from one layer to the next. For such networks the layers shall form a Markov chain where any variables in one layer is independent of all other variables in the same alyer given its parents. We shall write $X \prec Y$ if X is in a lower layer than Y, and $X \approx Y$ if X and Y are in the same layer. For each variable there is a Markov kernel from the parents of the variables to the variable and the joint distribution on all variables is determined by these Markov kernels. There exists a syncronization of a network if and only all directed paths from one variable to another are of equal length. By introduction of dummy variables it is easy to syncronize any network and therefore the restriction to syncronious networks gives no loss of generality.

The previous description of a Bayesian network is static in the sense that it relates a graph to some variables with a probabilistic structure but with no time evolution. We shall now introduce *dynamic Bayesian networks*. For each variable X in a static Bayesian network we associate an double infinite sequence $\mathbf{X}$ of variables. The joint distribution on all the sequences should be a stationary Markov chain with the Markov kernel being the product of the Markov kernels from the parents of a variable to the variable. The dynamic Bayesian network can be considered as a Bayesian network with infinitely many nodes/variables. The static Baysian network can then be considered as the stationary distribution of the dynamic Bayesian network.

For a stationay sequence $\mathbf{X}$ we define the *entropy rate* in the usual way, i.e.

$$\begin{aligned} H(\mathbf{X}) &= \lim \frac{1}{N} H\left(X_1^N\right) \\ &= \lim \frac{1}{N} \sum_{n=1}^{N} H\left(X_n \mid X_1^{n-1}\right) \\ &= H\left(X_1 \mid X_{-\infty}^0\right). \end{aligned}$$

Similarly we define the *conditional mutual information rate* by

$$I(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) = \lim_{N\to\infty} \frac{1}{N} I\left(X_1^N; Y_1^N \mid Z_1^N\right).$$

For sequences of variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ in a dynamic Bayesian network we have

$$I(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) = I(X; Y \mid Z),$$

where X, Y and Z are the associated variables in the static Bayesian network. The relation can be used to translate concepts for sequences into concepts for static Bayesian networks. Note that in general

$$I\left(X_1^N; Y_1^N \mid Z_1^N\right) \neq I(X; Y \mid Z).$$

Let X denote a sequence. Then we let $T(\mathbf{X})$ denote the *delayed sequence* such that $T(\mathbf{X})_n = X_{n-1}$.

III. Directed and residual information

Consider two sequences of random variables X^N and Y^N. Then the *directed information* from X^N to Y^N is defined by

$$\begin{aligned} &\tilde{I}\left(X^N \to Y^N\right) = \\ &\quad \frac{1}{N} \sum_{n=1}^{N} H\left(Y_n \mid Y_1^{n-1}\right) - H\left(Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \\ &\qquad = \frac{1}{N} \sum_{n=1}^{N} I\left(Y_n; X_1^{n-1} \mid Y_1^{n-1}\right) \end{aligned}$$

If $\mathbf{X} = (X_n)_{n\in\mathbb{Z}}$ and $\mathbf{Y} = (Y_n)_{n\in\mathbb{Z}}$ are stationary processes then the *directed information rate* from $\mathbf{X}$ to $\mathbf{Y}$ is defined as the limit

$$\tilde{I}(\mathbf{X} \to \mathbf{Y}) = \lim_{N\to\infty} \tilde{I}\left(X^N \to Y^N\right).$$

It easy to prove that the limit exists and equals

$$H\left(Y_1 \mid Y_{-\infty}^0\right) - H\left(Y_1 \mid X_{-\infty}^0 \cup Y_{\infty}^0\right).$$

The mutual information rate between two sequences of random variables X^N and Y^N is given by

$$\begin{aligned} &\frac{1}{N} I\left(X^N; Y^N\right) \\ &\quad = \frac{1}{N}\left(H\left(X^N\right) + H\left(Y^N\right) - H\left(X^N \cup Y^N\right)\right) \\ &\quad = \frac{1}{N} \sum_{n=1}^{N} \begin{pmatrix} H\left(X_n \mid X_1^{n-1}\right) + H\left(Y_n \mid Y_1^{n-1}\right) \\ -H\left(X_n \cup Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \end{pmatrix} \\ &= \frac{1}{N} \sum_{n=1}^{N} \begin{pmatrix} H\left(X_n \mid X_1^{n-1}\right) - H\left(X_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \\ +H\left(Y_n \mid Y_1^{n-1}\right) - H\left(Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \\ \begin{pmatrix} H\left(X_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \\ +H\left(Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \\ -H\left(X_n \cup Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right) \end{pmatrix} \end{pmatrix} \\ &\quad = \tilde{I}\left(X^N \to Y^N\right) + \tilde{I}\left(Y^N \to X^N\right) \\ &\qquad + \frac{1}{N} \sum_{n=1}^{N} I\left(X_n; Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right). \end{aligned}$$

We see that a mutual information rate is a sum of two directed information terms plus a term that we shall call the *residual information*. Thus

$$\tilde{I}_{res}\left(X^N; Y^N\right) = \frac{1}{N} \sum_{n=1}^{N} I\left(X_n; Y_n \mid X_1^{n-1} \cup Y_1^{n-1}\right).$$

The residual information measures how much the sequences deviates from being a Bayesian network of the form depicted in Figure 1. For stationary sequences we define the residual information as the limit

$$\begin{aligned} \tilde{I}_{res}(\mathbf{X}; \mathbf{Y}) &= \lim_{N\to\infty} \tilde{I}_{res}\left(X^N; Y^N\right) \\ &= I\left(X_1; Y_1 \mid X_{-\infty}^0 \cup Y_{-\infty}^0\right). \end{aligned}$$

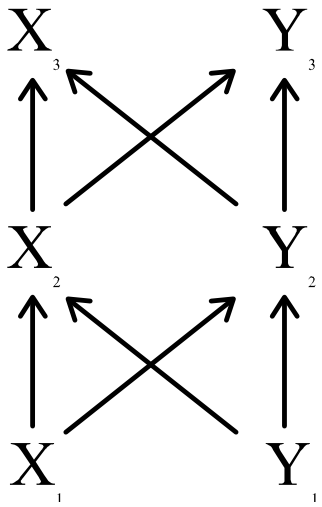

Fig. 1. Two sequences with zero residual information.

Thus we get

$$\frac{1}{N}I\left(X^N;Y^N\right) = \tilde{I}\left(X^N \to Y^N\right) + \tilde{I}\left(Y^N \to X^N\right) + \tilde{I}_{res}\left(X^N;Y^N\right)$$

and

$$I\left(\mathbf{X};\mathbf{Y}\right) = \tilde{I}\left(\mathbf{X} \to \mathbf{Y}\right) + \tilde{I}\left(\mathbf{Y} \to \mathbf{X}\right) + \tilde{I}_{res}\left(\mathbf{X};\mathbf{Y}\right). \quad (5)$$

Example 1: Let $\mathbf{X}$ and $\mathbf{Y}$ be sequences of variables in a dynamical Bayesian network corresponding to variables X and Y associated with single noden in the graph. Then

$$I\left(X;Y\right) = \begin{cases} \tilde{I}\left(\mathbf{X} \to \mathbf{Y}\right) & \text{if } X \prec Y \\ \tilde{I}\left(\mathbf{Y} \to \mathbf{X}\right) & \text{if } Y \prec X \\ I_{res}\left(\mathbf{X};\mathbf{Y}\right) & \text{if } X \approx Y \end{cases}.$$

With this defintion of directed information we have that there may be an information flow from X to Y allthough $\neg X \ll Y$. In this sense information may not flow along the direction of causation. We shall discuss this problem later and suggest a solution.

IV. Relativity of information flows

Both the definition of directed information and residual information depends on the syncronization of the sequences. We shall now see how the definitions rely on the syncronization. We have

$$\begin{aligned}\tilde{I}\left(T\left(\mathbf{X}\right) \to \mathbf{Y}\right) &= H\left(Y_1 \mid Y_{-\infty}^0\right) - H\left(Y_1 \mid T\left(X\right)_{-\infty}^0 \cup Y_{\infty}^0\right) \\ &= H\left(Y_1 \mid Y_{-\infty}^0\right) - H\left(Y_1 \mid X_{-\infty}^{-1} \cup Y_{\infty}^0\right) \\ &\leq H\left(Y_1 \mid Y_{-\infty}^0\right) - H\left(Y_1 \mid X_{-\infty}^0 \cup Y_{\infty}^0\right) \\ &= \tilde{I}\left(\mathbf{X} \to \mathbf{Y}\right).\end{aligned}$$

The identity

$$\tilde{I}\left(\mathbf{X} \to T\left(\mathbf{Y}\right)\right) = \tilde{I}\left(T^{-1}\left(\mathbf{X}\right) \to \mathbf{Y}\right)$$

is obvious. Thus $\tilde{I}\left(T^n\left(\mathbf{X}\right) \to \mathbf{Y}\right)$ is a decreasing function of n and $\tilde{I}\left(\mathbf{X} \to T^n\left(\mathbf{Y}\right)\right)$ is an increasing function of n. Therefore the numbers for which both $\tilde{I}\left(T^n\left(\mathbf{X}\right) \to \mathbf{Y}\right)$ and $\tilde{I}\left(\mathbf{X} \to T^n\left(\mathbf{Y}\right)\right)$ are positive is an interval. Only in this interval communication is possible in both directions and in practice this idea is often use to check syncronization. In the litterature directed information is defined by

$$I\left(\mathbf{X} \to \mathbf{Y}\right) = \tilde{I}\left(\mathbf{X} \to T^{-1}\left(\mathbf{Y}\right)\right),$$

see [5] and [8]. With this definition mutual information will not split up into two flows and a residual term. Instead one has to involve the operator T to get formulas relating mutual information and directed information.

Let k be a positive integer and let $\mathbf{X}$ be a stationary sequence. Then $k\mathbf{X}$ shall denote "block sequence" where the n'th variable is block $X_{(n-1)k+1}, X_{(n-1)k+2}, ..., X_{nk}$. We get

$$\begin{aligned}&\tilde{I}\left(k\mathbf{X} \to k\mathbf{Y}\right) \\ &\quad = H\left(Y_1^k \mid Y_{-\infty}^0\right) - H\left(Y_1^k \mid X_{-\infty}^0 \cup Y_{\infty}^0\right) \\ &\quad = \sum_{n=1}^{k} H\left(Y_n \mid Y_{-\infty}^{n-1}\right) - H\left(Y_n \mid X_{-\infty}^0 \cup Y_{\infty}^{n-1}\right) \\ &\quad \leq \sum_{n=1}^{k} H\left(Y_n \mid Y_{-\infty}^{n-1}\right) - H\left(Y_n \mid X_{-\infty}^{n-1} \cup Y_{\infty}^{n-1}\right) \\ &\quad = k\tilde{I}\left(\mathbf{X} \to \mathbf{Y}\right).\end{aligned}$$

Similarly we get

$$\tilde{I}_{res}\left(k\mathbf{X};k\mathbf{Y}\right) \geq kI_{res}\left(\mathbf{X};\mathbf{Y}\right).$$

The conclusion is that one can explain less mutual information as flow and more as residual information if one consider a block sequence instead of the original sequence.

V. Conditional flows

Conditional versions of directed information have been studied in [4], [5], [8] and [6]. Note that the defintion we shall use is a little different from what is found in the litterature. We define the *conditional directed information* by

$$\tilde{I}\left(X^N \to Y^N \| Z^N\right) = \tilde{I}\left(X^N \cup Z^N \to Y^N\right) - \tilde{I}\left(Z^N \to Y^N\right) \quad (6)$$

so that the information flowing from X^N to Y^N given Z^N is the information flowing from both variables minus the amount of information flowing Z^N. Thus

$$\begin{aligned}&\tilde{I}\left(X^N \to Y^N \| Z^N\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\begin{pmatrix} H\left(Y_n \mid Y_1^{n-1}\right) \\ -H\left(Y_n \mid X_1^{n-1} \cup Y_1^{n-1} \cup Z_1^{n-1}\right)\end{pmatrix} \\ &- \frac{1}{N}\sum_{n=1}^{N} H\left(Y_n \mid Y_1^{n-1}\right) - H\left(Y_n \mid X_1^{n-1} \cup Z_1^{n-1}\right) \\ &= \frac{1}{N}\sum_{n=1}^{N}\begin{pmatrix} H\left(Y_n \mid Y_1^{n-1} \cup Z_1^{n-1}\right) \\ -H\left(Y_n \mid X_1^{n-1} \cup Y_1^{n-1} \cup Z_1^{n-1}\right)\end{pmatrix}.\end{aligned}$$

Note that $\tilde{I}\left(X^N \to Y^N \| Z^N\right) \geq 0$.

For stationary sequences we define

$$\tilde{I}(\mathbf{X} \to \mathbf{Y} \| \mathbf{Z}) = \lim_{N\to\infty} \tilde{I}\left(X^N \to Y^N \| Z^N\right).$$

Then

$$\begin{aligned}\tilde{I}(\mathbf{X} \to \mathbf{Y} \| \mathbf{Z}) &= H\left(Y_1 \mid Y_{-\infty}^0 \cup Z_{-\infty}^0\right) \\ &\quad - H\left(Y_1 \mid X_{-\infty}^0 \cup Y_{-\infty}^0 \cup Z_{-\infty}^0\right).\end{aligned}$$

We observe that

$$\begin{aligned}&\tilde{I}(\mathbf{X} \cup \mathbf{Y} \to \mathbf{Z} \| \mathbf{W}) \\ &= H\left(Z_1 \mid Z_{-\infty}^0 \cup W_{-\infty}^0\right) \\ &- H\left(Z_1 \mid X_{-\infty}^0 \cup Y_{-\infty}^0 \cup Z_{-\infty}^0 \cup W_{-\infty}^0\right) \\ &= H\left(Z_1 \mid Z_{-\infty}^0 \cup W_{-\infty}^0\right) - H\left(Z_1 \mid X_{-\infty}^0 \cup Z_{-\infty}^0 \cup W_{-\infty}^0\right) \\ &+ H\left(Z_1 \mid X_{-\infty}^0 \cup Z_{-\infty}^0 \cup W_{-\infty}^0\right) \\ &- H\left(Z_1 \mid X_{-\infty}^0 \cup Y_{-\infty}^0 \cup Z_{-\infty}^0 \cup W_{-\infty}^0\right) \\ &= \tilde{I}(\mathbf{X} \to \mathbf{Z} \| \mathbf{W}) + \tilde{I}(\mathbf{Y} \to \mathbf{Z} \| \mathbf{X} \cup \mathbf{W}).\end{aligned}$$

We are now able to write a conditional mutual information in terms of conditional flows and residual information. We have

$$\begin{aligned}I(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) &= \lim_{N\to\infty} \frac{1}{N} I\left(X^N; Y^N \mid Z^N\right) \\ &= \lim_{N\to\infty} \begin{pmatrix} \frac{1}{N} I\left(X^N \cup Z^N; Y^N\right) \\ -\frac{1}{N} I\left(Z^N; Y^N\right) \end{pmatrix} \\ &= I(\mathbf{X} \cup \mathbf{Z}; \mathbf{Y}) - I(\mathbf{Z}; \mathbf{Y}).\end{aligned}$$

Each of these terms can be written as a sum of two flow and residual information. Thus,

$$\begin{aligned}&I(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) = \\ &\tilde{I}(\mathbf{X} \cup \mathbf{Z} \to \mathbf{Y}) + \tilde{I}(\mathbf{Y} \to \mathbf{X} \cup \mathbf{Z}) + \tilde{I}_{res}(\mathbf{X} \cup \mathbf{Z}; \mathbf{Y}) \\ &\quad - \tilde{I}(\mathbf{Z} \to \mathbf{Y}) - \tilde{I}(\mathbf{Y} \to \mathbf{Z}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{Y}).\end{aligned}$$

and we get the formula

$$\begin{aligned}&I(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z}) = \\ &\tilde{I}(\mathbf{X} \to \mathbf{Y} \| \mathbf{Z}) + \tilde{I}(\mathbf{Y} \to \mathbf{X} \cup \mathbf{Z}) - \tilde{I}(\mathbf{Y} \to \mathbf{Z}) \\ &\quad + \tilde{I}_{res}(\mathbf{X} \cup \mathbf{Z}; \mathbf{Y}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{Y}).\end{aligned}$$

The formula is closely related to [8, Prop. 3], but by introduing residual information we have formulas that do not involve the delay operator. It seems tempting to define $\tilde{I}_{res}(\mathbf{X} \cup \mathbf{Z}; \mathbf{Y}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{Y})$ as the residual mutual information of X and Y given Z, but it may lead to a negative quantity.

The basic equations for conditional mutual information can now be stated in terms of directed and residual information. Equation 3a states that

$$\begin{aligned}&\tilde{I}(\mathbf{X} \to \mathbf{Y} \| \mathbf{Z}) + \tilde{I}(\mathbf{Y} \to \mathbf{X} \cup \mathbf{Z}) - \tilde{I}(\mathbf{Y} \to \mathbf{Z}) \\ &+ \tilde{I}_{res}(\mathbf{X} \cup \mathbf{Z}; \mathbf{Y}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{Y}) \\ &= \tilde{I}(\mathbf{Y} \to \mathbf{X} \| \mathbf{Z}) + \tilde{I}(\mathbf{X} \to \mathbf{Y} \cup \mathbf{Z}) - \tilde{I}(\mathbf{X} \to \mathbf{Z}) \\ &+ \tilde{I}_{res}(\mathbf{Y} \cup \mathbf{Z}; \mathbf{X}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{X})\end{aligned}$$

The terms can be reorganized so that we get

$$\begin{aligned}&\tilde{I}(\mathbf{X} \to \mathbf{Y} \cup \mathbf{Z}) - \tilde{I}(\mathbf{X} \to \mathbf{Y} \| \mathbf{Z}) - \tilde{I}(\mathbf{X} \to \mathbf{Z}) \\ &+ \tilde{I}_{res}(\mathbf{X}; \mathbf{Y} \cup \mathbf{Z}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{X}) \\ &= \tilde{I}(\mathbf{Y} \to \mathbf{X} \cup \mathbf{Z}) - \tilde{I}(\mathbf{Y} \to \mathbf{X} \| \mathbf{Z}) - \tilde{I}(\mathbf{Y} \to \mathbf{Z}) \\ &+ \tilde{I}_{res}(\mathbf{Y}; \mathbf{X} \cup \mathbf{Z}) - \tilde{I}_{res}(\mathbf{Z}; \mathbf{Y}).\end{aligned} \tag{7}$$

We see that relation 7 together with Equation 5 and the definitions 6 and 4 implies the identities 3a and 3b.

VI. Lower bounds on directed information

As we have noticed that directed information is relative to the syncronization of the sequences. Nevertheless it is sometimes possible to provide lower bound to the directed information, which are independent of the syncronization. Here we shall just provide a simple example that will illustrate the idea.

Consider (sets of) variables X, Y, Z and W in a Bayesian network satisfying that statistical independence holds if and only the d-separation criteria is fulfilled. Assume that X and Y are independent given $Z \cup W$. Then we have the lower bound

$$I(W \to X \cup Y \cup Z) \geq I(X; Y \mid Z). \tag{8}$$

To se this we put $W' = W \cap a(X \cup Y \cup Z)$ and note X and Y are independent of $Z \cup W'$ [9]. Now,

$$\begin{aligned}\tilde{I}(W \to X \cup Y \cup Z) &= \tilde{I}(W' \to X \cup Y \cup Z) \\ &\quad + \tilde{I}\left(\complement W \to X \cup Y \cup Z \mid W'\right) \\ &\geq \tilde{I}(W' \to X \cup Y \cup Z) \\ &\geq I(W; X \cup Y \cup Z) \\ &\geq I(W; X \cup Y \mid Z) \\ &\geq I(X; Y \mid Z).\end{aligned}$$

The last inequality is easily proved by using the Venn diagram method [10].

The bound (8) can be written as

$$\tilde{I}(W \to X \cup Y \cup Z) \geq I(X; Y \mid Z) - I(X; Y \mid Z \cup W).$$

Under weak conditions this lower bound holds even when X and Y are not independent given $Z \cup W$. More refined lower bounds are an area for future investigations.

VII. Acknowledgement

The author want to thank Jim Massey for useful discussions.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, 1948.

[2] H. Marko, "The bidirectional communication theory - a generalization of information theory," *IEEE Trans. Commun.*, vol. COM-21, pp. 1345–1351, Dec. 1973.

[3] J. L. Massey, "Causality, feedback and directed information," in *Proc. 1990 Int. Symp. on Info. Th. and its Appls., Hawaii, USA,*, pp. 303–305, 1990.

[4] G. Kramer, *Directed Information for Channels with Feedback*, vol. 11 of *ETH Series in Inform. Proc.* Konstanz: Hartung-Gorre, 1998.

[5] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inform. Theory*, vol. IT-49, pp. 4–21, Jan. 2003.

[6] R. Venkataramanan and S. S. Pradhan, "Directed information for communication problems with common side information and delayed feedback/feedforward," in *Proc. Allerton conference on communication, control and computing*, Sept. 2005.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, Califonia: Morgan Kaufmann Publ., 1988.

[8] J. L. Massey and P. C. Massey, "Conservation of mutual and directed information," in *Proceedings of the 2005 IEEE International Symposium on Information Theory*, (Adelaide, South Australia, Australia), pp. 157–158, Sept. 2005.

[9] P. Harremoës, *Time and Conditional Independence*, vol. 255 of *IMFUFA-tekst*. IMFUFA Roskilde University, 1993. Original in Danish entitled Tid og Betinget Uafhængighed. English translation partially available at http://www.math.ku.dk/ moes/index.html .

[10] R. W. Yeung, *A First Course in Information Theory*. New York: Kluwer, 2002.